EKANTIPUR-15Y: A LONGITUDINAL BENCHMARK CORPUS AND SEMANTIC ANALYSIS OF NEPALI NEWS (2010–2025)

Diwash Mainali[*] *Utsav Mainali*[†]

ABSTRACT

This paper introduces Ekantipur-15Y, a long-scale longitudinal corpus of Nepali news articles spanning from 2010 to 2025. As Nepali is considered a low-resource language, the lack of a clean and temporally diverse dataset has been a barrier for the development of robust Natural Language Processing (NLP) models. We collected and cleaned 109,704 unique articles with approximately 14.3 million tokens from Ekantipur. The corpus is validated using Zipf's law confirming linguistic integrity and Heap's law demonstrating continuous growth of vocabulary without plateauing. Furthermore, the semantic analysis successfully detects the major historical events in the context of Nepal, including the 2015 Earthquake and the COVID-19 pandemic validating the accuracy of the dataset. Finally, a baseline is established for text classification, where a Linear Support Vector Machine (SVM) achieves an accuracy of *74.50%*, significantly outperforming Naive Bayes and Logistic Regression.

Index Terms— Nepali NLP, Low-Resource Languages, Longitudinal Corpus, Text Classification, Event Detection

1. INTROCUCTION

The advancement of Natural Language Processing (NLP) depends on a large high-quantity of quality training datasets [1]. For the high-resource language, it benefits greatly from the massive corpora but Nepali is a low-resource language limiting the advancement of modern deep learning models [2]. In recent research advancements, the sequence labelling tools like Bi-LSTM-CRF models are in use, they lack a substantial longitudinally diverse corpus required for the study of temporal language shift and training robust Large Language Models (LLMs) [3, 4].

For bridging the gap, Ekantipur-15Y is introduced, a comprehensive dataset of Nepali news articles covering the period from February 1, 2010 to November 25, 2025. Compared to a static dataset, this corpus enables longitudinal analysis in order to track the evolution of sentiment, vocabulary and other aspects as required by the researcher.

The contributions of this paper are threefold:

- Corpus Creation: We release a cleaned dataset of 109,704 articles with rich metadata including titles, summaries, and publication dates.
- Validation: We verify the linguistic naturalness of the corpus using Zipf's and Heaps' laws [5] and validate its temporal integrity by correlating keyword frequencies with known events like the 2015 Earthquake and the 2022 Elections.
- Benchmarking: We provide baseline classification performance using standard supervised learning models, establishing a benchmark of *74.50%* accuracy for future comparison [6].

2. RELATED WORKS

The development of the Ekantipur-15Y corpus acts as an important benchmark for the Nepali language traditionally regarded as a low-resource language from the computational standpoint [7]. This literature review explores research across natural language processing (NLP), web engineering and longitudinal media analysis to analyse the current state of the art and justify the need for a 15-year benchmark dataset.

2.1. NLP Methodologies for Low Resource Languages

Recent research on Nepali language processing focuses on fundamental sequence labelling tools like POS tagging and chunking using Bi-LSTM-CRF model which uses both work and character embedding reaching an accuracy upto 99.20% alongside handling semantic representation, morphological richness and out-of-vocabulary words effectively [7, 8].
Traditional models like HMM and SVM are clearly outperformed by deep learning models [7, 9] reaching higher accuracy along with better handling of unknown words. LLM depends heavily on a large training dataset and metadata of low-resource language management is computationally costly and very error-prone (10%) in comparison to high-resource languages [10]. The recent ev-

idence showed that instruction tuning in LLMs can enhance zero-shot summarization effectively than scaling model sizes.

2.2. Framework for News Extraction and Archiving

A string based content extraction algorithm provides a significant speed boost (about 60x speed) over the DOM based parsers, but they are less robust than DOM based method for dynamic and irregular web content [11]. For large-scale scraping, it must comply with the legal frameworks like the Computer Fraud and Abuse Act (CFAA) and ethical norms of robots.txt to maintain proper transparency of the researcher via a proper user-agent string. [12].

