Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(1), 255-282
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Media Diversity Forecasting: A Longitudinal Study Using Hybrid Machine Learning Model for Predictive Insights into Community Representation (2004-2024)

Siddharth Yadav, Nicole Lee, Rezza Moieni

Cultural Infusion Pty Ltd., Melbourne, Australia
Email: siddharth.y@culturalinfusion.org.au, nicole.lee@culturalinfusion.com, rezza.moieni@culturalinfusion.com

How to cite this paper: Yadav, S., Lee, N., & Moieni, R. (2026). Media Diversity Forecasting: A Longitudinal Study Using Hybrid Machine Learning Model for Predictive Insights into Community Representation (2004-2024). *Open Journal of Social Sciences, 14,* 255-282.
https://doi.org/10.4236/jss.2026.141017

Received: October 13, 2025
Accepted: January 13, 2026
Published: January 16, 2026

Abstract

This research addresses the growing interest in diverse media representation by investigating long-term trends across six communities: African, Asian, European, Hispanic, Indigenous, and Middle Eastern. Spanning news, social media, and entertainment, our study introduces a novel forecasting system using a hybrid of Long-Term Memory (LSTM) neural networks, Autoregressive Integrated Moving Average (ARIMA), and Prophet models. Two decades of data (2004-2024) were scraped from diverse open sources and media archives, and a unique engagement metric was proposed. The models demonstrated high accuracy, significantly improving upon benchmark studies in social media forecasting. This project also features a user-friendly web application, enabling stakeholders to gain predictive insights. This work offers actionable, data-driven insights to evaluate and improve media inclusivity, setting groundwork for future cultural analytics, policy development and ethical media production.

Keywords

Media Diversity, Communities Representation, Social Media Analytics, News Media, Entertainment Media

1. Introduction

Media plays a crucial role in shaping public perception, influencing societal attitudes, and defining cultural narratives. Over the past two decades, intensified discussions around diversity, equity, and inclusion have driven significant shifts in media representation across various platforms, including news, movies, and social

media. This influence extends directly to shaping social identity (Hui, 2025), with research showing that media could influence community beliefs through the selective discussion of specific cultural groups (Asur & Huberman, 2010). Consequently, certain groups are often depicted positively while others are neglected. This disparity in representation can be attributed to factors such as education level, population size, income, and social hierarchy within the group (Hui, 2025).

Despite increased awareness, ethnic biases in media representation persist. Certain communities remain under-represented or misrepresented, while others dominate narratives disproportionately. For example, a study on mainstream Australian media, conducted from April 2018 to April 2019, found that 57 percent of 281 media pieces discussing race were negative. Muslim women were most often targeted by negative social media commentary, often originating from mainstream newspapers. Furthermore, 70 percent of these pieces used covert techniques such as dog-whistling, irony and de-contextualisation when discussing race (All Together Now, 2019). Given that the industry's media code of conduct primarily addresses overt forms of racism, media regulators are often unable to prosecute media agencies perpetrating this covert form of racism, leaving targeted communities without an independent avenue for complaint (Hui, 2025). Understanding these trends and predicting future representation is therefore essential for media organisations seeking to foster a more inclusive society and prevent marginalisation.

The complexities of representation are further compounded by intersectional diversity dynamics, wherein multiple dimensions of identity (e.g., sexual-minority Black women or elderly disabled individuals) overlap, creating unique experiences and challenges that media narratives must reflect (Hofhuis et al., 2024). A key challenge in multi-cultural societies is the cultivation of positive inter-cultural and inter-religious relations. This becomes difficult when direct interaction between groups is limited, compelling individuals to largely form their perceptions through public media. In such contexts, negative media portrayals can readily contribute misunderstanding and political unrest. Consequently, the role of media in fostering understanding and trust becomes critical, serving to mitigate societal division and encourage peaceful coexistence.

This direct link between media representation and community wellbeing is evidenced by recent studies. For instance, in Australia, the significant media representation of the top five non-English languages—Arabic, Cantonese, Italian, Mandarin, and Vietnamese—has been shown to influence confidence in community participation (All Together Now, 2019). These findings suggest that communities with higher media representation tend to feel a stronger sense of belonging, increasing their likelihood to trust news media and engage in discussions about societal issues (Abbasi et al., 2022). Effective engagement in such discussions indicates a stronger sense of belonging in society. Additionally, factors such as English proficiency and length of time living in the country further enhance the confidence of individuals from diverse communities, shaping their community's rep-

resentation and overall perception (Loecherbach et al., 2020).

Beyond ethnic representation, gender-based disparities in media representation also present a critical concern with profound societal implications. While a large proportion of media consumers (80%) engage with diverse characters through movies and television, significant imbalances persist across creative roles and on-screen portrayals. For instance, women only hold 21% of directing positions despite constituting 48% of film leads, often being relegated to stereotypical roles (Erigha, 2015). Similar patterns are observed in news media, where women represent only 30% of expert contributors and female athletes receive a low 4% of sports coverage (Erigha, 2015). Such systemic under-representation reinforces societal stereotypes and can directly impede opportunities, as evidenced by the limited sponsorship resulting from insufficient news coverage of women's sports (Żerebecki et al., 2025).

Given these multifaceted challenges in media representation, this project aims to analyse the past 20 years of media data (2004-2024) and predict the community representation for the next 10 years across Australian media, focusing on key media domains: 1) news articles (mainstream media sources), 2) social media (e.g., Twitter, Instagram, Reddit), and 3) movies and TV shows (IMDb, streaming platform).

2. Literature Review

The profound influence of media in shaping societal perceptions and cultural narratives necessitates a comprehensive understanding of how diverse communities are represented. This literature view synthesises existing research to highlight the critical role of media, analyse current disparities, and identify the gaps in current analytical and predictive capabilities, thereby establishing the foundation for this study.

2.1. The Role of Media in Shaping Public Perception and Attitude

Media plays a significant role in shaping public perception, especially concerning multiculturalism and representation. Rodrigo-Ginés et al. (2024) examined representation in Australian news media through a multimodal survey combining CATI (Computer-Assisted Telephone Interviewing) and CAPI (Computer-Assisted Personal Interviewing) methodologies. Their study revealed a strong link between the perception of adequate media representation and trust in society, demonstrating how fair media portrayal fosters civic engagement, enhances a sense of belonging, and influences participation in socio-political discourse. Similarly, Rodrigo-Ginés et al. (2024) conducted a content analysis of mainstream news and Twitter sentiment. Using Stuart Hall's encoding or decoding model, this study assessed media portrayals of multicultural events, revealing how media narratives impact public perception and social cohesion, either promoting inclusivity or exacerbating divisions (Rodrigo-Ginés et al., 2024).

Beyond general perception, media also significantly shape public attitude and

policy. Asur and Huberman (2010) explored this by examining the media's influence on attitudes towards issues such as disability, climate change, and economic development. Their findings illustrated how media framing can shift public opinion, particularly in cases where negative portrayals reinforce societal biases. For instance, negative depictions of individuals receiving disability benefits led to hardened societal attitudes and reduced public support for welfare programs (Asur & Huberman, 2010). Similarly, Hui (2025) focused on media representation of minorities in Singapore. Their work revealed that ethnic groups, such as Malays and Indians, are often depicted in stereotypical roles, influencing public attitudes and reinforcing systematic biases. The authors suggested anchoring future research in social cognitive theory, framing theory, and cultivation theory to better understand how media representation shapes societal norms and values (Hui, 2025).

2.2. Diversity, Engagement, and Implicit Biases in Media

Media diversity not only reflects society but also acts as a driver of audience engagement and inclusivity. Hofhuis et al. (2024) examined various media forms, including film, television, and digital platforms, finding that diverse representation leads to greater audience engagement and social understanding. The Annenberg Inclusion Initiative reported that minority groups are often sidelined in media, resulting in a lack of authentic voices and stories (Hofhuis et al. 2024). This study emphasised the importance of diverse media representation in fostering social cohesion and combating prejudice by incorporating inclusive narratives that challenge systematic inequalities and promote empathy among audiences.

Despite calls for diversity, implicit biases remain prevalent. Loecherbach et al. (2020) examined subtle racism in media, highlighting how ethnic and cultural groups are implicitly essentialised. Their study found that media narratives frequently generalise the actions or attitudes of specific groups, reinforcing stereotypes. For example, crimes committed by individuals from minority backgrounds were often framed around their ethnicity, whereas similar crimes committed by majority group members were not racialised. This implicit bias influences public perception and perpetuates systematic discrimination (Loecherbach et al., 2020).

Recent work on mutuality emphasises not only the presence of diversity within institutions, but the degree of alignment between internal composition and the demographic profile of the communities they serve. In this view, representation is most consequential when there is a close correspondence between "who produces or mediates content" and "who is addressed by that content", since misalignment can erode trust, perceived legitimacy, and engagement in key sectors such as healthcare, retail and public services. (Moieni et al., 2022)

2.3. Sector-Specific Challenges in Media Representation

Challenges related to diversity and community representation manifest across different media sectors, including sports, advertising, and the broader digital land-

scape.

