# Influence of Contextual Information on Bengali-English Forward and Backward Transliteration Using Binary Coding

Paper ID: 75

Date: December 08, 2023

Md. Fahad *

- Department of CSE, Hajee Mohammad Danesh Science & Technology University, Dinajpur, Bangladesh

Amrita Das Tipu *

- Department of CSE, Hajee Mohammad Danesh Science & Technology University, Dinajpur, Bangladesh

Ashis Kumar Mandal

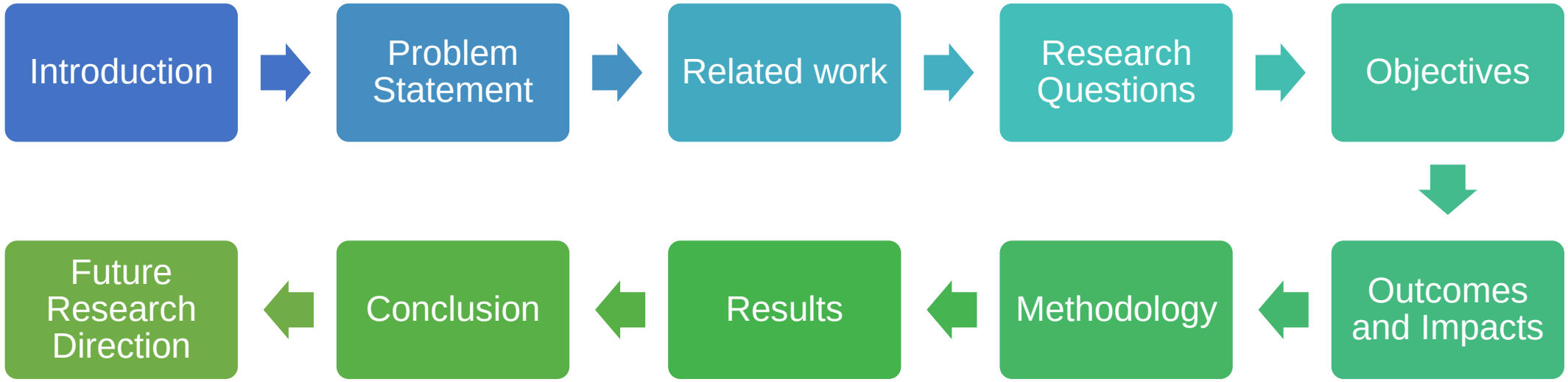
- Department of CSE, Hajee Mohammad Danesh Science & Technology University, Dinajpur, Bangladesh
- Department of Computer Science, University of Saskatchewan, Saskatoon, Canada

Basabi Chakraborty

- Research and Regional Cooperation Division, Iwate Prefectural University, Takizawa, Japan
- School of Computing, Madanapalle Institute of Technology and Science, Madanapalle, India


# Authors Information

# Outline


Introduction
Problem Statement
Related work
Research Questions
Objectives
Outcomes and Impacts
Methodology
Results
Conclusion
Future Research Direction

# Introduction

Transliteration

- Converts the written form of a language
- From one language to another
- Retains phonetic meaning

Table 1: Bengali to English transliteration

| Original | Transliterated |
|---|---|
| শিক্ষাবিদ | shikkhabid |
| বাংলা | bangla |
| ভাষা | bhasha |
| আমার | amar |

# Related Work

Table 2: Literature review

| AUTHOR | CONTRIBUTION | LIMITATION | DIFFERENCE |
|---|---|---|---|
| Ekbal et al. [1] | 6 n-gram based probabilistic models | Only used 6000 NEs, private dataset | Evaluated on common words, provides algorithm for TU identification |
| Sarkar and Chatterjee [2] | One-hot coding for representing TU, used traditional ML model SVM and KNN | Evaluated on only 1000 NEs, private dataset | Our study uses binary coding in place of one hot coding for the TUs, uses 6 contextual models |
| Dasgupta et al. [3] | Joint source channel model, SMT model | Backward transliteration approach | Both forward and backward transliteration, TU identification algorithm |
| Tipu et al. [4] | Binary coding for TU representation, evalua ted with process time | Only one feature representation, and no backward transliteration | This study uses 6 contextual models and evaluates both forward and backward transliteration |

# Problem Statement

- Limited publicly available datasets
- Lack of standardization for Bengali language
- Insufficient research using grapheme-level context

# Research Questions

How do contextual feature selection techniques affect transliteration outcomes?

