Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 2 (2025)

Analyzing Subaltern Narratives in Indian Social Media Using Machine Learning

1. Dr. Deepika K., Assistant Professor, Indian Institute of Mass Communication, Aizawl,
2. Dr. Kuber Nag, Independent Researcher, Ph.D, IIT Hyderabad,

Abstract
More than ever before, marginalised communities are gaining a platform on social media to tell their stories, question authority figures, and find common ground. Subaltern narratives in Indian social media will be investigated in this study by integrating machine learning with qualitative discourse analysis. Using advanced Natural Language Processing (NLP) techniques including text categorisation, topic modelling, and sentiment analysis, the study uncovers similarities and emotional undercurrents throughout social media platforms such as Reddit, Twitter, and Facebook. Based on annotated subaltern content, the authors propose a new model that could detect instances of gender, caste, religion, and tribe marginalisation. In addition, the intrinsic power dynamics and sociopolitical background of the online talks are dissected using narrative analysis and Critical Discourse Analysis (CDA). The findings shed light on the manner in which marginalised communities in India navigate digital platforms, revealing patterns of resistance, identity formation, and structural silence. By drawing attention to the need for ethically sound and culturally sensitive digital research, as well as the potential of machine learning to amplify under-represented perspectives, the work contributes to postcolonial theory and computational social science.

Keyword: Culture-Centered Approach (CCA), Information and Communication Technology (ICT), Generalized Linear Models (GLM)**.**

I. INTRODUCTION

The subalterns and society unspoken involvement in creating new identities and groups is the focus of this article. It explores the subaltern online presence in search of respect, independence and equity. Additionally, it provides a quick overview of how the subalterns understand the new digital cinema and social media. Agricultural, industrial and information societies are all heavily discussed in and the authors clearly want their readers to grasp their relevance to marginalized groups. An information society is superior and more meaningful terminology and it gives subalterns a lot of room to show the world their knowledge systems, culture, ethics and language [1]. By ignoring the impending worries of the digital divide access to quality information, information dominance and exploitation this study fails to critically believe in the liberating capacity of the information society, notwithstanding these advantages. The discourse around Indian culture, philosophy and history failed to include subaltern voices or involvement for an extended period of time in socially acceptable and esteemed locations and institutions. But digital social media have helped people become more self-aware, and they've been able to approve a shared ontology through Twitter handles, Face book pages and WhatsApp groups that's deeply rooted in democratic traditions [2]. Various online subaltern groups have unquestionably found social media avenues through which to bring their real-world experiences to light and offer the world new perspectives and associated models.

It highlights the challenges that subaltern actors and social media narratives face when trying to be represented in mainstream organizational theory [3]. Organizational studies of social media can be revolutionized through writing approaches that incorporate feminist postcolonial literary traditions. An auto ethnographic politics of decolonizing the neoliberal reproduction of social media in postcolonial spaces is performed by the authors of this article as they reflect on their experiences as academic-activists partnering with subaltern communities in the global South on social media

processes. It discusses how much postcolonial theory on social development and culture ignores White and Brown privileges in racial, caste, class and gender categories [4]. In our auto ethnographic dialogue, it challenges the dominance of neoliberal tropes in social media communication while simultaneously challenging the practice of auto ethnography as a means of producing privileged identities within imperial sites of mostly academic institutions. Collaborations for postcolonial social media that remove spaces for subaltern articulation force.