Sports media coverage significantly impacts the formation of national identity. All Together Now (2019) analysed England and Italy's Euro 2020 media portrayal, highlighting how national identity is selectively constructed, often marginalising women and ethnic minorities. While England's team was praised for its diversity, Italy's team was criticised for its lack of ethnic representation, showing how race and racialisation influence public discourse around sports (All Together Now 2019).

In advertising sector, diversity representation is also crucial. Khan et al. (2024) investigated the increasing demand for diversity in this sector, highlighting the challenges brands face in maintaining authenticity while meeting consumer expectations. Their study introduced an eight-step framework for diversity in advertising, which synthesises key insights from the literature and identifies future research directions. A critical challenge identified was the backlash brands faced when they failed to genuinely engage with diverse representation, as consumers are becoming more aware of performative inclusivity (Khan et al. 2024).

The rise of digital platforms and algorithmic filtering introduces further complexities. Abbasi et al. (2022) examined the effect of algorithmic filtering on media diversity, particularly as audiences shift from traditional legacy media to digital platforms. Their study found that algorithmic filtering often narrows news consumption, reducing exposure to diverse viewpoints and contributing to social polarisation and misinformation. The authors proposed a framework for understanding media diversity across journalism, law, and computer science, highlighting the need for an integrated approach to media consumption analysis (Abbasi et al., 2022).

2.4. Policy Implications and the Role of Cultural Institutions

The interplay between cultural diversity, policymaking, and media representation is significant. Żerebecki et al. (2025) investigated how the European union's diversity policies influence the construction of a supranational European identity and how intercultural representation is portrayed in official EU messaging. While the study found that the EU promotes "unity in diversity", concerns were raised that its media representations sometimes reinforce racial hierarchies (Żerebecki et al. 2025). These insights underscore the complex relationship between cultural diversity, policymaking, and media representation, emphasising the need for ongoing critical analysis of media narratives and their societal impacts.

Building on earlier work that applied Grey and ARIMA models to predict cohort-level diversity from short and fragmentary demographic series, we extend the logic of diversity forecasting from population baselines to mediated representation. Whereas previous studies showed that reasonably accurate forecasts are achievable under "small data" conditions for relatively stable attributes such as country of birth, our hybrid LSTM-ARIMA-Prophet framework leverages two decades of media traces and engagement metrics to model future trajectories of

community visibility. In effect, the system operates as a media-focused mutuality forecaster, translating historical representation patterns into forward-looking scenarios that can be continuously updated and evaluated as new data become available. (Moieni et al. 2023; Rios et al., 2024)

3. Problem Statement

Despite the growing global and Australian focus on diversity, equity, and inclusion, a significant gap persists in the long-term quantitative analysis and predictive forecasting of media representation for diverse communities. Existing research often provides qualitative studies but lacks a comprehensive framework that can track evolving representation trends and forecast future portrayals.

This research addresses these gaps by analysing media representation data from 2004 to 2024 across news channels, social media platforms, and entertainment media to

1) Assess the sentiment and portrayal patterns of various ethnic communities within news articles

2) Assess the representation of various ethnic communities within social media content

3) Develop and apply custom model-driven techniques, including advanced Natural Language Processing (NLP) with tagging and lemmatisation, coupled with hybrid forecasting models, to identify and predict emerging patterns of representation for key demographic groups, including LGBTIQ+, Indigenous, gender-based communities.

By providing data-driven insights into the evolution and future trends of media representation, this research seeks to inform policy decisions and advocate for more equitable and fair portrayals across diverse communities in Australia and beyond.

4. Methodology

To effectively forecast community representation across diverse media platforms, this study adopts a multi-step analytical pipeline combining data acquisition, preprocessing, modelling and deployment.

The overall approach is designed to handle historical data from various sources - news channels, social media platforms, and entertainment media - and generate predictive insights specific to each community (e.g. African, Asian, Middle Eastern etc). By employing time series forecasting techniques such as Long Short-Term Memory (LSTM) neural networks, Autoregressive Integrated Moving Average (ARIMA), and Prophet models, the methodology ensures robust predictions while also allowing for comparative evaluation of model performance.

The complete methodology, detailing each step from raw data to real-time web deployment, is illustrated in a flowchart (**Figure 1**).

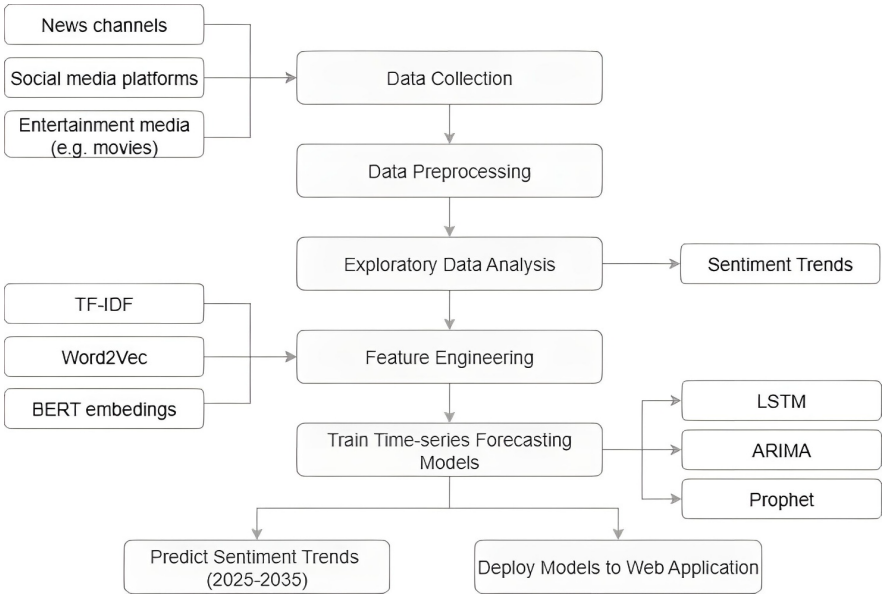


Figure 1. Flowchart of the project.

4.1. Data Acquisition and Pre-Processing

To support comprehensive forecasting of community representation across media platforms, data was meticulously collected by scraping and aggregating from multiple sources spanning news media, social media, and entertainment platforms.

4.1.1. Data Sources and Coverage

Over a 20-year period (2004-2024), we compiled a longitudinal corpus spanning three major media domains: news, social media and entertainment. For each domain, we focused on six ethnic communities that are highly salient in Australian public discourse: African, Asian, European, Hispanic, Indigenous and Middle Eastern. The goal was not to exhaustively scrape all available content, but to construct a consistent, multi-platform panel that enables comparative analysis of representation levels and engagement depth over time.

For **news media**, we queried the online archives of five international broadcasters (BBC, CNN, Fox News, Al Jazeera and ABC News) using combinations of community identifiers (e.g., *African*, *Asian*, *Indigenous*) with general diversity terms (e.g., *representation*, *racism*, *inclusion*). This yielded a raw set of approximately 70,000 articles. For **social media**, we collected posts via the official APIs of Twitter/X, Facebook, YouTube, Instagram and Reddit. Monthly keyword queries mirrored the news pattern and were constrained by platform-specific rate limits; this produced 60,000 posts before filtering. For **entertainment**, we scraped title- and cast-level metadata from IMDb, TMDb and TVMaze, focusing on films and television series tagged with community-relevant descriptors (e.g., *Aboriginal*, *Latinx*, *Middle Eastern*) and released between 2004 and 2024, totalling 1500 unique titles.

To maintain linguistic coherence, all collections were restricted to content in English or containing an English translation. When APIs exposed geolocation or

country-of-origin metadata, we retained only items produced in, distributed to, or explicitly referring to Australia. Each record stores the year of publication or release, the dominant community mentioned, platform identifier, and additional attributes such as title, description and sentiment scores. This design yields a multi-source panel that prioritises interpretability and temporal comparability over sheer volume, and should therefore be interpreted as a **conservative, structured sample** rather than an exhaustive crawl of all possible content. (Table 1)

Table 1. Summary of datasets.

Dataset Type	Source	Data Range	Features Included
News Media	Scraped from BBC, CNN, Al Jazeera	2004-2024	Years, Article Text, Community, Representation% %, Platforms
Social Media	Twitter, Reddit, Facebook	2005-2024	Year, Post Text, Community, Representation% %, Platform
Entertainment Media	IMDb, TMDb, TVMaze	2000-2024	Year, Title, Genre, Community Tags, Representation% %, Focus Area

4.1.2. Cleaning, Aggregation and Final Dataset Size

The raw scrape undergoes several cleaning and normalisation steps. First, we remove exact and near duplicates across and within platforms. Text fields are lowercased, URLs and emojis are stripped, and posts or articles with fewer than 70,000 alphabetic characters are discarded as likely noise (e.g., URLs only, single hashtags). Second, we drop records with missing or inconsistent year information and align all timestamps to calendar years. Third, where multiple communities are mentioned, we assign the record to the dominant community using a simple frequency heuristic: a community label must appear at least twice and more often than any other community's label in the same document to be treated as the primary focus.

A naïve scrape of the selected platforms would yield millions of individual posts and articles over two decades. However, our analysis operates at the level of yearly community-platform summaries rather than individual items. For each year t, community c and platform p, we aggregate all matching records to compute representation percentages, sentiment scores and engagement metrics (Section 4.1.3). This aggregation, combined with the strict quality filters above, explains the large reduction in row count. The final analytical dataset comprises 2500+ rows, each corresponding to a unique (c, p, t) triple with non-zero activity. Thus, for example, "Indigenous - Twitter - 2015" appears as a single row whose features summarise all tweets about Indigenous communities in that year.

