IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATED JOURNALISM USING ARTIFICIAL INTELLIGENCE

[1]D. PARAMESWARI, [2]A. POORNIMA, [3]S. PRAVEEN, [4]V. SHARAN

[1]PROFESSOR & HOD, [2]IV - CSE, [3]IV - CSE, [4]IV - CSE

[1]DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING,

[1]JERUSALEM COLLEGE OF ENGINEERING, CHENNAI, INDIA.

Abstract: This research endeavour aims to develop an intelligent news aggregation system skilled at extracting diverse news material from selected blogs and websites using web scraping methodologies. The gathered data undergoes thorough processing and is systematically stored within a structured database framework. To facilitate efficient data retrieval and manipulation, a robust API is devised.

The frontend interface is meticulously designed to ensure user-friendly interaction, presenting aggregated news in a unified manner. Furthermore, the system smoothly integrates social networking sites, automating post generation through predefined templates for effortless sharing of news updates across various channels. By amalgamating state-of-the-art technologies, that project offers a centralized hub catering to news enthusiasts, providing curated content, streamlined database administration, and extended outreach through strategic social media integration.

Keywords: News Aggregation, Data Collection, Web scraping, API Development, Content Summarization, Database Storage, frontend Interface, Data Integration, Information Retrieval, Digital Information, Content Curation, Diverse Sources and Searchable News.

I. INTRODUCTION

This initiative arises in response to the necessity for a streamlined and centralized method for news aggregation, consumption, and distribution. It incorporates a multifaceted approach, including web scraping, API development, database management, frontend design, and social network automation, to construct a comprehensive news ecosystem. Through meticulous web scraping of relevant news content from carefully selected sources, utilizing advanced programming languages and libraries, the project ensures the systematic extraction of essential information such as titles, content, and publication dates. Subsequently, this data is structured into a database, establishing a robust information repository. The creation of a purpose-built API facilitates seamless communication within the system, enabling the frontend to deliver an intuitive and engaging user interface while also facilitating external integrations, thus fostering a dynamic and expandable ecosystem. Crafted with cutting-edge technologies, the frontend enhances the user experience by providing a visually appealing and user-friendly platform for navigating, searching, and exploring curated news content. Moreover, it extends its reach into the realm of social networks by automating the generation and posting of news updates through predefined templates, ensuring a consistent and efficient distribution of carefully selected content across various channels in the ever-changing landscape of information dissemination. This endeavour not only aims to serve as a centralized hub for news enthusiasts but also seeks to empower individuals and organizations with an adaptable tool for staying informed and effortlessly sharing relevant content.

II. OVERVIEW

This paper is dedicated to create an advanced news aggregation and dissemination system, blending cutting-edge technologies to provide a comprehensive solution for information consumption. Commencing with ethical web scraping practices, the system extracts news articles from a selected website, adhering to terms of service and ethical guidelines. A pivotal aspect is the development of a robust API, utilizing frameworks like Flask or Django, to establish seamless communication between the backend and a user-friendly frontend. Crafted with modern web frameworks such as React, the frontend offers an intuitive interface for users to navigate and engage with curated news content. Processed data is efficiently stored in a chosen database system, ensuring optimal organization and swift retrieval. Noteworthy features include templates for social media posts and automation scripts employing social media APIs, facilitating strategic dissemination. Deployment on a web server, complemented by monitoring tools and routine maintenance checks, guarantees accessibility, responsiveness, and system stability. Upholding ethical considerations, including compliance with privacy regulations and security standards, is integral throughout the development process. In essence, this project

aspires to deliver a dynamic and user-centric news ecosystem, serving as a centralized hub for curated content while extending its impact through seamless integration with prominent social media platforms.

III. LITERATURE SURVEY

This paper titled, A scientometric review of automated journalism Analysis and visualization by Xu Z and Lan X, published in 2020, offers a comprehensive analysis of the landscape of automated journalism research. Through a scientometric approach, the authors delve into the status and trends of research in this rapidly evolving field. By meticulously examining a wide range of scholarly literature, the paper sheds light on the adoption, advancements, and challenges of automated journalism. Through systematic analysis and visualization techniques, the authors uncover valuable insights into the distribution of research efforts, key contributors, and emerging themes within the domain. This work not only serves as a valuable resource for scholars and practitioners interested in automated journalism but also provides a roadmap for future research endeavours in this area[1].

Algorithmic journalism - Current applications and future perspectives by Efthimis Kotenidis and Andreas Veglis, published in 2021, offers a detailed exploration of the landscape of algorithmic journalism. With a focus on both current applications and future prospects, the authors delve into the multifaceted challenges and opportunities associated with automated journalism. Through comprehensive analysis and synthesis of existing literature and case studies, the paper elucidates the current state-of-the-art practices in algorithmic journalism, ranging from content generation and curation to audience engagement and data-driven storytelling. Furthermore, the authors meticulously examine the ethical, legal, and societal implications of algorithmic journalism, highlighting the need for transparency, accountability, and fairness in automated news production processes. In addition to critically evaluating the existing implementations, the paper also presents insightful reflections on the potential future directions and advancements in the field. By addressing key challenges and proposing innovative solutions, this work serves as a valuable resource for researchers, practitioners, and policymakers seeking to navigate the complex landscape of algorithmic journalism and chart a course for its responsible and sustainable evolution[2].

Automated journalism - A meta-analysis of readers perceptions of human-written in comparison to automated news Graefe A. and Bohlken N., published in 2020, presents a meta-analysis that synthesizes evidence regarding readers' perceptions of automated news compared to human-written news. The study likely compiles and analyses findings from various research studies conducted in this domain to provide a comprehensive understanding of how readers perceive automated news in terms of credibility, quality, and readability. Through a systematic review and statistical synthesis of empirical data, the paper likely examines factors such as trustworthiness, accuracy, objectivity, and engagement associated with both human-written and automated news articles. By aggregating results from multiple studies, the paper aims to uncover overarching trends and patterns in readers' attitudes towards automated news, offering valuable insights for news organizations and policymakers. Additionally, the meta-analysis may identify gaps in existing research and propose avenues for future studies to further explore the impact of automation on journalism and reader perceptions. Overall, this work contributes to the scholarly discourse on automated journalism by providing evidence-based insights into how readers evaluate and interact with automated news content, thereby informing discussions on its integration and adoption in newsrooms[3].

This paper titled "Data-driven news generation for automated journalism" by Leppanen L, Munezero M, Granroth Wilding M, and Toivonen H., published in 2017, delves into the realm of automated journalism with a particular emphasis on leveraging data-driven approaches for news generation. The work likely examines the methodologies, techniques, and challenges associated with constructing a natural language generation (NLG) system tailored specifically for journalistic purposes. This paper probably explores the field by investigating various aspects, including data collection, preprocessing, analysis, and NLG algorithm design. It may discuss the intricacies of acquiring and processing diverse data sources to automatically generate news articles. Moreover, the authors likely delve into the complexities of designing NLG algorithms capable of transforming structured data into coherent and readable news stories, all while adhering to journalistic standards. Throughout the paper, the authors likely analyze the effectiveness and limitations of data-driven approaches in automated journalism. They may present empirical results or case studies to demonstrate the capabilities and potential challenges of NLG systems in producing news content. Additionally, the paper may discuss key considerations such as accuracy, objectivity, and narrative coherence in automated news generation. By addressing these challenges and proposing potential solutions, the paper likely advances our understanding of automated journalism and provides valuable insights into the opportunities and limitations of data-driven approaches in this domain. It serves as a foundational work that informs future research and development efforts aimed at enhancing the capabilities of NLG systems for journalistic purposes[4].

