Nepali Context-Aware Spelling Tool

Team Members:

Anish Raj Manandhar	**[THA077BCT010]**
Nabin Shrestha	**[THA077BCT026]**
Prayush Bhattarai	**[THA077BCT035]**

Supervised By:
Er. Shanta Maharjan

Co-Supervised by:
Er. Pravin Acharya

Department of Electronics and Computer Engineering
Thapathali Campus

June, 2024

Presentation Outline

- Motivation
- Objectives
- Scopes
- Applications
- Proposed Methodology
- Expected Results
- Tentative Timeline(Gantt Chart)
- Estimated Project Expense
- References

Motivation

- Gap in Research and Development
- Limited Resources
- Inadequate Existing Tools for Contextual Solutions

Objective

- To develop a sophisticated spelling checker that can detect and correct spelling errors based on the context of the entire sentence.

Scopes

Capabilities

- Accurately detect spelling errors by analyzing the context within sentences.
- Display the correct spelling options.

Challenges:

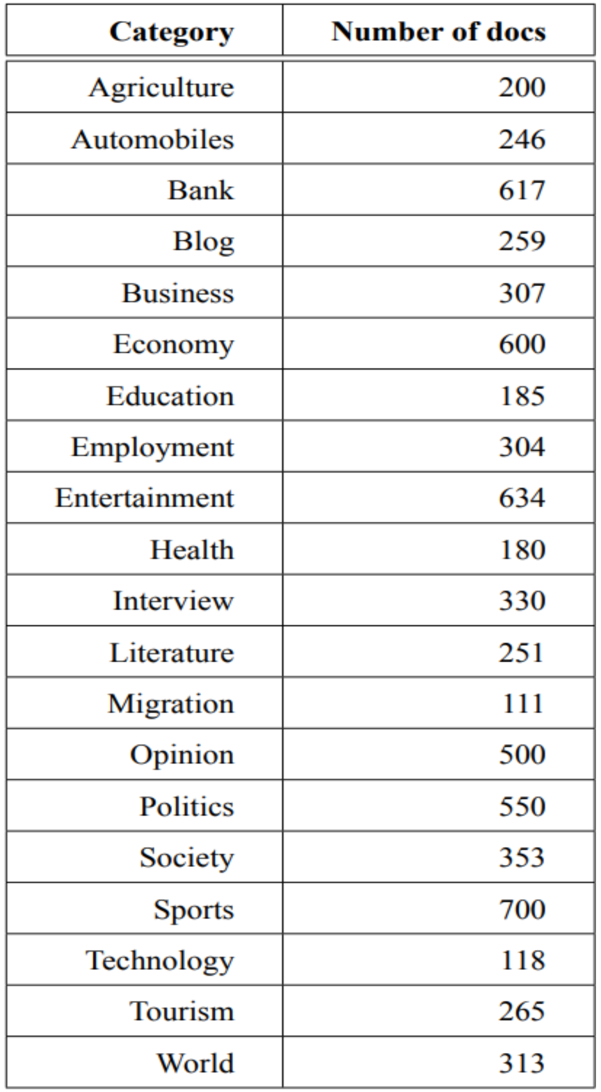
- Need for a diverse and comprehensive text corpus for training.
- Collecting and preprocessing large amounts of high-quality data.

Applications

- Media and Publishing
- OCR Projects
- Reliable TTS Systems
- Search Engines
- Text Processor Systems

