International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2) 2026 pp. 37-47

https://doi.org/10.70917/ijcisim-2026-2147
Article

Logistic Regression – Based Machine Learning Model for Predicting Student Awareness of Caste, Gender and Social Inequality: An Empirical Analysis of Arundhati Roy's writings

Shyla M.[1], Dr. J. Sheila[2]

[1] Research Scholar, Department of English, Noorul Islam Centre for Higher Education, Kanyakumari District, Tamil Nadu, India., shyla4894@outlook.com

[2] Assistant Professor, Department of English, Noorul Islam Centre for Higher Education, Kanyakumari District, Tamil Nadu, India.

Abstract: The caste and gender discrimination that permeates every area, socioeconomic class hinders the development of Indian educational systems. Gender inequality also caste inside academic institutions in India is a complicated and multifaceted reality that impacts all facets of lives, such as earnings, schooling and work opportunities, in addition to physical, societal, and financial challenges and culture. The multifaceted situation of gender and caste are common in Indian society. The analysis explores potential of intelligent system in shaping awareness out of caste, gender and social inequality using Arundhati Roy's writing. We Propose an AI driven Frame work that integrates natural language processing techniques to analyze Roy's works and identity themes related to social justice. The input data collected among students are preprocessed with lemmatization and TF-IDF vectorization. Features are then subjected to Correlation- based feature grouping to capture relevant patterns. A logistic regression classifier predicts the outcomes in two domains: Caste &Gender Inequality and Social Injustice in society. Comparison is made with Roy's writing to know the awareness and impact made by her among the students.
Keywords: NA

1. INTRODUCTION

In India, disparities within males and females in terms of ailment, ailment, politics, also the economy are described like gender and caste disparity. India has a different ranking on a composite basis and on each of these characteristics by a number of international gender and caste inequality indices, which are contentious. India's proportion of men to women, females' lifelong health, schooling attainment, also financial circumstances were every impacted due to male female disparity and its Socio foundations. The complex problem of gender and caste inequality in India impacts each of genders. A few argue some policies pertaining to caste and gender equality disadvantage men. However, women and caste are unfavorable in several important aspects when looking at India's population as a whole.

The persistence of inequalities in access to further schooling is probably the biggest issue facing India's higher education system. Disparities in availability of higher learning lead to go socioeconomic disparities within society, which exacerbate educational disparities. Actually, it's a cycle of Inequalities: discrepancies in higher schooling availability access lead to disparities information about the job market can result in employment disparities also workforce involvement, can result in incomes disparities. Take-home wages, which in turn fuels political and socioeconomic disparities. Once more, socioeconomic and political disparities are represented in the method of schooling, results in learning disparities. The disparity within genders caste is among the most important features of injustice. In India, as in many other nations, in every area, including higher schooling, females are frequently perceived as falling behind males, while recent trends within a number of countries have shown the opposite. Women's participation has significantly improved in the post-independence era. Losses in both society and individual welfare are reflected in disparities in educational opportunities. The fact that the economic returns on investments made in the learning of the less

fortunate segments are said to be greater compared to those of the equivalents suggests schooling disparities will cause the country's output to decline significantly.; additionally, wholesome tactics promote from the standpoints of Fairness in society and even economic prosperity, equity should be viewed positively. Since overall gains in the gains can exceed the defeats in efficiency.

The purpose of LR, a supervised ML approach, to predict categorical outcomes based on input features in classification tasks, such as binary and multiclass problems. It simulates connections between a group of independent variables and a dichotomous trait of interest. In order to get predicted probabilities that consistently lie between 0 and 1, the technique fits a logistic (sigmoid) operation to a linear blend of the input characteristics. This model may interpret the outcome as the likelihood that observation belongs to a specific class thanks to this mapping.

This Paper is about the intersection about caste gender and social inequality in educational domain using AI classification algorithm

2. LITERATURE REVIEW

One of the key fields of artificial intelligence is text processing. Text processing is a natural [1] language processing (NLP) aid in data analysis and information extraction. Numerous tools and techniques are used in text processing to assist obtain the needed information. This research article aims to give a general overview of text processing and natural language processing. It talks about the fundamentals of NLP and potential applications. It also gives a quick explanation of the algorithms that make up stemming and Lemmatization.

Lemmatization is the process of identifying a word's normalized form. It is the same as looking for a transformation to apply to a word to get its normalized form. In order to produce the normalized form, the strategy presented in this study focuses on word endings: which word suffixes should be [2] added or removed? This article compares the results of two word lemmatization algorithms, one based on induction techniques based on ripple down rules and the other based on if-then rules. It explains why lemmatizing phrases from Slovene free text is a perfect fit for the Ripple Down Rules (RDR) method. The RDR approach learns from a corpus of lemmatized Slovene words and generates straightforward rules. Compared to outcomes of rule learning attained in earlier work, the RDR approach produces simple rules with higher classification accuracy when studying from a corpus of Slovene words that have been lemmatized.