2.3. Longitudinal and Semantic Media Analysis

Longitudinal NLP studies analyse temporal shifts in sentiment and language use. Over the past two decades a global rise in negative and anger-related language can be observed in news headlines along with the strong correlation between public emotion and news coverage during events like COVID-19 [13, 14, 15]. These studies distinguish between word types and tokens with a contextualised model like BERT to capture deeper insights into polysemy and meaning variation [16, 17].

Although Supervised Machine Learning (SML) outperforms dictionary based methods, it depends on a high-quality, human-annotated and verified dataset and existing benchmarks lead to an underestimation of human performance due to limited low low-quality reference summaries [17]. Similarly, longitudinal analysis suffers from reporting bias where exceptional events are often overrepresented in events compared to stable linguistic patterns [16].

2.4. Synthesis

Despite the efficiency of modern sequence labelling models, Nepali NLP is currently constrained by static datasets that fail to capture language evolution [7]. To bridge this gap, Ekantipur-15Y provides the necessary longitudinal infrastructure, enabling researchers to move beyond snapshot analysis and model the complex, temporal dynamics of public discourse over a decade and a half.

3. METHODOLOGY

In the whole process libraries like Pandas [18], Matplotlib [19], Seaborn [20] and NLTK [21] are used.

3.1. Data Collection

The data were collected from one of the major Nepali digital news media outlets, *Ekantipur*, using a custom scraper permitted by the `robots.txt` rules. The dataset spans from February 1, 2010, to November 25, 2025, and includes the following metadata for each article: *news_id*, *news_title*, *news_short_description*, *published_at*, *modified_at*, *news_url*, *writer*, *publisher*, *news_summary*, and *news_content*.

3.2. Post Processing and Normalization

The news article contained advertisements, occasional English text, and HTML tags, along with inconsistent whitespace. The cleaning pipeline is constructed, which does the following:

1. **Content Cleaning:** Each article is cleaned by using the regex that removes the recurring English text noises using a case-insensitive regex. Following, the HTML tags and trailing whitespaces are removed. This ensures the valid string data for every row.

2. **Tokenization and Word Counts:** The Nepali language uses the Purna Viram (।) as a sentence delimiter. Direct processing of this symbol can be computationally heavy, so all occurrences were replaced with spaces to ensure proper sentence boundaries. The cleaned text was then used to compute word counts and tokens, enabling statistical analysis such as average sentence length and the Type–Token Ratio (TTR).

3. **Temporal Normalization:** Article published dates are loaded into pandas *datetime* objects, allowing for year-wise aggregation for plotting the distribution over the time interval.

3.3. Corpus Statistics

The final dataset consists of 109,704 unique articles containing approximately 14.3 million tokens. The vocabulary size (unique) is 560,872 with a TTR of 0.0390.

Statistic	Value
Total Documents	109,704
Total Tokens (Words)	14,385,778
Vocabulary Size	560,872
Average Article Length	131.13 words
Type–Token Ratio (TTR)	0.0390

Table 1. Key Statistics of the Ekantipur-15Y Corpus

The average sentence length is 131.13 words, showing a high level of syntactic complexity. This suggests that the dataset contains well-formed sentences rather than fragmented headlines.

4. CORPUS CHARACTERISTICS & ANALYSIS

4.1. Zipf's Law Validation

Zipf's law is used to verify whether the article represents human language or random noise [22, 5]. The frequency distribution of the words was analyzed, and a log-log plot of word frequency versus rank showed a near-perfect linear descent ($f \propto 1/r$), confirming the dataset's linguistic integrity and naturalness [23].