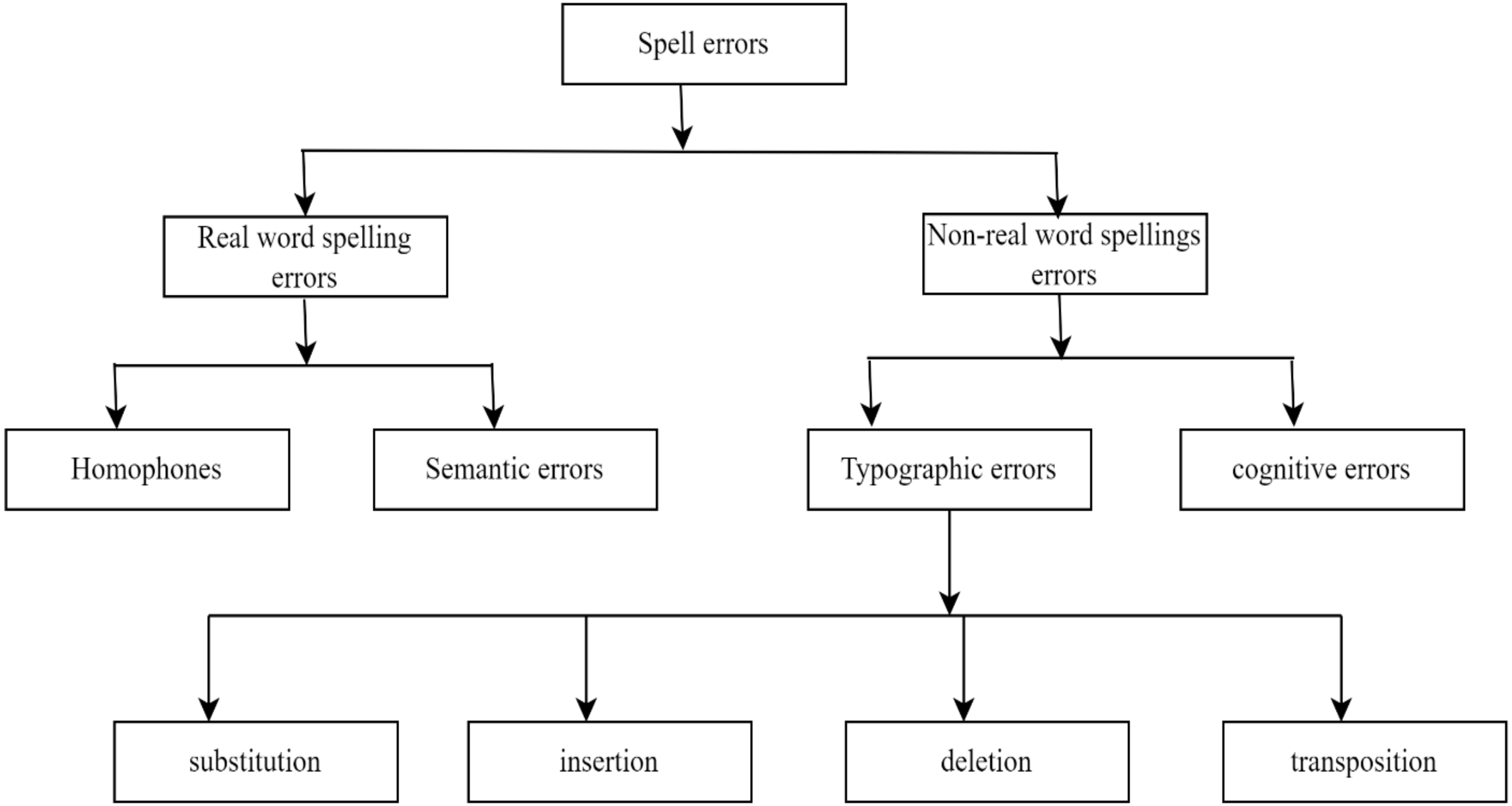
Dataset

- Collection of Nepali news articles categorized into 20 distinct categories
- extracted from the most trusted Nepali newspapers, such as Kantipur and Gorkha Patra
- 73,000 newspaper articles



Category	Number of docs
Agriculture	200
Automobiles	246
Bank	617
Blog	259
Business	307
Economy	600
Education	185
Employment	304
Entertainment	634
Health	180
Interview	330
Literature	251
Migration	111
Opinion	500
Politics	550
Society	353
Sports	700
Technology	118
Tourism	265
World	313