An essential component of IR is the sentence similarity task. IR will increase if this work is performed more effectively. Lemmatization and stemming [3] were used to aid accomplish this. Many people are still unsure of the optimal option for sentence similarity challenges, though. Thus, the purpose of this study is to ascertain whether preprocessing method—lemmatization or stemming—is most effective for sentence similarity tasks. Based on numerous research on the subject, the authors of this study would want to perform a SLR on stemming also lemmatization. Numerous factors influence the choice of preprocessing strategy, according to earlier research that attempted to evaluate and compare both approaches using a variety of evaluation tools.

Considering the broad a tool for lemmatizing English Adornment Morph, the tool concentrates on English inflectional morphology. Several known lexical resources are incorporated into the BioLemmatizer to better [4] customize it to the biological area. It outlines a collection of guidelines to convert a phrase to a lemma if it's not found inside the lexicon and uses a word lexicon to extract lemmas. The BioLemmatizer's use of a hierarchical lexicon search approach is a novel feature that allows the proper lemma to be found even in cases when the supplied a portion of speech data is erroneous. 96% of precision is attained by the BioLemmatizer.

The clustering of Finnish text texts was compared using stemming and lemmatization. We reasoned that lemmatization—which involves separating the compound words—would be a better normalizing strategy than simple stemming [5] because the language of Finnish is heavily inflective agglutinous. A relevance of four points evaluation scales were used to evaluate the documents' relevance. This scale was then compressed into a binary one, meaning that all relevant materials were deemed relevant, while only extremely pertinent documents were deemed pertinent. This notion was confirmed by experiments using four hierarchical clustering techniques.

Regression/classification performance has been thought to be enhanced by feature selection in conjunction with additional structural information about the features. In order to make Graph-guided fused lasso (GFlasso), which makes feature selection [6] and graph structure exploitation simpler when features exhibit particular graph structures, has recently been developed. Nevertheless, attribute grouping in GFlasso's approach depends on pairwise sample correlations, which may result in additional estimation bias. The suggested formulations can be solved using the Convex function differences (DC) interpreting and the ADMM. The efficiency of the suggested strategies is demonstrated by our experimental results on two genuine datasets and synthetic data.

By removing unnecessary and duplicated data, FS is an efficient method of lowering dimensionality. Majority conventional FS methods score and rank each feature separately before performing FS by either keeping highly-ranked features or removing lower-ranked features. This review address an innovative approach of FS [7] starts with feature grouping and then scores feature groups instead of individual features. As far as we this is the initial research that are aware concentrates on grouping-based FS approaches, despite existence of studies on clustering and FS algorithms. Generating groupings of comparable features with differences across groups and choosing representative features from each cluster is the general notion behind FS via grouping.

Dense feature groups are identified by the framework using density estimation, and for feature selection, the features of each dense group are handled as a coherent whole. DRAGS (Dense Relevant Attribute Group Selector) an efficient algorithm developed inside this framework. Additionally, we present a universal metric for evaluating feature selection [8] algorithm stability. Our empirical study based on microarray data shows that dense feature groups are stable under random sample holdout, and the DRAGS method successfully finds a collection of feature groups that exhibit both high classification accuracy and stability. Dense feature groups are identified by the system using kernel density estimation, and the features of each dense group are treated as a cohesive unit for feature selection. Additionally, we present a universal criterion for evaluating feature selection algorithm stability. Dense feature groups are stable under random sample holdout, according to our empirical analysis based on microarray data, and the DRAGS technique identifies a collection of feature groups that demonstrate both high classification accuracy and stability.

During the learning process, it is frequently beneficial to group comparable features together for high-dimensional data. Better generalization may result from lowering the estimation [9] variance and enhancing feature selection stability. Additionally, it can aid with data interpretation and comprehension. A novel sparse-modeling technique called OSCAR technique for reversal encourages such feature grouping by using a regularizer and a regularizer on the feature coefficients. Nevertheless, its optimization process is highly costly computationally.

The process involves identifying similar local picture patterns in two photographs. Both the left and right photos are used to extract linear edge segments. Each segment's orientation and position within the image, along with its connections to other segments, define it. Thus, each image is used to create a relational graph. A group of nodes in a communication chart represents a set of possible assignments for each segment in a picture. Compatible assignments [10] based on segment relationships are represented by arcs in the graph. Finding groups of nodes in this graph that are compatible with one another becomes the same as stereo matching.

The majority of feature selection methods used today concentrate on gradually adding or removing individual characteristics in relation to the candidate feature subset or subsets. When such methods are used, information like cooperative contribution between traits could be lost because just individual inclusion or exclusion [11] of features is taken into account. As a result, if the important information included in the initial feature set can still be preserved, the last chosen there could be a lot of redundancy across features in attribute set. Nowadays, most feature selection techniques focus on progressively adding or eliminating certain attributes with regard to the subset of potential attributes or subsets. Due to the fact that only each feature's inclusion or removal is taken into account when using such methods, Information could be lost, such as the association between traits or the collaborative input. Therefore, if the essential information in the original feature set can still be preserved, the final chosen feature subset may have high amounts of inter-feature redundancy.

As a misuse-based intrusion detection system, three-layer RNN structures using classified characteristics [12] source of attacks type of RNN outcome is suggested. The input features are divided into four categories: host-based, time-based, content-based, and fundamental traffic features. The attack techniques are classified as Denial-of-Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R). This study uses 41 features per connection from the International Knowledge Discovery and Data Mining Group (KDD) for this purpose. The usual class (no attack) corresponds to an extra output from the RNN. The nodes of two RNN hidden layers are partially connected to each other.

