Journal of Scientometric Research, 2025; 14(1):331-341.
https://www.jscires.org

Research Article

Predicting Reviewers' Decisions in Scientific Submissions through Linguistic Analysis

Mayte H. Laureano, Hiram Calvo*, Tania Alcántara, Omar Garćıa-Vázquez, Marco A. Cardoso-Moreno

Computational Cognitive Sciences Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Av. JD Bátiz e/MO de Mendizábal s/n, Mexico City, GAM, MEXICO.

ABSTRACT

This research investigates the efficacy of various computational models and feature sets in the task of classifying scientific text reviews into distinct categories. Utilizing a combination of Word Space Models (WSM) and the Linguistic Inquiry and Word Count (LIWC) dictionary, the study endeavors to categorize reviews initially into five classes before simplifying the classification scheme into a binary system ('accept' and 'reject'). Despite the relatively straightforward nature of the employed feature sets, the binary classification approach demonstrated a notable improvement over a basic baseline that non-discriminatively assigns reviews to the most populous category. We obtain a recall of 0.758, compared with a baseline of 0.585 to the majority class and 0.62 and 0.66 of BERT and RoBERTa respectively. This performance can be considered significant given the diverse and subjective nature of the review content, contributed by 80 distinct individuals, each with their unique writing style and evaluative criteria. This work contributes to XAI through linguistic analysis revealing, for example that a minimal subset of features, specifically two out of the seventy provided by LIWC, can yield insightful distinctions in review classifications (0.649 recall). The analysis further identifies specific lexemes, such as 'not', 'since' and 'had', which offer deeper insights into the linguistic constructs employed by reviewers.

Keywords: Text Classification in Scholarly Information, Review Classification in Scientific Literature, Feature Selection for Scholarly Databases, Computational Linguistics for Scientific Information, Peer Review Analysis with AI.

Correspondence:
Hiram Calvo
Computational Cognitive Sciences Laboratory, Center for Computing Research, Instituto Politécnico Nacional, Av. JD Bátiz/MO de Mendizábal s/n, Mexico City, GAM, MEXICO.
Email: hcalvo@cic.ipn.mx
ORCID: 0000-0003-2836-2102

Received: 25-07-2024;
Revised: 15-10-2024;
Accepted: 19-11-2024.

INTRODUCTION

Predicting the decisions of reviewers in scientific submissions based on the textual content of their reviews presents a novel challenge in the domain of computational linguistics and data science. This task involves understanding the nuanced and subjective elements of review texts to anticipate the final verdict of acceptance or rejection. There is, however, a need for a comprehensive model that integrates advanced text analysis techniques with machine learning algorithms to predict the decisions of scientific reviewers. By focusing on the linguistic and semantic features inherent in review texts, we endeavor to explore the complex interplay of factors guiding reviewers' decisions, aiming to contribute to a more transparent, efficient and fair peer review process. Despite the inherent complexity due to various factors such as the multidimensional criteria used by reviewers, subjectivity, bias and the lack of standardized evaluation metrics, recent advances in machine learning and natural language processing have opened new avenues for exploration in this field.

Existing studies have laid the groundwork by focusing on aspects such as identifying the factors predicting the quality of peer reviews,[1] the influence of biases[2] and the exploration of methods to enhance the fairness and reliability of the peer review process.[3,4] However, these studies primarily concentrate on the broader aspects of peer review quality and the systemic features of the peer review process, rather than directly predicting reviewers' decisions from text reviews.

Recent efforts have shifted towards leveraging textual analysis and machine learning algorithms to predict the helpfulness of reviews, a proxy to understanding the underlying decision-making process of reviewers.[5,6] These approaches demonstrate the potential of computational methods in extracting valuable insights from the text, highlighting the feasibility of predicting reviewers' decisions in scientific submissions.

This paper investigates the application of computational models and Natural Language Processing (NLP) techniques to predict the outcomes of peer review based on the textual content of reviews. Utilizing Word Space Models and the Linguistic Inquiry and Word Count (LIWC) dictionary, the study effectively categorizes

OPEN ACCESS

Manuscript TECHNOMEDIA

DOI: 10.5530/jscires.20251456

Publishing Partner : Manuscript Technomedia.[www.mstechnomedia.com]

reviews into binary outcomes (accept or reject). The research showcases a significant improvement in predictive accuracy over simpler baseline models, demonstrating the potential for these tools to streamline editorial decision-making by quickly filtering out likely rejections.

Furthermore, the analysis provides deeper insights into the linguistic patterns that correlate with reviewer decisions, identifying specific words and phrases that are predictive of outcomes. This contributes to the broader fields of computational linguistics and Explainable AI (XAI) by offering methods to objectively analyze subjective content and by simplifying complex decision processes into understandable models. Such advancements not only aid in enhancing the efficiency of editorial workflows but also in training new reviewers and in guiding authors on how to better tailor their submissions to increase the likelihood of acceptance.

Related work

Predicting decisions from texts is an interdisciplinary field that merges machine learning and Natural Language Processing (NLP) techniques to analyze and predict outcomes from textual data. This area has seen significant applications across various domains such as judicial decisions, medicine and text analytics.[7-10]

In the judicial domain, researchers have explored the use of Support Vector Machines (SVMs) to predict the decisions of the European Court of Human Rights, demonstrating the potential of machine learning in understanding complex legal texts and outcomes.[7] Similarly, the work by Visentin *et al.*[8] introduced a new ensemble classifier that provides a statistically rigorous approach for predicting judicial decisions, emphasizing the importance of combining multiple models for enhanced prediction accuracy.

