Effectiveness of natural language processing techniques in categorizing scientific articles by research methodology.

TAREK SAKHI, University of Twente, The Netherlands
FAIZAN AHMED, University of Twente, The Netherlands
YERAY D.C. BARRIOS FLEITAS, University of Twente, The Netherlands

Motivation: With the ever-growing number of published scientific articles, it becomes increasingly challenging for researchers to find, review and use relevant research. **Aim:** This study explores the potential of using unsupervised text classification models, specifically a zero-shot classification model (GPTNLI) and a similarity-based (Lbl2vec) classification model, to streamline the literature review process. **Method:** These models predict the methodological approach based on simple information like the title, keywords and abstract, thereby allowing for an extra filter during scientific database searches. To accomplish this, an extensive and well-structured definition is established for each class. **Result:** The finding demonstrates that the GPTNLI model using GPT4, outperforms the other models in accuracy and f1 scores while showing reduced variability in its performance. Through using a binomial test it is shown that the model's performance statistically outperforms a random-guess strategy. **Conclusion:** Although the study has its limitations; For instance, the use of small test datasets and lack of cost-benefit analysis, the results are promising. Future research could improve the performance of the models by incorporating more sections of the study, further fine-tuning and adding self-learning capabilities.

Keywords: Unsupervised Text Classification, Literature Review, GPT, Lbl2vec

1 INTRODUCTION

In recent decades, scientific production has grown dramatically across all research fields [30]. This exponential growth in scientific publications led to enormous amounts of new information. This results in infoxication, the inability to find what you are looking for due to the volume and dispersion of information. Literature reviews are valuable in this context, as they help organize knowledge and facilitate progress in various domains.

Correlated to the rise of publications, the screening process of literature reviews has become increasingly tricky. This study aims to explore the potential of unsupervised text classification models, notably a zero-shot (GPTNLI, Generative Pre-training Transformer Natural Language Inference) and a similarity-based model (Lbl2vec), in enhancing the efficiency of the screening process in a literature review. This study aims to determine whether these models can predict a study's methodological approach using just the abstract. In doing so, the goal is to allow for an additional filter layer during scientific database searches, thereby reducing the time spent on the literature screening process.

TScIT 37, July 7, 2023, Enschede, The Netherlands

Several studies have employed Natural Language Processing (NLP) and machine learning techniques to assist the automation of literature reviews [6, 14, 28]. Furthermore, the application of text categorisation in scientific articles has been studied [13, 32, 36]. Most of those studies revolve around categorising the study topics and classifying the studies using supervised methods. Due to the lack of an available large labelled training dataset for this study's use case, unsupervised NLP techniques are required. In spite of the extensive investigation conducted on evaluating various unsupervised text classification methods by Shopf et al [25], the present study aims to concentrate explicitly on applying such models to classify distinctive attributes, precisely the methodological approach. It should be noted that while Shopf et al employed unsupervised text classification methods, their focus was primarily on topic classification within a given text. Therefore, this study, with its distinctive blend of utilizing unsupervised text classification and categorizing specific characteristics of academic papers, represents a contribution to the field.

The following research questions (RQ) are answered to achieve that goal:

RQ1. Can an unsupervised text classification model, which classifies the research approach of a study, enhance the efficiency of the literature review screening process?

RQ2. In terms of predicting the research approach of a study based solely on the abstract, which type of unsupervised text classification model delivers the most accurate results?

The paper starts by providing a comprehensive review of the relevant literature, establishing a literature base for this study. Afterwards, the research methodology is discussed, providing insight into the data collection process, the models used and the evaluation methods. Then the results are presented, comparing the performance of the zero-shot and similarity-based models. Furthermore, the results are analysed considering the implications to the field. Finally, the study is concluded, summarising the findings, acknowledging the limitations and suggesting possible future research.

2 BACKGROUND

2.1 Current literature review process

According to [27], many methodologies exist to conduct a literature review. There are several frameworks to give structure to a study. Some standard methods are integrative reviews [34], systematic reviews [9] and meta-analysis [1]. The study [27] divides all the reviews into three main approaches: systematic, semi-systematic and integrative. However, this study aims to create a model accommodating various literature reviews.

Therefore it is required to align the overarching phases of these approaches. The study by [27] synthesizes these methodologies into four overarching phases.

(1) Designing the review
(2) The screening process
(3) Analysis
(4) Writing up the review

This study focuses on the screening process. By incorporating the model presented in this paper, researchers can effectively refine their search queries based on a specific methodological approach. This capability allows them to specify their requirements and narrow the search results accordingly.

2.2 Automatization of Literature Review

The increasing volume of scientific papers has motivated researchers to develop (semi-)automated methods to assist the literature review process. Several studies [6, 14, 28] have explored machine learning and natural language processing to achieve this goal. The tools can reduce the workload and time required to screen papers manually.