Based on their inherent qualities, high-dimensional data is divided into feature groups based on its characteristics. To concurrently determine the significance [13] two types of weights are incorporated into the clustering process for both feature groups and individual features in each cluster. The optimization process is described by a new optimization model, and a novel clustering method known as FG-k-means is used to optimize the optimization model. The novel method is an enhancement of k-means by adding two additional steps to automatically compute the two types of subspace weights. A new data generation method is presented to generate high-dimensional data with clusters in subspaces of both attribute group and individual attribute.

Since approximate algorithms yield good solutions in a reasonable amount of time, they are commonly used. However, feature grouping has emerged as a potent method for lowering dimensionality in high-

dimensional data. In order to enhance the model, some writers have recently concentrated on creating techniques that integrate attribute Sorting and choosing. In order to enhance search efficiency, we present a feature selection technique [14] in this study that makes use of feature grouping. We suggest the VNS metaheuristic as feature selection method. As far as we are aware it is the initial instance of attribute grouping using the Markov blanket. Through studies on many datasets with high dimensions from different fields, Mining text and tiny array - evaluate the effectiveness of VNS. We contrast VNS with competitive and widely used methods.

Based on their intrinsic characteristics, high-dimensional data features are divided into feature groups. To concurrently determine the significance two types of weights are added to the clustering process for both individual features and feature groups within each cluster. The optimization procedure [15] is described by a new optimization technique also the optimization model is optimized by a novel clustering technique called FG-k-means. In order to instinctively determine the new technique extends k-means by two stages for the two types of subspace weights. The novel method is an enhancement of k-means by adding two additional steps to automatically compute the two types of subspace weights. A new data generation method is presented to generate high-dimensional data with clusters in subspaces of both feature groups and individual attribute.

Intelligent transportation systems are one of the many application areas for tracking and detection of objects. We introduce an object recognition and tracking method that combines the feature tracking and grouping technique with the background subtraction strategy. First, we introduce an enhanced background subtraction technique that improves the feature detection [16] and grouping outcome by using a low-level feature tracking as a trigger. Next, we introduce a dynamic multi-level feature grouping method that yields high-quality trajectories and can be applied in real-time. Lastly, we showcase experimental findings from video segments of a challenging transportation application.

Machine learning (ML) algorithms may perform worse in classification when irrelevant features are present. A technique for selecting a subset [17] pertinent attribute that accurately describe dataset is called feature selection (FS). In this field, evolutionary algorithms (EAs) are often employed search techniques. For optimization problems, cooperative co-evolution (CC), an EA variation that employs dividing and conquering strategy, is a wise choice. Due to a variety of restrictions, like not taking feature interactions into account, handling an even number of attributes, also statically partitioning the data set, the current solutions perform poorly. In order to guarantee the likelihood of combining elements that interact into a single subcomponent and to dynamically deconstruct Big Data datasets, this work introduces a novel RFG also its 3 versions. It is known as Collaborative co-evolutionary-based attribute choosing with random feature grouping since it may utilized within CC-based FS procedures. Six popular machine learning classifiers were employed for experiment analysis on seven distinct datasets from the Princeton University library also UCI ML archive, both with and without FS.

Owing to the significance and growing intricacy of these kinds of challenges, techniques that yield more accurate and comprehensible outcomes are required. Boosting is one of these techniques; it works progressively through the application of a classification methodology [18] on sample data sets that have been recalculated. As previously shown, boosting can also be viewed as a functional estimating strategy. In order to identify which model had the best fit and capacity for discrimination when a specific property was present or absent, the current study compared the logistic regression models estimated by the maximum likelihood model (LRMML) also the LR model estimated using the Boosting algorithm, especially the LRMBB.

Logistic regression and neural network classification methods are taken into consideration while identifying seismic recordings in seismically active mines. These methodologies are applied to the classification of seismic records using an effective methodology. An examination [19] the recipient functioning traits curl and many execution measures associated with the contingency matrix are used to compare the suggested method accompanying seismic activity from two Ontario mines. Overall classification accuracy was similar for neural system and LR techniques. Measure of classification quality was the models' capacity to replicate the evaluating data set regularity - amount of dissemination. This reference dissemination was satisfactorily replicated by the logistic and neural network models.

Logistic regression and classification tree (CT) analysis can be used to develop classification models forecasting if a chemical will belong to two or more pre-established groupings, like toxicological categories. One distinction between 3 approaches is that, while CT analysis is a not using a parametric method, discrimination evaluation and LR [20] need to make certain presumptions regarding the underlying data. Another distinction, that CT evaluation alone generates average probability of various collectives, whereas discrimination evaluation and LR can be used to determine group membership likelihood of specific substances. A comparison of the CMs created by applying the three approaches to a data collection on eye discomfort serves as an example of how they are applied.

Using basic benchmarks such as evaluating the whole loss function and incorrect classification metrics, choosing a learning algorithm to use for a specific application based on performance is still an ad hoc procedure. In this work, we tackle the difficulty of choosing a model by contrasting logistic regression [21] and random forest's overall effectiveness of general classification of data collection with different underlying structures: adding more noise variables; adding more explanatory variables; adding more observations; and raising the noise and explanatory factors' variance.