The integration and classification of legal documents have also been explored, with studies focusing on specific court cases, such as those of the French Supreme Court[9] and the US.[10] This indicates a growing interest in applying text analysis techniques to legal documents for predictive purposes.

The study "Does Reviewer Training Improve the Quality of Peer Review? A Randomized Controlled Trial"[11] examines if training enhances peer review quality in biomedical research. Through a randomized controlled trial, it found that training interventions, including workshops and feedback, did not significantly improve review quality as judged by editors and authors but increased reviewers' confidence. This suggests formal training boosts confidence but has a limited effect on the actual quality of peer reviews.

A study by Li *et al.*[12] investigated the impact of reviewers' words on their helpfulness ratings, developing a novel prediction model to address this issue. Similarly, Kavousi and Saadatmand[13] analyzed review texts to evaluate and predict ratings, focusing on the influence of lexical and sentimental features on prediction accuracy. These studies highlight the potential of text analysis in predicting reviewer helpfulness, setting a foundation for further exploration in predicting reviewer decisions for scientific work submissions.

Historically, the focus has been on improving the efficiency and fairness of the peer review process, with less attention given to understanding the predictive elements of reviewer decisions based on textual analysis. Studies in related fields have demonstrated the feasibility of using machine learning and Natural Language Processing (NLP) techniques to glean insights from text data, suggesting a promising avenue for enhancing the peer review process.[14]

The challenge, however, lies in the nuanced nature of scientific reviews. Unlike medical decisions, which are often based on historical data and well-established guidelines, reviewer decisions are influenced by a complex interplay of factors including the novelty, rigor and relevance of the work, as well as the reviewers' own expertise and biases. This complexity necessitates a sophisticated approach to predict reviewer decisions, one that can account for the varied and intricate nature of review texts.

This research aims to bridge the gap in the literature by exploring the potential of text analysis and machine learning models to predict the decisions of scientific reviewers. By analyzing the textual content of reviews, we seek to identify patterns and features that correlate with reviewer decisions, thereby offering insights that could streamline the review process and enhance the overall quality of scientific publications. In doing so, we contribute to the broader discourse on improving the transparency and efficacy of peer review, a critical mechanism in the advancement of scientific knowledge.

Preliminaries

The most commonly used machine learning techniques for analyzing text data and classifying it into decisions include:

Naive Bayes: A probabilistic classifier that uses Bayes' Theorem to calculate the probability of each tag for a given text, predicting the tag with the highest probability.[15,16]

Logistic Regression: A supervised learning method that uses logistic functions to predict the probability of a text belonging to a specific class.[17]

Random Forest: An ensemble learning technique that combines several decision trees to improve the accuracy of predictions.[17]

Deep Learning: A family of algorithms that use neural networks to learn representations of text data and make predictions.[18]

These algorithms are often used in conjunction with Natural Language Processing (NLP) techniques such as TF-IDF

vectorization, word embeddings and named entity recognition to transform text into a format that can be understood by machine learning models.[19]

LIWC Characteristics Analysis

To analyze the reviews, we employed the Linguistic Inquiry and Word Count (LIWC) tool,[20] categorizing words into psychologically meaningful groups. Table 1 is a summary of key LIWC characteristics.

Feature selection

The algorithm for feature reduction using subset selection, as detailed in,[21] was utilized to condense the sets of attributes. This supervised method searches for optimal subsets of attributes that yield effective classification performance. The process of subset search is conducted using the *BestFirst* strategy.

The *BestFirst* approach begins with an initial empty attribute set and sequentially incorporates attributes. It also marks *backtracking* checkpoints to revert to a previous state if the Addition of new attributes fails to enhance classification results. In our implementation, the algorithm was configured to consider adding up to five nodes before it backtracked to a previously saved point.

Experiments and results

For our experiments, we used review data from the MICAI 2023 conference, organized by the Mexican Society for Artificial Intelligence (SMIA). This conference is a prominent international gathering in the field of Artificial Intelligence (AI). Its significance lies in its focus on both theoretical and practical aspects of AI, providing a platform for researchers, practitioners and educators to present and discuss the latest advancements, trends and challenges in the field. The conference covers a broad range of AI topics, including machine learning, computer vision, robotics, natural language processing and AI applications in various domains.

The 2023 edition of MICAI received 117 submissions, from which 60 were accepted. Each work had an average of 2.17 reviews. There were 80 reviewers and each reviewer wrote an average of 3.1 reviews. In total, he has 248 reviews. Each review had the following structure:

Table 1: LIWC Characteristics Categorization.

Emotional Processes	Cognitive Processes
Affect (Pos/Neg Emotions)	Cognitive Mechanisms
Anger	Certainty
Anxiety	Causation
Sadness	Inhibition
Positive Emotions	Insight
Negative Emotions	Tentativeness
Social and Personal Concerns	**Language Style Markers**
Social Processes	Language Dimensions
Biological Processes	Time Orientation
Relativity	Pronouns
Personal Concerns	Verbs
Achievements and Goals	Sensory Processes
Cultural Concerns	Quantifiers

Table 2: Distribution of classes.

Classification	Count	Two classes
Accept	80	145
Weak Accept	65	-
Borderline	35	-
Weak Reject	35	-
Reject	33	103

- Q1 (Please write a short summary of the paper).
- Q2 (Describe this paper's strengths).
- Q3 (Describe areas to improve in this paper).
- Q4 (General comments for authors).
- Q5 (Please select your overall recommendation).