How to increase transliteration accuracy?

# Objectives

- Automate TU decomposition step
- Explore the impact of contextual feature selection
- Investigate transliteration using a grapheme-based approach

# Outcomes and Impacts

**Expected outcome:**

- Correct transliteration

**Possible impacts:**

- Automatic and reliable system
- Unhindered cross-lingual communication

# Methodology

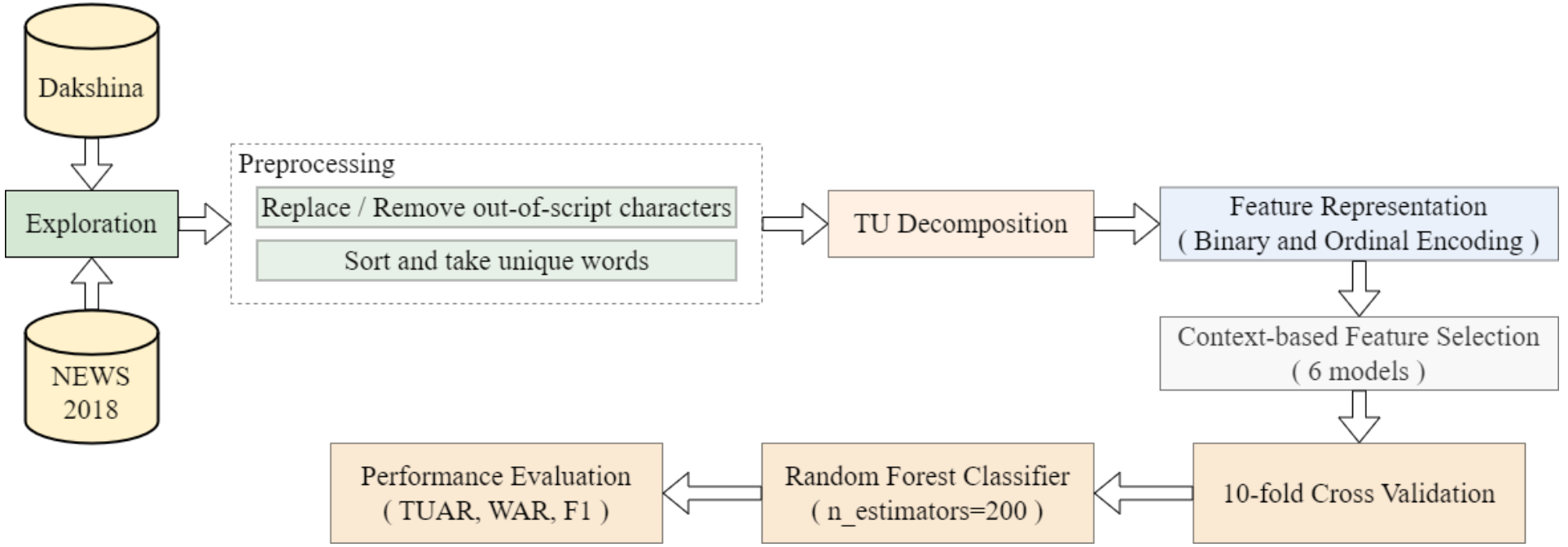



Figure 1: Proposed Methodology

# Methodology

- Types of words in dataset
  - Dakshina [5]: Dictionary, NE, Technical
  - NEWS 2018 [6]: NE
- **Invalid words**: Words containing null/empty, punctuations, numbers, out-of-script characters
- Sort and take unique Bengali words

