Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Intelligent Reviewer Matching and Research Ethics Screening based on Rule-based and Multilabel Classification Algorithms (ReMatch+)

Mohd Norhisham Razali[1], Mohd Rafiz Salji[2], Norizuandi Ibrahim[3] , Rozita Hanapi[1] ,Haiqal Shazrin Anuar[1], Syazleena Yahya[4]

[1]*Faculty of Business and Management, Universiti Teknologi MARA Cawangan Sarawak, Kampus Samarahan, Sarawak, Malaysia*
[2]*Faculty of Information Science, Universiti Teknologi MARA Cawangan Sarawak, Kampus Samarahan, Sarawak, Malaysia*
[3]*Faculty of Computer Science and Mathematics, Universiti Teknologi MARA Cawangan Sarawak, Kampus Samarahan, Sarawak, Malaysia*
[4]*Institut Tanah dan Ukur Negara (INSTUN), Kementerian Sumber Asli dan Kelestarian Alam, Tanjung Malim, Perak, Malaysia*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Introduction: This paper proposed the combination of rule-based algorithms and machine learning techniques to address the issues of accuracy and efficiency of reviewer matching and ethical issue detection in research ethics approval processes.

Objectives: The model addresses these by optimizing two key processes: reviewer matching and ethical issue prediction. Three core experiments were conducted. First, various rule-based algorithms, including Keyword Matching, TF-IDF, BM25, LSA, and Bag-of-Words (BoW), were used to align reviewer expertise with research fields, with effectiveness evaluated based on matching scores and thresholds. Second, the system's performance in reviewer matching was validated using precision, recall, and F1 scores against ground truth data. Third, a multi-label classification approach was employed to train machine learning models to detect ethical issues such as Privacy and Confidentiality, Informed Consent, and Conflict of Interest.

Methods: Various classification techniques that combining TF-IDF and BoW with models like Support Vector Machines (SVM), Random Forests (RF), and Decision Trees (DT), were compared using metrics like subset accuracy and per-label accuracy.

Results: The results demonstrate the positive outcomes of using rule-based and machine learning approaches with TFIDF-SVM performs the best overall, achieving average per-label accuracy (0.89) and subset accuracy (0.38)

Conclusions: Future work could explore the inclusion of semantic-rich models, such as transformers, to further enhance the performance of both reviewer assignment and ethical issue detection.

Keywords: Ethical Issue Detection, Machine Learning, Multilabel classification.

INTRODUCTION

Ethical integrity is vital as research grows more complex. Institutions struggle to assign expert reviewers and identify ethical issues in proposals. Manual methods are slow, inconsistent, and difficult to scale amid rising research volume and diversity. Consequently, there is a pressing need for automated systems that can enhance the efficiency and reliability of research ethics approval processes [1], [2]. The literature reveals several gaps and challenges in the current methods of reviewer assignment and ethical issue detection in research. Reference [3] highlighted the limitations of relying on pre-trained language models (PLMs), which may fail to capture detailed domain-specific knowledge and biases. Reference [4] identified challenges in optimizing parameters and managing diverse research topics, particularly for conferences with broad themes. Reference [5] emphasized fairness and diversity issues in reviewer assignments, especially for interdisciplinary research, and noted the inefficiency of existing systems in managing the growing demand for reviewers. Furthermore, reference [6], [7] emphasized ethical challenges, including biases, errors in decision-making, and lower performance in certain fields like social sciences. Author [8]

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

identified the need for algorithms that balance workload, fairness, and strategy-proofness, while Reference [9] noted scalability and ethical concerns in the automation of peer review processes. Similarly, the Research Ethics Commission struggles with the slow, manual review of ethical protocols, which impacts research schedules [2]. Furthermore, although machine learning techniques such as Naïve Bayes+BoW have been explored for automating protocol reviews, challenges with precision and dataset size persist, highlighting a gap in the scalability and accuracy of current automation efforts. Hence, this paper proposed a technique that addresses these problems by integrating rule-based algorithms and machine learning techniques to ensure detailed and accurate reviewer matching while incorporating ethical issue detection. The system is designed to optimize two key processes: reviewer matching and ethical issue prediction. The integration of rule-based matching algorithms enables systematic alignment of reviewer expertise with research topics, while ML techniques facilitate the detection of multiple ethical issues across diverse domains. The study is structured around three core experiments. First, various rule-based algorithms including Keyword Matching, TF-IDF, BM25, LSA, and Bag-of-Words (BoW) were employed to assess the effectiveness of reviewer-research matching, with thresholds set to determine the quality of matches. Second, the system's performance in reviewer assignment was evaluated using metrics such as precision, recall, and F1 score, based on ground truth data. Third, a multi-label classification approach was employed to train ML models to detect eight ethical issues, such as Privacy and Confidentiality and Conflict of Interest. Six ML techniques combining TF-IDF and BoW with classifiers such as Support Vector Machines (SVM), Random Forests (RF), and Decision Trees (DT) were compared to identify the optimal approach for this task. This research aims to contribute to the growing body of literature on automating research management processes by demonstrating the potential of combining rule-based and ML approaches. By enhancing the efficiency and accuracy of reviewer assignments and ethical issue detection, the proposed system provides a scalable solution for institutional ethics review boards and research management units.

RELATED WORKS

Automating research ethics approval focuses on two key tasks: (1) reviewer assignment and (2) ethical issue detection. We review existing work on automated reviewer matching and ML applications for ethical evaluation.