Initially, Marshall et al. [14] provide an overview of current machine-learning methods that can be used for evidence synthesis. That study provides a comprehensive overview of the methods' readiness, strengths, weaknesses and usage. A highly influential study [33] developed an open-source machine learning-aided pipeline called ASReview. It utilizes active learning to improve the efficiency of a literature review. The tool can be used for many tasks, including systemic reviews and meta-analyses.

More recent studies, like [7], assessed the performance of using the OpenAI GPT API in accurately and efficiently identifying relevant titles and abstracts for clinical reviews. The results showed high accuracy and the potential to streamline the clinical review process. Additionally, [22] used NLP techniques to analyze the NLP-focused literature, providing meta-level knowledge about the current state of the field and a guide to the use of essential NLP tools. It is fully automated, allowing for easy reproducibility, continuation and updating of the research.

Another influential study by [2] introduced Research Screener, a semi-automated tool that significantly reduces workload and review time in systematic reviews and meta-analyses. When only screening around 50% of the articles, this tool will likely identify all relevant papers.

2.3 Text classification techniques

Several researchers have been using (NLP) techniques to assist in automating the literature review process. The classification of scientific articles based on their characteristics, such as their topic, methodology or domain, using NLP techniques could significantly help streamline and automate the process. Before delving deeper into text classification techniques applied to scientific literature, this study first takes a moment to explore the fundamental techniques employed in text classification as a whole.

There are comprehensive overviews [4, 6, 28] of using various NLP techniques for text classification. In classifying scientific literature, research was conducted to investigate the effectiveness of topic modelling, document classification and trend analysis. For example, [13] proposed an automated evaluation method for abstracts of articles. At the same time, earlier studies used machine learning techniques such as k-Nearest Neighbour, Latent Dirichlet Allocation (LDA), Support Vector Machines, and Naïve Bayes [10, 32, 36]. More recent studies are utilizing transformer models [29]. Instances using transformers have shown [15] to outperform other models.

Multiple strategies are employed to achieve text classification within the outperforming transformer-based approach. For instance, in [29], the authors fine-tuned the BERT model and compared the performance to the XLNet model for automatic document classification. Another study by [38] utilized the Arabic BERT model, a fine-tuned BERT model for the Arabic language, in two different ways: as a transfer learning model and as a feature extractor. Lastly, [31] proposed a framework that employs DistilBERT as an encoder layer to obtain context-sensitive dynamic word vectors.

Transformer models often require extensive labelled data, which can be challenging when addressing the problem with limited to no labelled training data. Although few-shot classification techniques, such as Mask-BERT [12] and ContrastNet [3], allow classification with minimal labelled data.

2.4 Text classification techniques applied on scientific papers

Researchers have developed a text classification system for scientific papers using a combination of NLP and machine-learning techniques. [11] proposed a system that uses LDA and K-means clustering to cluster similar topics using each paper's Term frequency-inverse document frequency(TF-IDF) values. Similarly, [32] developed a classification model for articles that utilize different techniques, including Support Vector Machines (SVM), Naïve Bayes and k-Nearest Neighbours(k-NN). It showed the feasibility of using NLP and SVM for automatic article classification, achieving an accuracy of a little over 91%.

Another study by [23] applied NLP techniques to identify journal publication trends and topic clustering. [36] showed the automatic classification of papers published in Scopus. They apply k-NN and Linear Discriminant Analysis to achieve an accuracy of 88.44%.

Finally, [13] presented a framework for classifying and evaluating papers based on their abstracts. They use various techniques to model, classify and segment the text data to run a sentiment analysis. The framework was validated on oil production anomaly abstracts, showing promising results.

2.5 Unsupervised text classification techniques

Unsupervised text classification techniques are a powerful NLP tool that does not require labelled data for training. It is a suitable option when it is costly or too time-consuming to annotate a large data set. The latter is the case in this research. Schopf et al [25] presents a comprehensive overview of the state-of-the-art methods and divides the methods into two categories: zero-shot classification and similarity-based approaches.

Zero-shot classification uses pre-trained models to predict unseen classes, not requiring the utilization of examples. Several

methods exist to achieve zero-shot classification, according to [25], entailment approaches produce state-of-the-art results [37]. Those methods consider zero-shot classification an entailment problem, providing a textual description of the labels. For instance, TARS [8] approaches it as a binary classification problem, which uses the textual description to determine whether a provided text is that label.

The other category is similarity-based approaches, which compute the similarity between the semantic embeddings of the text and the textual label description. The computed similarity is leveraged to determine the correct label. Schopf et al. [25] focuses on the Lbl2Vec [26] method due to the improved accuracy over other similarity-based approaches. Lbl2vec starts by creating joint embeddings of labels and documents. The labels are defined using keywords. Then the centroid of the label vectors is used to determine the most similar label for each document using cosine similarity. Afterwards, using the previously assigned candidate documents, the average vector is computed to represent the label centroid. Lastly, new documents and word vectors are compared to the label centroids to determine the most similar label. Lbl2vec started off utilizing embeddings generated by Word2Vec. However, [26] used transformer-based embedding to obtain improved performance.