We concatenated all of these questions in a string and used it as the input to our classifier to predict one out of 5 possible outcomes (per reviewer). Table 2 shows the class distribution. See the Appendix for examples of reviews.

Now we aim to characterize the reviews attached to each one of the possible outcomes. First, we will analyze the possible classification into 5 classes. Then, we will reduce our number of classes to two, mapping Accept and Weak Accept to Accept (class 1) and Borderline, Weak Reject and Reject to Reject (class 0). Column 'Two classes' in Table 2 shows the distribution of classes when the 5 classes are reduced to 2.

Classification into 5 classes

Table 3 shows the results of classification into 5 classes (weak accept, accept, weak reject, borderline and reject) given the input of reviews, represented as a WSM vector of 2445 different words. All evaluations were performed using a 10-Fold Cross Validation. Results are only slightly above the baseline, which is assigning all results to the largest class (weak accept), resulting in a recall of 0.323; whereas the greatest recall is obtained by Naive Bayes, improving only 3.2% on the baseline-recall only, as other measures do not apply for the baseline. Note that *Borderline* seems difficult to characterize, as both Naive Bayes and Random Forest completely fail to classify its reviews.

Reducing the feature set, as described in Section "Feature selection" in "Preliminaries", we obtain 8 features that can be used to classify reviews into 5 classes. Selected features are: *2023, not, position, publication, since, aids, fails* and *had*.

Results are presented in Table 4. In terms of recall, there is an improvement compared to the baseline (0.323). It is important to note, however, that the *Weak Reject* and *Borderline* classes are predominantly unaddressed by all classifiers.

In order to see the effect of each of these features on each class, Table 5 shows the coefficients calculated by Logistic Regression for each class, while Figure 1 shows the distribution of words' values associated to each class. The word 'not', for example, is positively associated with class *reject*, while negatively associated with *accept*.

5-Class classification is possible; however, performance is rather low. That is why we decided to group 5 classes into 2: rejected (0) and accepted (1), as it was shown in Table 2. Next sections detail results obtained using two classes.

Table 3: Five-Class Classification using WSM features (2445).

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Baseline				**0.323**		**(Weak Accept)**
Naive Bayes	0.450	0.363	0.371	0.450	0.407	Weak Accept
	0.569	0.230	0.468	0.569	0.514	Accept
	0.086	0.099	0.125	0.086	0.102	Weak Reject
	0.000	0.070	0.000	0.000	0.000	Borderline
	0.364	0.098	0.364	0.364	0.364	Reject
Weighted Avg.	**0.355**	0.214	0.309	**0.355**	**0.329**	
Logistic Regression	0.650	0.643	0.325	0.650	0.433	Weak Accept
	0.277	0.137	0.419	0.277	0.333	Accept
	0.057	0.047	0.167	0.057	0.085	Weak Reject
	0.114	0.075	0.200	0.114	0.145	Borderline
	0.061	0.051	0.154	0.061	0.087	Reject
Weighted Avg.	0.315	0.267	0.287	0.315	0.271	
Random Forest	0.763	0.714	0.337	0.763	0.467	Weak Accept
	0.338	0.219	0.355	0.338	0.346	Accept
	0.029	0.005	0.500	0.029	0.054	Weak Reject
	0.000	0.005	0.000	0.000	0.000	Borderline
	0.061	0.000	1.000	0.061	0.114	Reject
Weighted Avg.	0.347	**0.289**	**0.405**	0.347	0.264	

Table 4: Five-Class Classification using reduced WSM features (8).

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Naive Bayes	0.613	0.506	0.366	0.613	0.458	Weak Accept
	0.692	0.322	0.433	0.692	0.533	Accept
	0.000	0.000	?	0.000	?	Weak Reject
	0.000	0.000	?	0.000	?	Borderline
	0.212	0.014	0.700	0.212	0.326	Reject
Weighted Avg.	0.407	0.250	?	0.407	?	
Logistic Regression	0.650	0.512	0.377	0.650	0.477	Weak Accept
	0.692	0.322	0.433	0.692	0.533	Accept
	0.000	0.000	?	0.000	?	Weak Reject
	0.000	0.000	?	0.000	?	Borderline
	0.182	0.000	1.000	0.182	0.308	Reject
Weighted Avg.	**0.415**	0.250	?	**0.415**	?	
Random Forest	0.588	0.512	0.353	0.588	0.441	Weak Accept
	0.692	0.322	0.433	0.692	0.533	Accept
	0.000	0.000	?	0.000	?	Weak Reject
	0.000	0.009	0.000	0.000	0.000	Borderline
	0.182	0.014	0.667	0.182	0.286	Reject
Weighted Avg.	0.395	0.253	?	0.395	?	
MLP	0.625	0.542	0.355	0.625	0.452	Weak Accept
	0.615	0.295	0.426	0.615	0.503	Accept
	0.000	0.019	0.000	0.000	0.000	Weak Reject
	0.000	0.005	0.000	0.000	0.000	Borderline
	0.182	0.009	0.750	0.182	0.293	Reject
Weighted Avg.	0.387	**0.257**	?	0.387	?	

Table 5: Logistic Regression coefficients for selected words for five-class classification.