Categories of Error



Spell errors

Real word spelling errors

Non-real word spellings errors

Homophones

Semantic errors

Typographic errors

cognitive errors

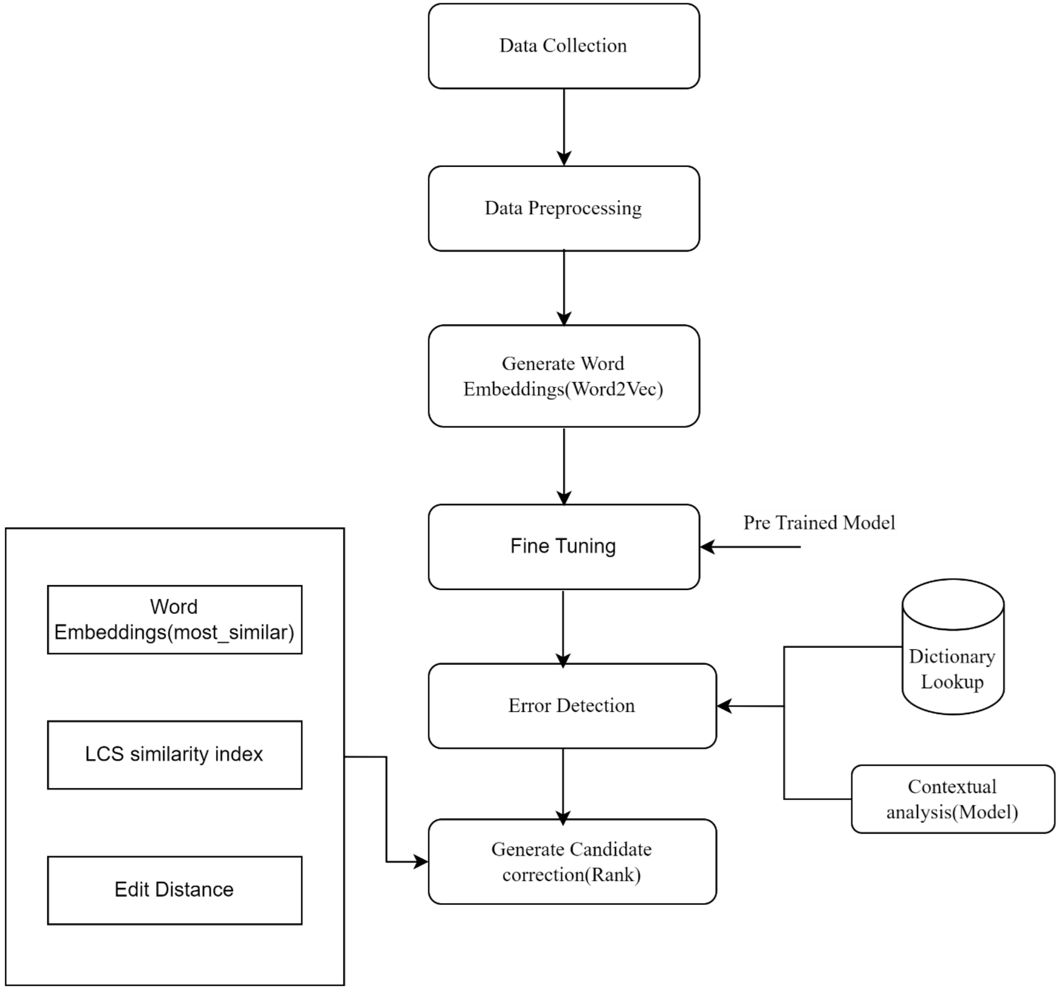
substitution

insertion

deletion

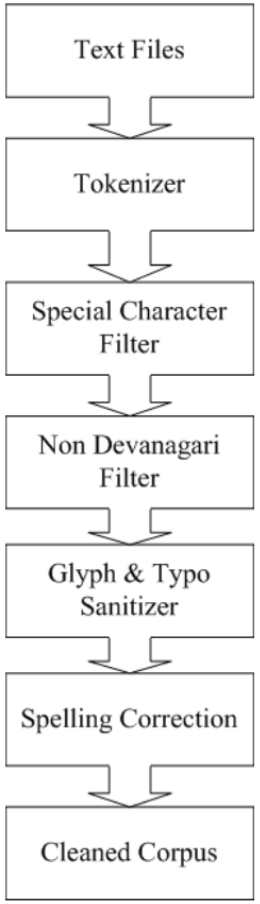
transposition

Proposed Methodology-[1]
(System Block Diagram)



Proposed Methodology-[2] (Preprocessing pipeline)

- Tokenizer
  - Based on end-of-sentence marker
- Special Character Filter
  - Filter characters not used in Nepali
  - Eg:←�…¬ = > < @ # $ ^ & * | \ / ` ~ _ { } [ ]
- Non-Devanagari Filter
  - Heuristics-based algorithm
- Glyph and Typo Sanitizer
  - Use glyph and typo mapping table
  - Regular Expression



