TAMIL NEWS CLUSTERING

M.S. Faathima Fayaza

179318T

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2020

TAMIL NEWS CLUSTERING

M.S.Faathima Fayaza

179318T

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science specializing in Data Science Engineering and Analytics

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or diploma in any other university or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part, in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: ……………… Date: …………….

Name: M.S. Faathima Fayaza

The above candidate has carried out research for the Master of Science thesis under my supervision.

Signature of the supervisor: …………………… Date: ……………..

Name: Dr. Surangika Ranathunga

ABSTRACT

The web has an abundance of online news articles that are updated frequently. Readers face difficulty in discovering content of interest from the overwhelming news sources and feel tired browsing various websites. This situation is valid in the case of Tamil online news as well, and the number of online news articles published in the Tamil language is on the rise. To address this issue, news aggregators and clustering techniques come into play. Even though there are many news aggregators available for languages like English, the only news aggregator that supports Tamil is Google news, which is a noticeable shortage. Google news mainly covers the Indian news and gives high weightage to the words that appear on the headline rather than those appearing in the body of the news when searching for the news [1].

This research focuses on clustering Tamil online news articles into related topics. There are several clustering techniques and similarity measures used to cluster the documents in the literature for other languages. Tamil is an agglutinative language, meaning that the techniques used for English documents might not readily work for Tamil. The purpose of this research is to study the techniques available for other languages and develop a mechanism to cluster the Tamil online news articles according to their content similarity.

As the first step of this study, ten different datasets were created by collecting news from nine different news providers. Data was collected on nonadjacent days to get diversified data. TF-IDF and word embedding techniques were used to create vector representations of data. One pass algorithm and affinity propagation algorithm were used to cluster the news articles, since the number of clusters cannot be predefined and there is a high number of single news clusters. We achieved the best solution when applying word embedding with one pass algorithm. As another contribution of this research, we were able to create a Tamil word embedding model with 21,077,843 words.

Keywords: Clustering, TF-IDF, Word embedding, One pass algorithm, Affinity propagation, Cosine similarity, Crawler

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Dr. Surangika Ranathunga, my research supervisor, for her support and guidance in selecting and conducting this research. I wish to thank her for the patient guidance, enthusiastic encouragement and useful critiques of this research work. Her continuous supervision greatly helped me in keeping the correct phase in research work. I especially appreciate the frequent feedback, which helped me to correct and fine-tune it to this level.

I would like to thank all staff from the Department of Computer Science and Engineering for various support rendered by them throughout this effort.

I am in much debt to my friends who lent a helping hand through various means. Also, I would like to extend thanks to everyone who helped me behind the scene to pursue this degree.

Last but not least, my heartfelt gratitude goes to my parents Mr & Mrs Meeraa Shahibo, my husband Mr Sifan and siblings for their great support and encouragement to make my dream a reality.

Finally, special thanks to my daddy and my baby for being my mentor and the strong pillars to me throughout this effort.