	2023	not	position	publication	since	aids	fails	had
Accept		-0.89			-1.05	1.69		
Weak Accept	1.95		2.39		0.87			
Borderline				-0.93	0.59	1.05		
Weak Reject		0.42		-0.98		-0.94		
Reject		0.83					2.48	2.54

Table 6: Two-Class Classification Using WSM Features (1007).

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Naive Bayes	0.505	0.386	0.481	0.505	0.493	0
	0.614	0.495	0.636	0.614	0.625	1
Weighted Avg.	0.569	0.450	0.572	0.569	0.570	
Logistic Regression	0.350	0.152	0.621	0.350	0.447	0
	0.848	0.650	0.647	0.848	0.734	1
Weighted Avg.	**0.641**	0.443	**0.636**	**0.641**	**0.615**	
Random Forest	0.311	0.152	0.593	0.311	0.408	0
	0.848	0.689	0.634	0.848	0.726	1
Weighted Avg.	0.625	**0.466**	0.617	0.625	0.594	7

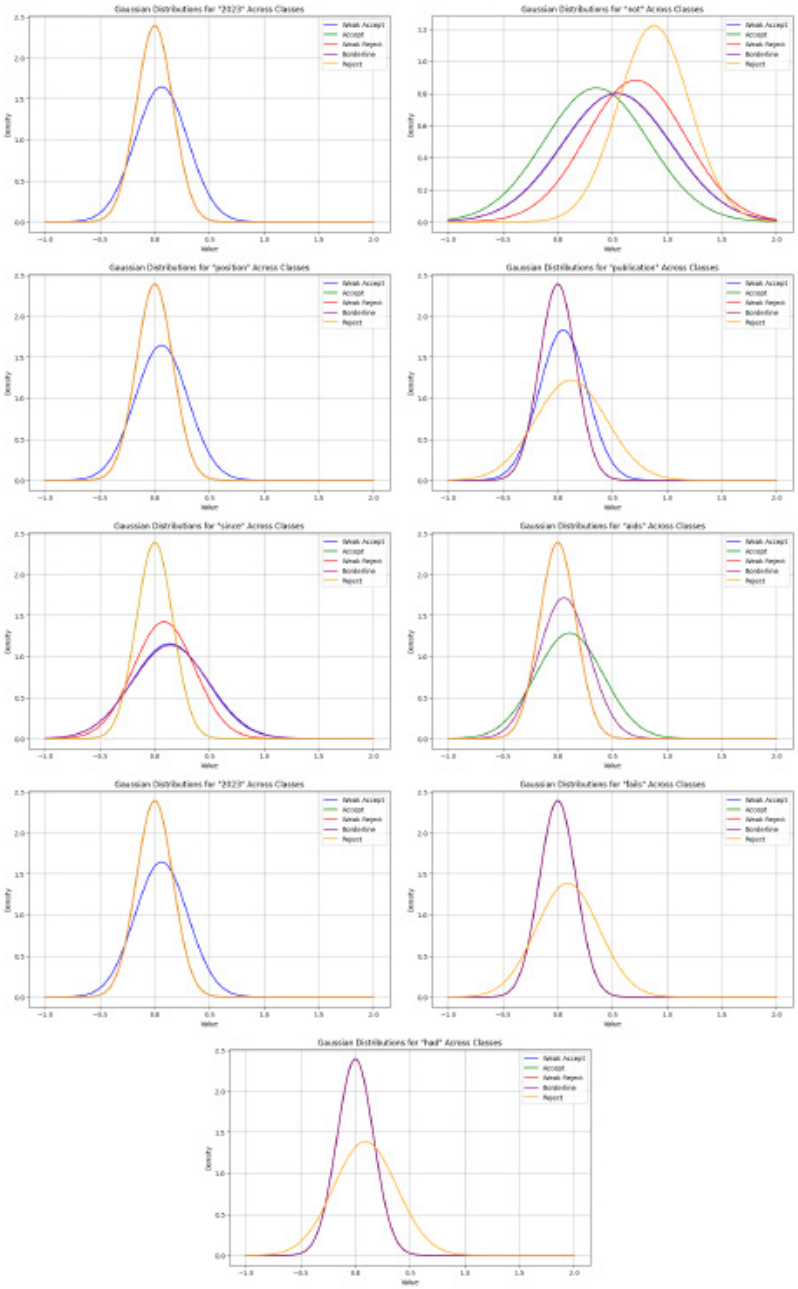

Figure 1: Feature analysis for WSM reduced set for 5-Class classification.

Two-Class Classification Using WSM Features

Table 6 shows result for the two-class classification using WSM features. A vocabulary of 1007 different tokens was used.

The algorithm for feature reduction using subset selection, as detailed in Section "Feature selection" of "Preliminaries", was utilized to condense the set of 1007 WSM attributes. Attributes were reduced to 25. Results of classification based on these attributes are shown in Table 7.

Two-Class Classification Using LIWC Features

As introduced in Section LIWC characteristics analysis, we used LIWC to characterize each one of the reviews. Table 8 shows a fragment of the input data to the classifier after the LIWC analysis.

Using the LIWC features alone, was not powerful alone to represent reviews for classification, as can be seen in Table 9. All tests were performed using 10-fold cross validation.

The algorithm for feature reduction described in Section "Feature selection" in "Preliminaries" Surprisingly, for this dataset, the 70 LIWC features were reduced to only 2:

Table 7: Two-Class Classification Using Reduced WSM Features (25).