3 METHODOLOGY

Initiating the study involves carefully selecting the appropriate characteristic to forecast. Following this step, the research establishes concrete definitions for each class lending structure to the subsequent steps. Afterwards, the raw dataset is transformed into a suitable training dataset, and a test dataset is created. The focus then shifts to designing and implementing Natural Language Processing (NLP) models tailored to meet the study's requirements. Next, the study defines suitable metrics for model evaluation to ensure their practical performance assessment. Finally, an experimental design explicitly formulated to address the posed research question completes this paper's methodological framework.

3.1 Selection of characteristic

The characteristic predicted in this study is methodological, more specifically, the approach, as defined in Table 1. This label exhibits a finite multi-class structure, proposing a suitable implementation for our model. The 'mixed' class introduces an additional layer of complexity in this classification task. Since the mixed class embodies both quantitative and qualitative characteristics, rendering it a unique blend of both classes.

Label	**Classes**
Approach	Qualitative, Quantitative, Mixed

Table 1. Classes for each label

Knowledge of the approach is crucial as it dictates the nature of the findings. Quantitative studies offer numerical insights, whilst qualitative studies provide more in-depth insight into particular cases. Furthermore, mixed studies use a blend of both approaches. Having the ability to filter on the approach can substantially aid researchers in reducing search results and finding the correct studies during the screening process of a literature review.

3.2 Defining the characterisics

Classifying a study's approach is a subjective process. In order to enhance the reproducibility and objectivity of this research, propositional criteria are introduced for quantitative and qualitative, based on the definitions introduced in [35].

Quantitative Research Approach:

- The research question of the study is designed to quantify and statistically measure outcomes, correlations or differences between variables.
- The data collected in the study is numerical and quantifiable.
- The study applies statistical methods to analyse the data collected.
- The data collection process uses (semi-)structured methods, for instance, surveys or questionnaires.
- The findings are mainly presented in a numerical form, such as tables, graphs or measurements.

Qualitative Research Approach:

- The research question of the study is designed to explore, interpret or generate understanding about a phenomenon.
- The data collected is non-numerical, for instance, text, video or audio.
- The study applies interpretive or subjective methodologies for data analyses.
- The data collection process uses unstructured or semi-structured methods, for example, interviews, observations or analysis of documents.
- The findings are presented in a narrative or descriptive form, providing detailed insights.

The mixed class implies that the study carries attributes of both quantitative and qualitative nature. The methodology adopted to classify a study as mixed is as follows:

Let us denote the total number of true quantitative criteria in each study as N_{QUANT_TRUE} and the total number of true qualitative criteria in each study as N_{QUAL_TRUE}.

The first step involves evaluating the difference in the counts of true qualitative and quantitative criteria in each study, defined as:

$$\text{DIF} = N_{\text{QUAL\_TRUE}} - N_{\text{QUANT\_TRUE}} \tag{1}$$

Next, the 33rd ($P33$) and 66th ($P66$) percentiles are computed of DIF across all studies.

Finally, the studies are categorised based on their respective DIF values relative to these percentile thresholds. Specifically:

- A study is classified as QUALITATIVE if DIF $\leq$ P33.
- A study is classified as MIXED if P33 $<$ DIF $\leq$ P66.
- A study is classified as QUANTITATIVE if DIF $>$ P66.

This classification strategy thus ensures a reproducible and accurate differentiation of the studies into quantitative, qualitative, and mixed categories.

3.3 Dataset acquisition and preparation

Based on the criteria outlined in the previous section, access to the full text of a paper would achieve optimal results. It would contain the information necessary to answer the criteria. However, obtaining such information is cumbersome without writing a program to scrape a corpus. Given the time limitation of this study, such an approach is not feasible. However, some datasets offer information such as a paper's abstract, authors, keywords and title. This study employs such a dataset.

The dataset [21] comprises of 6865 studies, including their titles, abstracts, author keywords and index keywords. Researchers from the University of Twente have manually curated the data. While the original data contained more columns than the used dataset, a few columns were removed due to not providing additional information when answering the criteria. These columns were the year the paper was published, the number of citations, the number of references, document type and id. Additionally, columns containing many empty values have been removed. These columns were the science category and WoS category.

Furthermore, it is worth noting that the transformed data includes the author and index keywords despite having missing values. There were two reasons for this decision. Firstly, these columns contained fewer missing values than the omitted WoS category and science category. Secondly, those characteristics could still provide valuable insight. Finally, the studies in the dataset encompass a particular research domain, focusing on team effectiveness.