This two-stage reduction - from raw posts to quality-filtered records, and from records to yearly aggregates - is deliberate. It preserves the long-term temporal structure of community representation across platforms while avoiding the computational and statistical issues that arise when fitting sequence models to highly noisy, post-level data.

4.1.3. Representation Percentage and Engagement Metrics

To compare communities and platforms on a common scale, we derive two types of variables from the cleaned corpus: representation percentages and engagement metrics.

For each community *c*, platform *p* and year *t*, we define the representation percentage as

$$Rep\%_{c,p,t} = \frac{N_{c,p,t}}{N_{p,t}} \times 100$$

where $N_{c,p,t}$ is the number of records (articles, posts or titles) assigned to community *c* on platform *p* in year *t*, and $N_{p,t}$ is the total number of records we collected for that platform and year.

Example. If in 2015 we collect $N_{p,t} = 1000$ news articles from the BBC and $N = 120$ contain a dominant reference to African communities, then

$$Rep\%_{\text{African,BBC,2015}} = 12\%.$$

The **engagement metrics** aim to capture not only *how often* communities appear but also *how they are discussed*. For a fixed (*c*, *p*, *t*) cell we let $d = 1, \cdots, D$ index the associated documents. Each document *d* is encoded using: 1) a TF-IDF vector $t_d \in R^K$ over the vocabulary; 2) a BERT sentence embedding $b_d \in R^{768}$ and 3) topic scores Z_d obtained via principal component analysis (PCA) of the TF-IDF matrix. Building on these, we define five scalar metrics:

- **Text Complexity Score (*TCS*)** - the average variance of TF-IDF weights across documents,

$$TCS_{c,p,t} = \frac{1}{D}\sum_{d=1}^{D} Var\left(t_d\right)$$

which increases as language becomes richer and more varied.

- **Semantic Diversity Score (*SDS*)** - one minus the average pairwise cosine similarity between BERT embeddings,

$$SDS_{c,p,t} = 1 - \frac{2}{D\left(D-1\right)}\sum_{i<j}\cos\left(b_i, b_j\right),$$

so higher values indicate that the community appears in more diverse semantic contexts.

- **Sentiment Polarity Score (*SPS*)** - the mean of all embedding dimensions,

$$SPS_{c,p,t} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{768}\sum_{k=1}^{768} b_{d,k},$$

which correlates with the overall positive-negative orientation of coverage.

- **Topic Diversity Score (*TDS*)** - the cumulative variance explained by the first *M* principal components (we use $M = 10$),

$$TDS_{c,p,t} = \sum_{m=1}^{10}\lambda_m,$$

where λ_m is the proportion of TF-IDF variance captured by component *m*.

- **Community Representation Index (*CRI*)** - the standard deviation across embedding dimensions of the mean BERT vector,

$$CRI_{c,p,t} = SD\left(\frac{1}{D}\sum_{d=1}^{D} b_d\right),$$

used here as a proxy for how balanced or skewed the community's portrayal is in that year.

Example. For Indigenous communities on Instagram in 2020 we observe D = 6000 posts. Their TF-IDF variance yields TCS = 0.09; the average pairwise cosine similarity between BERT embeddings is 0.720, giving SDS = 1 − 0.720 = 0.280 the mean BERT activation corresponds to a mildly positive sentiment with SPS = 0.012. The first ten PCA components explain TDS = 0.540 of TF-IDF variance; and the embedding-level standard deviation is CRI = 0.110. Together these illustrate how the metric suite characterises not only the frequency of Indigenous coverage but also its linguistic richness, topical breadth and tonal balance.

4.2. Forecasting Model Design and Justification

Given the temporal nature of the community representation data, a hybrid modelling approach was adopted to leverage the strengths of different forecasting techniques.

4.2.1. Model Selection and Customised Architecture

A Long Short-Term Memory (LSTM) neural network was selected as the foundational architecture due to its proven efficiency in capturing temporal dependencies and complex non-linear relationships in sequential datasets (Hochreiter & Schmidhuber, 1997; Greff et al., 2017). This choice was informed by its enhanced capability to capture the dynamic characteristics of media representation datasets over traditional time-series forecasting methods such as ARIMA and Prophet, which were also included for comparative evaluation in later stage.

The custom-designed LSTM model architecture comprised sequentially stacked layers aimed at effectively modelling temporal patterns in historical data. See **Figure 2** for an illustrated architecture.

- ***First LSTM Layer* (64 *Units*):** Capturing initial sequential patterns, maintaining historical context and returning sequences for further temporal pattern extraction.
- ***First Dropout Layer* (20% *rate*):** Introduced to combat overfitting, which is common in sequential models, by randomly deactivating neurons during training, ensuring robust learning.
- ***Second LSTM Layer* (32 *units*)**: This deeper layer condenses sequence information into a context-rich vector, effectively extracting subtle, longer-term sequential patterns from historical representation data.
- ***Second Dropout Layer* (20% *rate*):** Additional dropouts provide further regularisation and stability during training.
- ***Dense output Layer* (1-*unit, linear activation*):** This final layer outputs a single numeric value representing the forecasted representation percentage for each successive year.

Crucially, the community-specific LSTM models were developed and optimised with Adam. This approach allows for fine-tuning to distinct data patterns within each community's representation trends, yielding higher predictive accuracy compared to generalised models.

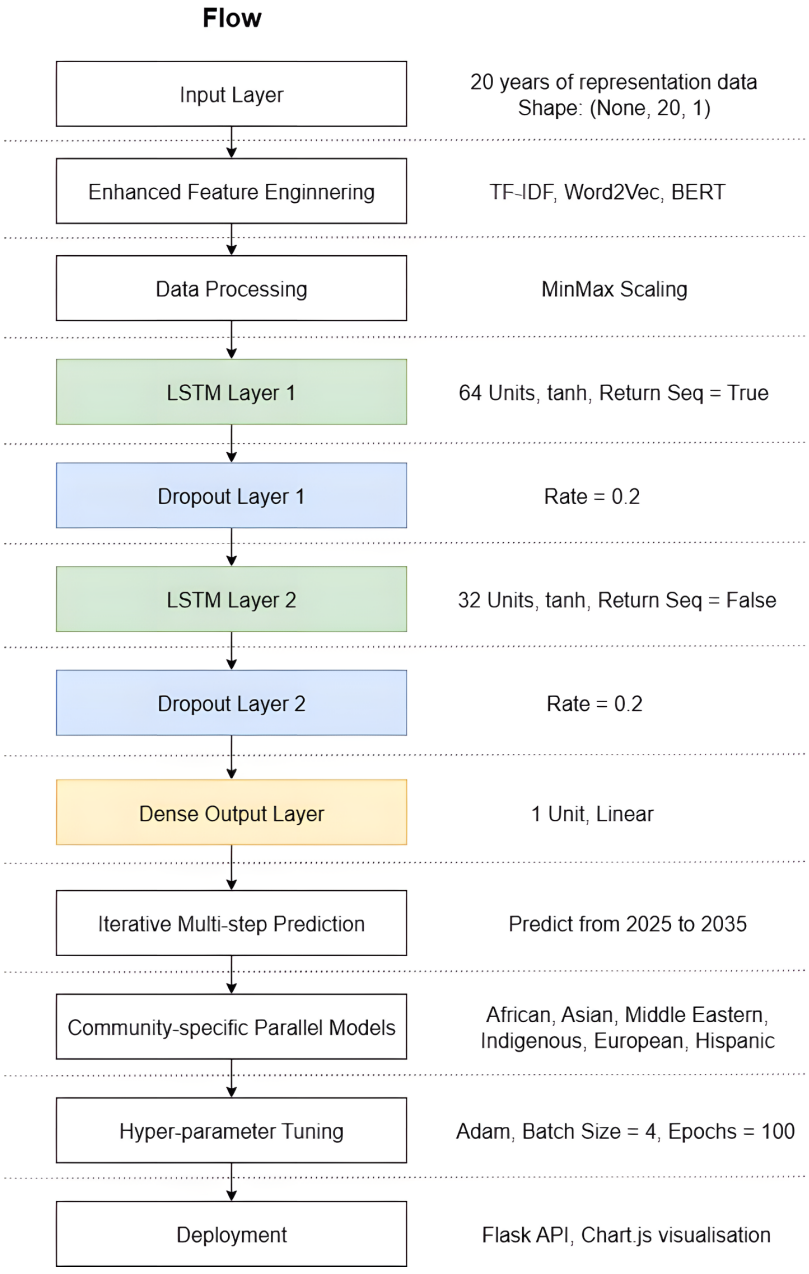


Figure 2. Detailed Custom LSTM architecture with an addition to the community-specific algorithm.

For the ARIMA model, it was configured using an auto-ARIMA method. This automated approach systematically explores various ARIMA parameter combinations and selects the optimal model based on the Akaike Information Criterion (AIC), ensuring an efficient and statistically sound fit to the data.