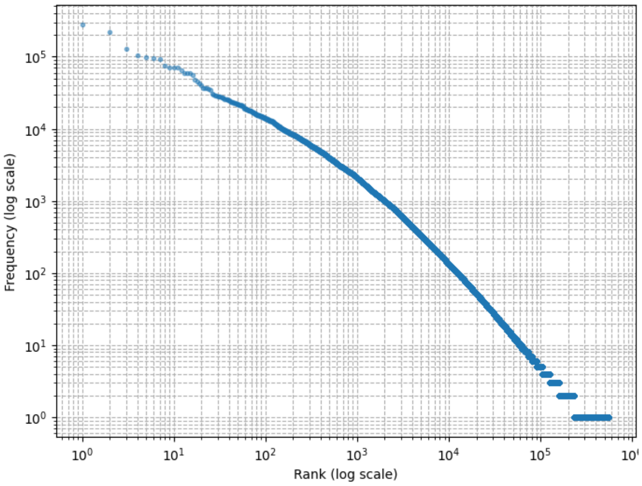


Fig. 1. Word Frequency V/S Rank graph

4.2. Vocabulary Growth (Heaps' Law)

With respect to the corpus size, the vocabulary growth is observed. The rise without plateauing after processing 14 million tokens can be observed, indicating the corpus is lexically rich and introduces neologisms, named entities, and domain-specific terms over the 15-year period [24, 5].

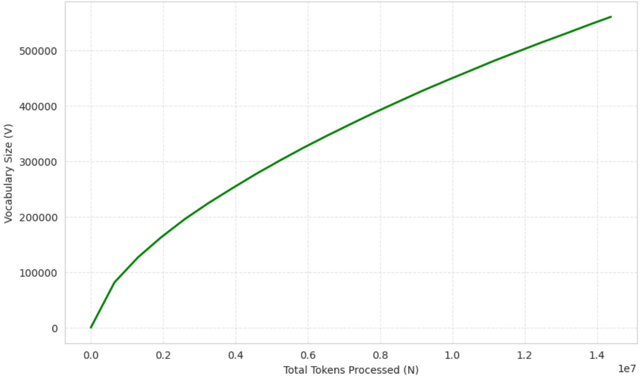


Fig. 2. Vocabulary growth graph

4.3. Semantic Topic Analysis

Using Latent Dirichlet Allocation (LDA) provided by Gensim [25], the latent semantic structure is extracted, revealing topics according to the structural priorities of Nepali journalism [15, 24, 26].

Topic	**Top Keywords**
Topic 1	अध्यक्ष, निर्वाचन, पार्टी, नेता, प्रधानमन्त्री, केन्द्रीय, छलफल, निर्णय, कांग्रेस, बैठकमा
Topic 2	हजार, काम, स्वास्थ्य, निर्माण, सडक, बढी, कारण, क्षेत्रमा
Topic 3	नेपाली, काम, सरकारले, प्रधानमन्त्री, उल्लेख, राजनीतिक, भनाइ
Topic 4	हजार, लाख, गरिएको, रुपैयाँ, निर्णय, व्यवस्था, करोड, दिएको, रकम
Topic 5	प्रहरी, जिल्ला, प्रहरीले, पक्राउ, मृत्यु, घर, जना, जनाएको, कार्यालय, जानकारी

Table 2. Top Keywords Identified for Each Topic

4.3.1. Governance and Political Parties

(Keywords: Chairman, Election, Politician, Prime Minister, Central, Meeting, Decision, Congress)
This cluster of topics is related to political parties, the central government, meetings, leadership disputes, and political leaders. The majority of news articles are more inclined to one of the major political parties, i.e., Congress.

4.3.2. Development and Infrastructure

(Keywords: Construction, Road, Health, Work, Area)
These topics aggregate news articles related to physical

and social infrastructure, showcasing the ongoing development of the country.

4.3.3. Executive Department

(Keywords: Government, Prime Minister, Political, Statement, Politics)
It differs from political parties, as this topic focuses more on the executive branch of government, primarily covering government decisions, the prime minister's statements, and nationwide political administration.

4.3.4. Finance and Economy

(Keywords: Rupees, Lakh, Crore, Amount, Management)
This cluster of topics mainly characterizes the financial aspects like currency, budget allocation, financial reports, and other economic aspects of the country.

4.3.5. Crime and Law enforcement

(Keywords: Police, Arrest, Death, District, Investigation)
These topics aggregate news related to police activities, particularly focusing on arrests, casualties, and security incidents.

The difference between Topic 1 and Topic 3 is a major finding, suggesting that party affairs and government affairs are distinguished.

4.4. Temporal Burstiness and Analysis of Events

In order to verify the historical records of the news article, the analysis of the temporal distribution is made over the interval from 2010 to 2025.