II. LITERATURE SURVEY

The social media storage and computing advancements have changed the way technology will impact civilization in the future. According to the International Data Corporation there will be an estimated of data created and consumed as a result of wearable devices and the IoT[5]. The proliferation of ML and AI into domains such as commerce, academia, healthcare the legal system, public safety and many more has been boosted by the availability of stream data. It subaltern crises and the big data disruption have changed the global landscape in recent years [6]. To address research gaps and educate policy it is important to do thorough scholarship that focuses on marginalized communities. It need to take a close look at the main areas where Indian is falling short in its pursuit of ML and AI in subaltern data collecting and data security [7]. When datasets fail to take into consideration the ethnic and demographic diversity of the Indian population it could affect the generalisability and ML and AI resilience. Simultaneously, there have been few legislative frameworks that adequately safeguard personal data used in ML and AI [8]. Using NLP computers can understand interpret and generate language. One branch of NLP is concerned with deciphering sentiment or the underlying emotional tone in textual input. This method has been criticized by HCI researchers despite its widespread use in computational linguistics [9]. Quantifying and categorizing complex NLP and emotion is the goal of sentiment analysis which aims to assign subjective or polarity ratings often within standardized scales or nominal sentiment categories [10]. But studies have shown that classification and sorting are political and reductionist which in turn can keep inequality alive. Artifacts like algorithms and ML technologies are political according to previous research [11]. They are influenced by both society norms and the developer's or group's own politics. Specifically, sentiment analysis tools are both created and influenced by subaltern making them socio media. One side social media has creating these tools, while the other has data on user interactions used to train them which in turn influences their outputs [12]. Examining how the subaltern world of Web is being used in the Indian setting is the main objective of this research project. A more pluralistic social media sphere has opened up because to Web which has allowed the subaltern to express their interests [13]. Historically the state and higher castes particularly Brahmins have controlled the mainstream social media. Individually the subalterns exhibit pluralism. Nevertheless the primary focus of this research is on the caste critics, particularly the in one of the case studies [14]. In the second case study, the focus is on feminists and their online activities [15]. The article's connection of violence to the preservation of social media order in Indian culture gives it continued significance despite the material's antiquity. In Indian society the social media that shapes the relationships between people at different levels of hierarchy is the caste system. They emphasize the importance of violence or the use of force to uphold social order. Looking at a statement from the by and comparing it to cases of caste-based violence in the present day highlights how relevant [16]. Begin by offering a general overview and analysis of the Subaltern Studies Group putting its project and internal conflicts and tensions in the context of the Gramscian tradition that the group asserted when it first formed [17]. It will go over the reception and consumption of Subaltern approaches thus far setting the stage with some political historical and cultural background that could be useful for future research on Latin America [18]. It reevaluates the inherent tensions in Subaltern Studies in light of my experiences in Latin America and proposes several avenues for further research and discussion that could help the field expand its

current scope [19]. By issuing the warning the Indian government and WhatsApp's founders have clearly concluded that technology dictates society's orientation and that the social media is the message. This has allowed the assertion of a majoritarian Hindu identity to proceed unchallenged [20]. It is already a Subaltern strong claim to say that technology is to blame but the lynching further cast doubt on the public's perception of technology's role. Because of this the majority of social media outlets that target migrant populations are conflicted. On one hand, there are government-sponsored publications that indoctrinate migrant workers from on high [21]. On the other hand there are magazines that provide moral education and guidance to migrant workers. Accordingly, subaltern determining the nuance and texture of various dugong subjectivities is extremely challenging if not impossible. The breadth of this marginalized group's place-desires and spatial imaginations is even more difficult to fathom [22].

III. METHODOLOGY

Indigenous communities have long endured the problems of communication and material exclusion, as well as the systematic silencing of their voices in public discourse. The culture-centered approach (CCA) and other critical-cultural frameworks contend that listening to the stories of marginalised groups is a crucial step in critically examining systemic gaps. Listening to marginalised voices is becoming more important in the field of communication and cultural studies as a means of elevating subaltern narratives in decision-making arenas. Scholarly communication becomes importance when prevailing powers systematically overlook disadvantaged voices and discourses to promote their own hegemonic objectives. These groups have a long history of depicting indigenous people as helpless victims of colonisation and a target for reform.

A. Data Preprocessing:

1) Transformation of Data:

Transformation of data follows after EDA. Data mining ease and data simplicity should take precedence. At this stage, which is comparable to data cleansing, characteristics and attributes are generated from the available data by incorporating human inspection. Applying any necessary treatments for cleaning, smoothing, or normalisation is the following stage.