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Naive Bayes	0.699	0.200	0.713	0.699	0.706	0
	0.800	0.301	0.789	0.800	0.795	1
Weighted Avg.	**0.758**	0.259	**0.757**	**0.758**	**0.758**	
Logistic Regression	0.650	0.193	0.705	0.650	0.677	0
	0.807	0.350	0.765	0.807	0.785	1
Weighted Avg.	0.742	0.285	0.740	0.742	0.740	
Random Forest	0.553	0.193	0.671	0.553	0.606	0
	0.807	0.447	0.718	0.807	0.760	1
Weighted Avg.	0.702	**0.341**	0.698	0.702	0.696	
MLP	0.573	0.193	0.678	0.573	0.621	0
	0.807	0.427	0.727	0.807	0.765	1
Weighted Avg.	0.710	0.330	0.707	0.710	0.705	

Table 8: Fragment of Reviews after LIWC Analysis.

Reviews' Text	Decision	Class	adverbs	articles	auxverbs	conjs	funct ...
The authors propose an *ex vivo* approach to transfer knowledge...	Weak Accept	1	0.01	0.06	0.02	0.02	0.24
The authors propose a two-step transfer learning approach...	Accept	1	0.01	0.05	0.02	0.02	0.22
This paper proposes a method to speed up the training of...	Weak Reject	0	0.03	0.05	0.06	0.01	0.26
The paper proposes combining SVM with a barycentric coordinate system...	Borderline	0	0.01	0.05	0.02	0.03	0.23
The article attempts to reconstruct some words of the extinct Khazar language...	Reject	0	0.01	0.06	0.05	0.02	0.25
The work presents a historical perspective of the Khazar people...	Borderline	0	0.02	0.05	0.03	0.03	0.23

Table 9: Two-class classification using LIWC features (70).

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Naive Bayes	0.330	0.228	0.507	0.330	0.400	0
	0.772	0.670	0.619	0.772	0.687	1
Weighted Avg.	0.589	0.486	0.573	0.589	0.568	
Random Forest	0.447	0.207	0.605	0.447	0.514	0
	0.793	0.553	0.669	0.793	0.726	1
Weighted Avg.	**0.649**	0.409	**0.642**	**0.649**	**0.638**	
Logistic Regression	0.262	0.166	0.529	0.262	0.351	0
	0.834	0.738	0.614	0.834	0.708	1
Weighted Avg.	0.597	**0.500**	0.579	0.597	0.559	
MLP	0.417	0.290	0.506	0.417	0.457	0
	0.710	0.583	0.632	0.710	0.669	1
Weighted Avg.	0.589	0.461	0.580	0.589	0.581	

Table 10: Two-class classification using the reduced set of LIWC features (2).

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Naive Bayes	0.262	0.090	0.675	0.262	0.378	0
	0.910	0.738	0.635	0.910	0.748	1
Weighted Avg.	0.641	**0.469**	0.651	0.641	0.594	
Random Forest	0.408	0.283	0.506	0.408	0.452	0
	0.717	0.592	0.630	0.717	0.671	1
Weighted Avg.	0.589	0.464	0.579	0.589	0.580	
Logistic Regression	0.311	0.110	0.667	0.311	0.424	0
	0.890	0.689	0.645	0.890	0.748	1
Weighted Avg.	**0.649**	0.449	**0.654**	**0.649**	0.613	
MLP	0.417	0.193	0.606	0.417	0.494	0
	0.807	0.583	0.661	0.807	0.727	1
Weighted Avg.	0.645	0.421	0.638	0.645	**0.630**	

Table 11: Selected WSM+LIWC features.

Exclusion words (LIWC)	does	many
explain	effectiveness	not
ML	further	problem
benefit	generating	proper
conditions	intelligence	qualitative
dependent	did	specific

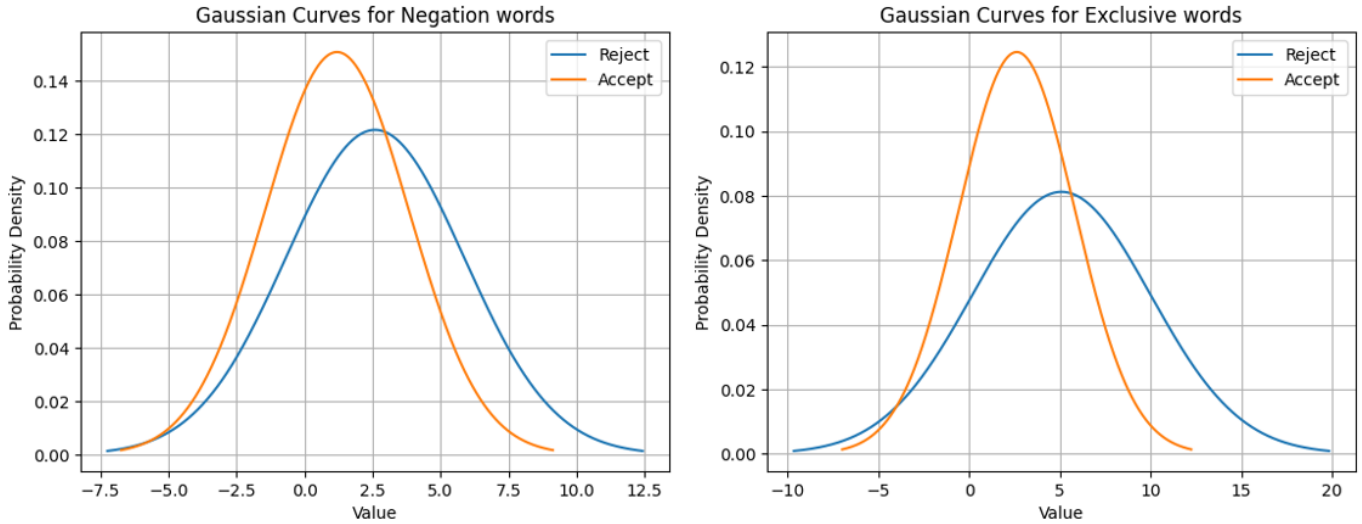

Figure 2: Gaussian curves for selected LIWC features to classify reviews.