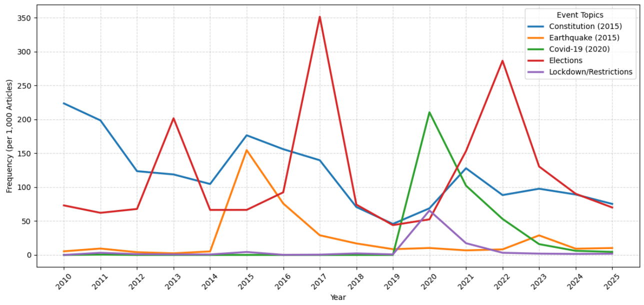


Fig. 3. Historical Events

- **Exogenous Shocks: The Signature of Crisis**
  The graph line representing earthquakes remained nearly zero from 2010 to 2014, validating the exclusion of false positive data. A vertical spike can be observed in 2015, where it was mentioned about 150 times per 1,000 articles. Afterwards, a "long-tail" effect can be observed ranging to 2017. After a while, a minor spike can again be observed in 2023 with an approximate mention in 30/1,000 mentions.
  This shows that the data is valid, since in 2015, a major earthquake of magnitude 7.8 hit [27]. A 4.9 magnitude earthquake was also observed in 2017 [28].

- **The Pandemic (COVID-19)**
  The graph line of COVID-19 and Lockdown remained nearly zero from 2010 to 2019. In 2020, both of them spiked significantly, but later the lockdown graph had a significant drop compared to COVID-19, reflecting that the virus persisted longer than the lockdown policy. As of current, the coverage of COVID-19 has been approaching almost zero.
  This validates the integrity of the data since the COVID-19 pandemic hit in 2020, causing a global lockdown [29].

- **Democratic and Election cycle**
  Three distinct peaks can be observed in 2013, 2017 and 2022. The magnitude of the 2017 election shows a higher spike(350/1,000 mentions) mainly due to the local, provincial, and federal elections at that time.
  There was no election held in 2021, but it shows a significant anomaly mainly due to the dissolution of the House of Representatives (December 2020) and the announcement of early elections for April 2021. The election was later voided by the Supreme Court, but there was significant media coverage. [30]
  This shows that the news articles cover not just physical events but also discursive anticipation and political intent.

- **The constitution**
  The graph line of the constitution has been the most discussed topic for the period of 15 years. Till 2014, the frequency of mention has remained almost consistent (>100/1,000 mentions) reflecting the discussion and drafting process by the assembly [31]. It peaked again in 2015, since the promulgation of the new constitution happened in September 2015 [32]. It again spiked in 2021, triggered due to the dissolution of parliament, and the legality of the Prime Minister was debated daily, where instead of constitutional formation, the constitutional defence is heavily discussed [**?**].
  This shows the government's instability in Nepal.

The above analysis shows that one of the major features of this corpus is its data on historical events. It is strongly able to cover all the historical events, making it a strong candidate for advanced analysis.

5. BASELINE CLASSIFICATION EXPERIMENTS

In order to establish a baseline for future research directions, a benchmark is exhibited by conducting a multi-class experiment using Scikit-learn [33].

5.1. Setup

The raw data lacked category labels, so a ”Silver Standard” test set was introduced using a strict keyword heuristic. The script was run to autolabel all the news articles. After auto-characterization, the sample data was verified. It is a fuzzy match, so some mismatch was expected, but it was able to categorize almost 90% correctly. After the categorization, the three models were used to compare: Multinomial Naive Bayes (MNB) [34], Logistic Regression (LR) [35] and Support Vector Machine (SVM) [36].

5.2. Findings

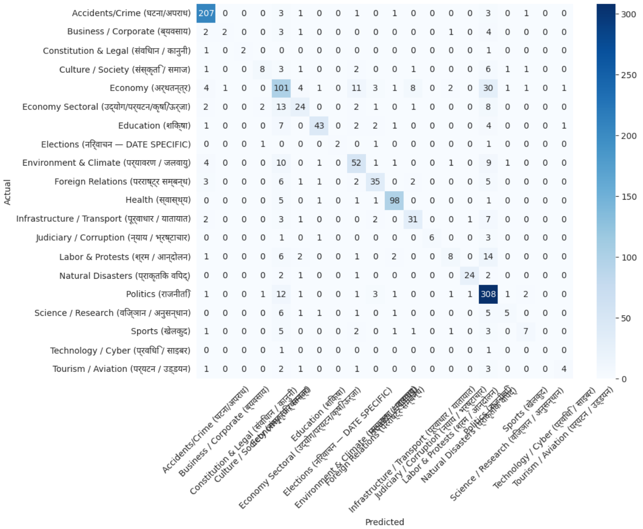
The SVM significantly outperformed other models with an accuracy of 74.50%.