We developed a model evaluation tool that can generate classifier models for certain dataset attributes and performance metrics, such as true positive rate, false positive rate, and accuracy under specific conditions. We found that when the variance in the explanatory and noise parameters grew, LR consistently performed better overall than random forest.

Useful technique for lowering dimensionality, eliminating unnecessary data, and improving learning accuracy is attribute choice. The effectiveness also efficacy of numerous attribute choice techniques are seriously hindered by the plague dimensions of the data.

Three feature selection [22] techniques are used in this paper quick correlation depend attribute selection in Pieces (FCFBiP), quick correlation based Feature Selection (FCBF), also speed correlation depend attribute choice (FCBF). When the three feature selections are compared, experimental results show that the FCFBiP is more efficient than FCBF and FCB.

3. METHODOLOGY

This study used a descriptive and analytical research design shown in Fig.1 based on an machine learning classification to investigate how Arundhati Roy's literary works impacts students' critical consciousness and social awareness.

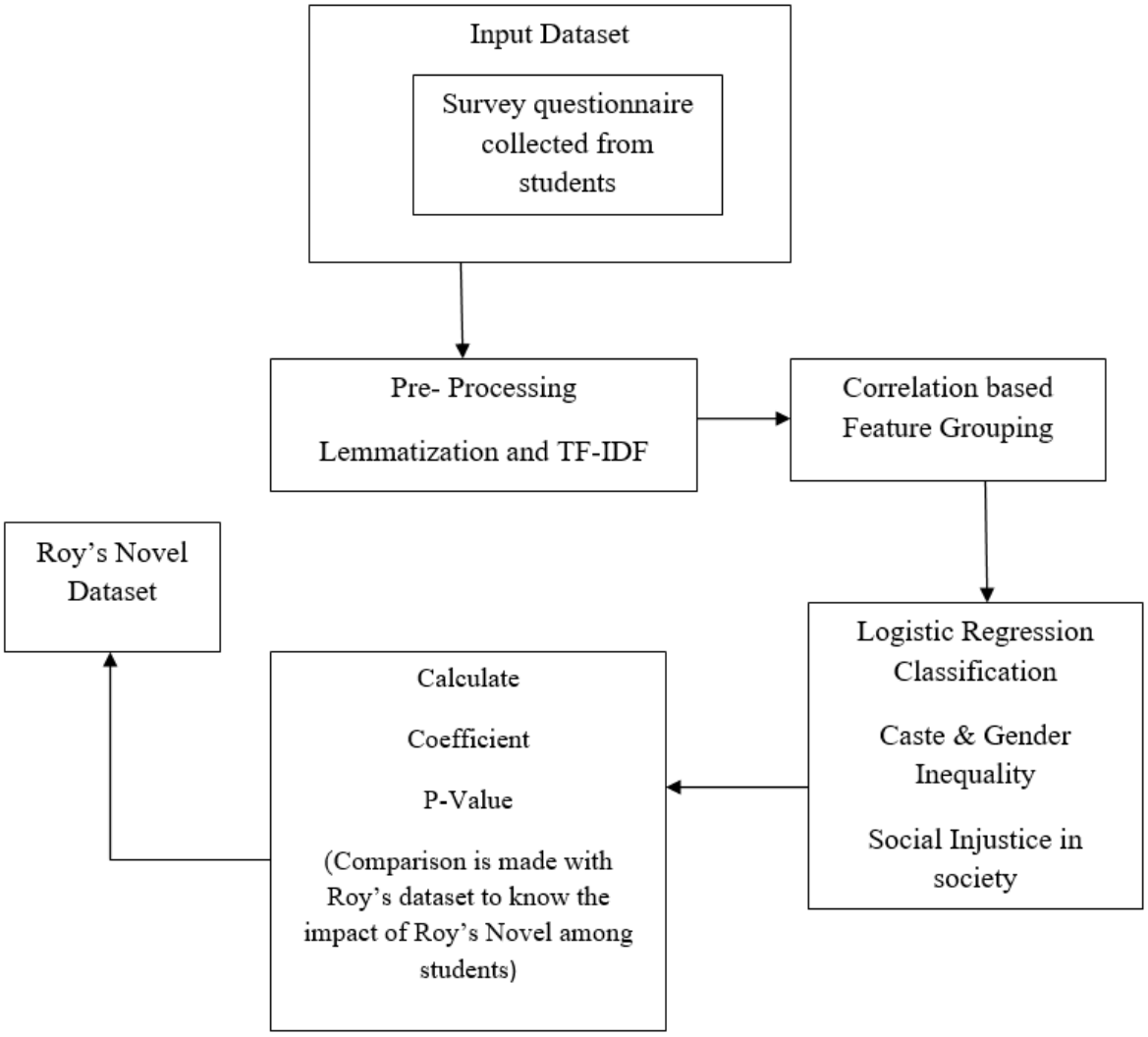


Fig.1 Frame work for the proposed model

3.1 Input Dataset

A structured survey was used to gather data in order to assess students' interpretive engagement with Roy's writings, with an emphasis on the goals of the study. The dataset includes Social interactions with students on dominance perceptions from Educational Institution in Kerala (N=300) and global socio- economic indices covering education, caste and gender. A thorough examination of gender dominance tendencies is made possible by these varied datasets, which aid in capturing various aspects of dominance in social and professional contexts.

