

LIST OF FIGURES

CHAPTER NO.	TITLE	PAGE NO.
2	REVIEW OF LITERATURE	
	Numerical Techniques to Represent Text Data	
Figure 2.1	1-Hot Encoding	7
Figure 2.2	Bigram Model	7
Figure 2.3	Bag of Words	8
Figure 2.4	Vector Semantics	9
Figure 2.5	Word2Vec Sample Sentence	10
Figure 2.6	Updation Criteria of Word2Vec Matrix	10
	Summarization	12
Figure 2.7	Text Summarization Approaches	13
Figure 2.8	Types of Extractive Summarization	16
Figure 2.9	PageRank Nodes and Edges	17
Figure 2.10	Singular Value Decomposition	
	Evaluation of text summarization	19
Figure 2.11	Types of Evaluations of Summaries	20
Figure 2.12	Algorithms for Intrinsic Evaluation	22
Figure 2.13	Types of ROUGE Measures	

	Evaluation of text summarization	
	Figure 2.14 Techniques for Web Scraping	24
	Twitter	
	Figure 2.15 Trends as shown on Twitter on 27-03-2020	28
3	ANALYSIS AND DESIGN	
	Figure 3.1 Component Diagram	32
	Figure 3.2 Swimlane Diagram	36
4	IMPLEMENTATION DETAILS	
	Figure 4.1 Extractive Text Summarization	39
	Figure 4.2 TextRank Algorithm	41
	Figure 4.3 Latent Semantic Analysis Algorithm	43
5	RESULTS AND DISCUSSION	
	Figure 5.1 Average F-Measure Scores	49
	Figure 5.2 F-Measure Scores Comparison	50
	Figure 5.3 Average Computation Time	51
	Figure 5.4 Bot Result on System Console on 13-03-2020	52
	Figure 5.5 Bot Result on Twitter on 13-03-2020	53
	Figure 5.6 Bot Result on Twitter on 27-03-2020	54

LIST OF TABLES

CHAPTER NO.	TITLE	PAGE NO.
5	RESULTS AND DISCUSSION	
	Table 5.1 Zero Value Count in each Method	49
	Table 5.2 High F-Measure Probability per Method	51

LIST OF EQUATIONS

CHAPTER NO.	TITLE	PAGE NO.
2	REVIEW OF LITERATURE	
	Equation 2.1 Word Frequency	14
	Equation 2.2 Term Frequency	14
	Equation 2.3 Inverse Document Frequency	15
	Equation 2.4 TF-IDF	15
	Equation 2.5 Recall	19
	Equation 2.6 Precision	20
	Equation 2.7 Brevity Penalty	21
	Equation 2.8 BLEU Score	22
3	ANALYSIS AND DESIGN	
	Equation 3.1 F-Measure	31
4	IMPLEMENTATION	
	Equation 5.1 Improved Word Frequency	42

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 Programming Interface
DOM	Document Object Model
GPU	Graphical Processing Unit
CPU	Central Processing Unit
RAM	Random Access Memory
XML	Extensible Mark-up Language
UML	Unified Modelling Language
JSON	JavaScript Object Notation
BBC	British Broadcasting Channel
URL	Uniform Resource Locator
TOI	Times of India

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 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 is a critical 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.

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, TextRank, latent semantic analysis based on the intrinsic evaluation. We also modified word frequency to obtain better results.

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 using the best method from those mentioned above, and then tweet them along with a link to the actual source.