Automatic Research Articles Reviewer Matching/Assignment

Reference [3] applied prompt-tuned PLMs to match papers/reviewers by research domain, using a greedy round algorithm with conflict checks. Their method improved interdisciplinary assignment quality and fairness (validated on public datasets), offering balanced coverage and low-resource adaptability. Limitations include PLMs' domain nuance gaps. Future work requires robust algorithms, bias mitigation, expanded datasets, and dynamic preference integration. Reference [4] automated conference paper assignments using TF-IDF with six ML algorithms (SVM, NB, LR, KNN, DT, RF). SVM achieved the highest F1-score (0.907), significantly reducing assignment time. While effective for narrowly-focused conferences, limitations include minimal gains from parameter tuning. Future work targets broader topic coverage, complementing approach in [3] to review process optimization. Reference [5] automates reviewer selection via citation network analysis, addressing peer review challenges from publication growth and interdisciplinary work. The system detects conflicts and aims to reduce bias. Future versions will integrate LLMs and improved algorithms to enhance fairness and interdisciplinary handling. Reference [8] formalized reviewer assignment as BFRAP, optimizing similarity scores under fairness/load constraints. Their FairColor algorithm outperformed baselines in speed/quality via individualized matching thresholds. Future work may expand to workload balancing, topic coverage, and strategy-proofness through novel graph formulations. Reference [6] assessed review score (RSP) and paper decision (PDP) prediction models, revealing performance drops (12% RSP, 23.31% PDP) for borderline cases and post-rebuttal updates. While useful, models exhibit high-confidence errors. Improvements require deeper analysis of borderline scores and review dynamics, with future work needed on rebuttal integration, reviewer confidence, and ethical implications. Reference [7] assessed ML models (Random Forest, XGBoost) for predicting REF2021 article quality using metadata from 84,966 articles. Performance varied by discipline (42% accuracy gain in STEM/economics vs. lower accuracy in humanities). Active learning improved precision but reduced coverage. While useful, limitations in metadata dependence and field-specific performance preclude replacing peer review.

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Reference [9] examines AI's impact across sectors, showing tools like GPT-3 and Dialogflow enhance productivity in business (customer service, healthcare) and academia (peer review). While transformative, challenges include privacy risks, bias, and scalability limits. The study advocates human-AI collaboration, calling for stronger ethical guidelines and domain-specific evaluations of emerging technologies."*

Ethical Issue Detection in Research

Ethical clearances are essential for research involving humans, animals, or the environment, but obtaining approvals is often slow due to manual processes and inefficient systems. Reference [10] propose an automated Ethical Clearance workflow system to streamline approvals, reduce delays, and improve efficiency. Testing at the University of the Western Cape showed promising results. Future efforts will refine the system and validate its effectiveness compared to manual methods, aiming to enhance research through ethical accountability and operational efficiency.

In Indonesia, the manual ethical review process by the Research Ethics Commission (KEP) struggles with increasing demand, causing delays. Reference [2] explored automating ethical protocol reviews using machine learning and deep learning. They found traditional methods like Naïve Bayes+BoW outperformed deep learning on small datasets, achieving precision, recall, and F-score values of 0.76, 0.80, and 0.78. NLP techniques improved classification accuracy but with some precision trade-offs. Automation speeds up reviews and aids decision-making without replacing reviewers. Future research will explore transfer learning to enhance performance and assess automation's broader impact. Reference [2], [10] demonstrate the potential of technology to improve ethical accountability and research efficiency.

METHODS

ReMatch+ FRAMEWORK AND EXPERIMENTAL SETUP

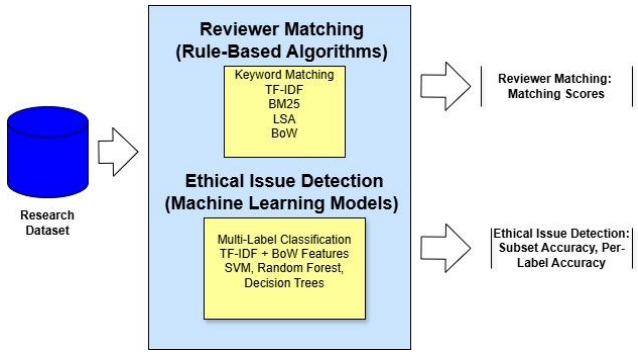

Figure 1. Framework of ReMatch+

Three set of experiments were conducted which are rule-based reviewer assignment, reviewer-research assignment performance evaluation, and ethical issue detection using machine learning. Each experiment is designed to address specific objectives and is implemented as follows:

Rule-Based Reviewer Matching

The first experiment focuses on aligning reviewer expertise with research fields using rule-based algorithms. Several studies have highlighted the effectiveness of rule-based approaches in automating reviewer assignments by leveraging text similarity techniques. References [11]–[16] provide empirical evidence supporting the use rule-based algorithms to align reviewer expertise with research topics. These studies demonstrate how structured matching methods improve the accuracy, fairness, and efficiency of reviewer selection while reducing manual workload. Additionally, they emphasize the importance of combining multiple similarity measures to ensure robust and contextually relevant reviewer assignments. The findings reinforce the role of rule-based systems in streamlining research ethics review processes and minimizing conflicts of interest.

The techniques used are keyword matching, keyword overlap, fuzzy match, TF-IDF (Cosine, Euclidean, Manhattan Distance, Hamming, Pearson Correlation), BM25 Match, Latent Semantic Analysis (LSA) and Bag-of-Words (BoW) (Cosine, Pearson Correlation).