Survey Questionnaire:

i. How do students understand Arundhati Roy's literary works' issues of gender, caste, and social injustice?

ii. How do the fiction and non-fiction works of Arundhati Roy affect students' comprehension of the systemic injustices in Indian society?
iii. What effects do Roy's narrative techniques—such as character development, symbolism, and non-linear storytelling—have on students' interest in sociopolitical issues?

iv. How can students' empathy and critical thinking be fostered through literature-based learning utilizing Roy's works?

v. How do students relate the modern Indian social environment to the cultural realities found in Roy's works?

3.2 Preprocessing

Lemmatization

AI is used by NLP to enable spoken communication between humans and machines. To do this, a variety of methods and procedures are used. Preprocessing Techniques: Cleaning and getting text input ready for analysis has always been crucial in NLP.

In Natural Language Processing lemmatization has a crucial text pre-processing method that breaks words down into the highest fundamental state. It considers the meaning and a segment of spoken of the word and guarantees that the base form is a valid word, in contrast to stemming, which only eliminates prefixes or suffixes. Lemmatization becomes more accurate as a result of not producing non-dictionary words.

Redundancy in the dataset is decreased by breaking words down to their most basic versions. Smaller datasets result from this, making it simpler to manage and process massive volumes of text for analysis or machine learning model training. By treating all related words equally, text becomes more consistent, which enhances NLP design capabilities.

TF-IDF

An analytical approach to retrieving details and natural language processing method known as TF-IDF (Term Frequency–Inverse Document Frequency) is used to evaluate a word's importance to a document in relation to a larger collection of documents.

Term Frequency determines how frequently a term appears in a manuscript. A greater frequency indicates higher relevance. A word is most likely relevant to the substance of the paper.

$$TF(t,m) = \frac{Number\ or\ time\ term\ t\ appears\ in\ the\ document\ m}{Total\ Number\ of\ terms\ in\ document\ m} \qquad (1)$$

Inverse Document Frequency raises the mass of unusual words when lowering the mass of common terms throughout multiple publications. If a term appears in fewer texts, it is more likely to be specific and important.

$$IDF(t,N) = \log\left[\frac{Total\ numer\ of\ document\ in\ corpus\ N}{Number\ of\ Documents\ containing\ term\ t}\right] \qquad (2)$$

TF-IDF is a helpful tool for tasks like search ranking and text classification because of this balance, which enables it to highlight phrases, both frequent inside a specific record and distinctive throughout the entire text document.

TF and IDF are combined to produce the TF-IDF result:

$$TF - IDF(t,d,N) = TF(t,m) \times IDF(t,N) \qquad (3)$$

3.3 Grouping features based on correlation

The input features are categorized according to the range of their correlation scores, and each group's features

are blended to produce a fresh representative vector of the feature.

The vector of feature of every data Xj (j = 1,..., M) in a dataset D = {X1, ..., XM} is characterized as vector of M-dimension {xi1, xi2, ..., xij, ..., xiM}, where j = 1,..., M,

Where,

M - Quantity of features.

As a result, every xij denotes jth feature score for the ith data in the collected data set. Likewise, in every feature, a vector Fj is created like {x1j, x2j,..., xNj}, which reflects the values of the jth feature for each of the dataset's N samples.

The following procedures can be used to group the feature vectors using the suggested CFG method:

(i) Determine the correlation scores:

The following formula is used to determine a feature vector Kj's correlation score with regard to class attribute x:

$$\text{Correlation Score } (Kj, X) = \text{cov}(Fj, X) / \sigma Fj \times \sigma X, \quad (4)$$

Where,

cov(Kj , X) - covariance between X and Kj

σKj and σX - Kj and X's standard deviations

(ii) Determine similar Features:

Features are grouped if their correlation score discrepancies fall below a predetermined threshold. As a result, based on the predetermined threshold, the quantity of collectives for each dataset is determined adaptively.

(iii) Group the relevant Features:

The features can be combined using a variety of combination functions, like the Vectors of binary features using OR or AND operations and for non-binary data, add, median, and average functions (numeric) vectors of feature. Using the given combination function, all of every group's remaining characteristics will be blended to create fresh feature vector representation for the group called FG (L), where l = 1,..., L,

Where,

L - Quantity of groups.

3.4 Logistic Regression Classification

Logistic regression is a supervised machine learning method for classification problems. Unlike linear regression, which predicts continuous values, it predicts the probability that an input belongs to a particular class.

To guarantee that the model is implemented correctly, it is crucial to comprehend the underlying assumptions of logistic regression. The primary assumptions are: Given that every piece of data is taken independent of others, there ought not to be anything reliance and correspondence among intake instances. It is predicated on the dependent variable being binary, meaning it can only have two possible values. The architecture makes the assumption that there is a linear relationship between the independent variables and the log chances of the dependent variable, which means that the predictors have a linear effect on the log chances.

Logistic Regression is used in binary classification, when the outcome can be classified into any of two groups, like No/Yes and False/True. It converts inputs into a likelihood score within 0 and 1 utilizing sigmoid function.