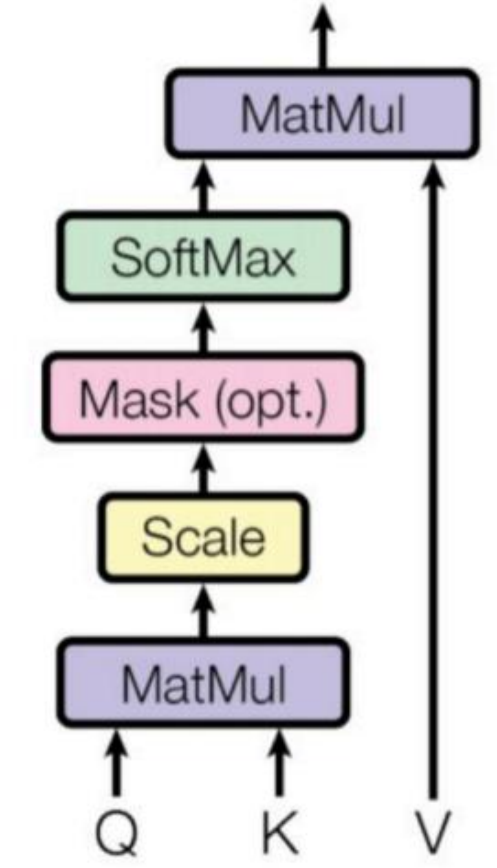
Proposed Methodology-[3] (Word2Vec)

- Transforms words into high-dimensional vector representations.
- Captures semantic relationships between words.
- Similar meanings are located close to each other in vector space.
- Vector representations capture meanings based on context.
- Training on collected corpus.
- Examples: "सुन्दर" and "राम्री"

Proposed Methodology-[4] (Self-Attention)

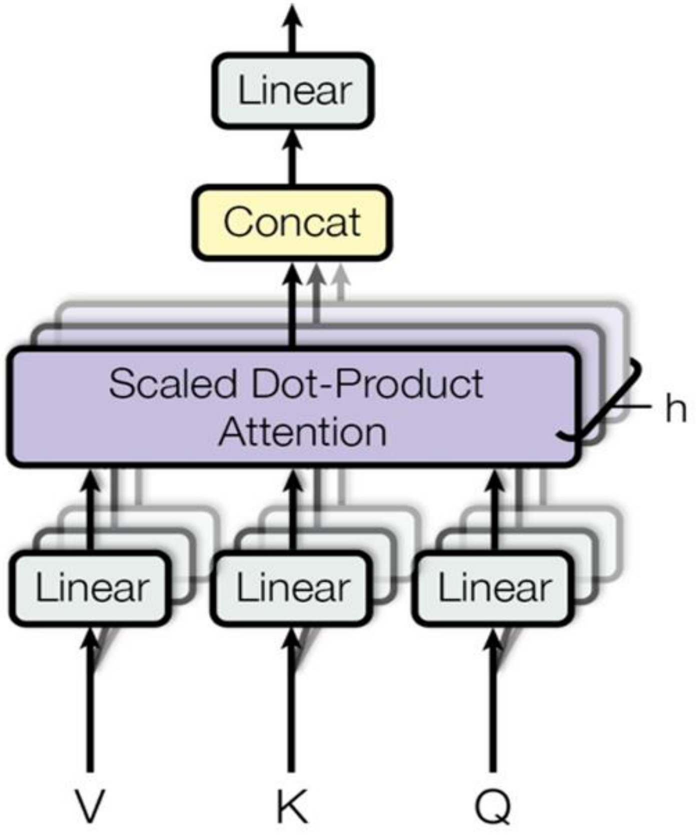
- Self-Attention $(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Where, Q = Query
K = Key
V = Value



Proposed Methodology-[5] (Multi-headed Attention)

- Combines multiple self attention module
- Learns different context attentions



Proposed Methodology-[6] (Error Detection)

- Identifying incorrect words in text.