In this paper "The bright future of news automation" by Carl Gustav Lind, published in 2018, the author offers a compelling exploration of the scope and potential of news automation. Lind delves into various facets of automation within the journalism industry, ranging from content generation to dissemination. By examining recent advancements in technology and their impact on journalistic practices, the paper provides valuable insights into how automation can streamline news production workflows and enhance the efficiency of newsrooms. Moreover, Lind discusses the broader implications of news automation for journalists, news organizations, and audiences alike. This includes considerations regarding job displacement, ethical concerns, and shifts in audience engagement and consumption patterns. Through a balanced analysis of the benefits and challenges of news automation, Lind offers a nuanced perspective on its role in shaping the future of journalism. Overall, the paper serves as a thought-provoking contribution to the ongoing discourse on the intersection of technology and media[5].

The paper "Automated Journalism 2.0 - Event-Driven Narratives" by Donald W. Reynolds, published in 2017, is a seminal work in the field of automated journalism. Reynolds explores the concept of creating stories from simple descriptions, particularly focusing on event-driven narratives. This paper represents a significant advancement in automated journalism, moving beyond basic reporting to generate compelling narratives using sophisticated algorithms and natural language generation techniques. Reynolds likely outlines the methodology employed to generate event-driven narratives, which involves parsing and analyzing data related to specific events and structuring it into coherent storytelling formats. By automating this process, news organizations can efficiently produce timely and relevant content, especially for breaking news events or sports recaps. Throughout the paper, Reynolds is expected to discuss both the opportunities and challenges associated with event-driven narratives in automated journalism, including considerations of accuracy, potential biases, and ethical implications. Overall, "Automated Journalism 2.0 - Event-Driven Narratives" provides valuable insights into the future of news storytelling, showcasing the transformative potential of automation in journalism[6].

The paper "Automation will save journalism" by Martin Kjellman, published in 2021, delves into the active effects of automation within newsrooms, presenting a thorough examination of how automation technologies are transforming traditional journalistic practices. Kjellman's work begins by addressing the pressing challenges confronting the journalism industry, including dwindling revenues and shifts in audience consumption patterns. It then meticulously explores how automation offers solutions to these challenges, particularly by streamlining news production processes and enhancing efficiency. Throughout the paper, Kjellman likely discusses various facets of automation's impact on newsroom dynamics, ranging from content creation and editing to fact-checking and distribution. By leveraging automated tools and algorithms, news organizations can not only optimize resource allocation but also improve the accuracy and timeliness of their reporting. Moreover, Kjellman likely examines the broader implications of automation on journalistic roles and workflows, considering factors such as job displacement, skill requirements, and organizational structures. By providing empirical evidence and real-world examples, the paper offers valuable insights into the transformative potential of automation in journalism, informing discussions on the future direction of the industry. Overall, "Automation will save journalism" serves as a comprehensive exploration of the active effects of automation within newsrooms, highlighting both its opportunities and challenges in shaping the future of journalism [7].

The paper authored by De William and Goodwill in 2018 introduces "Automated news in practice - connexon," which presents an innovative API designed to deliver news content tailored to individual user preferences. Through meticulous research and development, the paper outlines the creation and implementation of the Connexon API, focusing on various integral components. These include the formulation of advanced algorithms capable of analyzing user preferences derived from diverse sources such as past reading history and user-provided preferences. Additionally, the paper delves into the intricate process of data processing, involving the ingestion, cleaning, and preprocessing of news articles and user data to facilitate algorithmic analysis. The paper emphasizes the importance of personalization in news delivery, detailing methods to customize news recommendations for each user based on their distinct preferences and interests. Furthermore, it highlights the development of the Connexon API itself, with a focus on user authentication, preference input, and the retrieval of personalized news recommendations. Through rigorous testing and evaluation, the paper demonstrates the efficacy of the Connexon API in providing tailored news content compared to alternative methods. Moreover, it underscores the significance of optimizing user experience and performance to ensure seamless interaction and scalability. Overall, the paper offers valuable insights into the practical implementation of automated news recommendation systems, showcasing the potential for personalized news delivery based on user preferences [8].

Liu, Gao, and Hu (2021) conducted a comprehensive survey study published in the journal Telematics and Informatics, wherein they delved into the influence of automated journalism on media credibility. Their research aimed to elucidate the implications of automated journalism on the perceived trustworthiness and reliability of news sources. By employing a survey methodology, the researchers systematically gathered data to discern public perceptions regarding the credibility of news content generated through automated processes. Through their investigation, Liu et al. sought to uncover any potential shifts in audience trust towards media outlets employing automated journalism practices. The study likely involved the analysis of various factors such as perceived objectivity, accuracy, and transparency of automated news articles, along with comparisons to traditionally authored news content. The findings of their research may offer valuable insights into the evolving landscape of journalism and its intersection with technology, shedding light on the challenges and opportunities posed by automation in the realm of media credibility[9].

Neubarth and Mayr (2020) conducted a comparative analysis exploring the landscape of automated journalism initiatives from an international perspective. Published in Digital Journalism, their work aimed to provide insights into entrepreneurial endeavors utilizing automated journalism techniques across different regions. Through their research, Neubarth and Mayr likely examined various case studies and examples of automated journalism initiatives worldwide, investigating the motivations, strategies, and impacts of these ventures. The study may have involved a detailed analysis of the technological infrastructure, content production processes, and business models employed by different organizations engaging in automated journalism. By comparing and contrasting initiatives from diverse cultural and economic contexts, the researchers likely aimed to identify common trends, challenges, and opportunities in the field of automated journalism on a global scale. The findings of their comparative analysis could contribute to a deeper understanding of the factors influencing the adoption and diffusion of automated journalism practices across different media ecosystems, ultimately informing discussions on the future trajectory of journalism in the digital age[10].

Stevens and Witschge (2021) conducted a thorough review of the emergence of robot journalism within the Netherlands, as outlined in their work published in Journalism Studies. Their research aimed to provide a comprehensive examination of the adoption and impact of automated journalism within the Dutch media landscape. Through their review, Stevens and Witschge likely analyzed the development of automated journalism technologies, the integration of automated content production processes within newsrooms, and the implications for journalism practice and ethics. The study may have involved an exploration of various automated journalism initiatives within Dutch media organizations, ranging from algorithmically generated news articles to automated data-driven storytelling platforms. Additionally, the researchers might have investigated the reactions and responses of journalists, news consumers, and other stakeholders to the proliferation of automated journalism in the Netherlands. By offering insights into the specific socio-cultural, economic, and institutional contexts shaping the rise of robot journalism in the country, Stevens and Witschge's review likely contributes to broader discussions on the transformation of journalism practices in the digital age, with implications for both academia and industry stakeholders [11].