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Keyword Matching

Keyword Matching [17] matches applications and reviewers based on the presence of exact keywords in both the research title and the reviewer's expertise. A match is made if at least one keyword appears in both. Let K_r be the set of keywords extracted from the research title and K_c be the set of keywords from the reviewer's expertise.

The Keyword Matching Score is defined as:

$$M = \begin{cases} 1, & if\ K_r \cap K_e \neq \theta \\ 0, & if\ K_r \cap K_e \neq \theta \end{cases}$$

Keyword Overlap

Keyword Overlap counts the overlap of keywords between the research title and reviewer expertise. The technique computes the number of shared keywords and uses this overlap as a measure of similarity. We define K_r as the set of keywords extracted from the research title and K_c as the set of keywords from the reviewer's expertise, where $|K|$ denotes the cardinality (number of elements) of a given set K. The Keyword Overlap Score is defined as:

$$O = \frac{|K_r \cap K_e|}{|K_r \cup K_e|}$$

We define $|K_r \cap K_e|$ as the count of shared keywords between the research title and reviewer expertise, and $|K_r \cup K_e|$ as the total number of unique keywords across both sets.

Fuzzy Matching

Fuzzy Matching [18] based on FuzzyWuzzy library was used for fuzzy string matching, which allows for approximate matching of research titles and reviewer expertise. This is especially useful for handling variations in spelling, abbreviations, or typographical errors. The Fuzzy Matching Score calculates the similarity between two strings S_1 (research title) and S_2 (reviewer expertise). Using a metric like Levenshtein Distance (edit distance), the score is normalized to provide a percentage similarity. The fuzzy matching score between strings S_1 and S_2 is calculated using their Levenshtein distance LD(S1,S2), which measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to make the strings identical, normalized by the length of the longer string $\max(|S1|,|S2|)$. The Fuzzy Matching Score is calculated as:

$$F = \left(1 - \frac{LD(S_1, S_2)}{Max\ (|S_1|, |S_2|)}\right)$$

The similarity score F represents the percentage match between two strings, where F=100% indicates identical strings and F=0% denotes no similarity.

TF-IDF-Cosine Similarity

TF-IDF-Cosine Similarity [19] calculates the cosine similarity between the TF-IDF vectors of the research title and the reviewer's field of expertise. Cosine similarity is a measure of the angle between two vectors, and a smaller angle indicates higher similarity.

TF-IDF

We define D as the set of all documents (including research titles and reviewer expertise), where $F_{t,d}$ represents the frequency of term t in document d, n denotes the total number of documents, and n_t indicates the number of documents containing term t. Let D represent the collection of all documents (containing both research titles and reviewer expertise profiles), where for any term t and document d, $F_{t,d}$ denotes the frequency of term t in document d. The total document count is given by n, while n_t specifies how many documents contain term t.

TF-IDF:

$$TF\text{-}IDF_{t,d} = TF_{t,d}\ X\ IDF_t$$

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Cosine Similarity

Let V_1 be the TF-IDF vector representing a research title. Let V_2 be the TF-IDF vector representing a reviewer's field of expertise.

The Cosine Similarity is given by:

$$\text{Cosine Similarity } (V_1,V_2) = \frac{V_1 . V_2}{\|V_1\| \|V_2\|}$$

The cosine similarity between TF-IDF vectors V_1 (research title) and V_2 (reviewer expertise) is calculated using their dot product ($\Sigma V1_i \cdot V2_i$) divided by the product of their magnitudes ($\sqrt{\Sigma V1_i^2} \times \sqrt{\Sigma V2_i^2}$). This produces a similarity score ranging from -1 (perfect opposites) to 1 (perfect match), with 0 indicating no similarity. A threshold T determines matches: when the score ≥ T, the system returns a match (1); otherwise, no match (0). This method effectively captures semantic alignment between research topics and reviewer expertise based on term significance within the document collection.

TF-IDF-Euclidean Distance

The TF-IDF-Euclidean Distance method [20] measures the similarity between research titles and reviewer expertise by calculating the Euclidean distance between their TF-IDF vectors ($V_1 = [V_{11}, V_{12},...,V_{1n}]$ for the research title and $V_2 = [V_{21}, V_{22},...,V_{2n}]$ for the reviewer's expertise). A smaller distance value indicates greater similarity between the two documents. The Euclidean Distance between the two TF-IDF vectors V_1 and V_2 is calculated as:

$$d(V_1, V_2) = \sqrt{\sum_{i=1}^{n} (V_{1i} - V_{21i})^2}$$

The Euclidean distance between TF-IDF vectors V_1 and V_2 is calculated using their components (V_{1_i} and V_{2_i}) across all N unique terms in the combined vocabulary. Smaller distances indicate greater similarity between the research title and reviewer's expertise. A match is determined by threshold T: when the distance $d(V_1,V_2) \leq T$, the system returns a match (1); otherwise, no match (0) is returned.

TF-IDF-Manhattan Distance

TF-IDF-Manhattan Distance [21] uses Manhattan distance, which sums the absolute differences between the TF-IDF values of corresponding terms. The Manhattan Distance between the two TF-IDF vectors V_1 and V_2 is calculated as:

$$D_{\text{Manhattan}}(V_1, V_2) = \sum_{i=1}^{n} |V_{1i} - V_{2i}|$$

The Manhattan distance between TF-IDF vectors $V1$ (research title) and $V2$ (reviewer expertise) is calculated as the sum of absolute differences $|V1i - V2i|$ across all N unique terms in their combined vocabulary, where $V1i$ and $V2i$ represent the TF-IDF components for each term. Smaller distances indicate greater similarity between the research title and reviewer expertise. A threshold T determines matches: if the Manhattan distance $d(V1,V2) \leq T$, the system returns a match (M=1); otherwise, no match (M=0) is returned. This approach provides an effective measure of similarity based on term importance in both documents.