3.4 Models design

Due to this study's absence of labelled data and time constraints that prevent manually labelling a comprehensive training set. The study utilised two types of models, following Shopf et al. [26], Lbl2Vec, which leverages a transformer for semantic embedding and zero-shot classification.

Lbl2vec. For the implementation [18], this study employs an open-source library created by the authors of the Lbl2vec study [24]. This library was updated to support transformer embeddings.

The initial parameter of this model involved selecting the appropriate transformer model. For this purpose, the 'all-MPNet-base-v2' model is the best available general sentence transformer at the time of this study.

Another important choice is the initial keywords for each label. In order to achieve this, this study employs two strategies that involve prompting ChatGPT.
In the first approach $keywords_{knowledge-based-prompt}$ shown in [19], the methodology of Lbl2vec is explained, after which the model is asked to generate keywords for the quantitative and qualitative classes, based on its knowledge of these classes. Similarly, in the second approach $keywords_{criteria-based-prompt}$ shown in [17], the Lbl2vec methodology is explained, followed by providing the criteria for each class, then prompting ChatGPT to generate keywords for the classes, asking ChatGPT to draw upon its knowledge of the classes and the set criteria for each class. Lastly, a second prompt is formulated to obtain additional keywords and ensure the appropriate format is returned. Subsequently, the aforementioned procedure is iteratively performed for the criteria-based prompt, allowing for observing the random behaviour exhibited by ChatGPT. This repetition yields a second criteria-based prompt. All the keywords utilized in this context can be found in work cited [19].

The final parameter under consideration is n_{docs}, which signifies the number of documents used for model training. The following values are assigned to investigate the impact of incrementing that variable: 250, 1000, and 3000. The numerical values are selected in an arbitrary manner, guided by the requirement for a sufficiently large increment. The upper limit of 3000 is determined by computational constraints pertaining to the processing time.

Following the model training, it returns a dataframe containing the cosine similarities to the 'quantitative' and 'qualitative' classes, denoted as $label_0$ and $label_1$, respectively. As discussed before, a similar approach is used to classify the' mixed' class. A new column, dif, is generated following the formula $dif = (label_0 - label_1)/label_1$. Subsequently, the 33rd percentile $P33$ and the 66th percentile $P66$ are computed for dif. A study is classified as 'Qualitative' if $dif <= P33$, 'Mixed' if $P33 < dif < P66$, and 'Quantitative' if $dif >= P66$.

Zero-shot classification, GPTNLI. Similar to [8], this study approaches the zero-shot classification problem as a textual entailment problem [16]. The GPT-4 model uses a question-answering approach facilitated by prompts. This methodology led to the model's naming as the Generative Pre-training Transformer for Natural Language Inference, also known as GPTNLI, in this study. Abstractly the logic is as follows:

Given that a study is defined as the abstract of that study, the general knowledge of GPT-4 is utilised to assert if the study entails the satisfaction of a given criterion.

The GPT API yields an array with True or False for each criterion. Finally, the same methodology described in a previous section 3.2 determines the class.

This model solely leverages the abstract, excluding authors, keywords, and title, due to test iterations indicating a substantial enhancement in performance when only using the abstract.

3.5 Models evaluation

Due to the lack of labelled data, splitting the data into a training and testing set is not feasible. Therefore to evaluate the model's performance, a test dataset [20] is manually crafted using the methodology specified in a previous section 3.2. The manually labelled test data will then be utilised to calculate generic metrics to evaluate the performance. Due to the time constraints, the test data set consists of only 24 studies.

To measure the efficacy of the multi-class classification model, two indicators, namely *accuracy* and *F1-score*, will be evaluated. These metrics have been prominently employed in similar studies, as reported in [5, 39], facilitating a meaningful comparison.

In machine learning, *accuracy* is a commonly used metric for evaluating the performance of classification models.

It represents the proportion of correct predictions made by the model out of the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Where:

- TP represents the number of true positives: the cases where the model correctly predicted the positive class.
- TN represents the number of true negatives: the cases where the model correctly predicted the negative class.
- FP represents the number of false positives: the cases where the model incorrectly predicted the positive class.
- FN represents the number of false negatives: the cases where the model incorrectly predicted the negative class.

However, accuracy alone can be a misleading metric, particularly in cases where the data set is unbalanced, for instance, in a dataset where one class is more present than another. The *F1 score* is a metric that combines precision and recall, which are particularly useful in the context of unbalanced datasets. Precision is the proportion of true positive predictions out of the total positive predictions, while recall is the proportion of true positive predictions out of the total actual positives.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The F1 score is the harmonic mean of precision and recall, balancing these two metrics, defined as:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

A higher F1 score indicates a superior model.