Nguyen and Saito (2020) conducted a case study exploring the integration and impact of automated journalism within newsroom workflows in Japan, as detailed in their work published in Journalism Practice. Their research aimed to investigate how automated journalism technologies are utilized within Japanese news organizations and the implications for journalistic practices. Through their case study approach, Nguyen and Saito likely examined specific instances of automated journalism implementation in Japanese newsrooms, considering factors such as the types of stories automated, the roles of journalists in the production process, and the overall efficiency and effectiveness of automated content generation. The study may have involved interviews with journalists, editors, and other newsroom personnel to gain insights into their experiences and perspectives on the role of automation in news production. Additionally, the researchers might have analyzed changes in newsroom dynamics, routines, and decision-making processes resulting from the integration of automated journalism tools. By offering a nuanced understanding of how automated journalism is shaping newsroom workflows and practices in Japan, Nguyen and Saito's case study likely contributes to broader discussions on the transformation of journalism in the digital era, with implications for both academic research and industry practices[12].

Kaltenbrunner and Kocifaj (2020) conducted a comprehensive study published in Digital Journalism, where they focused on mapping the field of algorithmic news and its intellectual structure. Their work aimed to provide an overview of the academic discourse surrounding algorithmic news production, analyzing its key concepts, theories, and methodologies. Through their research, Kaltenbrunner and Kocifaj likely conducted a systematic review of existing literature on algorithmic news, identifying and categorizing relevant studies, frameworks, and research methodologies. By mapping the intellectual structure of the field, they likely aimed to uncover patterns of scholarly interest, interdisciplinary connections, and emerging research trends within the study of algorithmic news. The study may have involved the use of bibliometric analysis techniques to quantitatively assess the distribution and impact of publications, authors, and journals within the field. Additionally, the researchers might have conducted qualitative analyses to explore the conceptual frameworks and theoretical perspectives underpinning research on algorithmic news. By synthesizing and visualizing the intellectual landscape of algorithmic news research, Kaltenbrunner and Kocifaj's study likely provides valuable insights for scholars, practitioners, and educators seeking to understand the evolving dynamics of news production in the digital age[13].

Van Dalen's (2019) study meticulously dissects the control and influence dynamics inherent in algorithmic news aggregators, with Google News serving as a prominent case study. Through rigorous analysis, the paper elucidates the intricate mechanisms governing the selection and prioritization of news content in digital environments. It underscores the pivotal roles played by various actors in shaping the flow of information, emphasizing issues of power asymmetry, transparency deficits, and the need for accountability mechanisms. By shedding light on these dynamics, Van Dalen's work contributes to a deeper understanding of the complexities surrounding algorithm-driven content curation and its implications for media pluralism and democratic discourse. The insights gleaned from this study provide valuable guidance for policymakers, journalists, and technology stakeholders seeking to navigate the evolving landscape of digital news aggregation responsibly and ethically [17].

Witschge and Anderson (2019) explore the phenomenon of the datafication of news, delving into how news content is increasingly being transformed into data-driven formats. The paper meticulously examines the implications of this trend on journalistic practices and audience engagement. Through empirical research and theoretical analysis, it elucidates the challenges and opportunities arising from the datafication process, including shifts in news production, distribution, and consumption patterns. By shedding light on these dynamics, Witschge and Anderson provide valuable insights for understanding the evolving landscape of journalism in the digital age, offering guidance for practitioners and scholars alike on navigating the complexities of data-driven news environments responsibly and effectively [20].

IV. EXISTING SYSTEM

In the existing system digital journalists are tasked with a labour-intensive process of curating news material from multiple blogs and websites. This entails manually navigating through various online platforms to identify pertinent articles, summarizing their material by hand, and then sharing these summaries throughout various social media platforms. However, this manual approach poses efficiency challenges as journalists must individually sift through diverse sources, consuming valuable time that could be better utilized for more creative and investigative pursuits. Moreover, the lack of automation in this process makes it susceptible to delays and potential oversight of breaking news events.

V. PROPOSED SYSTEM

The proposed system introduces a more streamlined and automated approach to digital journalism. Utilizing advanced technologies, it aims to simplify the processes of news aggregation and dissemination. Rather than relying on manual efforts to gather information from various websites, the system employs intelligent algorithms for ethical web scraping, automatically retrieving news articles from selected sources. This information is then organized systematically within a user-friendly database, accessible through a well-designed interaction point constructed with contemporary web frameworks. Moreover, the system streamlines social media engagement by automating the posting of news summaries through predefined templates and social media APIs. This not only saves time for reporters in the digital realm but also guarantees the timely and strategic dissemination of news material. The primary objective is to enhance productivity, enabling reporters to concentrate on analysis and creativity while maintaining ethical standards and prioritizing user-friendly functionalities. This transition from manual to automated processes not only improves efficiency but also allows journalists to allocate more resources to high-value tasks, ultimately enhancing the overall flow of news processes news dissemination. In summary, the proposed system represents a noteworthy advancement in the realm of digital journalism, offering innovative methods to optimize the efficiency of news workflows production and distribution processes.

VI. PROBLEM STATEMENT

The automated approach to digital journalism represents a notable milestone advancement in the field. By leveraging cutting-edge technologies, the system strives to simplify the processes of news aggregation and distribution. Instead of relying on manual efforts to collect information from various websites, the system utilizes intelligent algorithms for ethical web scraping, automatically retrieving news articles from selected sources. Subsequently, this information is organized systematically within a user-friendly database, accessible through a well-crafted interface built with modern web frameworks. Moreover, the system simplifies social network engagement by automating the posting of news summaries using predefined templates and social network APIs. This not only conserves time for digital journalists but also ensures the timely and strategic distribution of news material. The overarching objective is to enhance efficiency, enabling journalists to concentrate on analysis and creativity. This transition from manual to automated processes not only improves efficiency but also empowers journalists to dedicate greater focus to content creation and analysis, thereby enhancing the overall workflow of news dissemination.

VII. ARCHITECTURE DIAGRAM

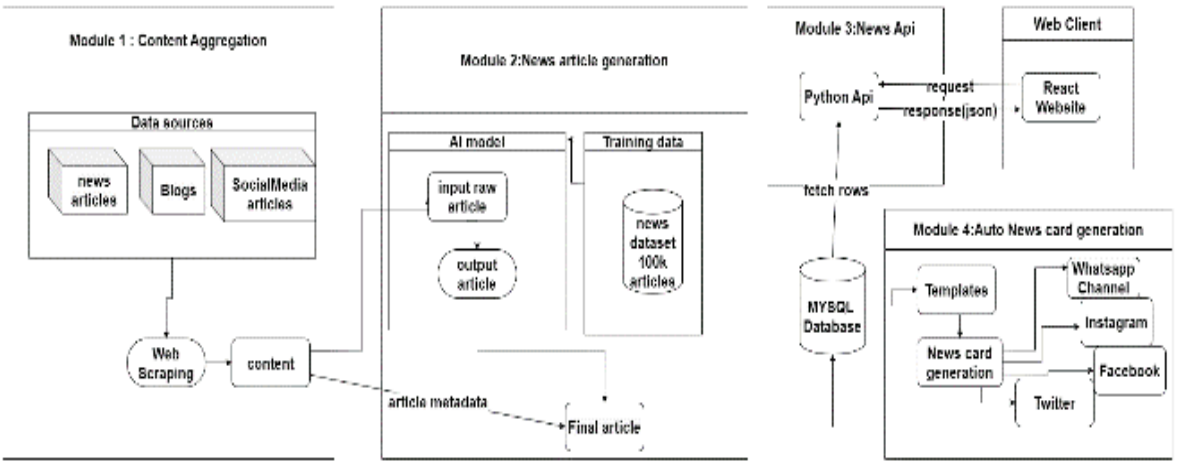

Fig: 1 Architecture Diagram

VIII. OUR APPROACH

The core algorithms of news automation are organized into four suites, each implementing one aspect of news production.