Model	Accuracy (%)
Naive Bayes	59.32
Logistic Regression	68.64
Linear SVM	74.50

Table 3. Classification Accuracy of Different Models

6. ERROR DISCUSSION & LIMITATIONS

The confusion matrix shows that the model classified Politics and Accidents with the highest precision. However, we cannot neglect the questionable confusion being shown on the other metrics. This finding is based on a fuzzy method and is not 100% accurate, suggesting future work on Nepali news classification should strongly consider hierarchical labeling correctly and take semantic overlapping domains into account.



	Accidents/Crime (घटना/अपराध)	Business / Corporate (ब्यवसाय)	Constitution & Legal (संवधिान / कानुनी)	Culture / Society (संस्कृत / समाज)	Economy (अर्थतन्त्र)	Economy Sectoral (उद्योग/पर्यटन/कृषि/ऊर्जा)	Education (शक्षिा)	Elections (नर्वािचन — DATE SPECIFIC)	Environment & Climate (पर्यावरण / जलवायु)	Foreign Relations (परराष्ट्र सम्बन्ध)	Health (स्वास्थ्य)	Infrastructure / Transport (पूर्वाधार / यातायात)	Judiciary / Corruption (न्याय / भ्रष्टाचार)	Labor & Protests (श्रम / आन्दोलन)	Natural Disasters (प्राकृतकि वपिद्)	Politics (राजनीती)	Science / Research (वज्ञिान / अनुसन्धान)	Sports (खेलकुद)	Technology / Cyber (प्रवधि / साइबर)	Tourism / Aviation (पर्यटन / उड्डयन)
Accidents/Crime (घटना/अपराध)	207	0	0	0	3	1	0	0	1	0	1	0	0	0	0	3	0	1	0	0
Business / Corporate (ब्यवसाय)	2	2	0	0	3	1	0	0	0	0	0	0	0	1	0	4	0	0	0	0
Constitution & Legal (संवधिान / कानुनी)	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Culture / Society (संस्कृत / समाज)	1	0	0	8	3	1	0	0	2	0	0	1	0	0	0	6	1	1	0	0
Economy (अर्थतन्त्र)	4	1	0	0	101	4	1	0	11	3	1	8	0	2	0	30	1	1	0	1
Economy Sectoral (उद्योग/पर्यटन/कृषि/ऊर्जा)	2	0	0	2	13	24	0	0	2	1	0	1	0	0	0	8	0	0	0	0
Education (शक्षिा)	1	0	0	0	7	0	43	0	2	2	1	0	0	0	0	4	0	0	0	1
Elections (नर्वािचन — DATE SPECIFIC)	0	0	0	1	0	0	0	2	0	1	0	0	0	0	0	1	0	0	0	0
Environment & Climate (पर्यावरण / जलवायु)	4	0	0	0	10	0	1	0	52	1	1	0	0	1	0	9	1	0	0	0
Foreign Relations (परराष्ट्र सम्बन्ध)	3	0	0	0	6	1	1	0	2	35	0	2	0	0	0	5	0	0	0	0
Health (स्वास्थ्य)	0	0	0	0	5	0	1	0	1	1	98	0	0	0	0	1	0	0	0	0
Infrastructure / Transport (पूर्वाधार / यातायात)	2	0	0	0	3	1	0	0	0	2	0	31	0	0	1	7	0	0	0	0
Judiciary / Corruption (न्याय / भ्रष्टाचार)	0	0	0	0	1	0	1	0	0	0	0	0	6	0	0	3	0	0	0	0
Labor & Protests (श्रम / आन्दोलन)	1	0	0	0	6	2	0	0	1	0	2	0	0	8	0	14	0	0	0	0
Natural Disasters (प्राकृतकि वपिद्)	0	0	0	0	2	1	0	0	1	0	0	0	0	0	24	2	0	0	0	0
Politics (राजनीती)	1	0	0	1	12	1	0	0	1	3	1	0	0	1	1	308	1	2	0	0
Science / Research (वज्ञिान / अनुसन्धान)	0	0	0	0	6	1	1	0	1	0	1	0	0	0	0	5	5	0	0	0
Sports (खेलकुद)	1	0	0	0	5	0	0	0	2	0	1	1	0	1	0	3	0	7	0	0
Technology / Cyber (प्रवधि / साइबर)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Tourism / Aviation (पर्यटन / उड्डयन)	1	0	0	0	2	1	0	0	1	0	0	0	0	0	0	3	0	0	0	4