3.6 Experimental design

The dataset described in Section 3.3 evaluates the different unsupervised models. Firstly, the experiment evaluates the Lbl2vec model using the 'all-MPNet-base-v2' sentence transformer model for generating vector embeddings. As described in Section 3.4 the Lbl2vec model is trained and evaluated in 9 epochs, changing two parameters. The first parameter, n_{docs}, signifies the number of documents used for model training. For each run, the studies used for training are randomly selected from the total dataset, specified in Section 3.3. The second parameter, *KEYWORDS*, marks the keywords used to create the label vectors. Refer to [19] for the different keywords generated by the prompts described in Section 3.3. Table 2 defines the following evaluation epochs.

Two transformer-based models are used when evaluating the zero-shot classification technique: GPT -4 and GPT-3.5. GPT-4 has state-of-art performance, and GPT-3.5 is a less accurate but faster model. There this part of the experiment consists of two evaluation epochs.

The models with identical parameters are executed four times to obtain reliable F1 scores and accuracy for each evaluation epoch. Subsequently, the average accuracy and F1 score, weighted equally, are calculated.

EPOCH	KEYWORDS	N_{DOCS}
1	$keywords_{knowledge-based-prompt}$	250
2	$keywords_{knowledge-based-prompt}$	1000
3	$keywords_{knowledge-based-prompt}$	3000
4	$keywords_{criteria-based-prompt}$	250
5	$keywords_{criteria-based-prompt}$	1000
6	$keywords_{criteria-based-prompt}$	3000
7	$keywords_{criteria-based-prompt2}$	250
8	$keywords_{criteria-based-prompt2}$	1000
9	$keywords_{criteria-based-prompt2}$	3000

Table 2. Analysis Epochs

In order to mitigate randomness, each model is executed four times across all parameter configurations. The mean values of accuracy and F1-score, along with their respective standard deviations, will be computed.

Finally, a binomial test validates the results, especially considering the small test dataset of just 24 studies. This test compares model performance against a random guess strategy, ensuring that the observed results aren't due to random chance but indicate genuine model effectiveness.

4 RESULTS

In Table 3 and Table 4, ACC is the accuracy and F1 the F1-score. The standard deviation, STDEV is computed for both these metrics. The accuracy is denoted as a percentage, and the F1-score is a regular decimal. All the values are actually the average values after 4 runs, as described in 3.6.

KEYWORD	N_DOCS	ACC	STDDEV_ACC	F1	STDDEV_F1
Knowledge	300	31,5225	8,962095644	0,3216	0,086224745
Knowledge	1000	24,9975	6,519444634	0,252825	0,068575135
Knowledge	3000	34,78	5,022947342	0,3541	0,045909549
Criteria	300	22,825	8,229574716	0,23575	0,083137497
Criteria	1000	17,3925	7,099438358	0,177	0,070954211
Criteria	3000	18,4775	2,175	0,194375	0,02183596
Criteria 2	300	31,52	7,42464814	0,331175	0,067343071
Criteria 2	1000	38,045	8,229574716	0,38065	0,08504879
Criteria 2	3000	43,48	6,151828996	0,4518	0,052097665

Table 3. Performance evaluation of the Lbl2Vec model, illustrating accuracy (ACC), its standard deviation (STDDEV_ACC), F1-score (F1), and its standard deviation (STDDEV_F1) with different keyword sets and varying number of documents (N_DOCS) used for training.

Model	ACC	STDDEV_ACC	F1	STDDEV_F1
GPT3,5	41,3025	6,52083344	0,427675	0,058809454
GPT4	53,2575	1,883605253	0,52605	0,024958015

Table 4. Comparative performance of GPT-3.5 and GPT-4 models. The table shows each model's accuracy (ACC), standard deviation of accuracy (STDDEV_ACC), F1-score (F1), and standard deviation of F1-score (STDDEV_F1) using Zero-shot classification approach.

Table 3 underscores the optimal performance of the Lbl2Vec model using the "Criteria 2" keyword set and 3000 documents

(N_{DOCS}) for training, reaching an accuracy of 43.48% and an F1-score of 0.4518. Interestingly, although generated from the same prompt, 'Criteria' and 'Criteria 2' provide contrasting performance outputs.

According to Table 4, the GPT-4 model exhibits superior performance with an accuracy of 53.26% and an F1-score of 0.52605. Although GPT-3.5 is less accurate, it offers faster processing capabilities.

A noteworthy trend is observed in the performance evaluation of the Lbl2Vec model (Table 3). As the number of documents used for training (N_{DOCS}) increases, there is a corresponding decrease in the standard deviation of both accuracy ($STDDEV_{ACC}$) and F1-score ($STDDEV_{F1}$). This trend suggests decreased model performance variability, contributing to more stable and consistent results with increasing training dataset size.

Similarly, the GPT-4 model (Table 4) displays a significantly lower standard deviation in accuracy and F1-score compared to the GPT-3.5 model. This decreased variability indicates a greater consistency in performance.