2) Reduction of Data:

When working with a dataset that contains a significant number of observations, reduction of data is an essential step before using any ML method. The dataset contains a large number of features or variables, some of which may not be useful or required. Examining only the features that significantly affect the learning, prediction, generalisability, and processing time of the model is crucial for minimising data. Problems with the model's performance may arise if overfitting features or noisy data are not addressed.

B. Feature Selection:

Prior to using principal component analysis, features must be pre-processed. The rationale for this is because RFE produces output in a wide range of units, some of which include percentage data and others which contain extremely large values. This means that different methods may provide different results when trying to extract the major components [23]. It is necessary to do feature pre-processing before feeding the data into the principal component analysis algorithm. To simplify things, one option is to use PCA, which takes a starting set of vector observations and produces a new set of vector data. Since it relies solely on the premise of actual linear correlations among certain classification factors, it is limited to numerical continuous fields where it can be applied. Although there is some loss of information, rotation can reduce the number of variables from several dimensions to a smaller, more relevant set. PCA with orthogonal rotation can simplify the process by reducing the number of classification variables and generating new, standardised,

uncorrelated components. These components are often more relevant and make use of the linear relationship between the classification variables.

C. Traning in the Model:

1) DT:

DT are ubiquitous in ML for their predictive modelling capabilities; structurally and functionally, they are quite similar to flowcharts. The tree's nodes act as evaluation checkpoints for individual attributes, with data flowing down branches that represent the results of those evaluations. One can reach the end of each branch in the tree by reaching the last node, which has the class label. With each iteration, the predictive power of the tree is improved by the iterative process of recursive partitioning. The initial dataset is divided depending on attribute values. After all additional splits have failed to improve prediction accuracy, this recurrent partitioning will end. DT classification is a great technique for exploring knowledge because it is domain neutral and doesn't require complicated parameter tuning. DT are important to classification learning because they can gracefully handle massive datasets and consistently produce high accuracy [24]. An example of the ease and efficiency of this long-standing system is the effortless classification of instances by following the tree from its base to its leaves, where a final classification.

2) **KNN**:

While KNN is versatile enough to handle both classification and regression, the latter is where it really shines. Known as a sort of lazy learning, KNN eliminates the need for a specific training phase. Finding a new data point's KNN using a distance metric, most often the Euclidean distance, allows for more accurate prediction-making. The class with the most nearby neighbours will then get it. The procedure is referred to as the 3-NN method when k equals 3. In this instance, the three nearest neighbours of the new item in green consist of two entities from the orange class and one item from the blue class. Consequently, the orange class will obtain the newly issued green item.

3) **LR**:

A statistical method known as LR can be used to generate a model that can predict the values of a categorical dependent variable from a set of explanatory variables. As a result, it describes the link function that is utilised in a regression model to determine the likelihood of an event:

$$\mu(v) = \frac{g^{\alpha_0+\alpha_1 v_1+\alpha_2 v_2+\cdots\alpha_k v_k}}{1 + e^{\alpha_0+\alpha_1 v_1+\alpha_2 v_2+\cdots\alpha_k v_k}}$$

when the predictive variable's value is v, the probability of success is denoted as $\mu(v)$. The predictive variables' coefficients are denoted by α_k, and the adjustment constant α_0 is utilised. Generalised Linear Models (GLM) must be explained before LR can be understood. Here are the three parts that make it up:

- Part that is completely at random and includes the distribution of the dependent variable's probability U.
- A linear relationship between the independent variables that represents the systematic component.
- The mathematical relationship between the systematic component and the random component can be described by a link function.

One variant of the GLM model is the binary LR model. The estimation of coefficients can be obtained using this function. Our fraud probability is then determined by plugging these values into Equation 1.

4) NB:

The NB classifier, which is based on Bayes' theorem, is an effective machine learning tool. What this means is that it runs on the premise of feature independence, which states that each feature or predictor works independently to aid in categorisation. Even with massive datasets, efficient and successful classification is possible because to this assumption, which reduces the computational complexity of the model [25]. Equation 2 summarises Bayes' theorem, which is the basis of this classifier:

$$N(C/D) = \frac{N(D/C)N(C)}{N(D)}$$

Given that $N(C)$ indicates the likelihood of event C, $N(D)$ signifies the likelihood of event D, and $N(D/C)$ reflects the likelihood of event D given that event C has occurred. Probabilistic thinking and decision-making are greatly enhanced by the theorem, which essentially offers a framework for updating probabilities in light of new evidence.