Table 3: Dataset Statistics

| Criteria | Dakshina | NEWS 2018 |
|---|---|---|
| Total Words | 130378 | 13623 |
| Valid Words | 113873 | 13614 |
| Unique Words ( Bengali, English ) | ( 25395, 25395 ) | ( 13312, 13312 ) |

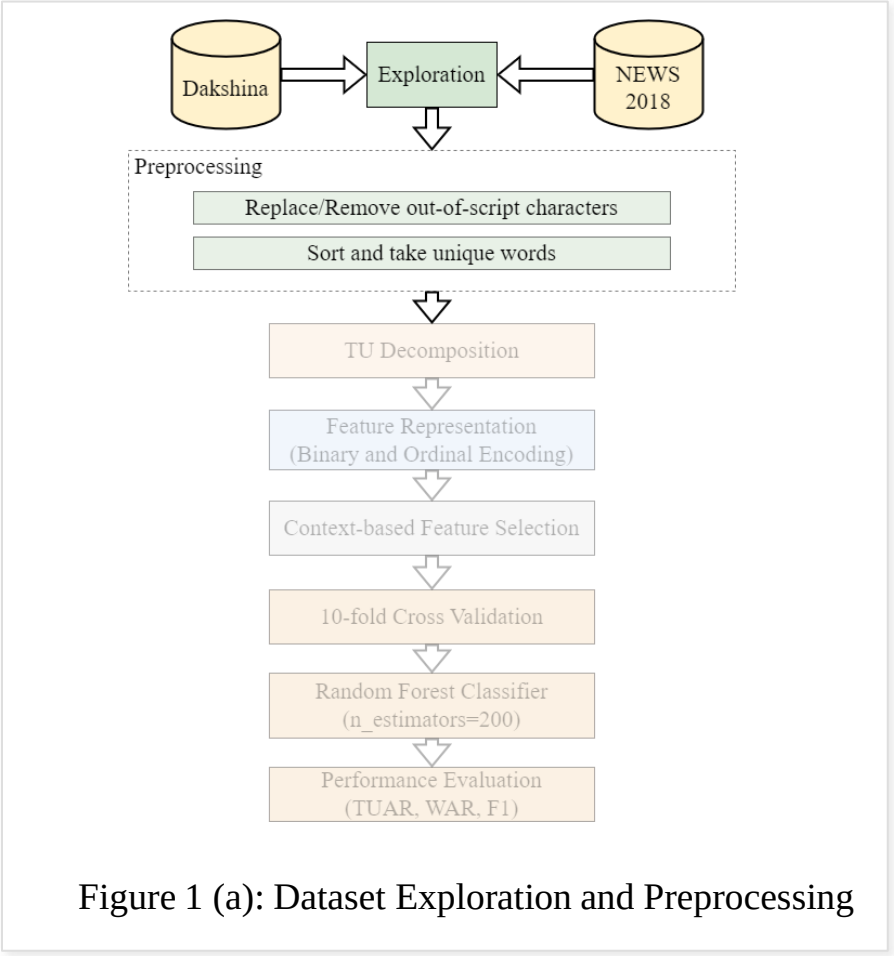



Figure 1 (a): Dataset Exploration and Preprocessing

# Methodology

- Transliteration Unit (TU) decompose:
  - আমার → [ আ | মা | র ]
  - amar → [ a | ma | r ]

Table 4: TU Statistics

| Criteria | Dakshina | NEWS 2018 |
|---|---|---|
| TU aligned Words | 15431 | 7535 |
| Average number of TU per word | 3.473 | 3.249 |
| Maximum number of TU per word | 9 | 8 |
| Minimum number of TU per word | 1 | 1 |
| Number of unique TUs ( Bengali, English ) | ( 1401, 1616 ) | ( 927, 1163 ) |

