Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ekantipur-15Y: A Longitudinal Benchmark Corpus and Semantic Analysis of Nepali News (2010 - 2025)

Diwash Mainali

diwash006@gmail.com

Utsav Mainali

Research Article

Keywords: Nepali NLP, Low-Resource Languages, Longitudinal Corpus, Text Classification, Event Detection

Posted Date: March 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8630749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ekantipur-15Y: A Longitudinal Benchmark Corpus and Semantic Analysis of Nepali News (2010 – 2025)

Diwash Mainali [1*] and Utsav Mainali [2†]

*Corresponding author(s). E-mail(s): diwash006@gmail.com;
Contributing authors: utsav.mainali@protonmail.me;
†These authors contributed equally to this work.

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.

Keywords: Nepali NLP, Low-Resource Languages, Longitudinal Corpus, Text Classification, Event Detection

1 Introduction

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 evidence 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 [13–15]. 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 (pūrṇa virām, ।) 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.

Table 1 Key Statistics of the Ekantipur-15Y Corpus

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

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 [5, 22]. 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].

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 [5, 24].

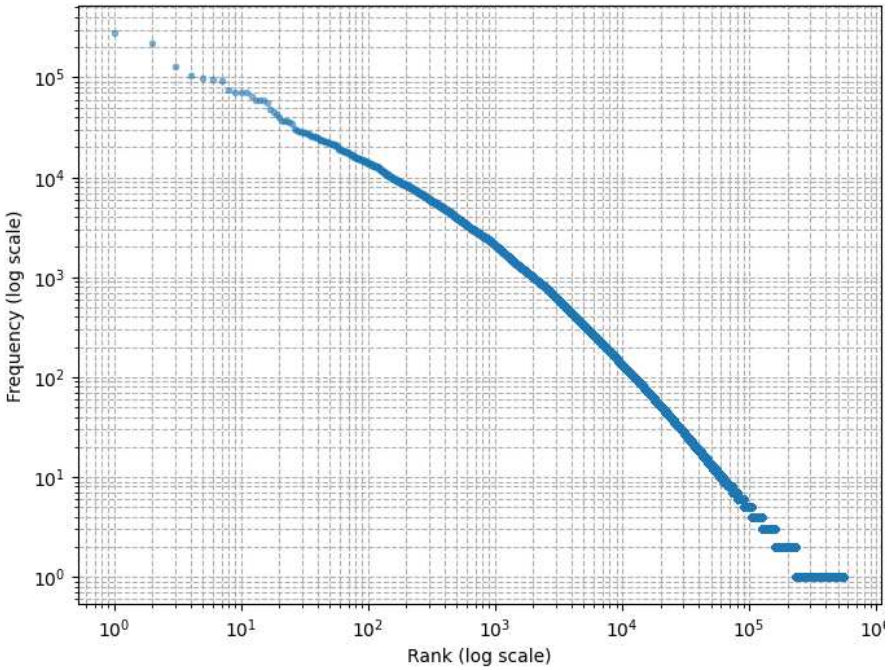


Fig. 1 Word Frequency V/S Rank graph

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].

Table 2 Top Keywords Identified for Each Topic in Romanized Nepali

Topic	Top Keywords
Topic 1	adhyaksha, nirvachana, parti, neta, pradhanamantri, kendriya, chalfal, nirnaya, kangres, baithakma
Topic 2	hajar, kaam, svasthya, nirmana, sadak, badhi, karan, kshetra
Topic 3	nepali, kaam, sarkarle, pradhanamantri, ullekh, rajanaitik, bhanai
Topic 4	hajar, lakh, gariyeko, rupaiya, nirnaya, vyavastha, karod, diyeko, rakam
Topic 5	prahari, jilla, praharile, pakrau, mrtyu, ghar, jana, janaeko, karyalay, jankari

4.3.1 Governance and Political Parties

(Keywords: Chairman, Election, Politician, Prime Minister, Central, Meeting, Decision, Congress)

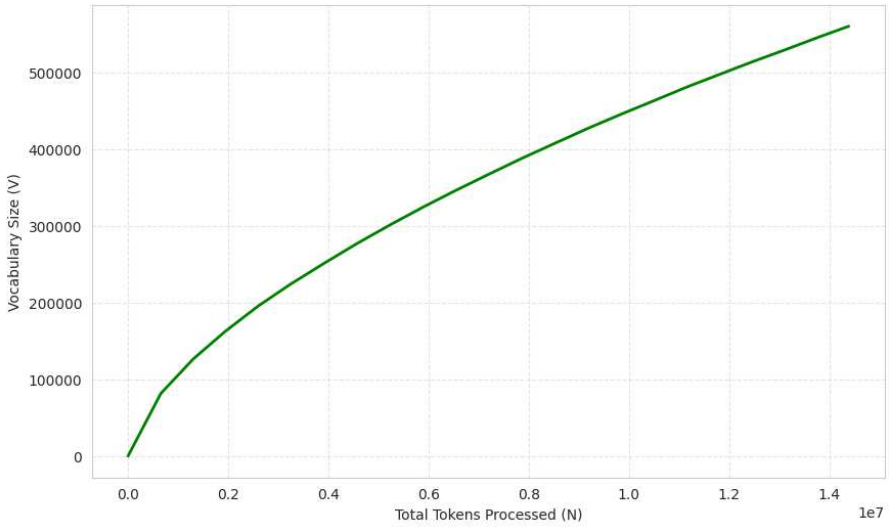


Fig. 2 Vocabulary growth graph

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.