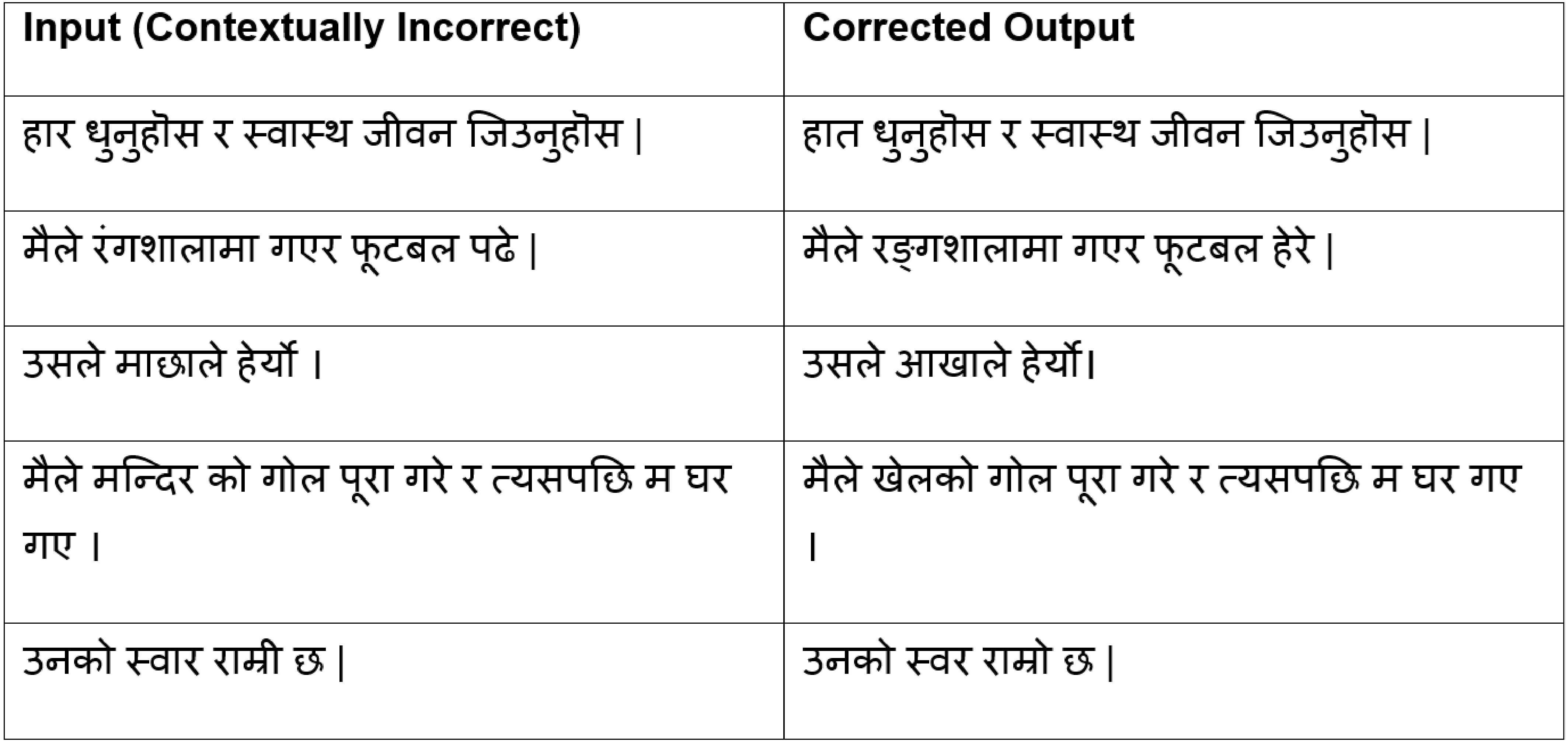
Methods:

- Dictionary Lookup
  - Checking if words exist in a predefined lexicon.
  - Words not found in the dictionary are flagged as potential errors.
- Contextual Analysis
  - Using models to determine if a word fits within the context.
  - Example: " उनको स्वार राम्री छ |" → " स्वार & राम्री " flagged as incorrect.

Proposed Methodology-[7]
(Contextual Ranking of Candidates)