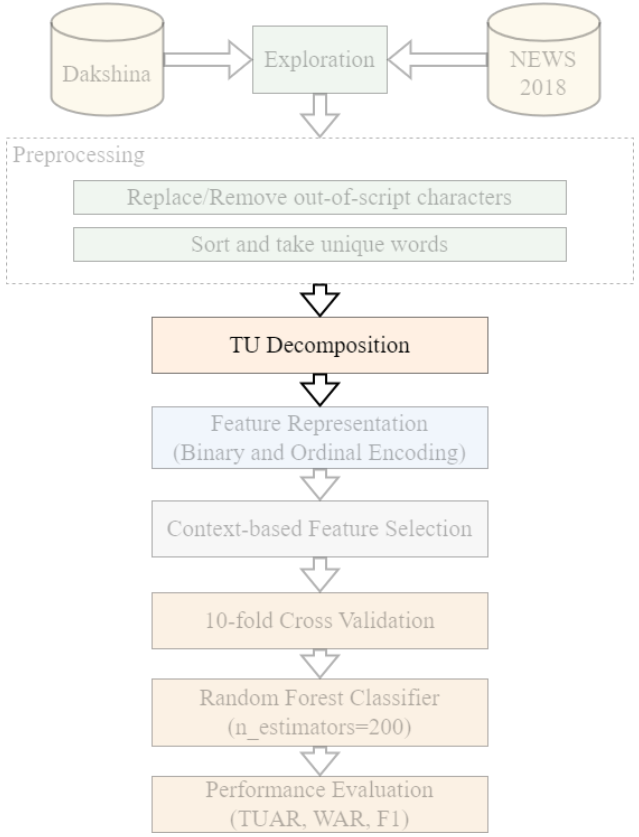



Figure 1 (b): TU decomposition

# Methodology

- Encode: converts string to numerical form

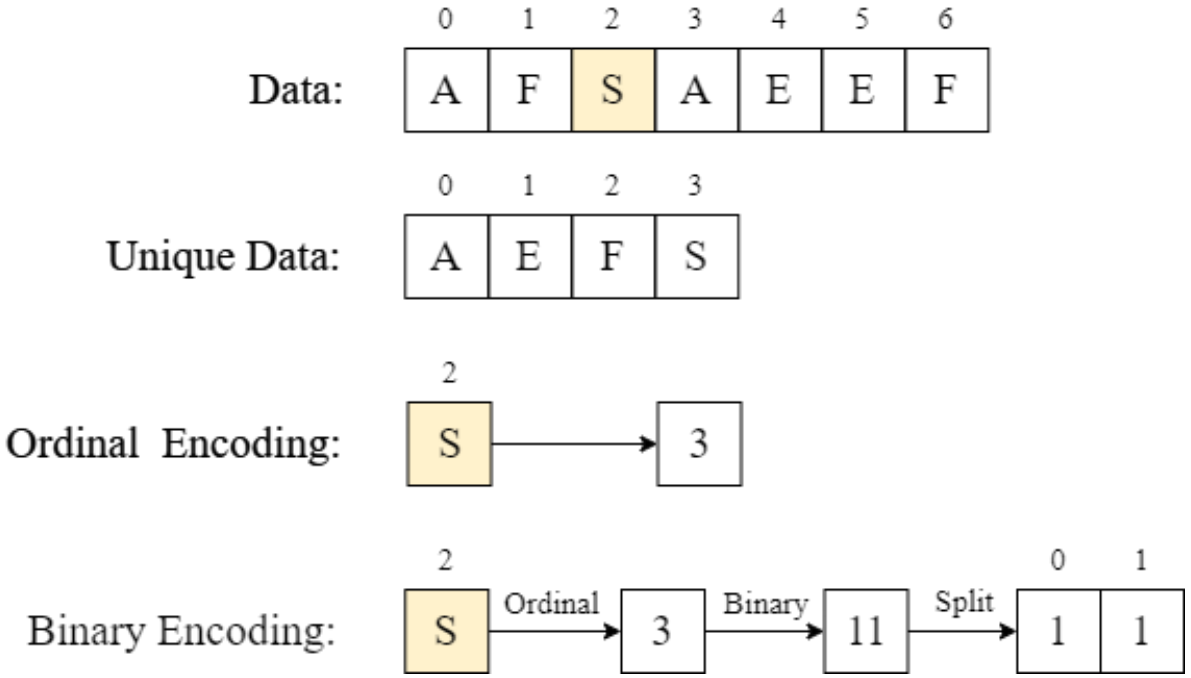



Figure 2: Encoding

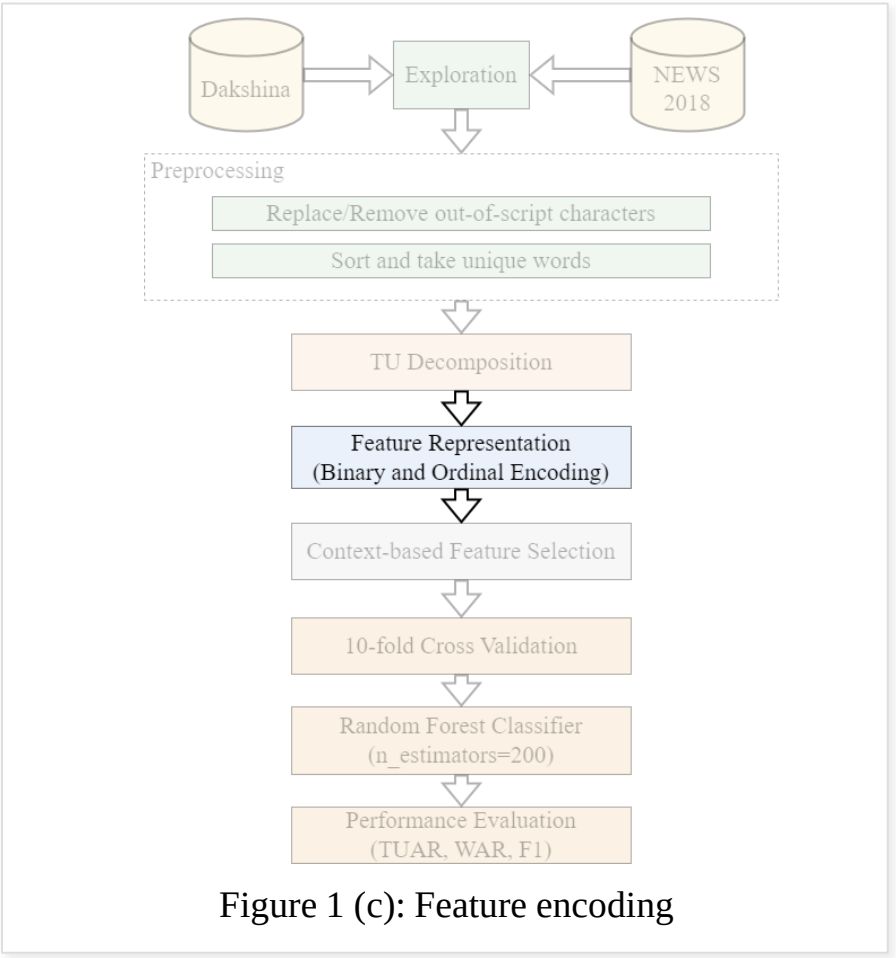



Figure 1 (c): Feature encoding

# Methodology

- Context
  - n-gram based
  - Preceding and/or succeeding STU
  - Preceding target TU (TTU)
- Feature selection
  - Model A (monogram, context-free)
  - Model B (Preceding STU)
  - Model C (succeeding STU)
  - Model D (preceding STU, TTU)
  - Model E (neighboring STU))
  - Model F (neighboring STU, TTU)