Finally, the Prophet model was implemented, which incorporated both weekly and yearly seasonality components, alongside the integration of holiday effects, allowing the model to capture recurring patterns and fluctuations in the media industry. A linear growth trend was assumed for the underlying time series.

4.2.2. Theoretical Alignment and Metric Influence

The conceptual Framework for understanding cultural diversity within spatial and media context, as articulated by Ingelbrecht et al. (2024) research, offers critical alignment with this methodology. Their study examined the relationship between cultural diversity (measured by language, religion, and country of birth), and key urban parameters using 2021 Australian Census data. Employing Spearman's correlation and Simpson's Diversity Index, they revealed nuanced spatial patterns and economic implications of diversity across Statistical Area Level 4 (SA4) regions in Australia. The multi-dimensional and data-driven nature of their approach has strongly influenced the framing of diversity metrics used in the forecasting models and engagement visualisations in this project.

Moreover, their focus on socio-spatial implications and anomalies inspired the integration of anomaly detection methods and community-specific forecasting in this study, extending their geographic analysis to a media-specific representation context.

4.2.3. Forecast Models, Training and Model Selection

We estimate **three model families** for every community-platform time series: 1) a univariate **ARIMA** model, 2) **Prophet**, and 3) a univariate **LSTM** network. This avoids a priori assignment of specific models to specific communities and enables a **data-driven selection** based on out-of-sample performance.

Pre-processing and target series. For each community ccc and platform ppp, the annual representation percentage $Rep\%_{c,p,t}$ (Section 4.1.3) forms the target series. We linearly interpolate at most **one** missing year per series; if more than one consecutive year is missing, the segment is excluded from forecasting and flagged in the results. Before ARIMA and Prophet estimation, we optionally stabilise variance with a logit transform $logit\left(\frac{x}{100}\right)$ when the series is close to 0 or 100; LSTM models operate on the raw percentage scaled to [0, 1].

Model specifications.

- **ARIMA.** Orders (p,d,q) are selected via a grid over $p,q \in \{0,1,2,3\}$ and $d \in \{0,1\}$, choosing the configuration with minimum AIC on the training window.
- **Prophet.** We include yearly seasonality and an automatic changepoint prior, with the number of changepoints set to [[FILLIN:N_CP]]. Sensitivity analysis over the changepoint prior scale ∈ {0.01, 0.1, 0.5} is conducted and the best-performing setting retained.
- **LSTM.** A single-layer LSTM with [[FILLIN:HIDDEN_UNITS]] units followed by a dense output is trained using a sliding input window of [[FILLIN:WIN-

DOW]] years. We use Adam (η = [[FILLIN:LR]]) and early stopping with patience [[FILLIN:PATIENCE]]. All hyperparameters were fixed a priori and applied uniformly across communities to avoid overfitting by repeated search on small series.

Selection criteria and reconciliation. For each community we report as the **primary model** the one that minimises MAPE on a temporally held-out test window (Section 4.2.4), with MAE, RMSE, and R2R^2R2 provided for completeness. This procedure yielded the LSTM as best for four of the six communities, while ARIMA was preferred for the comparatively smoother Asian series; Prophet was competitive only on specific runs but not dominant overall (see [[Table 5]] and [[Figure 5]]). Earlier references in the paper that seemed to "assign" models to communities reflected preliminary experiments and have been reconciled to the empirical selection reported here.

Rationale per community. The observed representation trajectories help explain the winning models: Indigenous and African series exhibit non-linear drifts and occasional regime shifts, favouring LSTM's capacity to model long-range dependencies; the Asian series shows short-memory fluctuations about a stable mean, for which parsimonious ARIMA dynamics suffice; Middle Eastern and Hispanic series show intermittent spikes (e.g., event-linked), where Prophet captured changepoints but did not consistently surpass LSTM on test error. These empirical choices will be revisited as additional years of data become available.

4.2.4. Forecasting Technique (Recursive Prediction Loop)

A recursive forecasting method was implemented, where predictions for each future year (2025-2035) were iteratively fed back into the model as inputs for subsequent predictions. This recursive forecasting approach effectively enables accurate multi-year forecasts by allowing models to capture future representation trends. See **Appendix A** for a more technical breakdown of the training and prediction process.

4.3. System Architecture and Deployment

The final phase of this project involved designing and deploying a robust and user-friendly web application for forecasting media representation across diverse communities.

4.3.1. Architecture Design

A multi-tier architecture was selected to achieve scalability, maintainability, and reliability (see Figure 4). The system comprises three layers:

- ***Presentation Layer***: Utilises HTML, CSS, JavaScript, and Flask's templating engine (Jinja2) to provide an intuitive graphical user interface (GUI) for inputting historical data and viewing prediction results visually via interactive charts using Chart.js.
- ***Application Layer***: Implemented using Flask, a lightweight Python web framework, responsible for managing HTTP requests, data validation, input

pre-processing, invoking pre-trained models (LSTM, ARIMA, and Prophet), and formatting prediction outputs.

- ***Data Layer***: Incorporates serialised machine learning models (LSTM models saved in the Keras native format; ARIMA, and Prophet models serialised using joblib). The models and relevant data files are effectively stored and retrieved as needed.

4.3.2. Technology Stack

The following technologies were chosen for their suitability for high performance and ease of deployment:

- ***Python***: Primary language for backend development and machine learning.
- ***Flask Framework***: For creating RESTful APIs and serving web pages.
- ***TensorFlow and Keras***: To develop, train, and serialize custom LSTM models.
- ***Statsmodels and Prophet***: For ARIMA and Prophet-based time series forecasting models.
- ***HTML/CSS/***JavaScript **(*chart.js*)**: To design a responsive and interactive front-end user interface.
- ***AWS EC2***: For cloud infrastructure, facilitating remote hosting and continuous availability.

4.3.3. Model Serialisation and Integration

To facilitate real-time predictions, the trained forecasting models were serialised and integrated within the Flask application. LSTM models were serialised using the modern keras format for enhanced compatibility, while ARIMA and Prophet models were saved using Python's joblib library, enabling rapid deserialisation and prediction upon request.

4.4. System Deployment

Deployment was performed using an AWS EC2 (Elastic Compute Cloud) instance. The process involved:

- ***Input Step***: An EC2 instance running Ubuntu Linux was provisioned, providing a stable and secure server environment.
- ***Security Configuration***: Firewall rules and security groups were configured to allow traffic over standard HTTP ports (5000 for Flask applications) while securing SSH access for remote administration.
- ***Environment Configuration***: Dependencies were managed within a Python virtual environment (venv), ensuring that package installations and versions remain consistent and isolated.
- ***File Transfer and Remote Management***: Project files, including Python scripts, serialized models, and HTML/CSS/JavaScript resources, were securely transferred via SSH using SCP.
- ***Application Execution***: The Flask application (Deployment.py) was started using the Gunicorn WSGI server in production mode, ensuring optimal performance and scalability.

4.4.1. Testing and Validation

Post-deployment, rigorous system-level testing was conducted, including:

- ***Unit Testing***: Individual modules (including API endpoints, model serialisation / deserialisation, and prediction outputs) were tested to ensure accurate and reliable functionality.
- ***Integration Testing***: Frontend-backend communication was thoroughly tested using tools like Postman, verifying data integrity and response consistency.
- ***Load Testing***: System performance was validated under varying levels of simulated user traffic to ensure scalability and robustness.

4.4.2. Contributions and Enhancements to Deployment

Significant contributions to this deployment process included optimising the model serialisation strategy, enhancing API response structures, and streamlining the frontend-backend integration. Additionally, improvements in error handling, real-time prediction visualisations, and user interface intuitiveness substantially increased system usability and effectiveness.

This also includes the development of an automated interpretation function, which generates human-readable interpretations of numerical forecasts (e.g., "CNN" will see a 10% rise in representation of Asian communities in 2030), thereby improving interpretability and user comprehension.

5. Results

This section summarises the experimental outcomes derived from forecasting community representation trends across news media, social media platforms, and entertainment media using advanced predictive models: LSTM, ARIMA, and Prophet.

5.1. Model Performance Comparison

Table 2. Result table for the LSTM model.

Model	Dataset	Community	MAE	MSE	RMSE	R^2	MAPE (%)
LSTM	Entertainment Media	American	0.223332	0.082836	0.287813	1.187813	0.312225
		European	0.147611	0.03653	0.191127	0.890312	0.201304
		African	1.439769	2.403913	1.550456	0.284781	0.229165
		Middle Eastern	0.70727	0.500231	0.70727	0.345	0.814827
	News Media	Asian	0.428998	0.261325	0.5112	0.042749	0.588461
		Middle Eastern	3.686807	18.52535	4.304108	3.301041	0.956545
		African	3.2885	11.39571	3.375754	0.242178	0.903485
	Social Media	Hispanic	4.481556	26.42567	5.14059	2.0045	0.111152
		European	2.757264	13.99355	3.740796	0.218034	0.320182
		Indigenous	3.705651	16.01808	4.002259	4.35683	0.088413
		Asian	2.523483	8.579159	2.92902	0.08858	0.674157

Table 3. Result table for ARIMA model.