TF-IDF-Hamming Distance

TF-IDF-Hamming Distance compares TF-IDF vectors by measuring the number of positions where the corresponding values V_1 and V_2 differ.

$$d_{Hamming}(v_1, v_2) = \sum_{i=1}^{n} \delta(v_{1i}, v_{2i})$$

The Hamming distance between vectors $v1$ (research title) and $v2$ (reviewer expertise) is calculated using the indicator function $\delta(v1i,v2i)$, which equals 1 if the ith components differ ($v_{1i} \neq v_{2i}$) and 0 if they match ($v_{1i}=v_{2i}$). The total distance is computed across all nn terms in the combined vocabulary. Smaller Hamming distances indicate

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

higher similarity, as they reflect fewer term mismatches. A threshold T determines matches: if dHamming(v1,v2)≤TdHamming($v1$,$v2$)≤T, the system returns a match (M=1); otherwise, no match (M=0) is returned. This method efficiently measures similarity by counting term-level discrepancies.

TF-IDF-Pearson Correlation

TF-IDF-Pearson Correlation calculates the Pearson correlation coefficient between the TF-IDF vectors to measure linear dependence between the research title and reviewer expertise. The Pearson Correlation Coefficient between v_1 and v_2 is calculated as:

$$r = \frac{\sum_{i=1}^{n}(v_{1i} - \bar{v}_1)(v_{2i} - \bar{v}_2)}{\sqrt{\sum_{i=1}^{n}(v_{1i} - \bar{v}_1)^2} \cdot \sqrt{\sum_{i=1}^{n}(v_{2i} - \bar{v}_2)^2}}$$

The Pearson correlation coefficient r measures the linear relationship between TF-IDF vectors v1$v1$ (research title) and $v2$ (reviewer expertise), where v_{1i} and v_{2i} are their respective components. The coefficient r ranges from −1–1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear relationship. Higher r values (closer to 1) reflect stronger similarity. A threshold T determines matches: if r≥T, the system returns a match (M=1); otherwise, no match (M=0) is returned.

BM25 (Best Matching 25)

BM25 (Best Matching 25) is a probabilistic information retrieval model that ranks documents (in this case, research titles) based on their relevance to a query (reviewer expertise). The model adjusts for term frequency and document length, making it a suitable approach for matching. The BM25 score between a research title and a reviewer's expertise is calculated as:

$$BM25(q,d) = \sum_{t \in q} IDF(t) \cdot \frac{f(t,d) \cdot (k_1 + 1)}{f(t,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})}$$

The BM25 ranking function measures the relevance between a research title (document dd) and a reviewer's expertise (query q), where t represents a term in q, f(t,d) is t's frequency in d, |d| is the document length, and $avgdl$ is the average document length across the dataset. The formula uses tunable parameters K1 (term frequency saturation, default 1.2) and b (length normalization, default 0.75). Higher BM25 scores indicate stronger relevance between the research title and reviewer expertise.

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a technique that uses dimensionality reduction to identify hidden relationships between words and documents. It uncovers patterns in the textual data, which may not be apparent through simple keyword matching. LSA uses Singular Value Decomposition (SVD) to reduce the dimensionality of the term-document matrix, identifying latent relationships between terms and documents. It represents both research titles and reviewer expertise in a lower-dimensional semantic space, revealing hidden patterns and relationships.

Let A∈R$^{m \times n}$ be the term-document matrix, where mm is the vocabulary size (number of unique terms) and n is the total number of documents (research titles and reviewer expertise combined), with each element a_{ij} representing the weight of term i in document j. Using Singular Value Decomposition (SVD), we factorize A into three matrices: A=$U\Sigma V^T$, where U∈R$^{m \times k}$ contains the left singular vectors, Σ∈R$^{k \times k}$ is a diagonal matrix of singular values, V∈R$^{n \times k}$ contains the right singular vectors, and k represents the number of latent semantic dimensions. This decomposition reduces dimensionality while preserving the underlying semantic relationships between terms and documents.

BOW-Cosine Similarity

BOW-Cosine Similarity represents text as a collection of words without considering the order. Cosine similarity is calculated between the BOW vectors of the research title and reviewer expertise. The Bag of Words model represents

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

text as a vector of term frequencies that ignore word order. For a vocabulary of m unique terms, the research title T is represented as a bag-of-words vector $t=[t_1,t_2,...,t_m]$, where each t_i denotes the frequency of term i in the title. Similarly, the reviewer expertise R is represented as vector $r=[r_1,r_2,...,r_m]$, where r_i indicates the frequency of term i in the reviewer's profile.

BOW-Pearson Correlation

The BOW-Pearson Correlation measures the linear relationship between term frequencies in bag-of-words vectors **t** (research title) and **r** (reviewer expertise) using Pearson's correlation coefficient ρ. The coefficient is calculated as $\rho(t,r) = Cov(t,r)/(\sigma_t\sigma_r)$, where $Cov(t,r) = (1/m)\sum(t_i-\bar{t})(r_i-\bar{r})$ represents the covariance between the vectors (with $\bar{t}$ and $\bar{r}$ being their mean values), and σ_t, σ_r denote their standard deviations. This method quantifies how strongly the term frequency patterns in research titles align with reviewer expertise on a scale from -1 (perfect inverse relationship) to 1 (perfect direct relationship), with 0 indicating no linear correlation.

