LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  |  |  |
| 2 | REVIEW OF LITERATURE   1. Text Summarization Approaches 2. Types of Extractive Summarization 3. Recurrent Neural Network Architecture 4. Types of Evaluations of Summaries 5. Algorithms for Intrinsic Evaluation 6. Types of ROUGE Measures | 5  6  9  10  12  13 |
|  |  |  |
| 3 | PROPOSED SOLUTION   1. Component Diagram 2. Swimlane Diagram | 18  22 |
|  |  |  |
| 4 | IMPLEMENTATION   1. Extractive Text Summarization | 23 |
|  |  |  |
| 5 | RESULTS AND DISCUSSION   1. Average ROUGE-2 Recall Scores 2. Average ROUGE-2 Precision Scores 3. F-Measure Comparison 4. Average F-Scores | 30  31  32  32 |

LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  |  |  |
| 5 | RESULTS AND DISCUSSION   1. Word Frequency Evaluation 2. TF-IDF Evaluation 3. Gensim’s TextRank Evaluation | 28  29  30 |

LIST OF EQUATIONS

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  |  |  |
| 2 | REVIEW OF LITERATURE   1. Word Frequency 2. Term Frequency 3. Inverse Document Frequency 4. TF-IDF 5. Recall 6. Precision 7. Brevity Penalty 8. BLEU Score | 7  7  8  8  11  11  11  12 |
|  |  |  |
| 3 | PROPOSED SOLUTION   1. F-Measure | 17 |

ABBREVIATIONS

|  |  |
| --- | --- |
| **ABBREVIATION** | **DESCRIPTION** |
|  |  |
| TF-IDF | Term Frequency – Inverse Document Frequency |
| BLEU | Bilingual Evaluation Understudy |
| BP | Brevity Penalty |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| LCS | Longest Common Subsequence |
| WOEID | Where On Earth IDentifier |
| NLP | Natural Language Processing |
| NLTK | Natural Language Toolkit (Python) |
| API | Application Program Interface |
| HTML | Hypertext Markup Language |
| XML | Extensible Markup Language |
|  |  |

ABSTRACT

With the boom of social media in the 21st century, it has grown to impact nearly every aspect of our life. This widespread nature of social media makes it a critical platform for the spread of information. This information more often than not is filled with flaws such as incomplete information, uncredited sources and may even go as far as being downright false. Platforms such as Facebook, twitter, Instagram and many more receive more than a billion active users every day with double of that active every month making them excellent platforms for the spread of information. There have always been suspicions and even some verified examples of these platforms being used to influence the lives of people by feeding them news from biased sources in order to influence the decisions they make in real life.[14]

Twitter being a key player in the propagation of information with 500 million tweets containing snippets of information being sent every day, proved to be the ideal platform on which we could implement a system to verify and then spread information. The decision was made to consider the trending topics on twitter to be the keywords on which more information must be found and propagated.

Bots proved to be the ideal form with which could implement said system as it reduces the probability of human error, removes work redundancy and is inherently unbiased. The bot would scrape data off articles from reliable sources, summarize them and then tweet them along with a link to the actual source.

Summarization of the selected articles however proved to be the greatest challenge. We decided to focus on extractive methods of summarization where sentences are lifted from the source verbatim instead of abstractive methods which generated sentences due to its high complexity and resource requirements. We compared a variety of methods such as word frequency, term frequency - inverse document frequency and a few others and have found that ideal extractive summarization is accomplished using a recurrent neural network.