IV. Results and Discussion

In the Indian subcontinent, a culture of oppression has been institutionalised through the social stratification system of caste. All groups now have a platform to debate, express, and form opinions on critical issues, due to the democratisation of voices made possible by the rise of social media. Regrettably, it also provides an opportunity for openly cattiest individuals to promote bigotry, intolerance, and casteism while posing as social media stars. Among the most ancient forms of social stratification, the caste system in India is defined by a sliding scale from most impure to most pure. Many communities have been oppressed and discriminated against as a result of this throughout many generations.

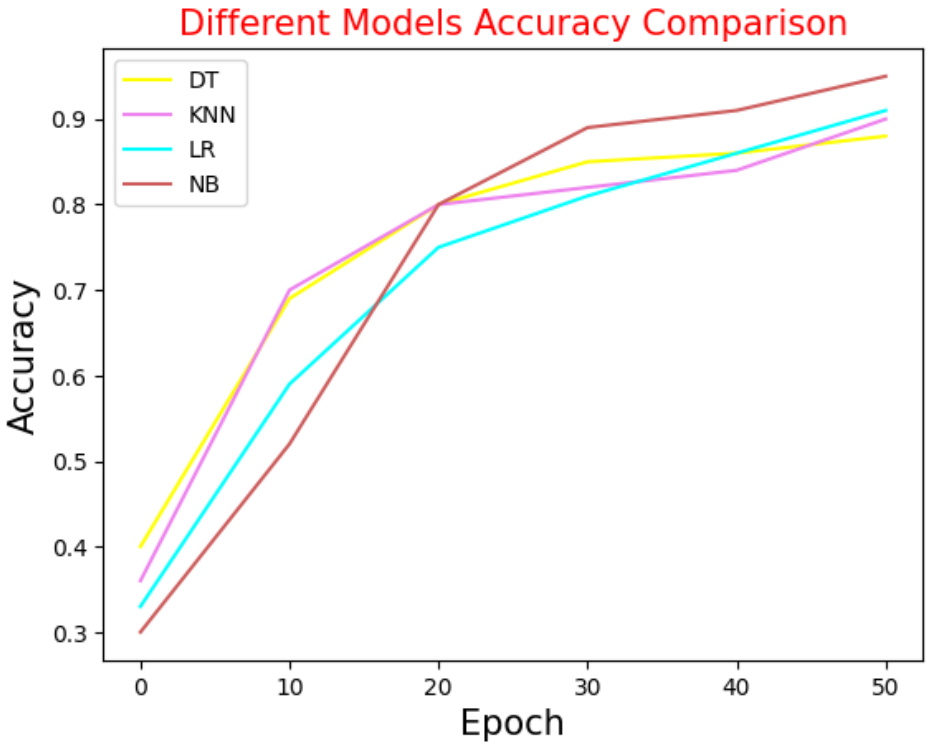

Fig. 1. Accuracy Comparison for Different Models

Figure 1 displays the Different Models Accuracy Comparison. The Models DT, KNN, LR, and NB are done and the NB models delivers the most accuracy when comparing to the other three Models.

TABLE I. PERFORMANCE COMPARISON

Models	Accuracy	Precision	Recall	F1-Score
DT	88.74	87.12	86.21	88.82
KNN	90.23	89.92	88.43	90.48
NB	**95.82**	**94.76**	**93.61**	**95.94**
LR	91.76	90.64	89.38	91.86

All four models' performance comparisons are summarised in Table 1. These models include DT, KNN, LR, and NB, among others. Accuracy, Precision, Recall, and F1-Score are just a few of the metrics where NB excels above its competitors. In jobs where precision and predictability are paramount, the results show that NB shine.