8.1. NOISE FILTERING:

In our approach to filtering noise and identifying news within the information spectrum derived from various sources, such as Twitter, blogs, and external news sites, we adopt a strategy that encompasses a range of content types, from spam and advertisements to breaking news. Recognizing the inherent challenge of unbalanced data, wherein noise may overshadow actual news events, we employ sub-algorithms specifically tailored to iteratively distinguish noise from the events and the news.

This iterative approach allows us to address the imbalance in the data effectively, mitigating the risk of misclassification. Moreover, to ensure comprehensive coverage and prevent the oversight of important stories, our algorithm is fine-tuned to prioritize the reduction of false negatives, even at the expense of allowing a certain degree of noise to persist. This design choice acknowledges the reality of noisy data streams while maintaining a focus on capturing significant events and news stories.

By optimizing our algorithm to prioritize recall over precision, we aim to maximize the detection of events and news items while tolerating a certain level of noise. Experimental results demonstrate the effectiveness of this approach, with over 78% of tweets, blog posts, and external news articles being successfully filtered with less than 1% false negatives. This high level of accuracy underscores the robustness and reliability of our noise filtering and news detection framework, contributing to more efficient and effective information retrieval and dissemination processes.

8.2. EVENT CLUSTERING:

In our study, we adopt a similar approach to previous research focusing on Twitter, where we tackle the task of news finding by framing it as the challenge of event detection through clustering. However, unlike previous studies that primarily concentrate on Twitter data, we broaden our scope to include data from blogs and external news sites, aiming to capture a more comprehensive picture of events and news stories.

Our clustering algorithm operates in two distinct phases: clustering and merging. During the clustering phase, unit clusters are generated, each comprising three tweets, blog posts, or news articles with similar content. These unit clusters are then merged with an existing pool of clusters. If a unit cluster is formed and does not result in a merge with an existing cluster, it is identified as an event. This design not only accelerates event detection but also streamlines periodic cluster updates, enhancing the efficiency of the overall process.

To evaluate the effectiveness of our algorithm, we benchmark it against a recently proposed event detection algorithm based on locality-sensitive hashing (LSH), which is tailored for handling big data. While LSH is an approximate search algorithm, our approach performs a full search across Twitter, blogs, and external news sites, supported by our robust big data processing infrastructure. This allows us to identify more events with greater accuracy and reliability. By expanding the scope of our analysis beyond Twitter to include blogs and external news sites, and leveraging a comprehensive clustering and merging algorithm, we demonstrate the advantages of our approach in identifying and detecting events from diverse sources. Our methodology not only enhances the coverage and accuracy of event detection but also highlights the importance of leveraging big data processing capabilities for effective news finding in the digital age.

8.3. NEWSWORTHINESS DETECTION:

The concept of newsworthiness is intricate, influenced by diverse factors such as human interest, the prominence of involved subjects (individuals, organizations, and locations), public attention, and personal perception. These components collectively shape the subjective determination of what qualifies as news. Acknowledging the dynamic nature of newsworthiness, our algorithm endeavors to capture these subjective aspects, primarily concentrating on the content of an incident.

Newsworthiness is not a fixed attribute but rather evolves with shifting and emerging news topics. To accommodate this dynamic nature, our algorithm models both short and long-term newsworthiness. By considering the temporal dynamics of news, we seek to offer a nuanced understanding of the relevance and significance of events within their respective contexts. The problem of newsworthiness detection is framed as follows: given an event cluster e and its associated outcomes {w1, ..., w(m)}, the algorithm

aims to predict the probability of the event being considered news, denoted as p(e). This probability is derived by aggregating the probabilities of various factors contributing to newsworthiness, including the likelihood of the event addressing news topics (pT(e)), involving newsworthy subjects (pO(e)), and attracting public attention (pA(e)). By incorporating these factors into our newsworthiness detection algorithm, we strive to offer a comprehensive and dynamic evaluation of the news value of events, facilitating more informed decision-making in the realm of automated news generation and dissemination.

8.4. VERACITY PREDICTION:

Social media has emerged as a crucial source of information for news outlets, yet it is fraught with misinformation and fake news. During the 2016 US presidential election, Facebook faced severe criticism for its inability to combat the rampant spread of false information favoring a particular candidate. To address this challenge, we have devised an algorithm within Tracer to assess the truthfulness of news, keeping our users informed about potential risks. While we have detailed our verification algorithm as a standalone application, we now reveal its integration with other components in Tracer when processing streaming data. Our key takeaway is that credibility does not always align with veracity; even celebrities and reputable news media occasionally fall victim to false statements. Journalists consider various factors, such as the news source's origin, credibility, legitimate identity, and the presence of multiple independent sources, in their quest to verify news. Consequently, we have incorporated these verification steps into our algorithm for enhanced efficiency. Specifically, we have trained multiple SVM regression models with diverse features to handle the early and developing stages of an event separately. These models generate scores indicating the level of veracity.

8.5. EVENT SUMMARIZATION:

In addition to being representative, the chosen summary for each event must also adhere to standards of readability and objectivity, aligning with the requirements for news headlines.

For instance, the summary "BREAKING: Donald Trump is elected president of the United States" is preferable to "OMG! Trump just won!! #Trump4Presdent" due to the latter containing personal emotions, shorthand notations, and misspellings. We approach this task as the selection of a suitable tweet from those present in an event cluster. "Lex Rank" stands out as one of the most commonly utilized algorithms for text summarization.

Lex Rank proves ineffective on tweet clusters as it tends to heavily favor repeated language (retweets) and deviates from the central topic of the event. Consequently, our algorithm opts for event summaries by relying on the cluster centroid while discouraging inaccurate and informal language. Every sentence transforms a vector, denoted as w→i, utilizing tf-idf representation. The tweet that most accurately reflects the event is the one closest to the centroid −C→. By incorporating the centroid, e mitigate the risk of low-quality text, as tf highlights significant terms and the penalty term enhances the readability and objectivity of summaries in the following manner

$$wi = \text{argmax}_{wi \in CE} (\text{sim}(wi, CE) - \lambda I(wi)) \quad (1)$$

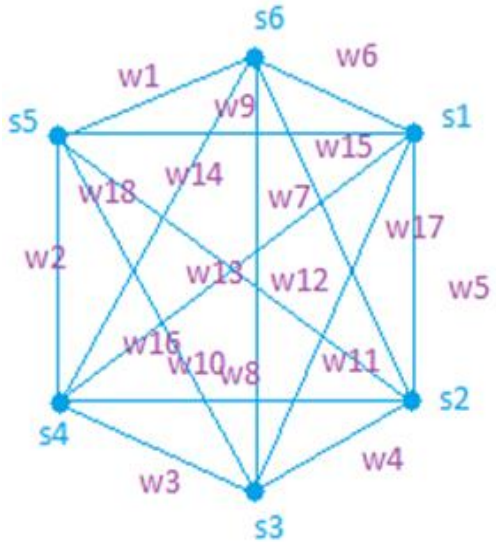

Fig: 2 Event Summarization

The parameter λ governs the intensity of the penalty. The (w_i) serves as an indicator function for identifying informal language in a tweet based on rules. These rules encompass cues like the existence of words not found in the vocabulary (excluding named entities), hashtags positioned amid a tweet, Twitter handles not associated with organizations or prominent users, and the presence of 1st & 2nd person pronouns.