Sigmoid function works as follows:

- A crucial component of logistic regression is the sigmoid operation, which converts the unprocessed output of the model into a likelihood score in the range of 0 to 1.
- This feature generates "S"- created by simply converting any real number between 0 and 1. This curve is referred to as the logistic curvature or sigmoid curvature. Since probabilities must lie between 0 and 1, the sigmoid function is perfect for this.

- The class label in logistic regression is determined by a point of threshold, often 0.5. Data is categorized as group one if the sigmoid output is the same or higher compare to threshold. The data is classified as group zero when it falls below the threshold.

The logistic regression model transforms the linear regression program's continuous value output into a categorical value output by applying a sigmoid function. It converts any set of not dependent factors with real values in range within0 and 1. The logistic function is the name for this function.

The matrix representing the input features:

$$X = \begin{bmatrix} x11 \dots \dots \dots \dots \dots \dots \dots . . x1m \\ x21 \dots \dots \dots \dots \dots \dots \dots \dots x2m \\ . \\ . \\ . \\ . \\ xn1 \dots \dots \dots \dots \dots \dots \dots . xnm \end{bmatrix} \quad (5)$$

The dependant variable only has binary values, such as 0 or 1.

$$Y = \begin{cases} 0; & \text{if class 1} \\ 1; & \text{if class 2} \end{cases} \quad (6)$$

Where, the likelihood of belonging to a class can be calculated as:

$$P(y = 1) = \sigma(z) \quad (7)$$

$$P(y = 0 = 1 - \sigma(z) \quad (8)$$

Where,

z - The linear regression's continuous value.

Knowing the probabilities of an event under specific conditions makes it simple to determine the coefficients. The following formula can be used to calculate these coefficients, also known as parameters in maximum likelihood estimation.

$$\beta = \ln\left(\frac{P}{1-P}\right) \quad (9)$$

Where, P - Likelihood that the event will occur

A p-value is a probability that represents the likelihood that the Negative conjecture is accurate. It is a range within 0 and 1. According to the Negative conjecture it has no relationship between independent and dependent variables.

$$P = \frac{1}{1+e^{(\beta 0+\beta 1Y1+\cdots \beta nYn)}} \quad (10)$$

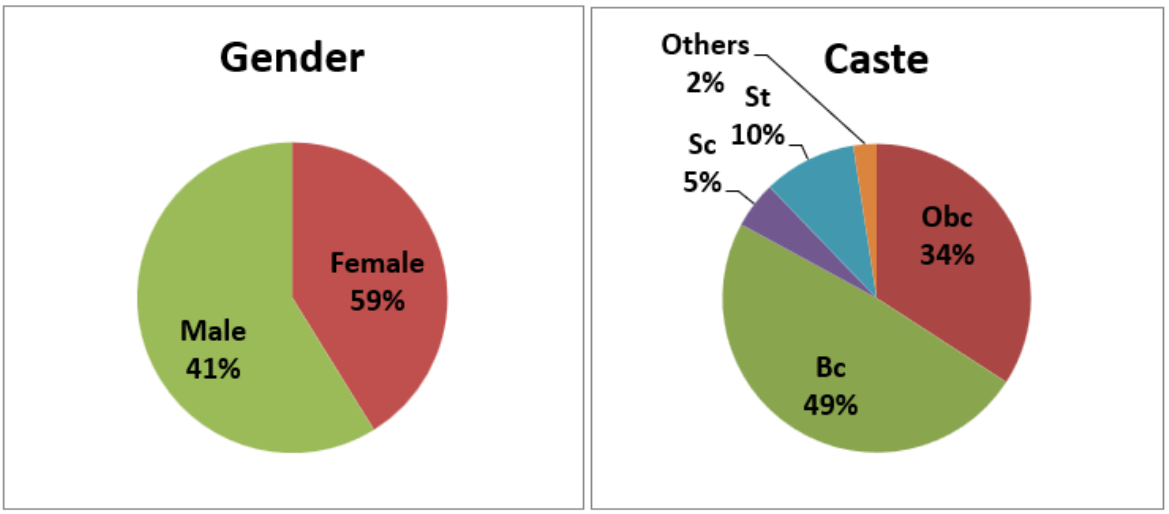
Where, β – coefficient of every event

Yn - Input variable

4. RESULTS AND DISCUSSION

Demographic View

The sample's demographic distribution demonstrates a varied representation of students, as shown by Fig. 2, ensuring a comprehensive perspective for the study. Undergraduates make up the majority of participants (62%), while postgraduates make up 38%. This indicates a good academic distribution and may represent a range of interpretive maturity levels.