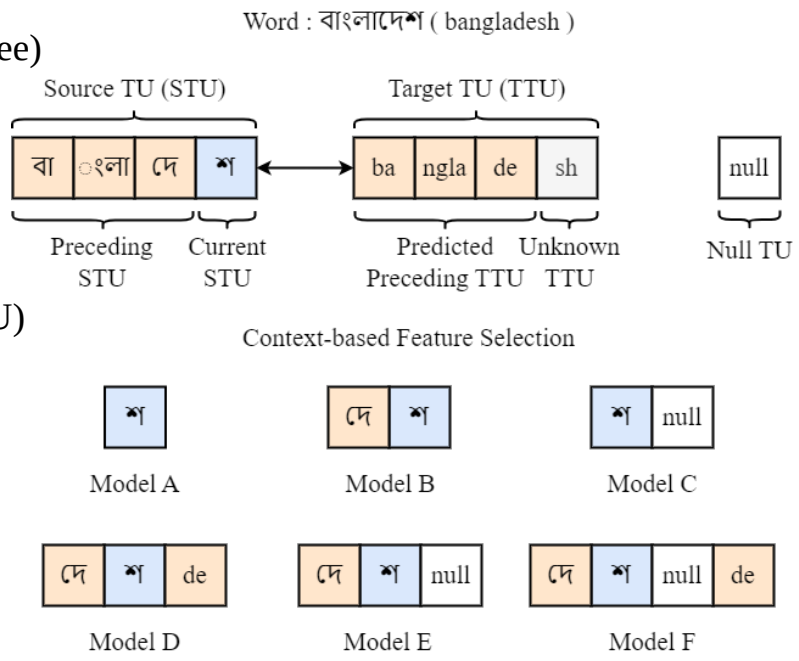



Figure 3: Context-based Feature Selection

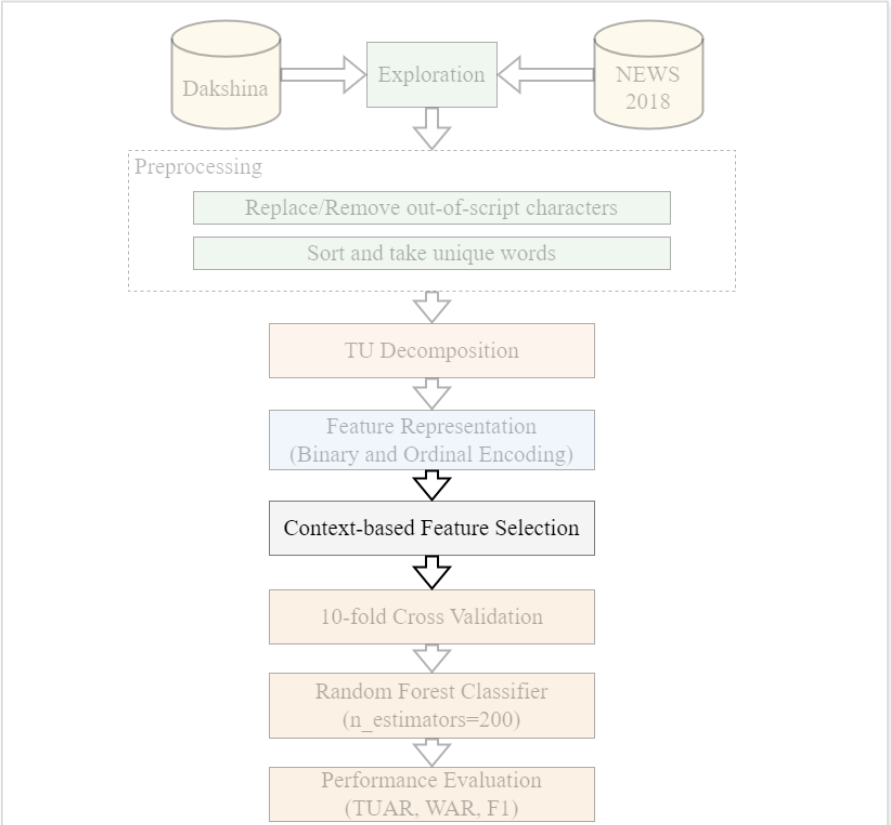



Figure 1 (d): Feature selection

# Methodology

- Machine learning model

**Random Forest Classifier**

- n_estimators = 200

- Performance Metrics

**Accuracy**

**F1 score**

TU level accuracy ( TUAR )

Word level accuracy ( WAR / Top-1 )