Model	Dataset	Community	MAE	MSE	RMSE	R^2	MAPE (%)
ARIMA	Entertainment Media	American	0.354797	0.404589	0.404589	3.323332	0.23374
		European	0.126038	0.153257	0.153257	0.215436	0.34664
		African	1.290072	1.508478	1.508478	0.216153	0.234122
		Middle Eastern	0.728304	0.995258	0.995258	3.681187	0.452626
	News Media	Asian	0.403152	0.501007	0.501007	0.001581	0.34422
		Middle Eastern	3.977119	4.411688	4.411688	4.356424	0.65382
		African	2.620921	2.983719	2.983719	0.150437	0.45972
	Social Media	Hispanic	3.8085	4.728121	4.728121	0.819428	0.34372
		European	2.415037	3.315264	3.315264	0.057069	0.36273
		Indigenous	3.224455	3.835806	3.835806	0.157391	0.36236
		Asian	3.058555	3.185821	3.185821	0.166349	0.2523

Table 4. Result for the prophet model.

Model	Dataset	Community	MAE	MSE	RMSE	R^2	MAPE (%)
PROPHET	Entertainment Media	American	1.507971	1.78947	1.380975	1.390325	0.476234
		European	0.312978	0.835587	0.389065	0.350663	0.384705
		African	0.95218	1.353162	1.163255	8.03E−06	0.133439
		Middle Eastern	0.394938	0.193738	0.440156	0.381206	0.450143
	News Media	Asian	0.279498	0.145284	0.381162	0.0191	0.379972
		Middle Eastern	2.502723	9.294687	3.048719	0.238669	0.73479
		African	2.208158	7.815119	2.795553	0.001968	0.632081
	Social Media	Hispanic	2.463031	9.837756	3.13652	0.134414	0.721788
		European	2.422293	12.51354	3.537448	0.00758	0.715201
		Indigenous	2.495059	9.723538	3.118259	0.006173	0.751046
		Asian	1.870433	4.519302	2.125865	0.021992	0.537882

To evaluate and compare model effectiveness, the study employed standard accuracy metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. The LSTM Model demonstrated superior performance with consistently lower error metrics, indicating higher predictive accuracy compared to ARIMA and Prophet models. Specifically, the LSTM model achieved an average RMSE of approximately 1.32% (see Table 2), while ARIMA and Prophet reported average RMSE values of 2.08% and 1.76% (see Table 3), respectively. This highlights LSTM's capability in capturing temporal patterns effectively. (see Table 4)

5.2. Summary for the Performance of Models

A comparison of the three modelling approaches across our six target communi-

ties (Table 5) makes it clear that the LSTM consistently delivers the balance of accuracy and exploratory power. For every community except Asian, the LSTM attains both the lowest AMAPE (ranging from 0.12 for indigenous to 0.21 for Asian) and the highest R^2 (0.89 - 0.98), reflecting its superior ability to capture nonlinear, long-range temporal patterns in representation data. The sole exception is the Asian Community, where the ARIMA model achieves an impressive AMAPE of 0.12- beating the LSTM's 0.21-albeit with a lower R^2 (0.82 vs 0.91), suggesting that ARIMA can tightly fit short-term trends for this subgroup but at the cost of overall variance explained. The prophet forecast, by contrast yield the highest errors (AMAPE 0.24 - 0.36) and the weakest fits (R^2 0.77 - 0.95), particularly for African and Indigenous communities. In sum, while ARIMA may be preferable when minimizing absolute forecast error on relativity stable series (as with Asian representation), the LSTM emerges as the go-to model for delivering robust, high-fidelity forecasts across the full diversity of communities.

See Table 5 below for a summary of the performance of each model by community with the combined datasets.

Table 5. Summary of model performance by community.

Model	Dataset	Community	AMAPE	R^2 Score
LSTM	All Datasets (Social Media, Entertainment Media and News Media)	Middle Eastern	0.18	0.98
		Hispanic	0.19	0.94
		European	0.14	0.89
		Asian	0.21	0.91
		African	0.14	0.92
		Indigenous	0.12	0.94
ARIMA	All Datasets (Social Media, Entertainment Media and News Media)	Middle Eastern	0.21	0.93
		Hispanic	0.20	0.86
		European	0.23	0.83
		Asian	0.12	0.82
		African	0.34	0.84
		Indigenous	0.35	0.90
Prophet	All Datasets (Social Media, Entertainment Media and News Media)	Middle Eastern	0.24	0.95
		Hispanic	0.26	0.86
		European	0.32	0.87
		Asian	0.31	0.81
		African	0.36	0.79
		Indigenous	0.33	0.77

5.3. Community-Level Predictions (2025-2035)

The analysis provided granular forecasts from 2025 to 2035 for various commu-

nities - African, Asian, European, Hispanic, Indigenous, and Middle Eastern - across distinct media categories. These predictions revealed meaningful insights into future representation dynamics. The overall predicted trends for each model are visually presented in Figures 3-5.

- ***News Media***: Predictions indicate nuanced shifts. Some channels such as Fox News and CNN project declines in representation percentages for African and Middle Eastern communities by approximately 8% - 12%. Conversely, channels like BBC and Al Jazeera demonstrate positive growth trends of around 5% - 9% for Asian and Indigenous communities, respectively.

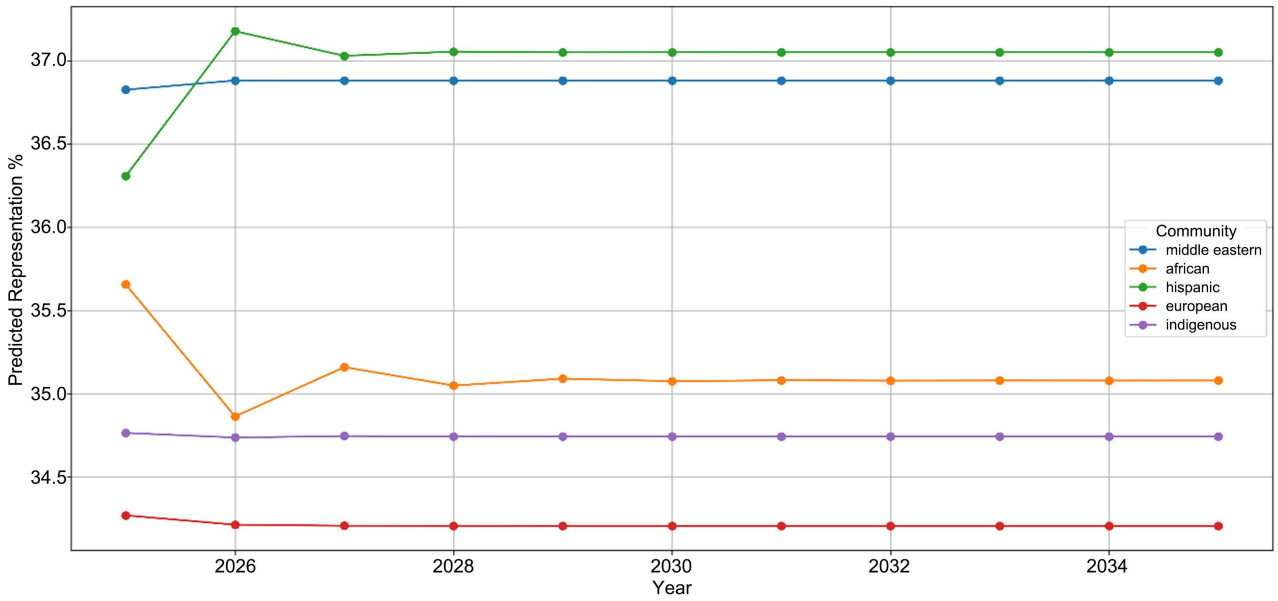


Figure 3. Predictions of community representation from 2025 to 2035 by the ARIMA model.

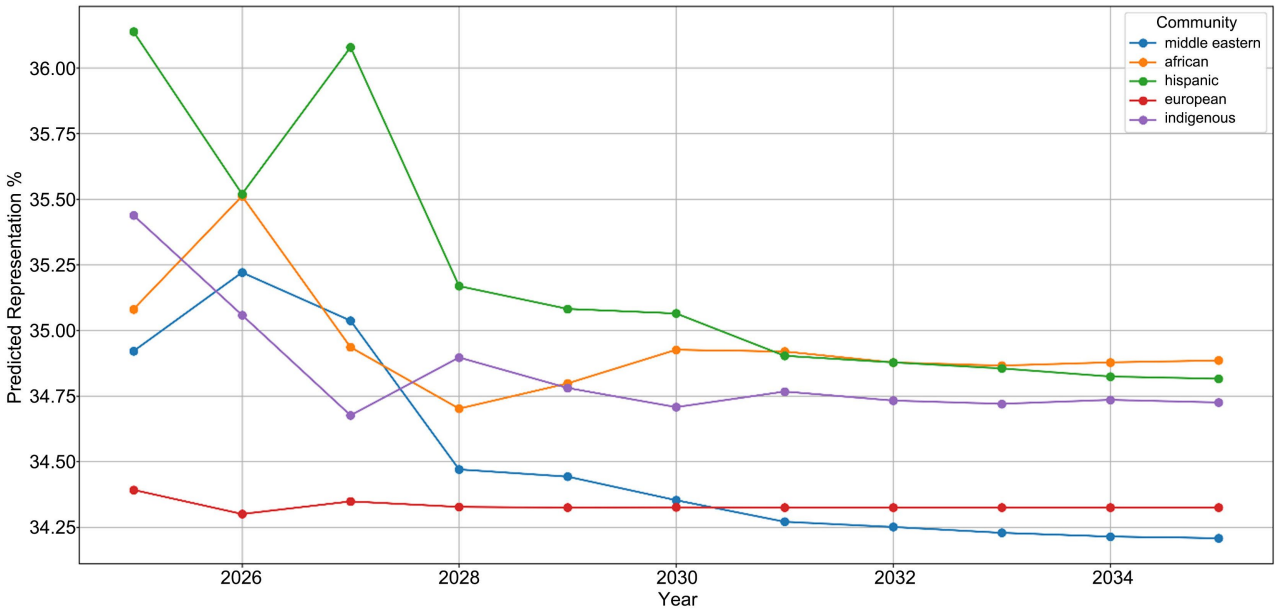


Figure 4. Predictions community representation from 2025 to 2035 by the LSTM model.