- Negation words (no, not, never, ...) 57 in dictionary.
- Exclusive words (but, without, exclude, ...) 17 in dictionary.

With this information, obtained results are shown in Table 10.

The Logistic Regression coefficients for an accepting review were: -0.49+0.02 * [negation words]+0.07*[exclusive words].

Figure 2 shows the Gaussian curves corresponding to the selected features for the classes 0 (reject) and 1 (accept). It can be seen that fewer negation words and fewer exclusive words likely conform an accepting review.

Using WSM+LIWC features

Given the previous results, we applied the feature selection procedure described in Section "Feature selection" in "Preliminaries" to the 1007+70 features (WSM+LIWC), resulting in 18 features (Table 11).

Table 12: Two-class classification using reduced set of WSM+LIWC features (18).

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Naive Bayes	0.573	0.145	0.738	0.573	0.645	0
	0.855	0.427	0.738	0.855	0.792	1
Weighted Avg.	**0.738**	0.310	**0.738**	**0.738**	**0.731**	
Random Forest	0.583	0.262	0.612	0.583	0.597	0
	0.738	0.417	0.713	0.738	0.725	1
Weighted Avg.	0.673	**0.353**	0.671	0.673	0.672	
Logistic Regression	0.553	0.145	0.731	0.553	0.630	0
	0.855	0.447	0.729	0.855	0.787	1
Weighted Avg.	0.730	0.321	0.730	0.730	0.722	
MLP	0.515	0.145	0.716	0.515	0.599	0
	0.855	0.485	0.713	0.855	0.777	1
Weighted Avg.	0.714	0.344	0.714	0.714	0.703	

Table 13: Summary of best results for each feature set.

Classifier	Features	TP Rate	FP Rate	Precision	Recall	F-measure
Baseline	-	-	-	-	0.585	-
LR	WSM (1007)	0.641	0.443	0.636	0.641	0.615
NB	Reduced WSM (25)	**0.758**	0.259	**0.757**	**0.758**	**0.758**
RF	LIWC (70)	0.649	0.409	0.642	0.649	0.638
LR	Reduced LIWC (2)	0.649	**0.449**	0.654	0.649	0.613
NB	Reduced WSM+LIWC (18)	0.738	0.310	0.738	0.738	0.731
BERT	-	-	-	-	∼0.62	-
RoBERTa	-	-	-	-	∼0.66	-

It is interesting to note that only 1 LIWC feature was selected (Exclusion words), whereas the other words are specific words (from WSM). Results are shown in Table 12.

DISCUSSION

Table 13 summarizes the best results for the different feature sets and classifiers used to classify reviews in two classes. This might yield the idea that the best results are obtained by using the reduced set of 25 words; however, it is important to consider that these words were selected considering the whole dataset and thus, if new information is added, it is probable that there will be OOV (out of vocabulary) words. Perhaps a more stable solution would be using LIWC, as unseen words can be still be covered by the LIWC dictionary.

For comparison purposes, we have implemented BERT and RoBERTa classifiers as well, based on the string input of reviews and two output classes. BERT was tested with several epochs and, despite reaching 0.1753 losses, classification results are around 0.6 and 0.62, while RoBERTa reached values around 0.66. Please note that this is not a direct comparison, as the rest of classifiers were evaluated using a 10-fold cross-validation strategy, while BERT was evaluated against a randomly selected 20% of the dataset. However, we found consistently similar values after several runs (around 10).

CONCLUSION

This study systematically examines the efficacy of various classifiers and feature sets in the context of review classification, initially across five defined classes and subsequently within a binary framework. The investigation commenced with an analysis of a Word Significance Measure (WSM) vector encompassing 2445 distinct terms, culminating in performance metrics modestly surpassing a baseline established by assigning all outcomes to the most populous class, 'weak accept', which achieved a recall of 0.323. Notably, the Naive Bayes classifier demonstrated a slight improvement over this baseline, enhancing recall by 3.2%. This increment, albeit minimal, underscores the inherent difficulty in accurately classifying reviews, particularly within the 'Borderline' category, a task at which both the Naive Bayes and Random Forest classifiers were unsuccessful.

Further analysis entailed reducing the feature set to eight significant terms, resulting in a marginal increase in recall, thereby indicating the challenges in classifying 'Weak Reject' and 'Borderline' reviews remain substantial. This outcome suggests

that both extensive and significantly reduced feature sets may not capture the nuanced sentiment embedded within review texts.

Adopting a binary classification approach, which consolidates the initial five categories into 'accept' and 'reject', streamlines the classification process and aligns with the practical requirements of review analysis. The exploration of Linguistic Inquiry and Word Count (LIWC) features, though limited in isolation, revealed a nuanced dimension when combined with WSM attributes, leading to a refined set of 18 predictive features. This feature set, particularly emphasizing negation and exclusive words, provides a framework for dissecting review sentiments, as delineated by logistic regression coefficients and Gaussian distribution analyses.