As mentioned, for effective screening, the accuracy of a model should be higher than that of a random guess. Using the best-performing parameters, let the accuracies of our models be defined as $acc_{lbl2vec} = 0.4348$ and $acc_{gptnli} = 0.5326$, and let $acc_{random} = \frac{1}{3} = 0.3333$ be the accuracy of a random model for a 3-class classification problem with equal distribution.

Our null hypothesis H_0 and alternative hypothesis H_1 are defined as:

H_0: The accuracies of our models are not significantly different from random guessing, i.e., $acc_{\text{lbl2vec}} = acc_{\text{random}}$, and $acc_{\text{gptnli}} = acc_{\text{random}}$.

H_1: The accuracies of our models are significantly higher than random guessing, i.e., $acc_{\text{lbl2vec}} > acc_{\text{random}}$ and $acc_{\text{gptnli}} > acc_{\text{random}}$.

If the p-value from the Binomial Test is less than the significance level, 0.05, the null hypothesis is rejected, thereby concluding that the model's accuracy is significantly greater than random chance.

The following python code is used the calculate the p-value:

```python
from scipy.stats import binomtest

# Model Lbl2vec n_docs=3000 and keywords=criteria2
p_value_lbl2vec = binomtest(10, 24, 0.3333,
    alternative='greater')

# Model GPTNLI model=GPT4
p_value_gptnli = binomtest(13, 24, 0.3333,
    alternative='greater')
```

Fig. 1. Python code snippet for Binomial test p-value calculation.

Executing this code, the Lbl2vec model has an accuracy of 43.48%, representing 10 successes (k) out of 24 trials (n), with a null hypothesis success probability (p) of 33.33%. Substituting these values into the code results in a p-value of 0.254.

Similarly, the GPTNLI model, with an accuracy of 53.26%, has 13 successes out of 24 trials. The p-value calculated with these values is 0.028.

Therefore the null hypothesis can be rejected for the GPTNLI model based on this test of only 24 trials. This implies that GPTNLI with GPT4 performs significantly better than a random guess strategy.

5 DISCUSSION

RQ1. *Can an unsupervised text classification model, which classifies the research approach of a study, enhance the efficiency of the literature review screening process?*

The results of this study provide compelling evidence in response to RQ1, showing that unsupervised text classification models can significantly enhance the efficiency of the literature review screening process. In particular, both the Lbl2Vec and the GPTNLI model demonstrated higher accuracy when using the GPT-4 models than random guessing in their optimal configurations.

The GPT-4-based GPTNLI model notably passed the binomial test, signifying statistically significant performance despite the limitations of the small dataset used in the study. It is worth noting that these models' predictive power was tested using only abstracts of the studies, which frequently lacked all the information necessary to determine the research approach. This shortfall was identified during the manual construction of the test dataset, as some criteria essential for GPT-4 GPTNLI's predictions were not always present in the abstracts. Despite these limitations, the model's performance reinforces its potential to extract valuable insights from limited text data.

While the test dataset's creation inevitably introduced some subjectivity, given it was crafted by a single researcher, this was mitigated by the precise definition of each class using explicit criteria. This approach reduces subjectivity and provides a more objective framework for classifying a study's characteristics, which can assist researchers during the screening process. By providing the number of objective criteria fulfilled for each class in a given study, the model further adds transparency and quantifiable metrics to the process, increasing reproducibility.

Therefore, although only the GPT-4 GPTNLI model demonstrated a statistically significant improvement over random guessing, both models have proven valuable for providing practical, quantifiable information during the screening process. This outcome underscores the feasibility and potential benefits of employing unsupervised text classification models in the literature review process, even when working with limited or abstract-only data.

***RQ2**. In terms of predicting the research approach of a study based solely on the abstract, which type of unsupervised text classification model delivers the most accurate results?*

Evaluating the performance of the Lbl2Vec and GPTNLI models provides several valuable insights. For the Lbl2Vec model, the results highlight the crucial role that keyword selection plays in achieving optimal performance. The model's performance varied substantially between the "Criteria" and "Criteria 2" keyword sets, despite these sets being generated from the same prompt. This discrepancy underscores that keyword generation is an essential factor in model performance, and a structured method such as self-learning keywords could potentially enhance this further.

In addition to keyword selection, the number of documents used for training (N_{DOCS}) also influenced the Lbl2Vec model's performance. An increase in N_{DOCS} corresponded to more stable results, suggesting that more extensive training sets lead to more consistent model performance. A consistent model is critical in practice, as researchers need to be able to rely on stable results.

However, the GPT-4 GPTNLI model outperformed all other configurations in terms of predicting the research approach based solely on the abstract of a study. Nonetheless, this superior performance comes at a higher cost, as the GPT-4 GPTNLI model is more expensive than others. As of writing, running the GPT-3.5 model is less expensive (4% of the price), and the Lbl2Vec model is free to run on a machine with an average GPU.