8.6. TREND MINING:

A trend on Google Trends refers to a subject or search term that undergoes a notable increase in query popularity at a particular moment, as determined by algorithms tracking search terms and user interests. To access trending topics, Trend Mining is utilized, crucial for comprehending audience inclinations, conducting demographic examination, and ensuring content stays up-to-date and pertinent. The procedure involves accessing Google Trends to collect search frequency data, selecting pertinent keywords or phrases, and assessing search engagement graphs to identify patterns, spikes, and shifts across regions. Further refinement involves experimenting with filters to adjust parameters such as temporal span, geographical area, or search classification. Comparisons between keywords or topics are conducted to determine popularity and patterns across different areas or demographics. Investigating associated inquiries and subjects reveals further understanding. Ultimately, interpretation of findings enables understanding of the

underlying reasons and implications for businesses or areas of interest, facilitating effective mining of Google Trends data to uncover meaningful perspectives aligned with specific objectives.

IX. METHODOLOGY

A Cross-lingual language model (XLM) to develop a transformer-based model capable of handling various linguistic expressions. They employed three distinct training methods:

9.1. SEQUENTIAL LANGUAGE MODELING (SLM):

SLM aims to predict the likelihood of the next word based on the preceding meaning in the sentence, denoted as P (wt|w1, w2, ...wt−1, θ).

9.2. MASKED TOKEN MODELING (MTM):

MTM involves randomly selecting 15% of the input BPE (byte-pair encoded) tokens. These chosen tokens are substituted with: (i) the [MASK] token 80% of the time, (ii) a random token 10% of the time, and (iii) left unchanged 10% of the time.

9.3. INTERLINGUAL LANGUAGE MODELING (ILM):

ILM extends the concept of MTM. In addition to solely utilizing monolingual text streams, it combines parallel sentences from various linguistic expressions. Random masking is then applied to all sentences. To predict a masked word in one linguistic, the model can consult the adjacent context within the same linguistic sentence or examine corresponding phrases in another language to gain clues about the masked word. This approach empowers the model to learn synchronization between various linguistic expressions.

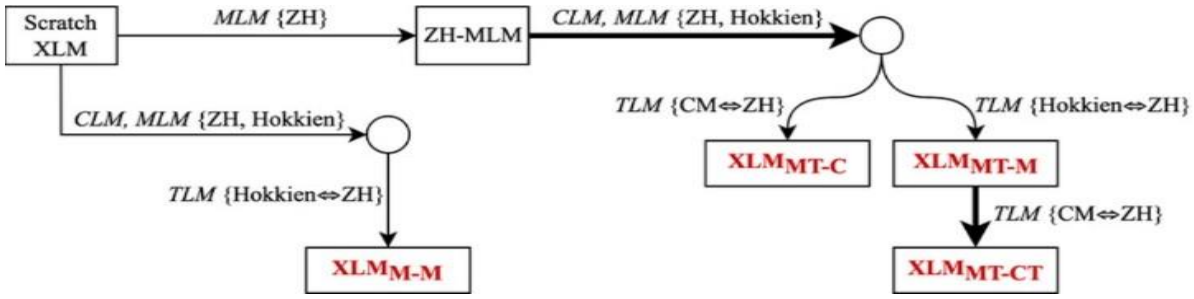

Fig 3. Training process of XLM X-Y

The XLM model employs a single encoder and a single decoder, incorporating language embeddings from various languages, in contrast to utilizing distinct encoders and decoders for constructing a multilingual system. Consequently, this model requires less memory compared to systems employing multiple encoders and decoders. In our experimentation, given the absence of parallel data, we exclusively utilize MLM. To implement this model, tweets, and news are regarded as two distinct styles, and we train distinct style embeddings for them, mirroring the approach of language embeddings employed in XLM.

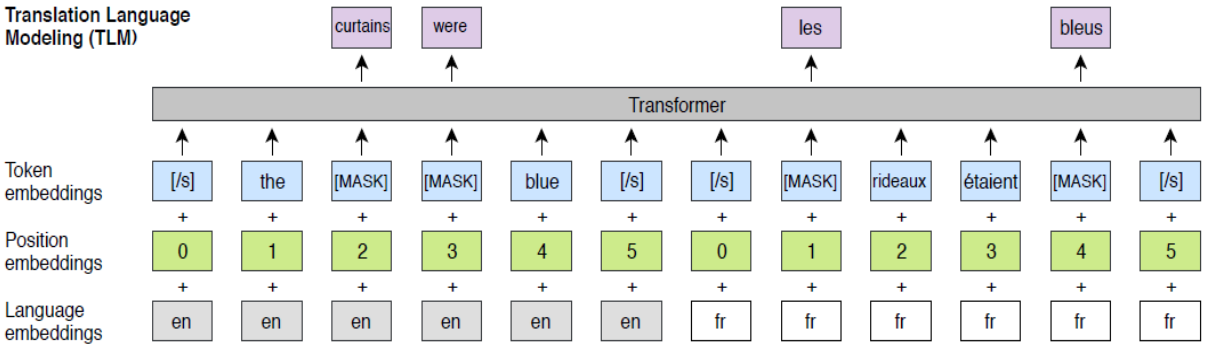

Fig: 4 XLM Masked Language Model (MLM) fine-tuning

9.4. RECURRENT NEURAL NETWORK:

In the realm of news generation, the utilization of Recurrent Neural Networks (RNNs) alongside a series-to-series model constitutes a sophisticated approach to content creation. The process initiates by representing each word in news articles through word embeddings such as Word2Vec or GloVe, thus transforming them into dense vectors denoted as x(t) at each time step (t). This embedding enables the model to capture the semantic connections between words, facilitating an enhanced comprehension of the article's content. The structure of the system adopts an encoder-decoder structure, wherein the encoder, a variant of RNN, analyses input news articles to grasp their contextual data. The latent state h(t) at each time step gathers and updates information about the article's content, culminating in the final hidden state *h*(final), which captures the essence of the overall background of the input news article. Subsequently, a separate RNN-based decoder utilizes this contextual depiction to produce the result news article. As the decoder processes the data, it updates its hidden states(t) at each timestamp, playing a pivotal role in generating coherent sequences. The result at each stage is determined by a SoftMax function, ensuring a likelihood spread across the vocabulary and enhancing the framework's ability to generate fluent and meaningful text. Throughout the training phase, the model learns by minimizing a sequence-to-sequence loss, typically measured using cross-entropy loss. This loss function quantifies the dissimilarity between the generated sequence and the target sequence, guiding the model to refine its ability to produce news articles that closely align with the desired content and framework. This approach underscores the iterative training regimen of the model, wherein it

continually refines its parameters to produce output that resembles human-generated news articles. In summary, leveraging RNNs and a sequence-to-sequence model for news generation involves several intricate steps, including word representation, encoder-decoder architecture, and training with a focus on minimizing sequence-to-sequence correspondence loss. This methodological framework underscores the sophistication and effectiveness of employing neural networks in the domain of content creation, paving the way for more advanced and nuanced automated journalism systems.