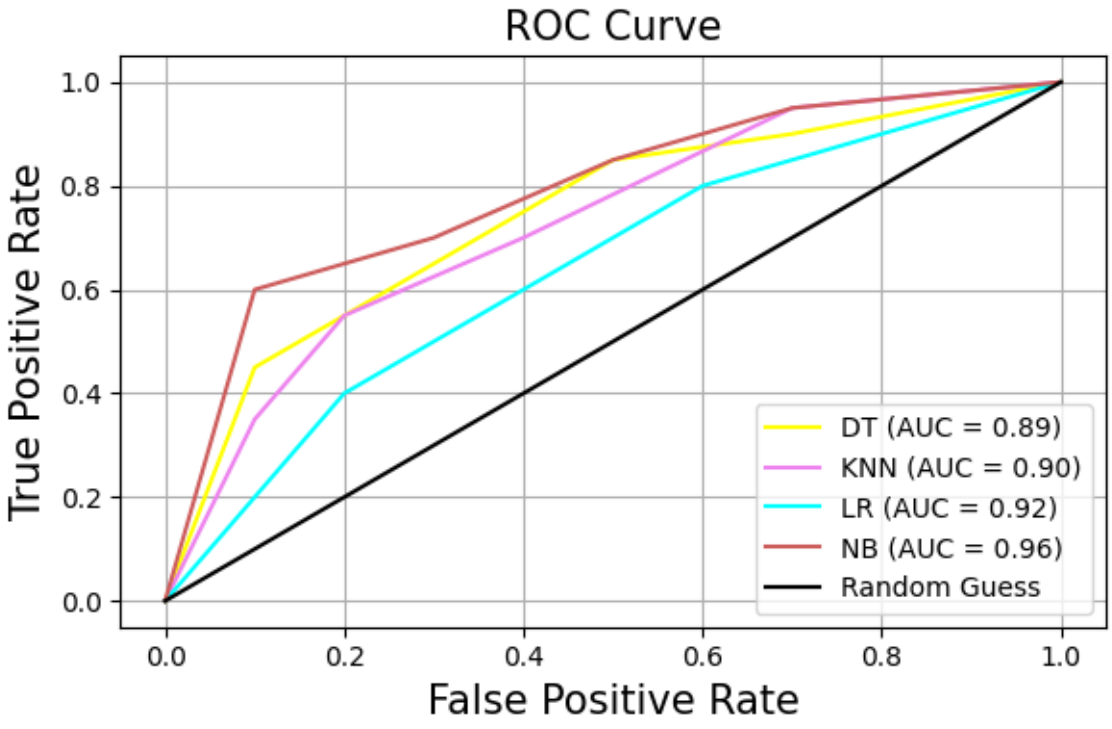

Fig. 2. ROC Curve

Figure 2 shows a ROC curve that compares the accuracy of DT, KNN, LR, and NB, four distinct classification models. The ROC curve provides a visual representation of the discriminative capacity of each model by plotting the TPR against the FPR at different threshold values. When compared to the other models, NB demonstrates the best classification skill on the dataset under consideration.