Research proposals and reviewer profiles are preprocessed using tokenization, stemming, and stop-word removal. Text similarity scores are calculated for each technique. Then, thresholds are predefined to classify matches as good or poor, based on the similarity score.The effectiveness of each technique is measured using average matching scores and threshold percentages, indicating the quality of the matches.

The dataset used in this experiment consists of two main components which are Applications Dataset and Reviewers Dataset. Application dataset includes the research applications, each consisting of a unique Application ID, Research Title, and the specific Fields (such as the research domain or keywords) that the application is related to. Additionally, each application has predefined Ground Truth Reviewers (Reviewer 1, 2, and 3), where each reviewer is associated with a Reviewer Code and Reviewer Name. Reviewer dataset contains details of potential reviewers, each associated with a Reviewer Code, Reviewer Name, and Research Field or expertise, which represents the domains or topics the reviewer is qualified to review.

Reviewer-Research Assignation Performance

The second experiment evaluated the system's ability to assign appropriate reviewers to research proposals using techniques detailed in Section 3.1. We evaluated and compared several techniques for automatically assigning reviewers to research titles based on similarity measures, assessing their performance against ground truth reviewer assignments. This evaluation used two pre-processed datasets: (1) a Ground Truth Dataset containing actual reviewer assignments for a set of research titles, with each row indicating the number of titles assigned to a reviewer across three ground truth scenarios; and (2) a Keyword Assignment Dataset containing keyword-based reviewer assignment results, with each row showing the number of titles assigned to a reviewer by each evaluated technique. Pre-processing ensured consistent naming and handled missing values (filling missing keyword totals with zeros and renaming columns for merging). The datasets were then merged by reviewer ID to compare actual and keyword-based assignments across techniques. The generated assignments are compared with ground truth data to compute evaluation metrics. The metrics of performance are as follows:

Precision: Measures the accuracy of the assignments—how many of the assigned reviewers were correct.

Recall: Measures the coverage of the assignments—how many of the relevant reviewers were correctly identified.

F1 Score: Balances precision and recall into a single metric, useful when both are important.

Ethical Issue Detection Using Machine Learning

The third experiment trains machine learning models (e.g., SVM, Random Forests, Decision Trees) for multi-label classification of ethical issues in research proposals. Prior work [22]–[25] shows that such models, combined with text representations like TF-IDF and Bag-of-Words, can effectively detect ethical concerns—including privacy, informed consent, and conflicts of interest—by learning patterns in text. This automation improves the efficiency, consistency, and objectivity of ethics review processes. The research proposals were annotated for eight ethical issues: Privacy and Confidentiality, Informed Consent, Vulnerable Populations, Harm and Risk, Conflict of Interest, Bias

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and Objectivity, Beneficence and Non-Maleficence, and Intellectual Property and Plagiarism, with each proposal potentially containing multiple labels represented as binary variables (1 = present, 0 = absent). Using the proposal titles as input features, we implemented two text representation methods—TF-IDF, which weights terms by their importance across documents, and Bag-of-Words (BoW), which uses raw term frequencies. These features were then classified using multi-label algorithms (Support Vector Machines, Random Forests, and Decision Trees) wrapped in a MultiOutputClassifier framework to simultaneously predict all ethical categories. Data is split into training and testing sets using a standard 80-20 split. The study employed three key metrics to evaluate model performance: Subset Accuracy (exact match rate between predicted and true labels for each proposal), Per-Label Accuracy (prediction accuracy for individual ethical categories), and Average Per-Label Accuracy (mean accuracy across all ethical issues). Formally, Subset Accuracy was calculated as the ratio of fully correct predictions to total proposals, while Per-Label Accuracy measured correct predictions per ethical issue ($Accuracy_i$ = correct predictions for issue i / total proposals). The Average Per-Label Accuracy represented the arithmetic mean of all Per-Label Accuracies ($\sum Accuracy_i/n$), providing a comprehensive performance summary. These metrics collectively assessed both precise multi-label classification capability (Subset Accuracy) and granular category-wise performance (Per-Label metrics).

EXPERIMENTS AND RESULTS

Experimental Setup

This section presents the findings from the evaluation of various rule-based techniques used to match suitable reviewers with research based on their expertise. The techniques employed include Keyword Matching, Fuzzy Matching, TF-IDF based methods, BM25, LSA, and BOW models, each yielding normalized scores that reflect their effectiveness in the matching process.

Table 1 summarizes the normalized scores achieved by each technique, highlighting their relative performance.

Table 1 Performance of Rule-Based Techniques

Technique	Normalized Score
Keyword	0.285
Keyword Overlap	0.285
Fuzzy Match	0.484
TF-IDF Cosine Similarity	0.19
TF-IDF Euclidean Distance	0.327
TF-IDF Manhattan Distance	0.486
TF-IDF Hamming Distance	0.97
TF-IDF Pearson Correlation	0.59
BM25 Match	0.60
LSA	0.33
BOW Cosine Similarity	0.28
BOW Pearson Correlation	0.635

TF-IDF Hamming Distance achieved the highest score (0.970), demonstrating superior performance in binary keyword matching, while BM25 (0.600) proved effective for ranking reviewers by topic relevance. Both Pearson Correlation methods (BOW: 0.635; TF-IDF: 0.590) showed strong linear relationships in term usage patterns. Fuzzy Match (0.484) and TF-IDF Manhattan Distance (0.486) provided reliable performance for approximate matching, whereas TF-IDF Cosine Similarity (0.190) underperformed, indicating potential limitations in its vector representation approach. These results highlight that binary and ranking-based methods are most effective for reviewer matching, while cosine-based approaches may require optimization.