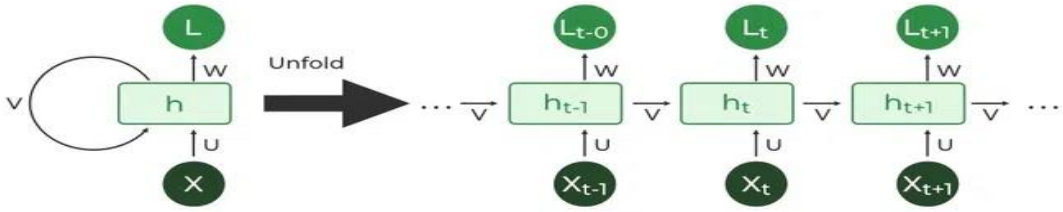

Fig: 5 Recurrent Neural Network

9.5. NEWS API:

The newly developed API efficiently addresses the need for targeted news consumption by offering a categorized list of news articles. Users can access this API by specifying a particular category of interest, ensuring a tailored news experience. The API employs a robust backend infrastructure, leveraging modern web technologies to handle incoming requests and deliver timely responses. Internally, the API utilizes algorithms or manual curation processes to sift through a diverse range of news sources, extracting relevant articles that align with the specified category. The platform supports a variety of categories, providing users with the flexibility to choose topics ranging from technology, business, and sports, to entertainment. Upon receiving a request, the API dynamically assembles a curated list of news articles within the specified category, ensuring that the content is up-to-date and relevant to user interests. This personalized approach enhances the user experience by delivering targeted and meaningful news content. The API's design prioritizes efficiency, scalability, and real-time updates to accommodate a growing user base and changing news landscapes. It adheres to industry standards for RESTful communication, offering a straightforward and accessible interface for developers and integrators.

9.6. WEBPAGE GENERATION:

Once the articles are fully ready for publication, an automated process generates web pages. This web page generation can be categorized into two sections:

• Creation of individual news article web pages
• Development of landing sites and category pages

The development of individual web pages for news articles commences by designing a fundamental HTML template. Rather than relying on web template frameworks such as Django or Flask, simpler HTML and CSS are utilized to ensure broad compatibility, even with older or mobile browsers. Each article is then incorporated into this template using JSON data stored on the web server. The template, containing HTML code stored as strings, is saved to a file for each article. This procedure is replicated for all stored articles. Additionally, every news article page includes a news ticker on the right-hand side for easy access to other intriguing articles. Following the creation of article pages, their titles are extracted to construct category pages and landing pages. The main page showcases a randomly chosen article from the available set. Each category page provides links to all news content within that specific category. Interconnections among individual pages have been established, resulting in the automated construction of a comprehensive site. The HTML files residing on the server are subsequently uploaded to the cloud server, inkbot.com, hosted on Amazon Web Services. As there is no caching of webpages, news articles persist on the server only until the next system execution.

X. RESULT

The overall process of this phase has accomplished the development of an automated system of news generation. The integration of automated journalism, combining web scraping, AI-driven personalized news generation, and multimedia dissemination on social media, signifies a groundbreaking advancement in information consumption and sharing. This innovative approach harnesses state-of-the-art technologies to deliver tailored news experiences based on individual preferences, facilitated by the rapid extraction and synthesis of pertinent information through web scraping and AI models. The incorporation of multimedia elements enhances the storytelling aspect, making news not only informative but also engaging. The real-time nature of this methodology ensures timely updates, while the proactive use of social media platforms broadens the reach to diverse audiences, fostering a dynamic news-sharing experience.

