|  |
| --- |
| **Table of Contents** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chapter 1: Introduction** | | | | **1** |
| 1.1 | Objective | | | 1 |
| 1.2 | Background of Study | | | 1 |
| 1.3 | Scope | | | 1 |
| 1.4 | Brief summary of Parts-of-Speech tagging | | | 2 |
|  |  | | |  |
| **Chapter 2: Background Research Review** | | | | **4** |
| 2.1 | Previous Research | | | 4 |
| 2.2 | Related Work | | | 5 |
|  |  |  | |  |
| **Chapter 3: Methodology** | | | | **6** |
| 3.1 | Data Source Selection | | | 6 |
| 3.2 | Domain Selection | | | 6 |
| 3.3 | Storing the Data into Database | | | 7 |
|  | 3.3.1 | Database Creation | | 7 |
|  | 3.3.2 | Web Crawling | | 9 |
| 3.4 | Data Annotation | | | 10 |
|  | 3.4.1 | Crowdsourcing | | 10 |
|  | 3.4.2 | Manual Annotation | | 11 |
| 3.5 | Data Modelling | | | 12 |
|  | 3.5.1 | Relational Database | | 12 |
| 3.5 | Test Set and Training Set Selection | | | 14 |
| 3.6 | Generating Rules for Identifying POS Tags | | | 14 |
|  |  |  |  |  |
| **Chapter 4: Findings** | | | | **16** |
| 4.1 | Sentence Identification | | | 16 |
| 4.2 | Position of Verb in Bangla Sentence | | | 17 |
| 4.3 | Data Analysis | | | 18 |
|  |  |  |  |  |
| **Chapter 5: Limitations** | | | | **19** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| **Chapter 6: Future Work & Conclusion** | | | | **21** |
| 6.1 | Future Work | | | 21 |
| 6.2 | Conclusion | | | 22 |
|  |  |  |  |  |
| **References** | | | | **23** |
| **Appendix - Acronyms and Abbreviations** | | | | **26** |

|  |
| --- |
| **List of Tables** |

|  |  |  |
| --- | --- | --- |
| **Table 4-A** | Overall Statistics | 18 |
| **Table 4-B** | Most frequent POS tags in corpus | 18 |

|  |
| --- |
| **List of Figures** |

|  |  |  |
| --- | --- | --- |
| **Fig. 1-A** | Pos Tagging | 2 |
| **Fig. 3-A** | Difference of word identity in sports domain | 6 |
| **Fig. 3-B** | ER Diagram of Database | 8 |
| **Fig. 3-C** | Work-Flow of web crawling | 9 |
| **Fig. 3-D** | Excluding expression for generating tokens | 10 |
| **Fig. 3-E** | Online crowdsourcing GUI | 11 |
| **Fig. 3-F** | article table | 12 |
| **Fig. 3-G** | article\_line table | 12 |
| **Fig. 3-H** | domain table | 12 |
| **Fig. 3-I** | pos table | 13 |
| **Fig. 3-J** | source\_name table | 13 |
| **Fig. 3-K** | user\_information table | 13 |
| **Fig. 3-L** | word table | 13 |
| **Fig. 3-M** | Table for sentence identification | 14 |
| **Fig. 4-A** | Sentence identification parse tree | 16 |
| **Fig. 4-B** | Sentence identification | 17 |
| **Fig. 4-C** | Position of verb in Bengali sentence | 17 |
| **Fig. 4-D** | Verb position in Bangla & English | 17 |
| **Fig. 5-A** | Variation of word expression in Bangla sentence | 19 |
| **Fig. 5-B** | Foreign words in Bengali form | 20 |
| **Fig. 5-C** | Noun case sentences in sports domain | 20 |
| **Fig. 6-A** | Word possibilities detection | 21 |