Furthermore, it is critical to recognize further limitations of the results. The GPT models, while performing superior, operate as "black boxes", and it is not allowed to fine-tune them further.
While GPT4 can be prompted to explain their answers, it is essential to note that this process does not always represent internal mechanisms. It is not a given fact that the explanation aligns with the actual internal reasoning of the model.

Given these considerations, an extensive cost-benefit analysis is recommended to determine the most suitable model based on cost and performance. The chosen model should balance affordability with predictive accuracy to ensure the efficient classification of research approaches during the literature review screening process.

Nevertheless, when costs are disregarded, the GPT-4 GPTNLI model exhibits the best and most stable performance. Hence, if cost is not a concern, the GPT-4 GPTNLI model is the most suitable option for enhancing the efficiency of the screening process, mainly when relying solely on the abstract of a study.

6 CONCLUSION

This study addresses the problem of enhancing the efficiency of the literature review screening process by utilizing unsupervised text classification models to predict the research approach of a study based solely on the abstract. The findings demonstrate that such models can significantly improve the screening process, providing valuable insights and reducing the workload for researchers. Specifically, the GPT-4 GPTNLI model showed statistically significant performance, surpassing random guessing and offering valuable insights for screening, although at a higher cost. The research provided valuable insights despite the study basing its conclusions on limited sample size, the inherent opacity of the GPNTNLI models typically seen in deep learning, and the need for a comprehensive cost-benefit analysis. Using only the abstracts for the analysis and with existing constraints, the study successfully illustrated that unsupervised text classification models could be employed in enhancing the efficiency of the literature review screening process when appropriately configured and applied. Utilizing the methodology employed in this study reduces the subjectivity inherent in predicting the methodological approach and can facilitate the prediction of other characteristics. Future work should address the limitations mentioned, explore the use of self-learning techniques for keyword selection, and incorporate more sections of the study in the classification process.

REFERENCES

[1] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2021. *Introduction to meta-analysis*. John Wiley & Sons. http://www.jennifervonk.com/uploads/7/7/3/2/7732985/meta_analysis_fixed_vs_random_effects.pdf

[2] Kevin E. K. Chai, Robin L. J. Lines, Daniel F. Gucciardi, and Leo Ng. 2021. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews* 10 (2021). https://link.springer.com/article/10.1186/s13643-021-01635-3

[3] Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022. ContrastNet: A Contrastive Learning Framework for Few-Shot Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 10 (June 2022), 10492–10500. https://doi.org/10.1609/aaai.v36i10.21292

[4] Varun Dogra, Sahil Verma, Kavita, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. 2022. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience* 2022 (June 2022), 1–26. https://doi.org/10.1155/2022/1883698

[5] Jan Gabriel O. Ebora, James Christian N. Español, and Dionis A. Padilla. 2022. Text Classification of Facebook Messages Using Multiclass Support Vector Machine. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 1–6. https://doi.org/10.1109/ICCCNT54827.2022.9984554

[6] Luyi Feng, Yin Kia Chiam, and Sin Kuang Lo. 2017. Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review. In *2017 24th Asia-Pacific Software Engineering Conference (APSEC)*. 41–50. https://doi.org/10.1109/APSEC.2017.10

[7] Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Michael Paget, and Christopher Naugler. 2023. Automated Paper Screening for Clinical Reviews Using Large Language Models. *ArXiv* abs/2305.00844 (2023). https://doi.org/10.48550/arXiv.2305.00844

[8] Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-Aware Representation of Sentences for Generic Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.285

[9] Joshua D Harris, Carmen E Quatman, MM Manring, Robert A Siston, and David C Flanigan. 2014. How to write a systematic review. *The American journal of sports medicine* 42, 11 (2014), 2761–2768. https://doi.org/10.1177/0363546513497567

[10] Ammar Ismael Kadhim. 2019. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review* 52, 1 (Jan. 2019), 273–292. https://doi.org/10.1007/s10462-018-09677-1

[11] Sang-Woon Kim and Joon-Min Gil. 2019. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences* 9, 1 (Aug. 2019). https://doi.org/10.1186/s13673-019-0192-7

[12] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Yuzhong Chen, Xi Jiang, Wei Liu, Dajiang Zhu, Tianming Liu, Sheng Li, Xiang Li, and Hongmin Cai. 2023. Mask-guided BERT for Few Shot Text Classification. arXiv:2302.10447 [cs.CL] https://doi.org/10.48550/arXiv.2302.10447

[13] Lucas Gouveia Omena Lopes, Thales Vieira, and William M. Lira. 2021. Automatic evaluation of scientific abstracts through natural language processing. *ArXiv* abs/2112.01842 (2021). https://doi.org/10.48550/arXiv.2112.01842

[14] Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews* 8 (2019), 1–10. https://link.springer.com/article/10.1186/s13643-019-1074-9

[15] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)* 54, 3 (2021), 1–40. https://doi.org/10.1145/3439726