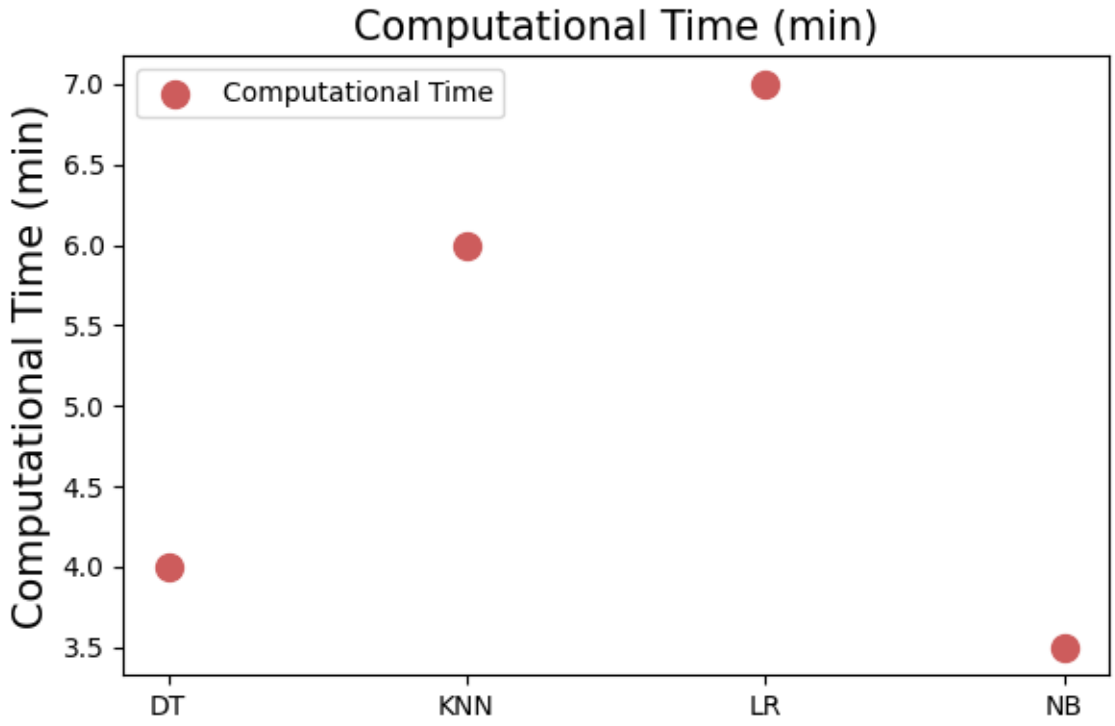

Fig. 3. Computational Time for Various Models

Figure 3 shows a scatter figure showing the amount of time that four ML models—DT, KNN, LR, and NB—took to compute. The y-axis shows the computational time in minutes, and each model is indicated by a red circular marker. With a computing duration of about 3.5 minutes, NB is the most efficient method. Especially in time-sensitive applications, the graph clearly shows how each model compares in terms of efficiency, highlighting the trade-off between computation and performance.