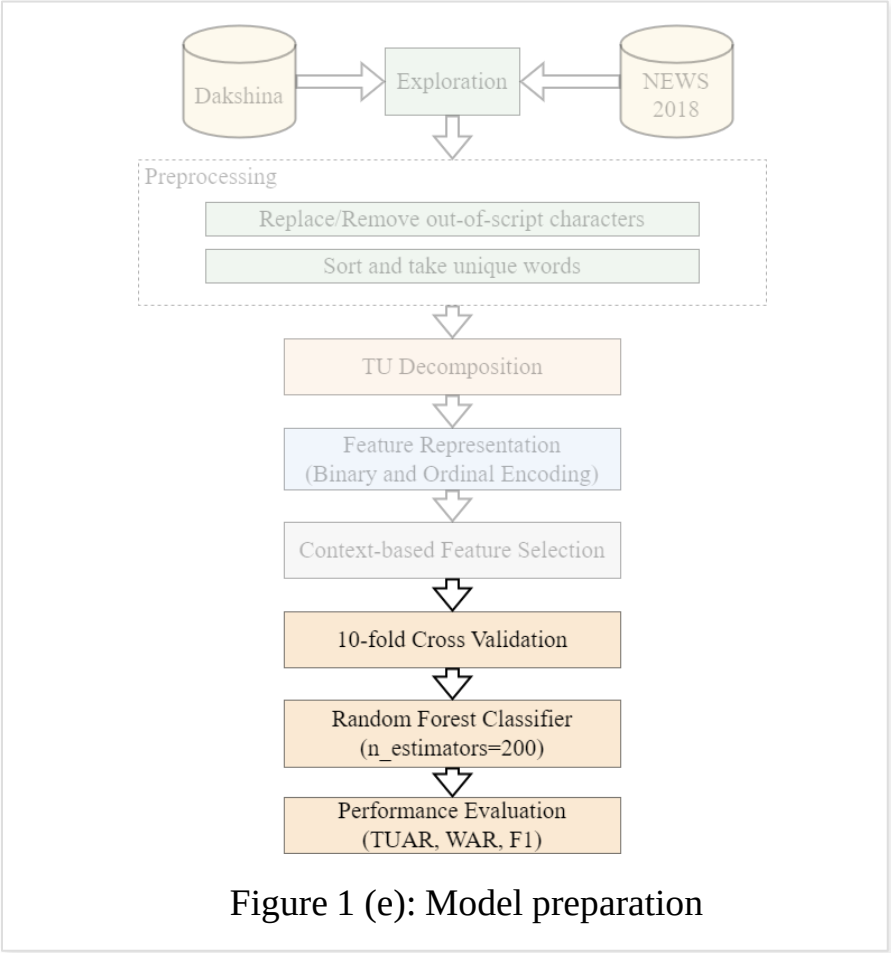



Figure 1 (e): Model preparation

# Results

- Baseline: Model A or context-free model
- For Dakshina and NEWS 2018 Datasets

Table 3: Performance on Forward Transliteration

| Model | Dakshina | | | NEWS 2018 | | |
|---|---|---|---|---|---|---|
| | TUAR | WAR | F1 Score | TUAR | WAR | F1 Score |
| Model A | 77.63 | 45.73 | 74.57 | 76.76 | 44.96 | 73.04 |
| Model B | 78.53 | 47.03 | 77.05 | 78.32 | 48.23 | 75.97 |
| Model C | **80.76** | **50.70** | 79.20 | **81.04** | **53.62** | **78.56** |
| Model D | 77.80 | 46.19 | 75.80 | 78.07 | 47.70 | 74.91 |
| Model E | 80.68 | 50.44 | **79.22** | 80.78 | 53.09 | 78.27 |
| Model F | 80.64 | **50.70** | 78.93 | 80.48 | 52.14 | 77.42 |

# Results

- Baseline: Model A or context-free model
- For Dakshina and NEWS 2018 Datasets

Table 4: Performance on Backward Transliteration

| Model | Dakshina | | | NEWS 2018 | | |
|---|---|---|---|---|---|---|
| | TUAR | WAR | F1 Score | TUAR | WAR | F1 Score |
| Model A | 73.57 | 37.25 | 68.37 | 68.90 | 33.56 | 63.87 |
| Model B | 75.71 | 40.52 | 73.68 | 70.03 | 32.78 | 67.44 |
| Model C | 76.46 | 40.17 | 74.47 | 72.44 | 36.10 | 70.59 |
| Model D | 73.91 | 37.71 | 71.04 | 68.09 | 30.03 | 65.23 |
| Model E | **78.41** | **44.28** | **76.91** | **73.12** | **37.84** | **71.09** |
| Model F | 77.73 | 43.17 | 75.79 | 72.55 | 36.88 | 70.36 |

# Contributions

- Contextual feature selection with binary coding for the Bengali language
- Provides algorithm for context-based feature selection
- Explores performance on forward and backward transliteration

# Conclusion

- Grapheme-based approach
- Binary encoding technique
- Influence of contextual information within words

# Future Research Direction

Investigate transliteration variations