XI. TESTING

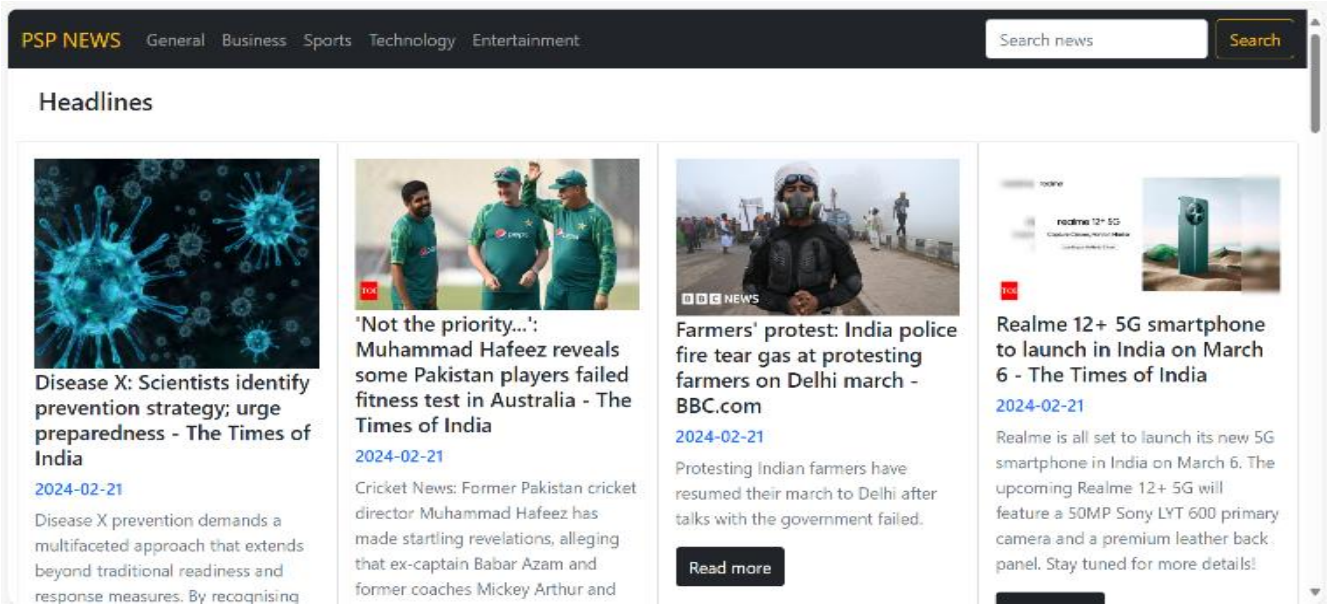

Fig: 6 Headline News Content

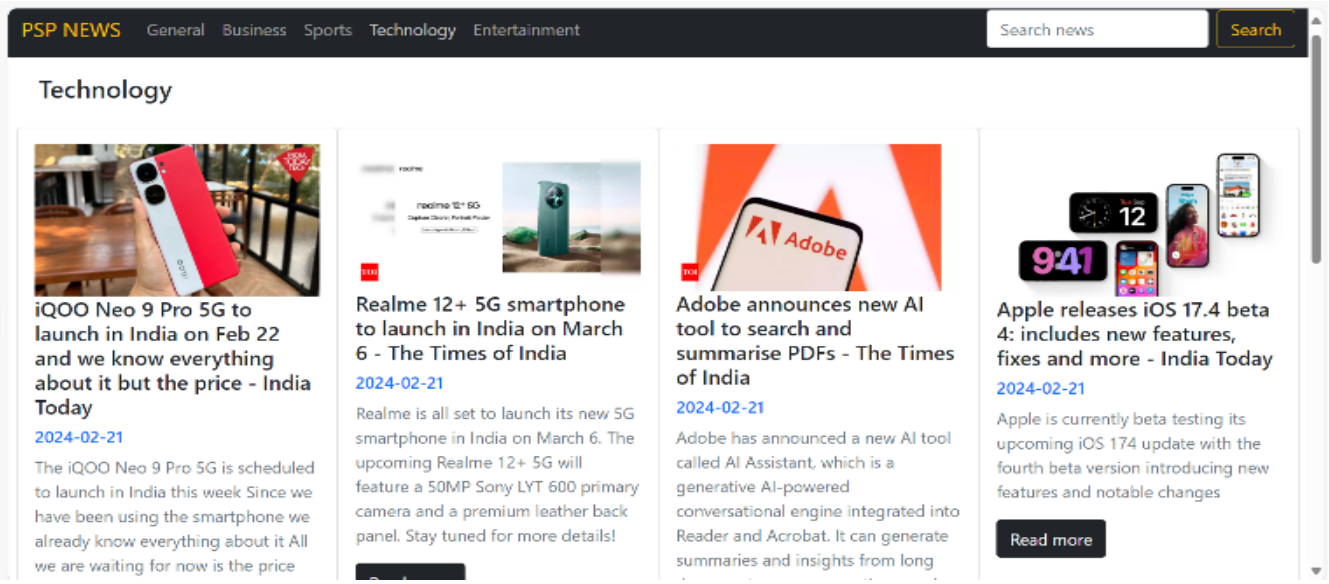

Fig: 7 Technology News Content

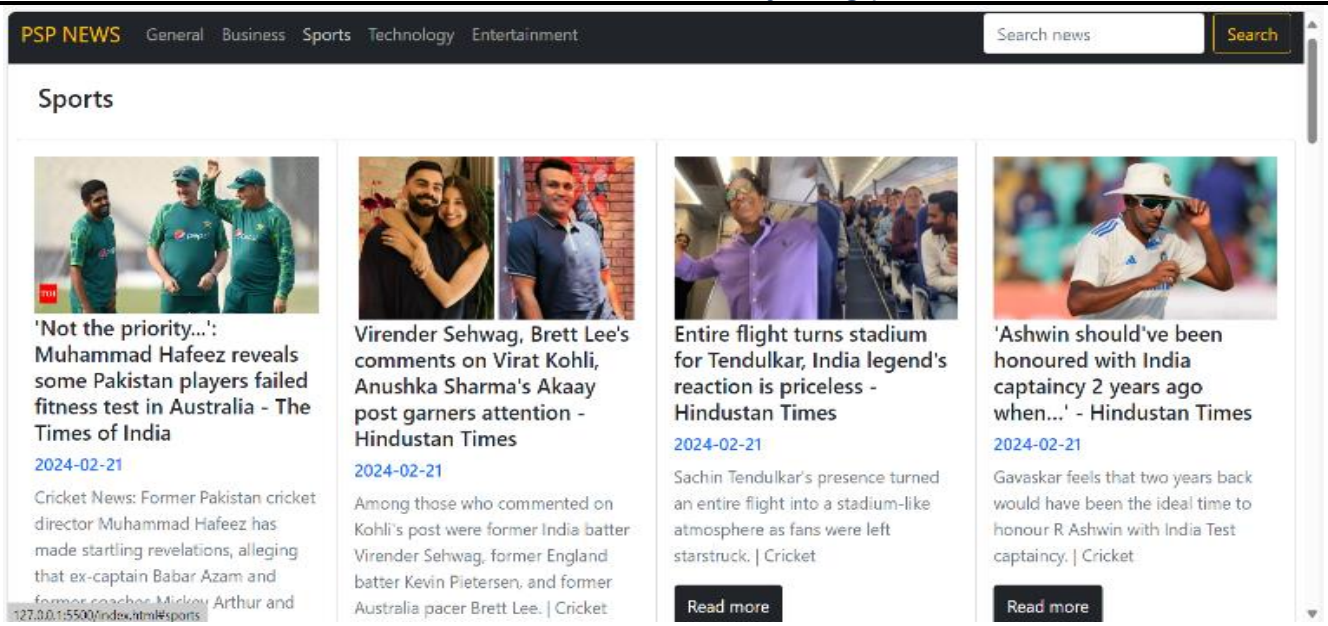

Fig: 8 Sports News Content

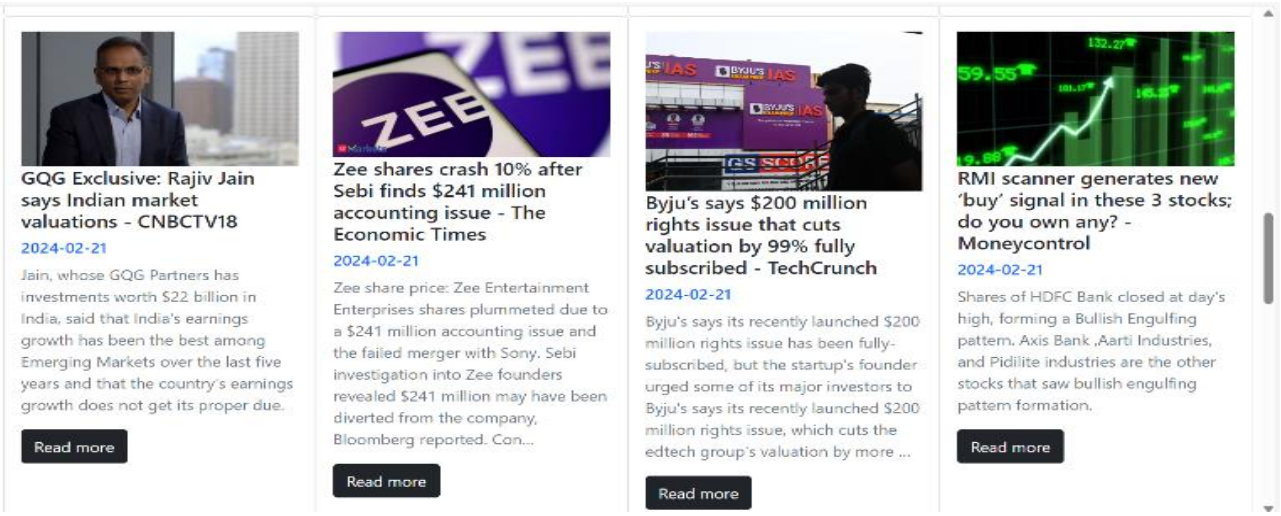

Fig: 9 Business News Content

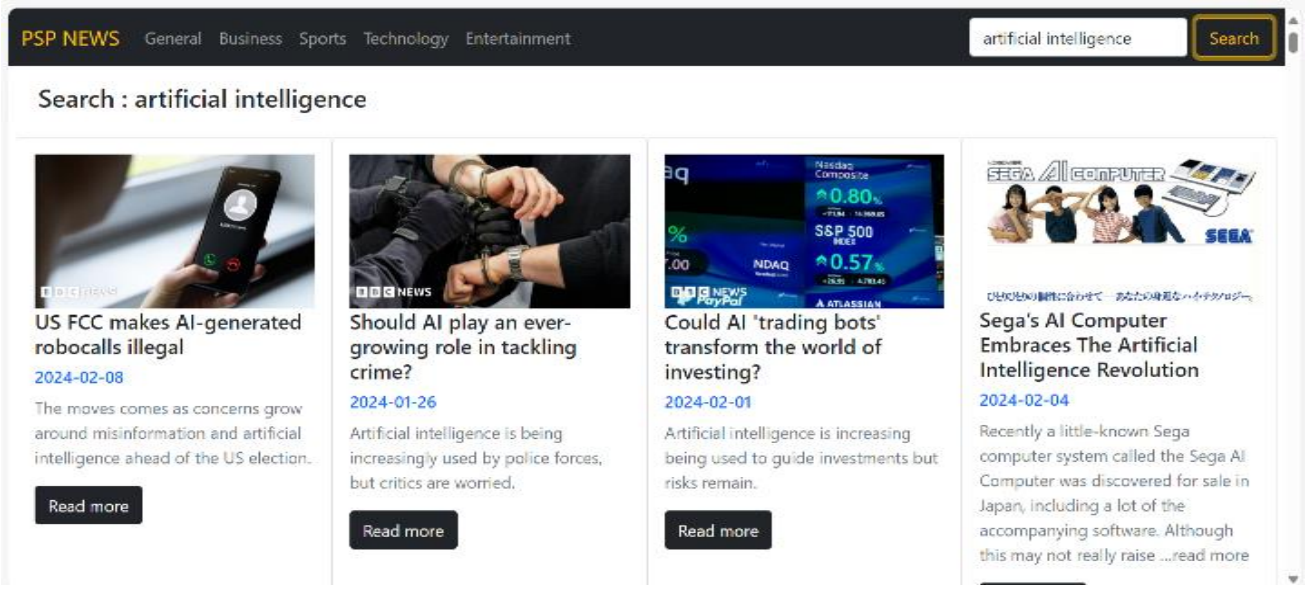

Fig:10 User Preference Based News Content

XII. CONCLUSION

As the project progresses, it encompasses a well-structured and systematic approach to aggregating, processing, and disseminating news content. By seamlessly integrating web scraping, API development, database management, front-end design, and social media interaction, the project seeks to create a streamlined solution for users seeking reliable and up-to-date news information. The use of web scraping tools ensures data retrieval from chosen websites, while the developed API facilitates efficient communication between the backend and front end. Storing the curated data in a chosen database enhances accessibility and organization. The front end, designed with frameworks like React or Angular, provides an intuitive interface for users to explore and engage with the aggregated news content. Social media integration adds a dynamic element, allowing users to share news seamlessly. Robust security measures, automated processes, thorough testing, and diligent documentation contribute to the overall reliability and sustainability of the system. In essence, this project represents a holistic and modernized solution for news aggregation, catering to the evolving needs of users in a digital age while upholding ethical and legal standards in data usage and collection.

XIII. FUTURE SCOPE

To enhance the data sourcing aspect by implementing automated fetching of news content. This system will utilize generative AI to dynamically generate news articles daily, ensuring accuracy and comprehensive coverage of relevant topics. The process will not be limited to a predetermined order of data sources. instead, the system will autonomously prioritize and retrieve news content from various sources based on relevance and importance. By leveraging advanced AI techniques, the goal is to deliver detailed and accurate news content consistently, catering to the evolving needs of users for up-to-date information. Additionally, the introduction of a recommendation system represents a notable stride towards optimizing news distribution. Combined, the implementation of an AI-driven web scraping model and a recommendation system positions the project at the forefront of technological innovation in the field of news aggregation. These progress not solely enhance the technical robustness of the system and furthermore elevate the overall user experience, ensuring the platform remains competitive and also adaptable amidst the rapidly changing landscape of digital news consumption.

REFERENCES

[1] Xu Z and Lan X. (2020), "A scientometric review of automated journalism Analysis and visualization ": A study to show the status and trends of research in automated journalism.
[2] Efthimis Kotenidis and Andreas Veglis. (2021), "Algorithmic journalism - Current applications and future perspectives": Discusses challenges and future implementations related to automated journalism.