Moreover, the comparison of traditional classifiers with advanced neural network-based models, specifically BERT and RoBERTa, highlights the advancements in text classification methodologies. Despite inherent variabilities in review datasets, these models demonstrated recall metrics approximately 0.62 for BERT and 0.66 for RoBERTa, indicating their robust- ness and adaptability. It is important to note, however, that direct comparisons between these advanced models and traditional classifiers are complicated by differences in evaluation methodologies. Despite the apparent simplicity of the utilized features, the classification outcomes, particularly within a binary framework, demonstrated a discernible improvement over the established baseline. This enhancement, however, was modest in the context of a five-class classification system, underlining the inherent challenges in employing such simplified methodologies for nuanced text analysis.

The endeavor to encapsulate the diverse perspectives and writing styles of 80 individuals underscores a significant challenge in review classification. The variability in individual criteria and modes of expression complicates the development of a universally applicable model. Nevertheless, achieving an improvement of 6.5% over the baseline represents a non-trivial advancement in this domain, suggesting that further refinement and exploration of these methods are warranted. Particularly, the reduction to a mere two features from the extensive suite provided by LIWC offers intriguing insights into the efficacy of focused linguistic features in review classification.

Furthermore, this research undertook an exploratory analysis within the realm of Explainable Artificial Intelligence (XAI), focusing on linguistic characteristics. This analysis illuminated a set of categories and specific lexemes that significantly contribute to the classification process. Notably, the words "not," "since," and "had" emerged as particularly salient, each offering unique insights into the reviewers' evaluative processes. For instance, the frequent use of "had" may indicate reviewers' emphasis on unmet requirements or actions necessary for acceptance.

These findings not only highlight the potential of simple, dictionary-based approaches for text classification but also underscore the complexity of capturing the multifaceted nature of human thought and expression. The modest yet significant improvements observed in this study advocate for continued investigation into linguistic feature-based models. Such efforts could further refine our understanding of textual analysis and enhance the accuracy and interpretability of review classification models, aligning with the broader objectives of XAI in providing transparent and understandable decision-making processes.

APPENDIX A

EXAMPLES OF REVIEWS

Example of an "Accept" review

The paper presents a case study with the intention of validate a conjecture called the minimum dominating set problem. The paper deal with a NP hard problem, this motivate the use of heuristics or something methodologies in order to decrease the computational effort and time. Even these results obtained by the use of a meta heuristic methods can be used in another problems and also in tacking theoretical challenges in very similar situations. * The approach used in this article is outstanding. It has a very good manner for show itself as a real interdisciplinary work, with a solid theoretical framework in the applied math. Moreover there are not errors that can be considered fatal or something like that. The description of the methodology it is considered good and the background mentioned into the article are worth. It is considered very well written and it has a good structure a way in order to presents. * There are two areas of opportunity, one hand is the need to highlight the fact that this is only a study case and at same time justify the number of vertices that was used in this case. In other hand, the explanation performed in the presentation of the results indicates that the method seems suitable in order to solve the problem with the presented scenarios. Finally in the section named back-ground it is suggested add some references in order to support the use of the genetic algorithms, this is only an accessory to the work, not because what is stated in paragraph 4 in this section be wrong or false, on the contrary, it demonstrates very good degree of expertise. All above mentioned in order to get better the contribution. * In general the article is very interesting as it has been above mentioned. Maybe the work could be include in the title the clarification about that this is a study case in order to reinforce or support the contribution, moreover this work as proposal is worth to taking it into account. Other aspect into the structure of the article is that in the section conclusions from paragraph 2 until the paragraph number four it is mentioned something like a discussion of the results but this can be implemented in the article as a new section or something that complements and argue the exact results in the Table 1. In the sub section 3.2 the first terms of the definition may be announced and put the letters in bold font, in order to emphasize the definitions or something like that. And this it could be improved in the whole section. By last, maybe the paper could

be oriented towards whatever application or project more large, that is, when orienting the work toward some practice field, over all in the artificial intelligence or related fields. Out of the above mentioned the work is a very good contribution in order to prove a conjecture in the discrete math field. *

Example of a "Reject" review

The authors propose the use of chatbots as an interaction interface for a glossary of terms in medical training is an effective tool to improve the understanding of specific and relevant concepts in health care and present an initial idea or proof of concept for realizing it. * The application of Chatbots in medical education is a very interesting and important area of research. The authors make clear the motivation for the proposed research. The motivation reflects a valid area of opportunity that aims to adapt technology with the current needs as well as medical services evolution. The state of the art makes a good work showcasing other works done in the field in a succinct but comprehensive manner.