Apply Deep Learning techniques

# Acknowledgement


- This research was funded by Japan Society of Promotion of Science (JSPS) KAKENHI Grant Number JP 20K11939.

# References


1. Sarkar, K., Chatterjee, S.: Bengali-to-english forward and backward machine transliteration using support vector machines. In: J.K. Mandal, P. Dutta, S. Mukhopadhyay (eds.) Computational Intelligence, Communications, and Business Analytics, pp. 552–566. Springer Singapore, Singapore (2017).
2. Ekbal, A., Naskar, S.K., Bandyopadhyay, S.: A modified joint source-channel model for transliteration. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 191–198. Association for Computational Linguistics, Sydney, Australia (2006).
3. Dasgupta, T., Sinha, M., Anupam, B.: Resource creation and development of an english-bangla back transliteration system. International Journal of Knowledge-based and Intelligent Engineering Systems 19, 35–46 (2015).
4. A. D. Tipu, M. Fahad, and A. K. Mandal, “A romanisation method for the bengali language with effcient encoding scheme,” in Lecture Notes in Networks and Systems, Springer Nature. in press.
5. Roark, B., Wolf-Sonkin, L., Kirov, C., Mielke, S.J., Johny, C., Demirsahin, I., Hall, K.: Processing South Asian languages written in the Latin script: the Dakshina dataset. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 2413–2423. European Language Resources Association, Marseille, France (2020).
6. Chen, N., Banchs, R.E., Zhang, M., Duan, X., Li, H.: Report of NEWS 2018 named entity transliteration shared task. In: Proceedings of the Seventh Named Entities Workshop, pp. 55–73. Association for Computational Linguistics, Melbourne, Australia (2018).

# Thank You

# Variations in words

| Original | Transliterated |
| --- | --- |
| ফুল | Phool, phul, fool, ful |