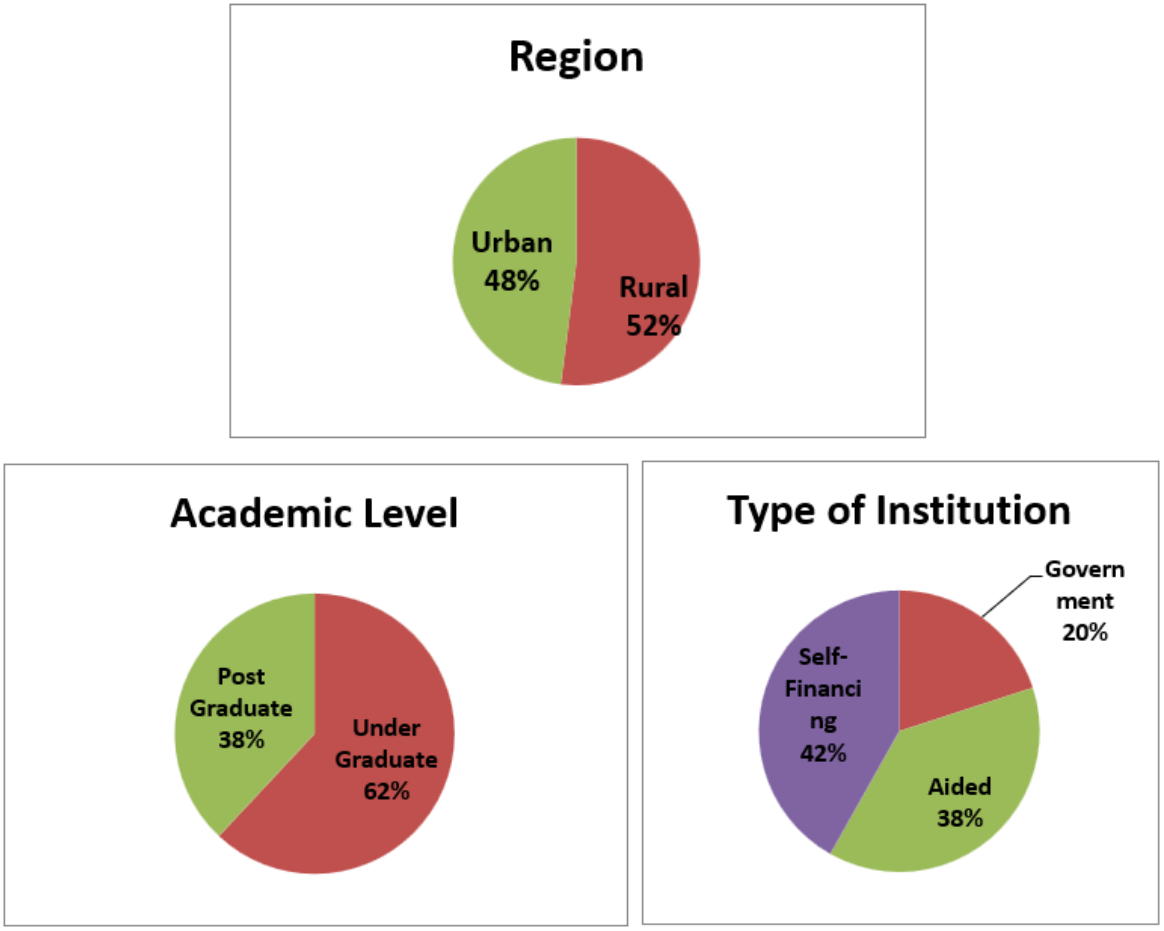


Fig. 2 Demographic view

In terms of gender, the sample is somewhat biased in favor of female respondents (59%), which was done on purpose to provide a potentially useful insight into gendered participation a pertinent factor in light of Arundhati Roy's emphasis on feminist ideology and her discussion of gender-based violence. With 48% of the sample coming from metropolitan Kerala and 52%% from rural regions, it is possible to gain a contextual understanding of how geographical variations affect reader interpretation. The study is under gone with 42 % 0f self- financing college, 38% Aided and 20% Government colleges.

The Table 1 shows the calculated probability score of logistic regression classification algorithm. The probability of caste gender inequality among students is higher. The Probability score of social injustice is also higher. The scores strongly predicts the existence of caste – gender inequality among students and the

Probability score of Logistic Regression	P (yes)	P (NO)
Caste and Gender inequality in Education Institution	0.89	0.57
Social Injustice in society	0.78	0.39

Table 1 Probability score of Logistic Regression

The results of Roy's unique dataset and logistic regression analysis of a survey questionnaire are compared in Table 2, which focuses on two major themes: social injustice in society and caste and gender disparity in educational institutions. Literacy and empowerment are looked at for the first theme. Racism, Discrimination and Social status are obtained for the second theme. The finding shows that the calculated coefficient and P-value from survey questionnaire collected among the students is slightly lower than the coefficient and P- value obtained from the Roy’s Novel dataset.

Comparison	Category	Roy's Novel Dataset		Logistic Regression result obtained from Survey Questionnaire	
		Coefficient B	P-Value	Coefficient B	P- value
Caste and Gender inequality in Education Institution	Literacy	1.45	0.56	1.42	0.55
	Empowerment	1.56	0.45	1.50	0.45
Social Injustice in society	Racism	1.48	0.32	1.42	0.30
	Discrimination	1.58	0.56	1.48	0.48
	Social Status	1.60	0.78	1.57	0.75

Table 2 Comparison of Student survey with Roy's Novel

This analysis clearly demonstrating unequivocally, students opinions are significantly shaped by their exposure to Roy's writings. Also it demonstrates how students are becoming more sensitive to Roy's examination of social injustice. This analysis strongly suggests the students deeper understanding of caste, gender, and societal inequities is positively correlated with stronger involvement with Roy's writings.