Matching Score Analysis

The performance of each technique was evaluated based on the established thresholds and average matching scores. Thresholds represent the minimum number of matches required for a proposal to be considered a good fit. Average matching scores indicate the overall quality of matches. For example, in the keyword-based technique, the threshold

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

was set at three keyword matches, meaning proposals with more than three matches were considered suitable. Average matching scores ranged from 0 to 6, with 0 being the worst and 6 the best.

Table 2. Thresholds and Average Matching Score

Technique	Threshold Descriptions	Average Matching Score Descriptions	Percentages above threshold	Average Matching Score
Keyword	Percentage of application with More Than 3 Keyword Matches	Number of keyword matches (0-6), where 0 indicates the worst match and 6 indicates the best match.	0.2116	1.71
Keyword Overlap	Percentage of application with More Than 3 Keyword Matches	Number of keyword matches (0-6), where 0 indicates the worst match and 6 indicates the best match	0.2116	1.71
Fuzzy Match	Percentage of application with Fuzzy Match score more than 50%	Fuzzy score percentages (0-100), where 0% indicates the worst match and 100% indicates the best match	0.195	0.4844
TF-IDF-cosine	Percentage of application with Cosine value more 0.3	Cosine Similarity value (0-1), where 0 indicates the worst match and 1 indicates the best match	0.1743	0.19
TF-IDF-euclidien	Percentage of application with Euclidean value below 6	Euclidean distance value (0-n), where the lower value indicates the best match.	0.3651	6.73
TF-IDF-Manhattan distance	Percentage of application with Manhattan value below 5	Manhattan distance value (0-n), where the lower value indicates the best match.	0.4149	5.14
TF-IDF-hamming	Percentage of application with Hamming value below 0.02	Hamming distance value (0-n), where the lower value indicates the best match.	0.249	0.03
TF-IDF-Pearson Correlation	Percentage of application with Pearson value above 0.3	Pearson correlation value (-1 to 1), where -1 indicates the worst match and 1 indicates the best match	0.1618	0.18
BM25 Match	Percentage of application with BM25 score above 6.0	BM25 value (0-n), where the higher value indicates the best match.	0.3734	6
LSA	Percentage of application with LSA score above 5.0	LSA value (0-n), where the higher value indicates the best match.	0.1618	0.33
BOW-cosine	Percentage of application with Cosine value more 0.3	Cosine Similarity value (0-1), where 0 indicates the worst match and 1 indicates the best match	0.3485	0.28
BOW-pearson	Percentage of application with Cosine value more 0.3	Pearson correlation value (-1 to 1), where -1 indicates the worst match and 1 indicates the best match	0.3278	0.27

Overall, the TF-IDF-Manhattan Distance technique holds the highest percentage of applications above the threshold (41.49%) and a solid average score (5.14). BM25 Match follows closely with 37.34% above the threshold and an

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

average score of 6. The BOW-Cosine and BOW-Pearson methods provide decent matches but with lower average scores.The Keyword and Keyword Overlap techniques, while showing a moderate percentage of applications above the threshold (21.16%), have lower average scores (1.71). Overall, TF-IDF-Manhattan Distance remains the best technique based on the highest percentage of strong matches, with BM25 Match as a strong alternative. The Keyword and Keyword Overlap techniques offer reasonable performance but may not be as effective as the top contenders.

Reviewer Assignation Performance Analysis

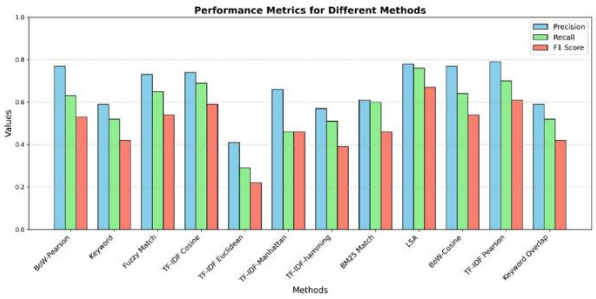

Figure 2. The Performance of Rule-based Algorithms

LSA emerged as the top-performing method, achieving the best balance with high recall (0.76), precision (0.78), and F1 score (0.67). TF-IDF Pearson followed closely with the highest precision (0.79) and strong recall (0.70), while TF-IDF Cosine maintained consistent performance across all metrics. Moderate performers like BoW-Pearson and BoW-Cosine showed good precision (0.77) but weaker recall, while Fuzzy Match delivered acceptable but unremarkable results. Poor performers included TF-IDF Euclidean (lowest scores: precision 0.41, recall 0.29) and basic keyword methods (F1 ≤ 0.42), which lacked sophistication for this task. TF-IDF Hamming and BM25 also underperformed (F1 ≤ 0.46). LSA, TF-IDF Pearson, and TF-IDF Cosine are the most reliable for reviewer assignments, whereas weaker methods require improvements to be viable.