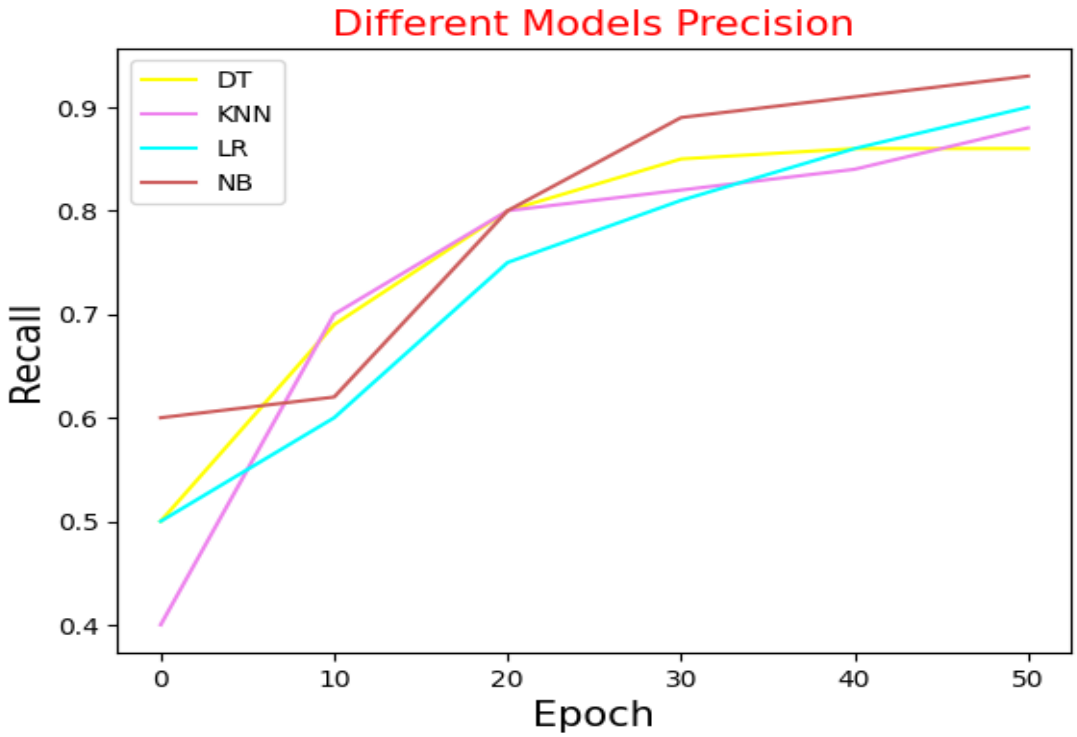

Fig. 4. Comparison of Different Models Precision

Over the course of various training epochs, Figure 4 shows a line plot depicting the recall performance of four ML models: DT, KNN, LR, and NB. On one side, they have the recall score, and on the other, we have the number of epochs. By the end of the 50th period, NB had achieved the best recall of about 0.93, continuing its streak of outperformance. In sum, the graph does a good job of showing how each model learns and how they compare in terms of recall performance over time.

V. CONCLUSION AND FUTURE DIRECTIONS

In particular, the subalterns grasp the information society's covert function in forging new communities and identities. For the sake of respect, independence, and equity, it probes thoroughly into the online personas of marginalised groups. Additionally, it provides a quick overview of how the subalterns understand the new digital media and cinema. With the use of ICTs, they may build new normative standards for how to act as citizens of the information age, and so transform human society away from outdated, space-and time-driven epistemological frameworks. Citizenship, which is emphasised as a legal entitlement bestowed to the nation-state, assumes geography. As soon as the new ideas of place and time are deemed unimportant for new forms of online citizenship, this paradigm becomes inadequate because of the predominance of ICT and sovereign ruling beyond defined pre-existing areas.

REFERENCES

[1] P. Reviewed and R. Journal, “Subaltern Self, Communities, And Digital Media In Indian Information Society,” subalt. Self, communities, digit. Media indian inf. Soc., vol. 014, no. 6, pp. 108–112, 2024.

[2] D. Vijay, S. Gupta, and P. Kaushiva, “With the margins: Writing subaltern resistance and social transformation,” Gender, Work Organ., vol. 28, no. 2, pp. 481–496, 2021, doi: 10.1111/gwao.12583.

[3] M. J. Dutta and A. Basu, “Subalternity, Neoliberal Seductions, and Freedom: Decolonizing the Global Market of Social Change,” Cult. Stud. - Crit. Methodol., vol. 18, no. 1, pp. 80–93, 2018, doi: 10.1177/1532708617750676.

[4] G. I. Ahmad, J. Singla, A. Ali, A. A. Reshi, and A. A. Salameh, “Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus: A Comprehensive Review,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 2, pp. 455–467, 2022, doi: 10.14569/IJACSA.2022.0130254.

[5] F. E. Mallon, “The Promise and Dilemma of Subaltern Studies: Perspectives from Latin American History,” Am. Hist. Rev., vol. 99, no. 5, p. 1491, Dec. 2022, doi: 10.2307/2168386.

[6] D. Gupta, A. Ekbal, and P. Bhattacharyya, “A deep neural network based approach for entity extraction in code-mixed Indian social media text,” Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval., pp. 1762–1767, 2019.

[7] G. Ganapathy-doré, “Artificial Intelligence in Indian Fiction." Postcolonial Text,” Artif. Intell. Indian Fict. Postcolonial Text, vol. 19, no. 3, 2024.

[8] A. Jamatia, A. Das, and B. Gambäck, “Deep learning-based language identification in English-Hindi-Bengali code-mixed social media corpora,” J. Intell. Syst., vol. 28, no. 3, pp. 399–408, 2019, doi: 10.1515/jisys-2017-0440.

[9] W. Sun, “Narrating translocality: Dagong poetry and the subaltern imagination,” Mobilities, vol. 5, no. 3, pp. 291–309, 2022, doi: 10.1080/17450101.2010.494837.