Fig. 4. Confusion Matrix

7. CONCLUSION & FUTURE WORK

This study presented Ekantipur-15Y, a validated benchmark corpus for the Nepali language consisting of over 14 million tokens. With the various statistical analysis we have confirmed the dataset’s quality along with the valid temporal analysis. However, there is a limitation to the study. A ”Silver Standard” test set was used for baseline experiments which relied heavily on strict keywords heuristics and a fuzzy matching method which is not 100% accurate.

Future research should focus on replacing the fuzzy heuristic labels with high-quality, human-annotated data to improve evaluation. Additionally, given the complexity of news categorization, future work should consider hierarchical labeling strategies to better handle overlapping domains. Finally, researchers are encouraged to apply Transformer-based models (e.g., BERT, RoBERTa) to this corpus, as they are expected to capture context and polysemy better than the traditional SVM baselines used in this study.

8. DATA AVAILABILITY STATEMENT

Due to copyright restrictions, the full text of the articles is not publicly shared. We instead release a Metadata Index containing Article IDs, URLs, Titles, Dates, and publisher via Zenodo (DOI: 10.5281/zenodo.18145188) [37]. This dataset allows researchers to identify the analyzed articles and reproduce the corpus by retrieving content directly from the source.

9. DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used **Gemini (Google)** in order to improve the readability, grammar, and language flow of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

10. REFERENCES

[1] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, Eds., Online, July 2020, pp. 6282–6293, Association for Computational Linguistics.

[2] Tej Bahadur Shahi and Ashok Kumar Pant, "Nepali news classification using naïve bayes, support vector machines and neural networks," in *2018 International Conference on Communication information and Computing Technology (ICCICT)*. Feb. 2018, IEEE.

[3] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kevin Knight, Ani Nenkova, and Owen Rambow, Eds., San Diego, California, June 2016, pp. 260–270, Association for Computational Linguistics.

[4] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal, "Diachronic word embeddings and semantic shifts: a survey," June 2018.

[5] Martin Gerlach and Eduardo G Altmann, "Stochastic model for the vocabulary growth in natural languages," *Phys. Rev. X.*, vol. 3, no. 2, May 2013.

[6] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao, "Deep learning–based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021.

[7] Pooja Rai, Sanjay Chatterji, and Byung-Gyu Kim, "Deep learning-based sequence labeling tools for nepali," *ACM Trans. Asian Low-resour. Lang. Inf. Process.*, vol. 22, no. 8, pp. 1–23, Aug. 2023.

[8] Ali Raza, Shafiq Ur Rehman Khan, Raja Sher Afgun Usmani, Ashok Kumar Das, and Shehzad Ashraf Chaudhry, "A novel temporal footprints-based framework for fake news detection," *IEEE Access*, vol. 12, pp. 172419–172428, 2024.

[9] Hamza Salem, Hadi Salloum, and Manuel Mazzara, "Mathematical model and algorithm for accurate main content extraction from news websites," *IEEE Access*, vol. 13, pp. 15694–15711, 2025.

[10] Muzammil Khan, Yasser Alharbi, Ali Alferaidi, Talal Saad Alharbi, and Kusum Yadav, "Metadata for efficient management of digital news articles in multilingual news archives," *SAGE Open*, vol. 13, no. 4, Oct. 2023.

[11] Erdinc Uzun, "A novel web scraping approach using the additional information obtained from web pages," *IEEE Access*, vol. 8, pp. 61726–61740, 2020.