Results of Ethical Issue Prediction

This study tests machine learning models (SVM, Random Forest, Decision Tree) for identifying ethical issues in research titles. Using TF-IDF and Bag-of-Words text representations, the system classifies titles into multiple ethical categories like Privacy and Informed Consent. Performance is measured through exact match accuracy and per-label accuracy, evaluating prediction quality for both single and multiple ethical concerns.

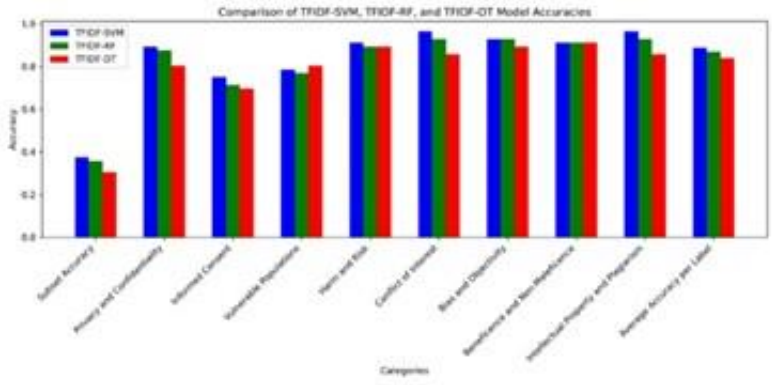

Figure 3 Performance of Ethical Issues Prediction based on TF-IDF Approaches

Journal of Information Systems Engineering and Management
2025, 10(54s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

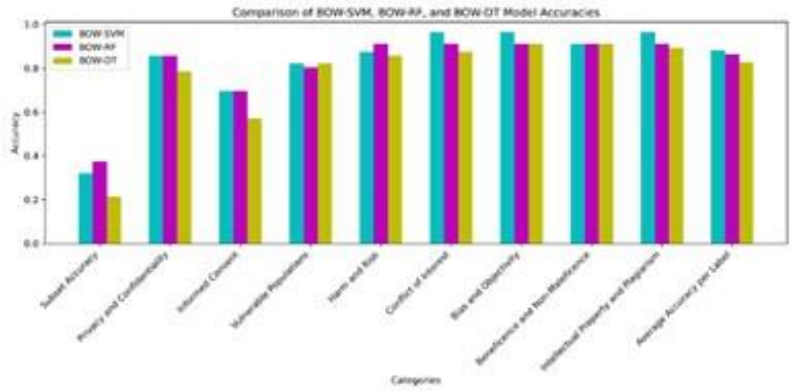

Figure 4 Performance of Ethical Issues Prediction based on BoW Approaches

TF-IDF with SVM performed best, with top accuracy (0.89 per-label, 0.38 subset) across most ethical categories. BOW with SVM was the next best option, especially for detecting bias. Random Forests worked better than Decision Trees, which had the weakest results. For real-world use, TF-IDF SVM is the top choice, with BOW SVM as a backup. With tuning, the other models might improve.

CONCUSION

This study demonstrates how combining rule-based algorithms and machine learning can effectively automate reviewer assignment and ethical issue detection in research proposals. The system's performance was validated through three experiments: rule-based techniques successfully matched reviewers to research topics, ML-based reviewer assignment achieved high accuracy, and TF-IDF-SVM emerged as the most effective model for detecting multiple ethical issues. The results show this integrated approach can significantly improve the efficiency and consistency of ethics review processes while maintaining accuracy. Future enhancements could incorporate advanced language models to further boost performance. This work provides research institutions with a scalable, automated solution for managing ethics reviews.

ACKNOWLEDGEMENT

This research was funded by Universiti Teknologi MARA Cawangan Sarawak under Dana Kecemerlangan Dalaman (DKCM) Grant No. 600-UiTMKS (RMU.5/2) (16/2023/KCMS) received in 2024.

REFRENCES

[1] A. C. Ribeiro, A. Sizo, and L. P. Reis, “Investigating the reviewer assignment problem: A systematic literature review,” *J. Inf. Sci.*, vol. 76, pp. 761–827, 2023, doi: 10.1177/01655515231176668.

[2] R. W. Sholikah, D. Purwitasari, and M. Z. Hamidi, “A Comparative Study of Multi-Label Classification for Document Labeling in Ethical Protocol Review,” *Techno.Com*, vol. 21, no. 2, pp. 211–223, 2022, doi: 10.33633/tc.v21i2.5994.

[3] L. Xu, D. Zeng, J. Dai, and L. Gui, “Combining Coverage with TMPS for Reviewer Assignment,” *IEIR 2022 - IEEE Int. Conf. Intell. Educ. Intell. Res.*, pp. 107–113, 2022, doi: 10.1109/IEIR56323.2022.10050060.

[4] S. C. H. Ngan, M. J. Lee, and K. C. Khor, “Automating Conference Paper Assignment Using Classification Algorithms Incorporated with TF-IDF Vectorisation,” *2023 IEEE Symp. Ind. Electron. Appl. ISIEA 2023*, pp. 1–6, 2023, doi: 10.1109/ISIEA58478.2023.10212219.

[5] D. Harvey *et al.*, “Prophy: An automated reviewer finder to improve the efficiency, diversity and quality of reviews,” *Inf. Serv. Use*, vol. 44, no. 1, pp. 37–42, 2023, doi: 10.3233/ISU-230196.