[16] T. Sakhi. 2023. GPTNLI implementation. https://github.com/TarekSakhi/unsupervised-nlp-research-project/blob/main/ResearchProjectGPTNLI.ipynb

[17] T. Sakhi. 2023. Keywords generation prompts. https://github.com/TarekSakhi/unsupervised-nlp-research-project/blob/main/prompts.md

[18] T. Sakhi. 2023. Lbl2vec implementation. https://github.com/TarekSakhi/unsupervised-nlp-research-approach/blob/main/ResearchProjectLbl2VecTransformer.ipynb

[19] T. Sakhi. 2023. Lbl2vec keywords. https://github.com/TarekSakhi/unsupervised-nlp-research-project/blob/main/lbl2vec_keywords.md

[20] T. Sakhi. 2023. Test dataset. https://github.com/TarekSakhi/unsupervised-nlp-research-project/blob/main/studies_test.csv

[21] T. Sakhi. 2023. Training dataset. https://github.com/TarekSakhi/unsupervised-nlp-research-project/blob/main/studies_training.csv

[22] Jan Sawicki, Maria Ganzha, and Marcin Paprzycki. 2023. The state of the art of Natural Language Processing - a systematic automated review of NLP literature using NLP techniques. *Data Intelligence* (2023). https://doi.org/10.1162/dint_a_00213

[23] Jonathan P. Scaccia and Victoria C. Scott. 2021. 5335 days of Implementation Science: using natural language processing to examine publication trends and topics. *Implementation Science* 16, 1 (April 2021). https://doi.org/10.1186/s13012-021-01120-4

[24] Tim Schopf, Daniel Braun, and Florian Matthes. 2021. Lbl2Vec: An Embedding-based Approach for Unsupervised Document Retrieval on Predefined Topics. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010710300003058

[25] Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. *arXiv preprint arXiv:2211.16285* (2022). https://doi.org/10.48550/arXiv.2211.16285

[26] Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Lbl2Vec: an embedding-based approach for unsupervised document retrieval on predefined topics. *arXiv preprint arXiv:2210.06023* (2022). https://doi.org/10.48550/arXiv.2211.16285

[27] Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research* 104 (2019), 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[28] Girish Sundaram and Daniel Berleant. 2022. Automating Systematic Literature Reviews with Natural Language Processing and Text Mining: a Systematic Literature Review. arXiv:2211.15397 [cs.IR] https://doi.org/10.48550/arXiv.2211.15397

[29] Parsa Sai Tejaswi, Saranam Venkata Amruth, Prakya Tummala, and M. Suneetha. 2022. Automatic Documents Categorization Using NLP. In *ICT Infrastructure and Computing*. Springer Nature Singapore, 215–225. https://doi.org/10.1007/978-981-19-5331-6_23

[30] Mike Thelwall and Pardeep Sud. 2022. Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies* 3, 1 (04 2022), 37–50. https://doi.org/10.1162/qss_a_00177 arXiv:https://direct.mit.edu/qss/article-pdf/3/1/37/2008360/qss_a_00177.pdf

[31] Jiajun Tong, Zhixiao Wang, and Xiaobin Rui. 2022. A multimodel-based deep learning framework for short text multiclass classification with the imbalanced and extremely small data set. *Computational Intelligence and Neuroscience* (Oct 2022). https://doi.org/10.1155/2022/7183207

[32] Dien Tran Thanh, Bui Loc, and Nguyen Thai-Nghe. 2019. Article Classification using Natural Language Processing and Machine Learning. 78–84. https://doi.org/10.1109/ACOMP.2019.00019

[33] Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence* 3, 2 (2021), 125–133. https://www.nature.com/articles/s42256-020-00287-7

[34] Robin Whittemore and Kathleen Knafl. 2005. The integrative review: updated methodology. *Journal of Advanced Nursing* 52, 5 (Dec. 2005), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

[35] Carrie Williams. 2011. Research Methods. *Journal of Business &amp Economics Research (JBER)* 5, 3 (Feb. 2011). https://doi.org/10.19030/jber.v5i3.2532

[36] Ortiz Yesenia and Segarra-Faggioni Veronica. 2021. Automatic Classification of Research Papers Using Machine Learning Approaches and Natural Language Processing. In *Information Technology and Systems*, Álvaro Rocha, Carlos Ferrás, Paulo Carlos López-López, and Teresa Guarda (Eds.). Springer International Publishing, Cham, 80–87. https://link.springer.com/chapter/10.1007/978-3-030-68285-9_8

[37] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. 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)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1404

[38] Fatima zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. *Journal of King Saud University - Computer and Information Sciences* 34, 10 (Nov. 2022), 8422–8428. https://doi.org/10.1016/j.jksuci.2021.02.005

[39] Fatima zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. *Journal of King Saud University - Computer and Information Sciences* 34, 10, Part A (2022), 8422–8428. https://doi.org/10.1016/j.jksuci.2021.02.005