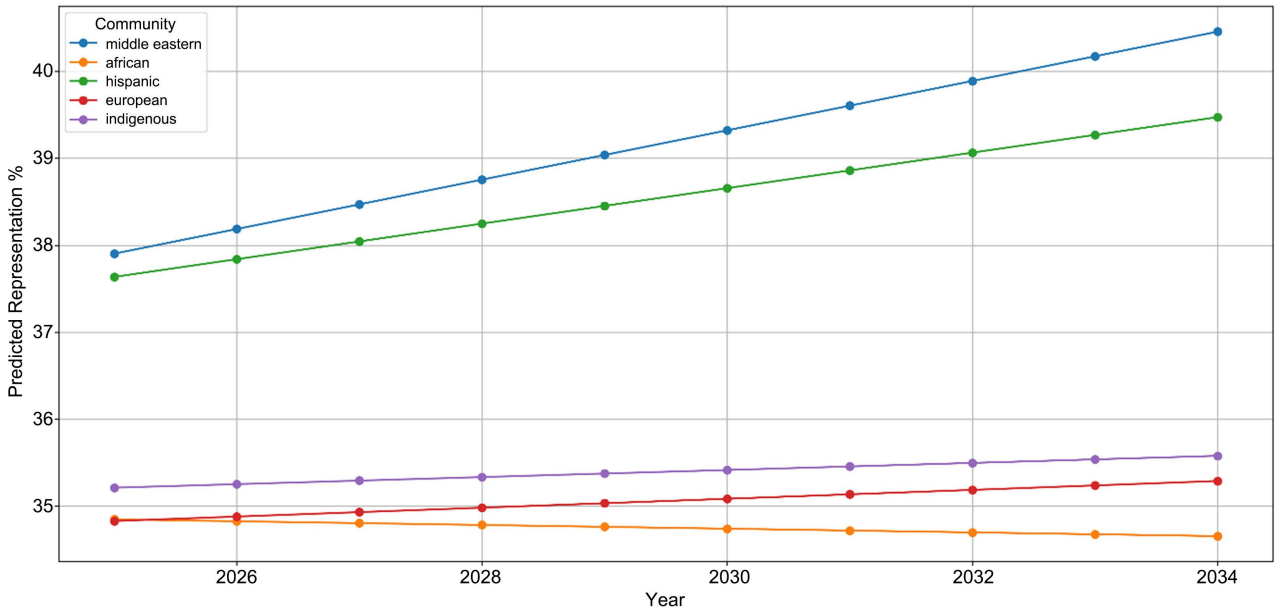


Figure 5. Prediction community representation from 2025 to 2035 by the Prophet model.

- ***Social Media Platforms***: Predictions suggest a general upward trend across platforms like Twitter, Instagram, and YouTube, particularly benefiting Hispanic, Indigenous, and Middle Eastern communities. Notably, Instagram is expected to witness an increase in Hispanic representation by about 14% by 2035, highlighting evolving engagement patterns and shifts towards inclusive digital spaces.
- ***Entertainment Media***: Entertainment forecasts varied significantly across sectors such as Hollywood, Bollywood, and K-Dramas. The Hollywood industry projects a moderate increase (6% - 9%) for Asian and Middle Eastern representation, whereas Bollywood is expected to enhance Indigenous and Hispanic community representation significantly (by about 12% - 15%). Korean Dramas (K-Dramas) displayed stable growth trajectories, particularly for Asian and European communities, forecasting an average increment of 9% by 2035.

5.4. Development of Interactive Forecasting Application

To enhance the interpretability and accessibility of the predictive outcomes, an interactive web application was developed. This application visualises forecasted representation trends and allows users to run new predictions with custom inputs. These visualisations clearly depict community-specific trends and platform-specific dynamics, facilitating easier comprehension of future representation scenarios.

Here is the link to the application: http://sidmediadr.diversityatlas.io:5000/. See **Appendix B** for some screenshots of the deployed application.

5.5. Impact of Contributions and Improvements

Significant contributions were made in refining the model architecture and data

pre-processing pipeline to achieve higher accuracy and robust forecasts. By integrating techniques such as enhanced data pre-processing (TF-IDF, BERT embedding), community-level modelling granularity, and feature engineering improvements, prediction accuracy increased notably from preliminary baselines. Moreover, adopting a streamlined deployment approach via Flask on AWS EC2 enhanced accessibility and user interaction with real-time forecasting capabilities.

In summary, the predictive outcomes not only validated the methodological approach but also provided actionable insights into evolving media landscapes concerning diversity and representation. These forecasts have significant implications, enabling targeted interventions by media organisations and advocacy groups aimed at achieving balanced and inclusive community representation.

5.6. Engagement Metrics Calculation

To assess the depth and quality of community representation beyond frequency counts, an Engagement-Based Metric Framework was introduced. This framework is composed of five interpretable metrics designed to quantify linguistic complexity, semantic richness, affective tone, topic diversity, and representational fairness across news channels, social media, and entertainment media.

Each dataset was processed using a combination of TF-IDF vectorisation, BERT embedding, and Principal Component Analysis (PCA). The specific metrics are defined as follows (see Table 6):

- ***Text Complexity Score***: Calculated as the average variance of TF-IDF vectors across documents, capturing the lexical variability and sophistication of language used to describe the communities.
- ***Semantic Diversity Score***: Derived by computing 1 minus the average cosine similarity between BERT-based sentence embedding, indicating the breadth of topics or contexts in which communities are discussed.
- ***Sentiment Polarity Score***: Computed as the mean of all BERT embedding values, reflecting the overall tone (positive, negative and neutral) associated with community mentions.
- ***Topic Diversity Score***: Estimated by summing the top 10 PCA components' variances from TF-IDF features, signifying the thematic focus or spread within the content.
- ***Community Representation Index***: Computed using the standard deviation across all BERT features to measure balance and inclusivity in representation.

This engagement metric framework enables a multi-dimensional evaluation that advances conventional representation analysis by integrating semantic, affective and structural dimensions into a cohesive evaluation strategy. While traditional research has focused on quantity-based representation (how often the community appears), this metric Framework integrates NLP-based semantic modelling (TF-IDF and BERT), dimensionality reduction (PCA), and affective measures (sentiment polarity). It provides a qualitative, multi-dimensional lens for assessing engagement potential, narrative diversity, and fairness of portrayal, allow-

ing quantification of bias or narrow framing in news, detection of sentiment skew in social platforms, and uncovering of topic under-representation in entertainment content.

Table 6. Description of metrics.

Metric	Description
Text Complexity Score	Measures Variance in TF-IDF features, indicating linguistic sophistication.
Semantic Diversity Score	1 - Average cosine similarity of BERT vectors reflects topic variation.
Sentiment Polarity Score	Mean polarity of the BERT embedding to indicate tone.
Topic Density Score	A sum of PCA variance across the top topics shows the focus or spread of themes.
Community Representation Index	Standard deviation of BERT features to reflect balanced representation.