* The abstract should describe the problematic and how the authors solve it, providing some concrete results, this is nowhere to be found. This article seems more like a work in progress as it presents a rough approach for developing the proposed software. There are no implementation details that allow the reader to understand the technical contribution nor details about how such a system could be used in the real world. In the same vein, the authors do not present a proper methodology for the gathering the requirements for such a system. The state of art does not cover how the current systems are implemented so that the proposal can differentiate and provide value. Moreover, as it stands now, the paper reads more appropriate for a software engineering conference, not an AI conference such as MICAI. * I would suggest the authors follow the Springer LNCS format and remove the abstract in Spanish. Even though there is an intention of contributing with the area, I recommend a more structured project, understanding the state of art and detecting a problem that can be covered by a scientific proposal that adds value to the current state. This paper reads like some kind of report on what the authors intend to make. For MICAI we expect more technical details and a contribution to the field of AI. Even if this is an application paper, the use of AI should be emphasized more. The authors are only using 6 of the allowed 14 pages limit and the paper barely scratches the surface. *

ACKNOWLEDGEMENT

The authors wish to thank the support of the Instituto Polite´cnico Nacional (COFAA, SIP- IPN, Grant SIP 20240610) and the Mexican Government (CONAHCyT, SNI).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Callaham M, McCulloch C. Longitudinal trends in the performance of scientific peer reviewers. Ann Emerg Med. 2011;57(2):141-8. doi: 10.1016/j.annemergmed.2010.07.027, PMID 21074894.
2. Eysenbach G. The impact of preprint servers and electronic publishing on biomedical research. Curr Opin Immunol. 2000;12(5):499-503. doi: 10.1016/s0952-7915(00)00127-8, PMID 11007350.
3. Hoang DT, Nguyen NT, Collins B, Hwang D. Decision support system for solving reviewer assignment problem. Cybernetics and Systems. 2021;52(5):379-97. doi: 10.1080/01969722.2020.1871227.
4. Available from: https://doi.org/10.1080/01969722.2020.1871227.
5. Recio-Saucedo A, Crane K, Meadmore K, Fackrell K, Church H, Fraser S, *et al.* What works for peer review and decision-making in research funding: a realist synthesis. Res Integr Peer Rev. 2022;7(1):2. doi: 10.1186/s41073-022-00120-2, PMID 35246264.
6. Li ST, Pham TT, Chuang HC. Do reviewers' words affect predicting their helpfulness ratings? locating helpful reviewers by linguistics styles. Inf Manag. 2019;56(1):28-38. doi: 10.1016/j.im.2018.06.002.
7. Teplitskiy M, Acuna D, Elamrani-Raoult A, Körding K, Evans J. The sociology of scientific validity: how professional networks shape judgement in peer review. Res Policy. 2018;47(9):1825-41. doi: 10.1016/j.respol.2018.06.014.
8. Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Comput Sci. 2016;2:93. doi: 10.7717/peerj-cs.93.
9. Visentin A, Nardotto A, O Sullivan B. Predicting judicial decisions: A statistically rigorous approach and a new ensemble classifier. In: 31st International Conference on Tools with Artificial Intelligence (ICTAI). Vol. 2019. IEEE PUBLICATIONS. IEEE PUBLICATIONS; 2019. p. 1820-4. doi: 10.1109/ICTAI.2019.00275.
10. Quemy A, Wrembel R. On integrating and classifying legal text documents. In: International Conference on Database and Expert Systems Applications; 2020. p. 385-99. doi: 10.1007/978-3-030-59003-1_25.
11. Springer.
12. Medvedeva M, Wieling M, Vols M. Rethinking the field of automatic prediction of court decisions. Artif Intell Law. 2023;31(1):195-212. doi: 10.1007/s10506-021-09306-3.
13. Callaham ML, Tercier J. The relationship of previous training and experience of journal peer reviewers to subsequent review quality. PLOS Med. 2007;4(1):e40. doi: 10.1371/journal.pmed.0040040, PMID 17411314.
14. Li ST, Pham TT, Chuang HC. Do reviewers' words affect predicting their helpfulness ratings? locating helpful reviewers by linguistics styles. Inf Manag. 2019;56(1):28-38. doi: 10.1016/j.im.2018.06.002.
15. Kavousi M, Saadatmand S. Estimating the rating of the reviews based on the text.
16. Data analytics and learning. In: Proceedings of the DAL. Vol. 2018. Springer; 2019. p. 257-67.
17. Baxt WG, Waeckerle JF, Berlin JA, Callaham ML. Who reviews the reviewers? feasibility of using a fictitious manuscript to evaluate peer reviewer performance. Ann Emerg Med, S.: Naive Bayes text classification. 1998;32(3):310-7. doi: 10.1016/s0196-0644(98)70006-x, PMID 9737492.
18. Raschka S. Naive bayes and text classification I – introduction and theory. arXiv preprint arXiv:1410.5329; 2014. Available from: https://arxiv.org/abs/1410.5329.
19. AI L. Text classifiers in machine learning: A practical guide. Text classification using naive Bayes. p. 7-text-classification-techniques-for-any-scenario. Available from: https://levity.ai/blog/text-classifiers-in-machine-learning-a-practical-guideMonkeyLearn. Available from: https://monkeylearn.com/text-classification-naive-bayes/Dataiku: 7 Text Classification Techniques for Any Scenario. Available from: https://blog.dataiku.com/.
20. Pennebaker JW, King LA. Linguistic styles: language use as an individual difference. J Pers Soc Psychol. 1999;77(6):1296-312. doi: 10.1037/0022-3514.77.6.1296, PMID 10626371.
21. Hall MA. Correlation-based feature subset selection for machine learning [PhD thesis]. Hamilton, New Zealand: University of Waikato; 1998.

Cite this article: Laureano MH, Calvo H, Alcántara T, Garćia-Vázquez O, Cardoso-Moreno MA. Predicting Reviewers' Decisions in Scientific Submissions through Linguistic Analysis. J Scientometric Res. 2025;14(1):331-41.