[10] F. Vasudeva and N. Barkdull, “WhatsApp in India? A case study of social media related lynchings,” Soc. Identities, vol. 26, no. 5, pp. 574–589, 2020, doi: 10.1080/13504630.2020.1782730.

[11] K. Rakshitha, R. H M, M. Pavithra, A. H D, and M. Hegde, “Sentimental analysis of Indian regional languages on social media,” Glob. Transitions Proc., vol. 2, no. 2, pp. 414–420, 2021, doi: 10.1016/j.gltp.2021.08.039.
[12] S. Masiero, “Decolonising critical information systems research: A subaltern approach,” Inf. Syst. J., vol. 33, no. 2, pp. 299–323, 2023, doi: 10.1111/isj.12401.
[13] et al Kumar, Akshi, “Hybrid deep learning model for sarcasm detection in Indian indigenous language using word-emoji embeddings,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 22, pp. 1–20, 2023.
[14] H. Kalır, “Can the Artificial Intelligence Speak? Subalternity of ‘Subontologies’ and the Death of the Programmer,” Acta Infologica, vol. 7, no. 1, pp. 173–185, 2023, doi: 10.26650/acin.1279545.
[15] S. Kumar et al., “Classification of Indian media titles using deep learning techniques,” Int. J. Cogn. Comput. Eng., vol. 3, no. November 2021, pp. 114–123, 2022, doi: 10.1016/j.ijcce.2022.04.001.
[16] R. Mahajan, R. Mahajan, E. Sharma, and V. Mansotra, “Are we tweeting our real selves? personality prediction of Indian Twitter users using deep learning ensemble model,” Comput. Human Behav., vol. 128, no. November, pp. 1–8, 2022, doi: 10.1016/j.chb.2021.107101.
[17] D. Kapoor, “Subaltern Social Movement Learning and the Decolonization of Space in India SUBALTERN SOCIAL MOVEMENT LEARNING,” Subalt. Soc. Mov. Learn. decolonization Sp. India." Int. Educ., vol. 37, no. 1, 2007.
[18] N. B. Defersha and K. K. Tune, “Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Approach,” Indian J. Sci. Technol., vol. 14, no. 31, pp. 2567–2578, 2021, doi: 10.17485/ijst/v14i31.1019.
[19] D. Anagandula and S. Bandari, “Machine Learning Based the Role of Social media in Promoting the Safety of Women in Indian Cities,” Turkish J. Comput. Math. Educ., vol. 13, no. 03, pp. 1268–1278, 2022.
[20] S. Agrawal and A. Awekar, “Deep learning for detecting cyberbullying across multiple social media platforms,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10772 LNCS, no. Table 2, pp. 141–153, 2018, doi: 10.1007/978-3-319-76941-7_11.
[21] P. Sv, J. Tandon, and H. Hinduja, “Indian citizen’s perspective about side effects of COVID-19 vaccine–A machine learning study,” Diabetes Metab. Syndr. Clin. Res. Rev., no. January, 2020.
[22] A. Saroj and S. Pal, “An Indian Language Social Media Collection for Hate and Offensive Speech,” Proc. Work. Resour. Tech. user author profiling Abus. Lang., no. May, pp. 11–16, 2020, [Online]. Available: https://www.aclweb.org/anthology/S19-2007/
[23] E. Caldeira, G. Brandao, and A. C. M. Pereira, “Fraud analysis and prevention in e-commerce transactions,” Proc. - 9th Lat. Am. Web Congr. LA-WEB 2014, no. December, pp. 42–49, 2014, doi: 10.1109/LAWeb.2014.23.
[24] F. T. Johora, R. Hasan, S. F. Farabi, J. Akter, and M. A. Al Mahmud, “Ai-Powered Fraud Detection in Banking: Safeguarding Financial Transactions,” Am. J. Manag. Econ. Innov., vol. 6, no. 6, pp. 8–22, 2024, doi: 10.37547/tajmei/volume06issue06-02.
[25] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, “Machine learning approaches for EV charging behavior: A review,” IEEE Access, vol. 8, pp. 168980–168993, 2020, doi: 10.1109/ACCESS.2020.3023388.