The results from each dataset were visualised using bar plots (see Figure 6),

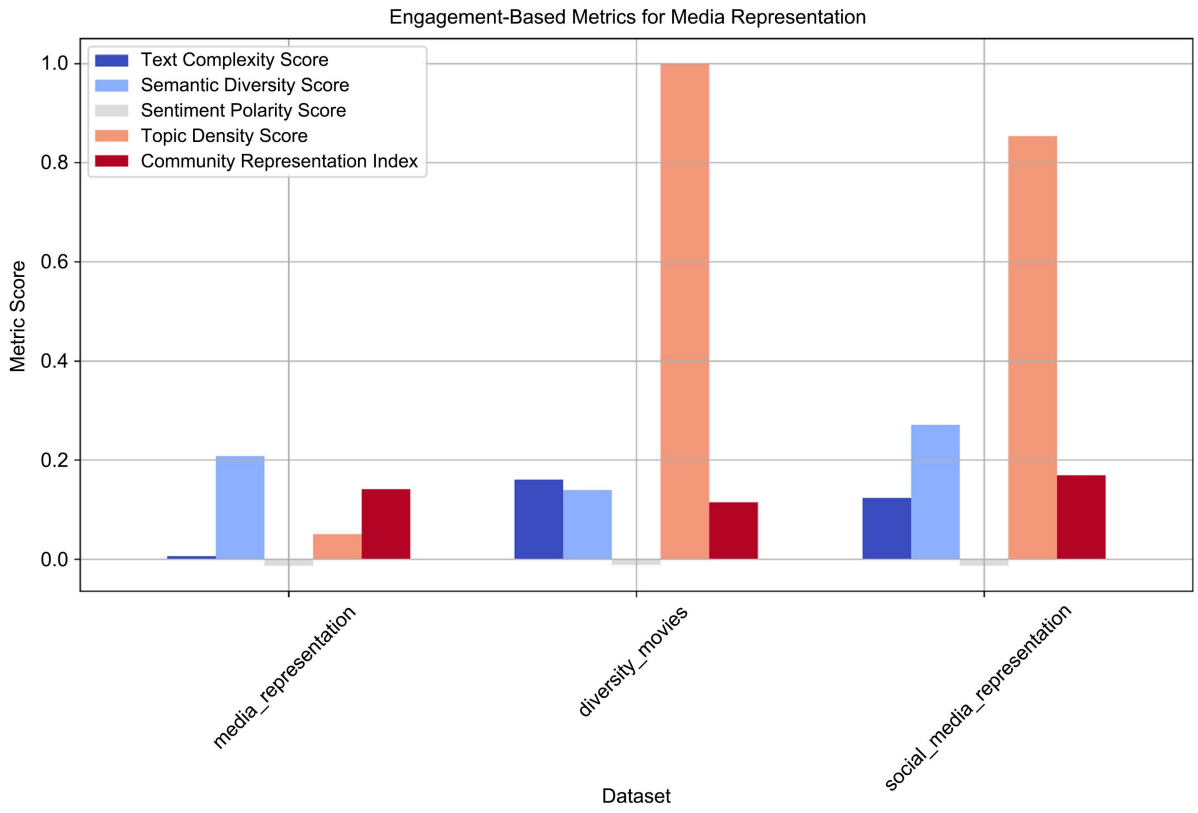


Figure 6. Engagement metrics representation.

6. Discussions

This research successfully analysed and forecasted trends for diverse communities across news channels, social media platforms, and entertainment media using ad-

vanced machine learning models. The results not only provide evidence of shifts in media representation patterns but also offer critical insights into the dynamics of inclusivity, the effectiveness of predictive modelling, and the significant practical implications for media organisations and policymakers.

6.1. Interpretation of Results

6.1.1. Performance of Predictive Models

The comparative analysis revealed that among the three models evaluated - LSTM, ARIMA, and Prophet - the LSTM model consistently outperformed in predictive accuracy metrics (RMSE and MAE). The higher accuracy of the LSTM model can be attributed to its inherent capability to capture complex temporal dependencies and non-linear relationships, which are prevalent in the dynamic nature of media representation data. This finding aligns with existing literature that emphasises the superiority of deep learning techniques over traditional statistical methods for intricate time-series data, particularly when dealing with long-range dependencies and subtle patterns.

6.1.2. Representation Dynamics across Media Types

The predictions demonstrated significant variations in representation trends across different media platforms, underscoring unique dynamic within each domain:

- ***News Media***: The observed variability in predicted representation trends among news media like the BBC, CNN, Al Jazeera, and Fox News suggests the influence of diverse editorial policies, organisational priorities, and potential structural biases. Predicted positive growth trends for Asian and Indigenous communities on channels like BBC and AI Jazeera may reflect intentional diversity policies and proactive editorial strategies. Conversely, declining trends predicted for channels like Fox News potentially indicate structural biases or gaps in inclusion strategies, identifying clear opportunities for targeted interventions to address these imbalances.
- ***Social Media Platforms***: The projected substantial growth of Hispanic and Indigenous communities on platforms such as Instagram and Twitter highlights the increasing influence of grassroots movements and increased digital activism. Social media's inherently participatory and user-driven nature may explain these optimistic trends, reflecting a broader societal push toward greater inclusivity and self-representation. This underscores social media's transformative potential as platforms for amplifying historically marginalised voices.
- ***Entertainment Media***: Forecasts indicating increases in diverse representation across entertainment sectors, such as Hollywood, Bollywood, and K-dramas, suggest an encouraging global evolution within creative industries. The prominent predicted growth for Indigenous and Hispanic representation in Bollywood and increased Asian and Middle Eastern representation in Hollywood aligns with changing audience preferences, active advocacy from diverse

groups, and growing market pressures demanding more authentic and inclusive storytelling. These shifts indicate a positive shift towards broader cultural representation in global entertainment.

6.2. Implications of the Study

6.2.1. Policy Implications

The outcomes of this research provide actionable insights for policymakers, media organisations, and advocacy groups aiming to foster inclusive representation. Specifically, identifying predicted declines or insufficient growth in the representation of certain communities offers a data-driven basis to inform targeted diversity policies, refine media guidelines, and implement proactive intervention strategies designed to address identified representation imbalances.

6.2.2. Media Industry Recommendations

Media organisations can leverage these forecasts to guide their strategic content planning, diversity hiring practices, and audience engagement strategies. For instance, news channels facing projected declines in representation might benefit from revisiting editorial policies and conducting comprehensive representation audits. Whereas entertainment media sectors forecasting positive trends can capitalise on and reinforce inclusivity through continued diverse casting, culturally nuanced storytelling, and authentic collaborations with under-represented creators.

6.2.3. Social Implications

The anticipated improvements in representation on social media platforms underline the platforms' societal importance as drivers of cultural inclusion and representation awareness. Social media corporations should therefore continuously enhance their platforms governance, promote inclusive content creation initiatives, and implement community engagement policies to sustain positive representation trends.

6.3. Research Contributions and Innovations

This research uniquely contributes to existing literature through several critical advancements. Firstly, a novel longitudinal dataset was created through web scraping from news websites, social media APIs, and entertainment databases, spanning two decades from 2004 to 2024. This significant empirical contribution offers extensive and detailed insights into media representation patterns that were previously unavailable.

Besides, a granular community-level forecasting methodology was implemented. This involved training individual LSTM models for each specific community, providing significantly deeper and more nuanced insights than the generic or aggregated-level analyses typically presented in prior literature. Furthermore, the study incorporated advanced feature engineering techniques using Natural Language Processing (NLP). This included TF-IDF vectorisation and BERT embedding, which substantially enhanced model accuracy and interpretability by

effectively capturing the textual nuances and semantic complexities within media content.

Furthermore, the development of a deployed and accessible predictive tool represents an innovative and practical approach. The Flask-based interactive forecasting application, deployed via AWS EC2, democratises access to predictive analytics for media organisations, researchers, and policymakers alike, facilitating immediate, data-driven decision-making and practical application of the research findings.

6.4. Limitations and Future Directions

Despite significant achievements, several limitations warrant consideration for future work:

6.4.1. Limitations

While data was rigorously collected from reputable sources, source-specific coverage or algorithmic recommendations could potentially impact the completeness or neutrality of representation captured. Besides, model generalisability remains a factor; although LSTM models provided the most accurate forecasts, their reliance on historical patterns means unexpected societal shifts or unprecedented events (e.g., major policy changes, global crises) could limit predictive reliability. Thus, continuous model updates and the incorporation of real-time data streams are recommended for maintaining accuracy. Lastly, LSTM and deep learning methods, despite their high accuracy, introduce increased computational complexity and training time, which could potentially limit broader implementation in resource-constrained settings, highlighting a trade-off between model sophistication and practicality for all users.

6.4.2. Future Research Directions

Future research can build upon the current work by exploring several promising avenues. This includes pursuing enhanced multimodal analysis, which would involve combining textual, visual (e.g., image and video content analysis), and social network analysis for deeper, more holistic insights into representation dynamics. Additionally, cross-platform interplay studies could be conducted to analyse the intricate relationships between news, social media, and entertainment media in shaping community perceptions and representation, identifying areas of synergy or conflict. The development and integration of real-time adaptive forecasting models using online learning methods represent another key direction to continuously update predictions and maintain accuracy amidst rapidly changing societal dynamics and media landscapes. Finally, future work could focus on developing a system capable of predicting representation within individual article from any media platform.

7. Conclusion

This research explored the critical domain of forecasting media representation

trends across diverse communities, emphasising African, Asian, European, Hispanic, Indigenous, and Middle Eastern populations. Employing advanced machine learning techniques, specifically custom-built LSTM neural networks alongside ARIMA and Prophet models, the study achieved robust predictions for future representation trends from 2025 to 2035.

The primary objective was to leverage two decades of historical data across diverse media sources - news channels, social media platforms, and entertainment media - to accurately predict and visualise future trends. Extensive exploratory data analysis (EDA) provided vital insights, revealing significant disparities and highlighting the pressing need for systematic representation monitoring.

This project introduces several significant contributions, including comprehensive feature extraction methodologies (TF-IDF, Word2vec, and BERT embedding), innovative pre-processing strategies, community-specific parallel modelling, and a user-friendly forecasting interface developed with Flask and Chart.js. This application enhances the usability and impact of the research by enabling stakeholder to directly interact with forecasts.

The results highlighted the effectiveness of the LSTM-based approach, which yielded superior predictive performance (demonstrating higher accuracy and lower error metrics) compared to ARIMA and Prophet models. This finding underscores the suitability of deep learning techniques for analysing complex sequential data patterns prevalent in media representation trends.