[6] G. L. Fernandes and P. O. S. Vaz-De-Melo, “Between acceptance and rejection: Challenges for an automatic peer review process,” *Proc. ACM/IEEE Jt. Conf. Digit. Libr.*, pp. 1–12, 2022, doi: 10.1145/3529372.3530935.

[7] M. Thelwall *et al.*, “Predicting article quality scores with machine learning: The U.K. Research Excellence Framework,” *Quant. Sci. Stud.*, vol. 4, no. 2, pp. 547–573, 2023, doi: 10.1162/qss_a_00258.

[8] K. Bouanane, A. N. Medakene, A. Benbelghit, and S. B. Belhaouari, “FairColor: An efficient algorithm for the Balanced and Fair Reviewer Assignment Problem,” *Inf. Process. Manag.*, vol. 61, no. 6, p. 103865, 2024, doi:

10.1016/j.ipm.2024.103865.

[9] L. Fiorillo and V. Mehta, "Accelerating editorial processes in scientific journals: Leveraging AI for rapid manuscript review," *Oral Oncol. Reports*, vol. 10, no. May, p. 100511, 2024, doi: 10.1016/j.oor.2024.100511.

[10] W. Mbabe, O. Ajayi, A. Bagula, L. Leenen, and N. Schoeman, "A workflow system for managing ethical clearance in research work," *2021 IST-Africa Conf. IST-Africa 2021*, pp. 1–9, 2021.

[11] Q. Xi and P. Jiang, "Design of news sentiment classification and recommendation system based on multi-model fusion and text similarity," *Int. J. Cogn. Comput. Eng.*, vol. 6, no. April 2024, pp. 44–54, 2025, doi: 10.1016/j.ijcce.2024.11.003.

[12] Jhih-Yi, Hsieh, A. Raghunathan, and N. B. Shah, "Vulnerability of Text-Matching in ML/AI Conference Reviewer Assignments to Collusions," pp. 1–30, 2024, [Online]. Available: http://arxiv.org/abs/2412.06606.

[13] M. Bani Saad, L. Jackowska-Strumillo, and W. Bieniecki, "Hybrid ANN-Based and Text Similarity Method for Automatic Short-Answer Grading in Polish," *Appl. Sci.*, vol. 15, no. 3, 2025, doi: 10.3390/app15031605.

[14] C. Papadimas, V. Ragazou, I. Karasavvidis, and V. Kollias, "Predicting learning performance using NLP: an exploratory study using two semantic textual similarity methods," *Knowl. Inf. Syst.*, 2025, doi: 10.1007/s10115-024-02293-2.

[15] L. Cagliero, P. Garza, A. Pasini, and E. Baralis, "Additional Reviewer Assignment by Means of Weighted Association Rules," *IEEE Trans. Emerg. Top. Comput.*, vol. 9, no. 1, pp. 329–341, 2021, doi: 10.1109/TETC.2018.2861214.

[16] M. Aksoy, S. Yanık, and M. F. Amasyali, *A comparative analysis of text representation, classification and clustering methods over real project proposals*, no. February. 2023.

[17] N. Gangoda and S. Thelijjagoda, "Resume Ranker : AI-Based Skill Analysis And Skill Matching System," *2024 Sixth Int. Conf. Intell. Comput. Data Sci.*, pp. 1–8, 2024, doi: 10.1109/ICDS62089.2024.10756304.

[18] C. Madhu and M. S. Sudhakar, "EmoDialect : Leveraging Fuzzy Matching and Dialect-Emotion Mapping for Sentiment Analysis," *IEEE Trans. Affect. Comput.*, vol. PP, pp. 1–17, 2024, doi: 10.1109/TAFFC.2024.3514862.

[19] K. Nadar, "Job Recommendation by Content Filtering using TF-IDF and Cosine Similarity," *2024 8th Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud)*, pp. 1585–1590, 2024, doi: 10.1109/I-SMAC61858.2024.10714763.

[20] A. Alsharef, "Exploring the Efficiency of Text-Similarity Measures in Automated Resume Screening for Recruitment," *2023 10th Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 36–42.

[21] B. S. Neysiani and S. M. Babamir, "New Methodology for Contextual Features Usage in Duplicate Bug Reports Detection," *2019 5th Int. Conf. Web Res.*, no. 4, pp. 178–183, 2019.

[22] X. Fang, X. Hu, Y. Hu, Y. Chen, S. Xie, and N. Han, "Fuzzy bifocal disambiguation for partial multi-label learning," *Neural Networks*, vol. 185, no. January, 2025, doi: 10.1016/j.neunet.2025.107137.

[23] T. Li, J. X. Jia, J. Y. Li, X. W. Xin, and J. C. Xu, "A novel random fast multi-label deep forest classification algorithm," *Neurocomputing*, vol. 615, no. September 2024, 2025, doi: 10.1016/j.neucom.2024.128903.

[24] P. N. Ahmad, A. M. Shah, K. Y. Lee, R. A. Naqvi, and W. Muhammad, "Optimizing slogan classification in ubiquitous learning environment: A hierarchical multilabel approach with fuzzy neural networks," *Knowledge-Based Syst.*, vol. 314, no. January, p. 113148, 2025, doi: 10.1016/j.knosys.2025.113148.

[25] M. I. Shukat, M. Usman, B. Fong, K. F. Tsang, and A. C. M. Fong, "Multilabel Learning for Massive Categorization of Resources in Metaverse Across SmartX Environments," *IEEE Consum. Electron. Mag.*, vol. PP, no. December 2024, pp. 1–11, 2025, doi: 10.1109/MCE.2025.3546050.