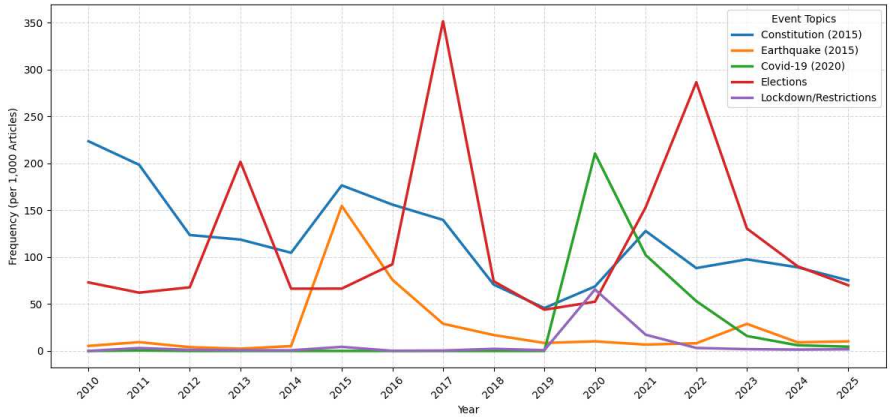


Fig. 3 Historical Events

- **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($\tilde{3}$50/1,000 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 [30].
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%.

Table 3 Classification Accuracy of Different Models

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

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.

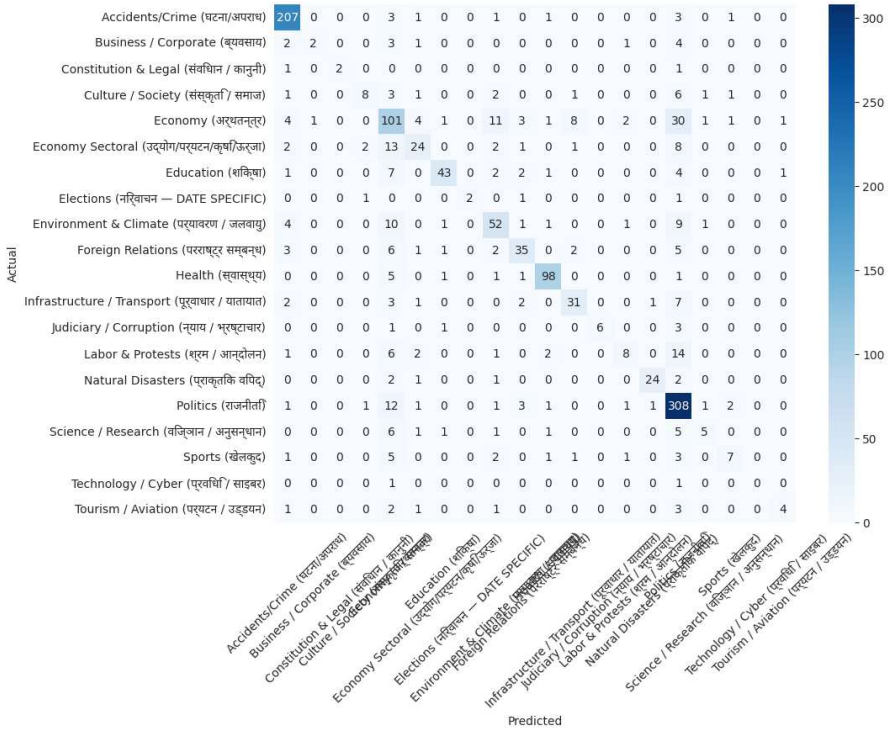


Fig. 4 Confusion Matrix

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.