[12] Jonathan Paige, "The legality and ethics of web scraping in archaeology," *Adv. Archaeol. Pr.*, vol. 12, no. 2, pp. 98–106, May 2024.

[13] Yanfang Li, "Zipf's law in china's local government work reports: A 21-year study using natural language processing and regression analysis," *PLoS One*, vol. 20, no. 5, pp. e0324713, May 2025.

[14] Slobodan Beliga, Sanda Martinčić-Ipšić, Mihaela Matešić, Irena Petrijevčanin Vuksanović, and Ana Meštrović, "Infoveillance of the croatian online media during the COVID-19 pandemic: One-year longitudinal study using natural language processing," *JMIR Public Health Surveill.*, vol. 7, no. 12, pp. e31540, Dec. 2021.

[15] Simon L Evans, Rosalind Jones, Erkan Alkan, Jaime Simão Sichman, Amanul Haque, Francisco Bráulio Silva de Oliveira, and Davoud Mougouei, "The emotional impact of COVID-19 news reporting: A longitudinal study using natural language processing," *Hum. Behav. Emerg. Technol.*, vol. 2023, pp. 1–16, Mar. 2023.

[16] David Rozado, Ruth Hughes, and Jamin Halberstadt, "Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models," *PLoS One*, vol. 17, no. 10, pp. e0276367, Oct. 2022.

[17] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto, "Benchmarking large language models for news summarization," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 39–57, Jan. 2024.

[18] The pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020.

[19] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[20] Michael L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, pp. 3021, 2021.

[21] Steven Bird and Edward Loper, "NLTK: The natural language toolkit," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Barcelona, Spain, July 2004, pp. 214–217, Association for Computational Linguistics.

[22] A Chacoma and D H Zanette, "Heaps' law and heaps functions in tagged texts: evidences of their linguistic relevance," *R. Soc. Open Sci.*, vol. 7, no. 3, pp. 200008, Mar. 2020.

[23] Mason Smetana and Lev Khazanovich, "Publication trend analysis and synthesis via large language model: A case study of engineering in PNAS," Oct. 2025.

[24] Charlie Pilgrim and Thomas T Hills, "Bias in zipf's law estimators," *Sci. Rep.*, vol. 11, no. 1, pp. 17309, Aug. 2021.

[25] Radim Rehurek and Petr Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[26] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[27] "Resource Guide: Nepal's Journey from Post-Earthquake Reconstruction to Resilience — worldbank.org," `https://www.worldbank.org/en/country/nepal/brief/post-earthquake-reconstruction-in-nepal`, [Accessed 17-12-2025].

[28] Disaster Risk Reduction Knowledge Service, "July 2, 2017 Earthquake Information of Nepal Disaster Risk Reduction Knowledge Service — ikcest-drr.data.ac.cn," `https://ikcest-drr.data.ac.cn/data/96d2b`, [Accessed 17-12-2025].

[29] Thakshana Vadivel, Kiran Belur, and Kala Paneerselvam, "Coronavirus disease 2019 (COVID-19) pandemic outburst: A web-based cross-sectional study of the knowledge, attitudes, and practices among undergraduate students at a tertiary care teaching hospital in tamil nadu," *Cureus*, vol. 17, no. 3, pp. e81425, Mar. 2025.

[30] Kristine Eck, "Nepal in 2021," *Asian Surv.*, vol. 62, no. 1, pp. 193–200, Feb. 2022.

[31] International IDEA, *Nepal's Constitution Building Process: 2006-2015*, International Institute for Democracy and Electoral Assistance, Stockholm, 2015.

[32] The Asia Foundation, “Nepal’s constitution and federalism: Vision and implementation,” Tech. Rep., The Asia Foundation, Kathmandu, Nepal, 2020.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes, “Multinomial naive bayes for text categorization revisited,” in *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence*, Berlin, Heidelberg, 2004, AI’04, p. 488–499, Springer-Verlag.

[35] David R Cox, “The regression analysis of binary sequences,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[36] Corinna Cortes and Vladimir Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[37] Diwash Mainali, “Ekantipur-15Y metadata,” 2026.