[3] Graefe A. and Bohlken N. (2020), "Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news": Summarizes evidence of credibility, quality, readability of automated news.
[4] Leppanen L, Munezero M, Granroth Wilding M and Toivonen H. (2017), "Data-driven news generation for automated journalism ": Explores field and challenges with building a journalistic natural language generation system.
[5] Carl Gustav Lind. (2018), "The bright future of news automation ": Discusses the scope of news automation.
[6] Donald W and Reynolds. (2017), "Automated Journalism 2.0 - Event-Driven Narratives ": Creation of stories from simple descriptions.
[7] Martin Kjellman. (2021), "Automation will save journalism ": Discusses active effects on automation in news room.
[8] De William and Goodwill. (2018), "Automated news in practice - connexon": An API that provide news based upon user's preference.
[9] Liu Y, Gao, H, & Hu Y. 2021. Exploring the impact of automated journalism on media credibility: A survey study in Telematics and Informatics.
[10] Neubarth K, & Mayr P. 2020. Automated journalism in international perspective: A comparative analysis of entrepreneurial initiatives in Digital Journalism.
[11] Stevens M, & Witschge T. 2021. The Rise of Robot Journalism: A Review of Automated Journalism in the Netherlands - Journalism Studies.
[12] Nguyen A, & Saito M. 2020. Investigating the role of automated journalism in newsroom workflows: A case study of news organizations in Japan, Journalism Practice.
[13] Kaltenbrunner A, & Kocifaj M. 2020. Algorithmic news in the making: Mapping a field of study and its intellectual structure, Digital Journalism.
[14] Diakopoulos N. (2014). "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures": Explores the role of journalists in scrutinizing the impact and accountability of algorithms in various societal domains.
[15] Carlson M. (2015). "The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority": Investigates the transformation of labor, storytelling methods, and the authority of journalism in the context of automated reporting.
[16] Carlson M, & Eberwein T. (2017). "Journalism in an Era of Big Data: Cases, Concepts, and Critiques": Examines the intersection of journalism and big data, presenting cases and critiques to understand its implications on journalistic practices.
[17] Van Dalen A. (2019). "Who Controls the Algorithmic News Aggregator? The Case Study of Google News": Analyzes the control and influence dynamics behind algorithmic news aggregators, using Google News as a case study.
[18] Thornley C. V. (2020). "Robot Journalism: A Survey of the Field and its Implications for Media Responsibility and Democracy": Conducts a survey of automated journalism, exploring its implications for media responsibility and democratic discourse.
[19] Carlson M, & Lewis S. C. (2015). "Boundaries of Journalism: Professionalism, Practices and Participation": Explores the evolving boundaries of journalism in the digital age, considering shifts in professionalism, practices, and audience participation.
[20] Witschge T, & Anderson C. W. (2019). "The Datafication of News": Investigates the transformation of news into data-driven formats, examining the implications of datafication on journalistic practices and audience engagement.