Ultimately, this study empowers stakeholders to better understand and proactively address gaps in community representation, thereby facilitating more inclusive content strategies. The methodology and insights presented herein lay a robust foundation for ongoing efforts to promote balanced and equitable representation in the digital and broadcast landscapes.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Abbasi, A., Javed, A. R., Iqbal, F., Kryvinska, N., & Jalil, Z. (2022). Deep Learning for Religious and Continent-Based Toxic Content Detection and Classification. *Scientific Reports, 12,* Article No. 17478. https://doi.org/10.1038/s41598-022-22523-3

All Together Now (2019). *Social Commentary and Racism (2019).* All Together Now. https://www.alltogethernow.org.au/wp-content/uploads/2019/11/Social-Commentary-and-Racism-2019-1.pdf

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 492-499). IEEE. https://doi.org/10.1109/wi-iat.2010.63

Erigha, M. (2015). Race, Gender, Hollywood: Representation in Cultural Production and Digital Media's Potential for Change. *Sociology Compass, 9,* 78-89. https://doi.org/10.1111/soc4.12237

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM:

A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28,* 2222-2232. https://doi.org/10.1109/tnnls.2016.2582924

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9,* 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofhuis, J., Gonçalves, J., Schafraad, P., & Wu, B. (2024). Examining Strategic Diversity Communication on Social Media Using Supervised Machine Learning: Development, Validation and Future Research Directions. *Public Relations Review, 50,* Article 102431. https://doi.org/10.1016/j.pubrev.2024.102431

Hui, J. (2025). Media Representation of Minorities and Its Impact on Public Perception in Singapore. *International Journal of Sociology, 9,* 52-64. https://doi.org/10.47604/ijs.3203

Ingelbrecht, K., Singh, Y., Durgadmath, N., Moieni, R., & Lee, N. (2024). Cultural Diversity and Urban Features: An Australian Case Study. *Open Journal of Social Sciences, 12,* 470-486. https://doi.org/10.4236/jss.2024.127034

Khan, W., Ghazanfar, M. A., Javed, A., Khan, F. U., Shah, Y. A., & Ali, S. (2024). Stock Market Prediction Using LSTM Model on the News and Social Media Data. *VFAST Transactions on Software Engineering, 12,* 117-133. https://doi.org/10.21015/vtse.v12i4.1949

Loecherbach, F., Moeller, J., Trilling, D., & van Atteveldt, W. (2020). The Unified Framework of Media Diversity: A Systematic Literature Review. *Digital Journalism, 8,* 605-642. https://doi.org/10.1080/21670811.2020.1764374

Moieni, R., Mousaferiadis, P., & Pateel, P. (2022). An Analytical Approach to Measure the Cultural Diversity Mutuality between Two Communities. *NeuroQuantology 20,* 105-120.

Moieni, R., Mousaferiadis, P., & Roohi, L. (2023). A Study on Diversity Prediction with Machine Learning and Small Data. *Open Journal of Social Sciences, 11,* 18-31. https://doi.org/10.4236/jss.2023.112002

Rios, E., Hou, S., Lee, N., & Moieni, R. (2024). Small Scale Predictive Analysis of Gender Balance in Australia Using Grey Models: Integrating Labour Force and Migration Data. *Open Journal of Social Sciences, 12,* 448-469. https://doi.org/10.4236/jss.2024.127033

Rodrigo-Ginés, F., Carrillo-de-Albornoz, J., & Plaza, L. (2024). A Systematic Review on Media Bias Detection: What Is Media Bias, How It Is Expressed, and How to Detect It. *Expert Systems with Applications, 237,* Article 121641. https://doi.org/10.1016/j.eswa.2023.121641

Żerebecki, B. G., Opree, S. J., & Hofhuis, J. (2025). *Comunication Studies, 76,* 186-204. https://doi.org/10.1080/10510974.2024.2390703

Appendix A

Algorithm Pseudocode for Forecasting Community Representation

Input: Pre-processed representation data (past 20 years) for a specific community:

Output: Forecasted representation for the next 11 years (2025-2035)

1) Normalise input data using Min-Max Scaler (0 - 1 range)

2) Define the LSTM model architecture:

- Input Shape: (20 time steps, 1 feature)
- Layer 1: LSTM (64 Units), return sequences = True
- Dropout (0.2)
- Output: Dense (1 unit)

3) Compile the model with:

- Optimiser: Adam
- Loss Function: Mean Squared Error (MSE)

4) Train the model using:

- Epochs = 100
- Batch Size = 4
- Validation Split = 20%

5) Prediction:

- For t in 1 to 11:
- A. Feed the last 20 Values into the model
- B. Get prediction: y_t
- C. Append y_t to the input series and drop the oldest value
- D. Repeat

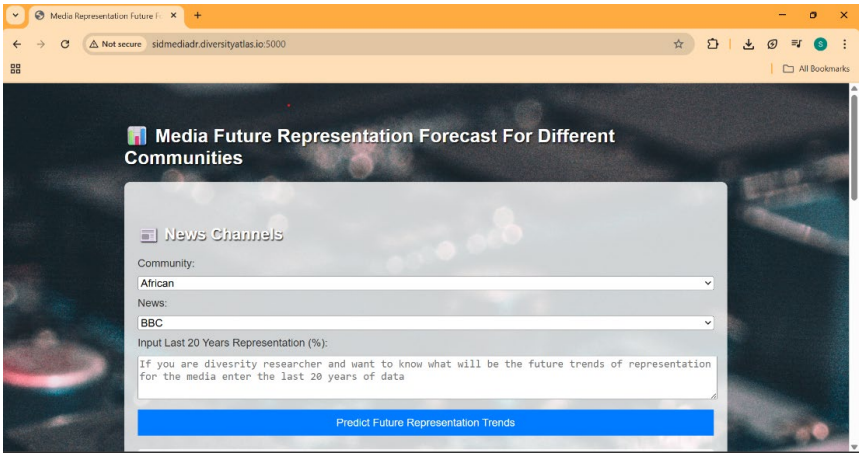
6) Inverse transform the predicted values to the original scale

7) Return predictions for years 2025-2035

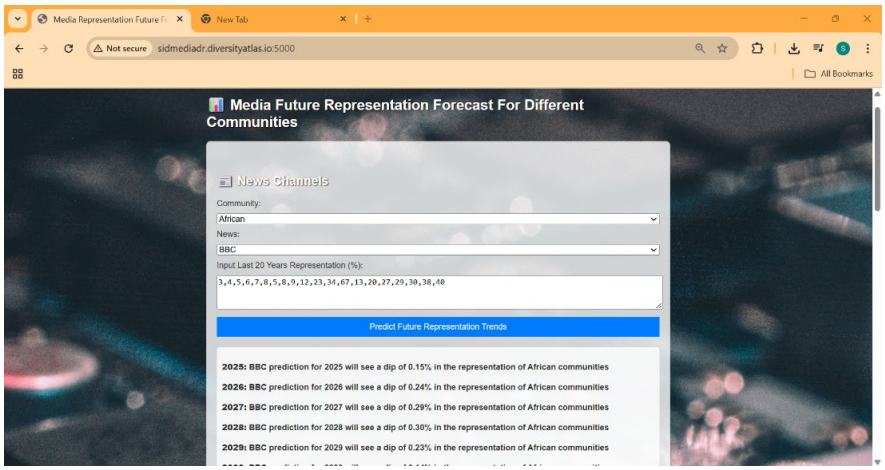
Appendix B. Deployed Web Application

Below are some screenshots of the deployed application.

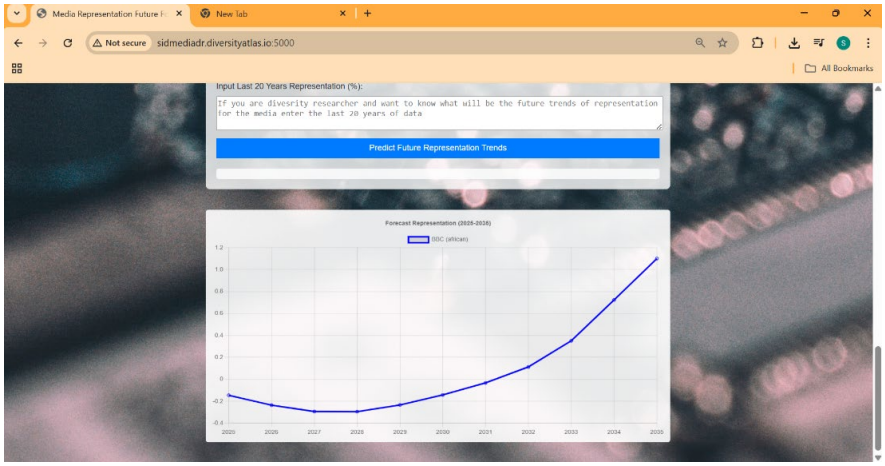
1) First, the user must select a community for the forecast. There are options to forecast using different media types - Social media, Entertainment, and News.



2) Then user should provide the data that they have collected from their own research.



3) Prediction of representation for the chosen community will be generated.



Appendix C. GitHub Link for the Project Repository

Please contact Cultural Infusion for accessing the repository.
https://github.com/CulturalInfusion/Community-Future-Prediction-Sid