References

[1] Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the NLP world. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.560 . https://aclanthology.org/2020.acl-main.560/

[2] Shahi, T.B., Pant, A.K.: Nepali news classification using naïve bayes, support vector machines and neural networks. In: 2018 International Conference on Communication Information and Computing Technology (ICCICT), pp. 1–5. IEEE, Mumbai (2018). https://doi.org/10.1109/iccict.2018.8325883 . http://dx.doi.org/10.1109/ICCICT.2018.8325883

[3] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Knight, K., Nenkova, A., Rambow, O. (eds.) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-1030 . https://aclanthology.org/N16-1030/

[4] Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E.: Diachronic word embeddings and semantic shifts: a survey (2018) arXiv:1806.03537 [cs.CL]

[5] Gerlach, M., Altmann, E.G.: Stochastic model for the vocabulary growth in natural languages. Phys. Rev. X. **3**(2) (2013)

[6] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: A comprehensive review. ACM Comput. Surv. **54**(3) (2021) https://doi.org/10.1145/3439726

[7] Rai, P., Chatterji, S., Kim, B.-G.: Deep learning-based sequence labeling tools for nepali. ACM Trans. Asian Low-resour. Lang. Inf. Process. **22**(8), 1–23 (2023)

[8] Raza, A., Ur Rehman Khan, S., Sher Afgun Usmani, R., Kumar Das, A., Ashraf Chaudhry, S.: A novel temporal footprints-based framework for fake news detection. IEEE Access **12**, 172419–172428 (2024)

[9] Salem, H., Salloum, H., Mazzara, M.: Mathematical model and algorithm for accurate main content extraction from news websites. IEEE Access **13**, 15694–15711 (2025)

[10] Khan, M., Alharbi, Y., Alferaidi, A., Alharbi, T.S., Yadav, K.: Metadata for efficient management of digital news articles in multilingual news archives. SAGE Open **13**(4) (2023)

[11] Uzun, E.: A novel web scraping approach using the additional information obtained from web pages. IEEE Access **8**, 61726–61740 (2020)

[12] Paige, J.: The legality and ethics of web scraping in archaeology. Adv. Archaeol. Pr. **12**(2), 98–106 (2024)

[13] Li, Y.: Zipf's law in china's local government work reports: A 21-year study using natural language processing and regression analysis. PLoS One **20**(5), 0324713 (2025)

[14] Beliga, S., Martinčić-Ipšić, S., Matešić, M., Petrijevčanin Vuksanović, I., Meštrović, A.: Infoveillance of the croatian online media during the COVID-19 pandemic: One-year longitudinal study using natural language processing. JMIR Public Health Surveill. **7**(12), 31540 (2021)

[15] Evans, S.L., Jones, R., Alkan, E., Sichman, J.S., Haque, A., Oliveira, F.B.S., Mougouei, D.: The emotional impact of COVID-19 news reporting: A longitudinal study using natural language processing. Hum. Behav. Emerg. Technol. **2023**, 1–16 (2023)

[16] Rozado, D., Hughes, R., Halberstadt, J.: Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. PLoS One **17**(10), 0276367 (2022)

[17] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B.:

Benchmarking large language models for news summarization. Trans. Assoc. Comput. Linguist. **12**, 39–57 (2024)

[18] team, T.: pandas-dev/pandas: Pandas. Zenodo (2020). https://doi.org/10.5281/zenodo.3509134 . https://doi.org/10.5281/zenodo.3509134

[19] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007) https://doi.org/10.1109/MCSE.2007.55

[20] Waskom, M.L.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021) https://doi.org/10.21105/joss.03021

[21] Bird, S., Loper, E.: NLTK: The natural language toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 214–217. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/P04-3031/

[22] Chacoma, A., Zanette, D.H.: Heaps' law and heaps functions in tagged texts: evidences of their linguistic relevance. R. Soc. Open Sci. **7**(3), 200008 (2020)

[23] Smetana, M., Khazanovich, L.: Publication trend analysis and synthesis via large language model: A case study of engineering in PNAS (2025) arXiv:2510.16152 [cs.DL]

[24] Pilgrim, C., Hills, T.T.: Bias in zipf's law estimators. Sci. Rep. **11**(1), 17309 (2021)

[25] Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**(2) (2011)

[26] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 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] Service, D.R.R.K.: 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] Vadivel, T., Belur, K., Paneerselvam, K.: 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 **17**(3), 81425 (2025)

[30] Eck, K.: Nepal in 2021. Asian Surv. **62**(1), 193–200 (2022)

[31] International IDEA: Nepal's Constitution Building Process: 2006-2015. International Institute for Democracy and Electoral Assistance, Stockholm (2015). https://www.idea.int/sites/default/files/publications/nepals-constitution-building-process-2006-2015.pdf

[32] The Asia Foundation: Nepal's constitution and federalism: Vision and implementation. Technical report, The Asia Foundation, Kathmandu, Nepal (2020). https://asiafoundation.org/wp-content/uploads/2020/10/Nepals-Constitution-and-Federalism_Vision-and-Implementation_English.pdf

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

[34] Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive bayes for text categorization revisited. In: Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence. AI'04, pp. 488–499. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30549-1_43 . https://doi.org/10.1007/978-3-540-30549-1_43

[35] Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958)

[36] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)

[37] Mainali, D.: Ekantipur-15Y Metadata. Diwash Mainali (2026). https://doi.org/10.5281/ZENODO.18145188 . https://zenodo.org/doi/10.5281/zenodo.18145188

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Baselinemodel.png
- highlevelcrawlerarchitecture.png