TABLE OF CONTENTS

DECLARATION i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF FIGURES viii

LIST OF TABLES ix

LIST OF ABBREVIATIONS x

CHAPTER 1 – INTRODUCTION 1

1.1 Overview 1

1.2 Problem and Motivation 2

1.3 Aim and Objectives 4

1.4 Methodology 4

1.5 Organization of the Thesis 4

CHAPTER 2 – LITERATURE REVIEW 6

2.1 Overview 6

2.2 Clustering 6

2.3 Document clustering 6

2.4 Challenges in document clustering 7

2.5 Representation of the Textual Documents 7

2.5.1 Bag of words 8

2.5.2 n-grams 8

2.5.3 Bag of phrases 8

2.5.4 Vector Space Model 9

2.5.5 Ontology-based representation 9

2.5.6 Continuous Bag-of-Words Model (CBOW) 10

2.5.7 Skip-gram Model 10

2.6 Preprocessing 11

2.6.1 Stemming and Lemmatization 11

2.6.2 Stemmers for Tamil language 11

2.6.3 Stop word removal 12

2.7 Document Similarity Measure 12

2.7.1 Euclidean Distance 13

2.7.2 Cosine Measure 14

2.7.3 Jaccard Coefficient 15

2.7.4 Inner Product Measure 15

2.7.5 Pearson Correlation Coefficient 15

2.7.6 Averaged Kullback-Leibler Divergence 16

2.7.7 Latent Semantic Analysis (LSA) 16

2.8 Clustering of Textual Documents 17

2.8.1 Hierarchical methods 17

2.8.2 Partitioning methods 19

2.8.3 Model based methods 21

2.8.4 Density based methods 21

2.8.5 Grid based methods 22

2.8.6 Affinity propagation 22

2.9 Document/News Clustering Discussion 22

2.10 News clustering systems 23

2.11 Clustering Results Evaluation 24

2.12 Web Crawling 26

CHAPTER 3 – METHODOLOGY 28

3.1 Overview 28

3.2 Tamil News Article Data Collection 28

3.2.1 Tamil news sources 29

3.2.2 Web crawling 30

3.2.3 News article content collection 32

3.2.4 News data storing 33

3.3 Data representation 34

3.3.1 Pre-processing 34

3.3.2 Term Frequency-Inverse Document Frequency (TF-IDF) vector 35
3.3.3 Word Embedding model 35
3.4 Tamil News Article Clustering 35
3.4.1 Similarity calculation 36
3.4.2 Tamil news Clustering 36
3.4.3 Baseline system setup 37
3.5 Alternative approach 37
3.6 Clustering based on title 38
3.7 Summary 38
CHAPTER 4 – EVALUATION AND ANALYSIS 39
4.1 Overview 39
4.2 Data Collection 39
4.3 Evaluation Method 41
4.4 Baseline (TF-IDF with One pass algorithm) 41
4.5 Effectiveness of clustering based on TF-IDF with Affinity propagation Algorithm 42
4.6 Effectiveness of clustering based on Word Embeddings with One pass Algorithm 44
4.7 Effectiveness of clustering based on Word Embeddings with Affinity propagation Algorithm 45
4.8 Effectiveness of clustering based on article title 46
4.8.1 Effectiveness of clustering based on article title using TF-IDF with One pass algorithm 49
4.8.2 Effectiveness of clustering based on article title using the TF-IDF with Affinity Propagation Algorithm 50
4.8.3 Effectiveness of clustering based on article title using Word Embedding with One pass algorithm 51
4.8.4 Effectiveness of clustering based on article title using Word Embedding with Affinity Propagation Algorithm 52
4.9 Effectiveness of document representing techniques and clustering techniques.....53
4.9.1 TF-IDF Vs Word embedding 53
4.9.2 One pass clustering algorithm Vs Affinity propagation 53

4.10 Overall Performance 54
4.11 Discussion 56
CHAPTER 5 – CONCLUSION AND FUTURE WORK 63
5.1 Conclusion 63
5.2 Future Work 63
REFERENCES 65
APPENDIX A: TAMIL SENTENCES WITH TRANSLITERATION AND TRANSLATION 69

LIST OF FIGURES

Figure 1.1: Google News Tamil search ………………………………………………...2

Figure 1.2: Google News Tamil content with unrelated image ………………………...3

Figure 3.1: Outline of the Tamil News Clustering system …………………..................28

Figure 3.2: Outline of the Tamil News Article Data Collection module ……………....29

Figure 3.3: News website ……………………………………………………………32

Figure 3.4: Sample extracted news content ……………………………………………33

Figure 3.5: News data file structure …………………………………………….……...33

Figure 3.6: Sample news data file ……………………………………………………...34

Figure 4.1: Effectiveness of Baseline …………………………………………………..42

Figure 4.2: Effectiveness of TF-IDF with Affinity propagation Algorithm.…………...43

Figure 4.3: Effectiveness of Word Embeddings with One pass Algorithm ……...........44

Figure 4.4: Effectiveness of Word Embeddings with Affinity propagation Algorithm…………………………………………………………………………….....45

Figure 4.5: Tamil news clustering based on the news title ……………………………47

Figure 4.6: Tamil news clustering based on TF-IDF with One pass algorithm………...49

Figure 4.7: Tamil news clustering based on TF-IDF with Affinity propagation algorithm…………………………………………………………………………….50

Figure 4.8: Tamil news clustering based on Word Embedding with One pass algorithm…………………………………………………………………………….51

Figure 4.9: Tamil news clustering based on Word Embedding with Affinity Propagation Algorithm……………………………………………………………………………52

Figure 4.10: Pairwise F-score values obtained for each data set………………………..55

LIST OF TABLES

Table 3.1: Tamil news sources……………………………………………………….30

Table 3.2: Seed URL list……………………………………………………………..31

Table 4.1: Details of the Datasets…………………………………………………….39

Table 4.2: Kappa statistic values of agreement between manual groupings……………40

Table 4.3: Statistical analyze of effectiveness of Baseline system (F-score values)……42

Table 4.4: Statistical analyze of TF-IDF with Affinity propagation Algorithm……..…43

Table 4.5: Statistical analyze of Word Embeddings with One pass Algorithm……......44

Table 4.6: Statistical analyze of Word Embeddings with Affinity propagation Algorithm…………………………………………………………………………46

Table 4.7: Statistical analyze of Tamil news clustering based on the news title……….47

Table 4.8: Statistical analyze of Tamil news clustering based on TF-IDF with One pass algorithm…………………………………………………………………………..49

Table 4.9: Statistical analyze of Tamil news clustering based on TF-IDF with Affinity propagation algorithm...………………………………………………...................50

Table 4.10: Statistical analyze of Tamil news clustering based on Word Embedding with One pass algorithm…………………………………………………………...51

Table 4.11: Statistical analyze of Tamil news clustering based on Word Embedding with Affinity Propagation Algorithm……...…………………………………………52

Table 4.12: Statistical analyze of document representing approaches…….……………53

Table 4.14: Statistical analyze of clustering approaches for Tamil news…...………….54

Table 4.15 pairwise F-score values obtained for each data set……………..………..54

Table 4.16 summary of pairwise F-score values obtained for each data set……………55

LIST OF ABBREVIATIONS

Abbreviation	**Description**
URL	Universal Resource Locator
WWW	World Wide Web
HTML	Hyper Text Markup Language
ANN	Artificial Neural Networks
VSM	Vector Space Model