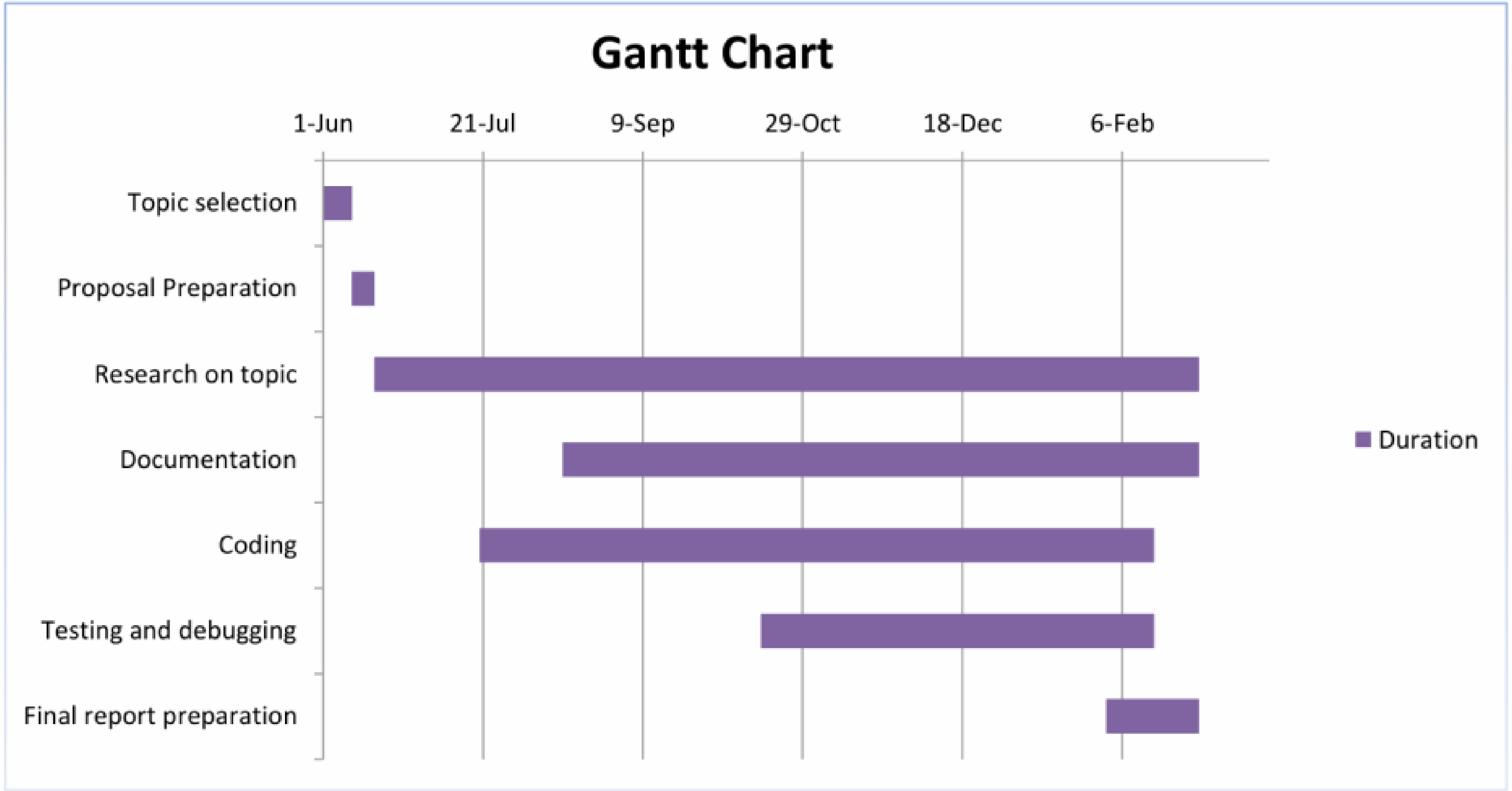
- Selecting the most appropriate correction based on surrounding text.
- Methods:
  - Cosine Similarity
  - Comparing context vector with candidate vectors.
  - Candidates with highest similarity scores are ranked higher.
  - Neural Networks
  - Using Transformer models to rank candidates.