5. CONCLUSION

The advancement of Indian educational systems is hampered by the gender and caste inequality that permeates all spheres of society and socioeconomic level. In India, caste and gender inequalities in higher education are complex and multidimensional realities that impact all facets of lives, comprising schooling, employment opportunities, earnings, culture, socio and financial difficulties. The results show that caste, gender inequality, and social injustice in Indian society may be successfully exposed by machine learning-enabled text analysis of literary works like Arundhati Roy's. The logistic regression model sheds light on the relationships between variables like literacy, empowerment, discrimination, and social position and these discrepancies, indicating that targeted educational initiatives based on this kind of analysis can raise awareness and lessen social injustices.

References

1. Pant, Vinay Kumar, Rupak Sharma, and Shakti Kundu. "An overview of stemming and lemmatization techniques." Advances in networks, intelligence and computing (2024): 308-321.
2. Plisson, Joël, Nada Lavrac, and Dunja Mladenic. "A rule based approach to word lemmatization." In Proceedings of IS, vol. 3, pp. 83-86. 2004.
3. Pramana, Rio, Jonathan Jansen Subroto, and Alexander Agung Santoso Gunawan. "Systematic literature review of stemming and lemmatization performance for sentence similarity." In 2022 IEEE 7th international conference on information technology and digital applications (ICITDA), pp. 1-6. IEEE, 2022.
4. Liu, Haibin, Tom Christiansen, William A. Baumgartner Jr, and Karin Verspoor. "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text." Journal of biomedical semantics 3, no. 1 (2012): 3.

5. Korenius, Tuomo, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. "Stemming and lemmatization in the clustering of finnish text documents." In Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 625-633. 2004.
6. Yang, Sen, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. "Feature grouping and selection over an undirected graph." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 922-930. 2012.
7. Kuzudisli, Cihan, Burcu Bakir-Gungor, Nurten Bulut, Bahjat Qaqish, and Malik Yousef. "Review of feature selection approaches based on grouping of features." PeerJ 11 (2023): e15666.
8. Yu, Lei, Chris Ding, and Steven Loscalzo. "Stable feature selection via dense feature groups." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 803-811. 2008.
9. Zhong, Leon Wenliang, and James T. Kwok. "Efficient sparse modeling with automatic feature grouping." IEEE transactions on neural networks and learning systems 23, no. 9 (2012): 1436-1447.
10. Horaud, Radu, and Thomas Skordas. "Stereo correspondence through feature grouping and maximal cliques." IEEE Transactions on pattern analysis and machine intelligence 11, no. 11 (1989): 1168-1180.
11. Zheng, Ling, Fei Chao, Neil Mac Parthaláin, Defu Zhang, and Qiang Shen. "Feature grouping and selection: A graph-based approach." Information Sciences 546 (2021): 1256-1272.
12. Sheikhan, Mansour, Zahra Jadidi, and Ali Farrokhi. "Intrusion detection using reduced-size RNN based on feature grouping." Neural Computing and Applications 21, no. 6 (2012): 1185-1190.
13. Chen, Xiaojun, Yunming Ye, Xiaofei Xu, and Joshua Zhexue Huang. "A feature group weighting method for subspace clustering of high-dimensional data." Pattern Recognition 45, no. 1 (2012): 434-446.
14. García-Torres, Miguel, Francisco Gómez-Vela, Belén Melián-Batista, and J. Marcos Moreno-Vega. "High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach." Information Sciences 326 (2016): 102-118.
15. Chen, Xiaojun, Yunming Ye, Xiaofei Xu, and Joshua Zhexue Huang. "A feature group weighting method for subspace clustering of high-dimensional data." Pattern Recognition 45, no. 1 (2012): 434-446.
16. Kim, ZuWhan. "Real time object tracking based on dynamic feature grouping with background subtraction." In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8. IEEE, 2008.
17. Rashid, ANM Bazlur, Mohiuddin Ahmed, Leslie F. Sikos, and Paul Haskell-Dowland. "Cooperative co-evolution for feature selection in Big Data with random feature grouping." Journal of Big Data 7, no. 1 (2020): 107.
18. De Menezes, Fortunato S., Gilberto R. Liska, Marcelo A. Cirillo, and Mário JF Vivanco. "Data classification with binary response through the Boosting algorithm and logistic regression." Expert Systems with Applications 69 (2017): 62-73.
19. Vallejos, J. A., and S. D. McKinnon. "Logistic regression and neural network classification of seismic records." International Journal of Rock Mechanics and Mining Sciences 62 (2013): 86-95.
20. Worth, Andrew P., and Mark TD Cronin. "The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects." Journal of Molecular Structure: THEOCHEM 622, no. 1-2 (2003): 97-111.
21. Kirasich, Kaitlin, Trace Smith, and Bivin Sadler. "Random forest vs logistic regression: binary classification for heterogeneous datasets." SMU Data Science Review 1, no. 3 (2018): 9.
22. Gopika, N., and A. Meena Kowshalaya ME. "Correlation based feature selection algorithm for machine learning." In 2018 3rd international conference on communication and electronics systems (ICCES), pp. 692-695. IEEE, 2018.