Expected Results

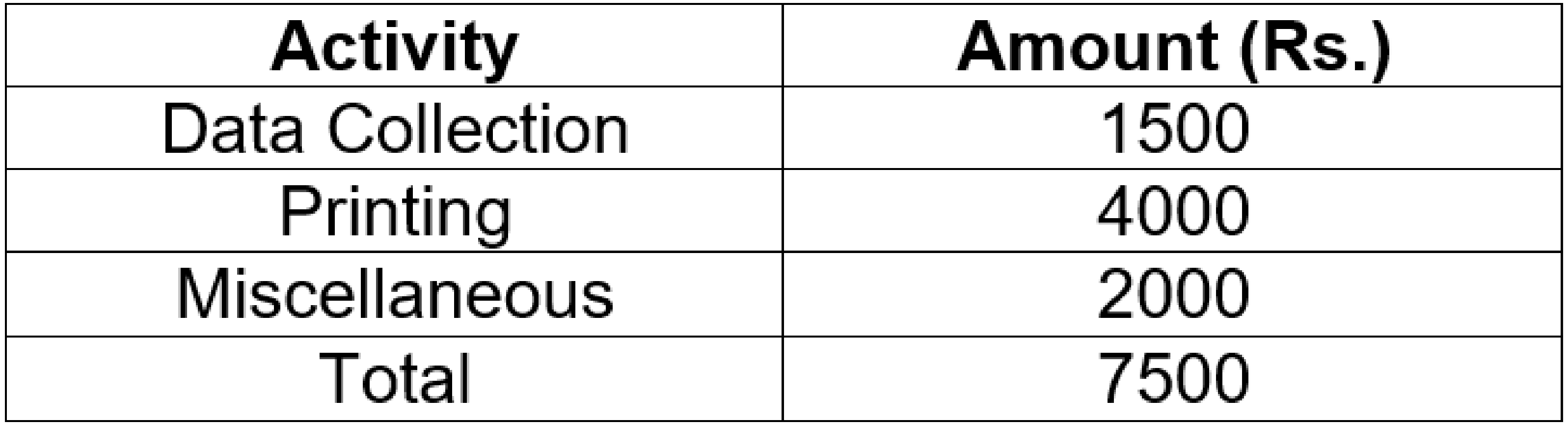


Input (Contextually Incorrect)	Corrected Output
हार धुनुहोस र स्वास्थ जीवन जिउनुहोस |	हात धुनुहोस र स्वास्थ जीवन जिउनुहोस |
मैले रंगशालामा गएर फूटबल पढे |	मैले रङ्गशालामा गएर फूटबल हेरे |
उसले माछाले हेर्यो ।	उसले आखाले हेर्यो।
मैले मन्दिर को गोल पूरा गरे र त्यसपछि म घर गए ।	मैले खेलको गोल पूरा गरे र त्यसपछि म घर गए ।
उनको स्वार राम्री छ |	उनको स्वर राम्रो छ |

Tentative Timeline(Gantt Chart)



Estimated Project Expense



Activity	Amount (Rs.)
Data Collection	1500
Printing	4000
Miscellaneous	2000
Total	7500

References-[1]

[1] A. M. Turing, Computing machinery and intelligence. Springer, 2009.

[2] P. Gupta, “A context-sensitive real-time spell checker with language adaptability,” 2020, 10.1109/ICSC.2020.00023. [Online]. Available: 10.1109/ICSC.2020.00023

[3] B. Prasain, N. lamichhane, N. Pandey, P. Adhikari, and P. Mudbhari, “Nepali spell checker,” 2023, https://doi.org/10.3126/jes2.v1i1.58461.

[4] S. Bista, Kumar, B. Keshari, L. Khatiwada, Prasad, P. Chitrakar, and S. Gurung, “Nepali lexicon development,” 2004-2007, https://www.yumpu.com/en/document/view/25135568/nepali-lexicon-development-pan-localization.

[5] X. Ziang, A. Anand, A. Naveen, J. Dan, and A. Y. Ng, “Neural language correction with character-based attention,” 2016, https://doi.org/10.48550/arXiv. 1603.09727. [6] N. Luitel, N. Bekoju, A. Kumar Sah, and S. Shakya, “Contextual spelling correction with language model for low-resource setting,” 2024, https://doi.org/10.48550/arXiv.1603.09727.

[7] A. PAL1 and A. MUSTAFI2, “Automatic context-sensitive spelling correction of ocr-generated hindi text using bert and levenshtein distance,” 2020, https://doi.org/10.48550/arXiv.2012.076527.

References-[2]

[8] Y. Bassil and M. Alwani, “A context-sensitive spelling correction using google web 1t 5-gram information,” 2020, https://doi.org/10.48550/arXiv.1204.5852.

[9] B. Rijal and S. B. Basnet, “Vector distance based spelling checking systemin nepali with language-dependent,” 2020.

[10] B. Rijal, S. Basnet, S. Awale, and S. Prasai, “Preprocessing of nepali news corpus for downstream tasks,” 2022, https://doi.org/10.3126/nl.v35i01.46553.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” CoRR, vol. abs/1409.3215, 2014. [Online]. Available: http://arxiv.org/abs/1409.3215 47

[12] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.