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

LIST OF FIGURES

LIST OF TABLES

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

CHAPTER 1 – INTRODUCTION

1.1 Overview

The evolution of the Internet has led to an exponential amount of information being available in electronic format. News articles are a source of information that can be found in different forms such as online newspapers, blogs or other types of news websites. These news articles are not only in English but in other regional languages as well.

These days, most of the younger generation is addicted to the Internet and follow online news updates. However, due to the availability of several news sources, users spend a certain amount of time and effort to browse through the Internet to gather complete news. It can be a tiresome task. This situation is applicable not only for English but also for other regional languages like Tamil, since Tamil news sources in electronic format are on the rise.

In recent times, the popularity of Tamil websites has shown a demand for Tamil news to be available online. So several Tamil news websites were introduced, which made a huge impact. Almost all the Tamil traditional news media launched online versions of their news updates. Further, some content-specific special websites such as business, technology, politics, sport, gossip, and cinema were introduced. So these days, the Internet has an overwhelming number of diverse free form Tamil news. Therefore, users need to access different websites to get updates in different areas.

Reading news online is an increasing trend because users can easily access most news sources free of charge even from geographically remote sites. Most of the time news sources cover the same news and publish similar articles. At the same time, most news providers update their sites instantly when an event occurs. Therefore, users frequently need to visit multiple sources to get a clear picture of the updated news. So if similar articles were accessed from a single site it would be more user-friendly. It can help news reviewers to find more information from a single spot and criticize the sources.

News aggregation techniques address this by gathering multiple news sources into a single point. News aggregators frequently collect news from different feeds and display aggregated content in a single interface. Google News[1], Yahoo News[2], Event Registry[3], Huffington Post[4] and Drudge Report[5] are some popular news aggregators. Google News covers more than 25,000 news sites and presents news from the past 30 days from different websites. It also supports thirty-five languages, including Tamil.

1.2 Problem and Motivation

As far as this author knows, Google News is the only aggregator that supports Tamil. Google News is not a specially designed news aggregator for Tamil. As mentioned earlier, it supports thirty-five languages, including Tamil, but it also has some ridiculous flaws. Even though Google News supports Tamil, searching in Tamil on Google News is challenging. Figure 1.1 shows an example where Google News attached an unrelated photo with a news article [Figure 1.1] giving less significant news more prominence on the page.

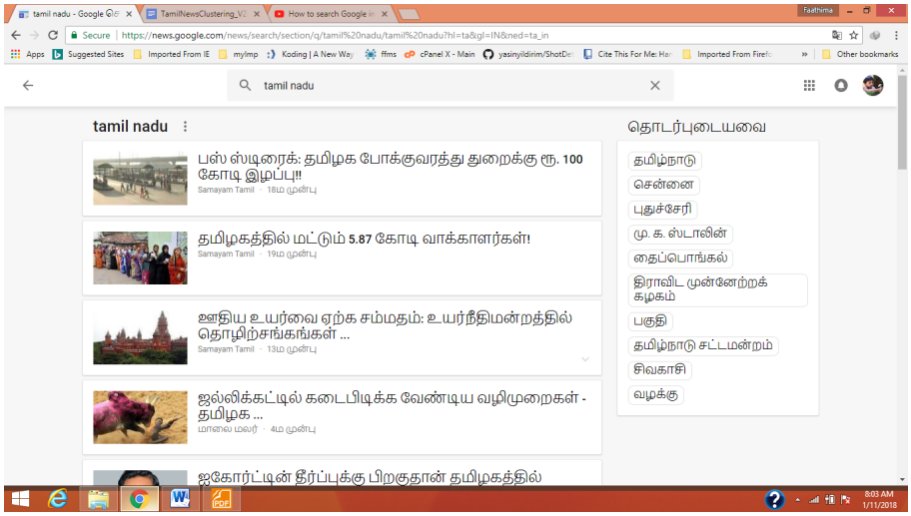

Figure 1.1: Google News Tamil search

[1] https://news.google.com/
[2] https://www.yahoo.com/news/
[3] http://eventregistry.org/
[4] http://www.huffingtonpost.com/
[5] http://www.drudgereport.com/

Figure 1.2 shows Google search results with content that have unrelated images. In Figure 1.2 the marked area is "Pakistan Threatens India with Nuclear Weapons". but an unrelated image is attached. Google News Tamil version contents are imbalanced, and they mostly cover Indian news, perhaps because it is mostly crawling Indian websites. However, native Tamil speakers live around the world in countries such as Singapore, Malaysia, and Sri Lanka. Further, Google News gives high weightage to the words that appear on the headline rather than those appearing in the body of the news when searching for news [1]. So, the terms chosen for the news headline play an importance role in the ranking. According to the author's knowledge, there are no news aggregators specifically designed for Tamil news.

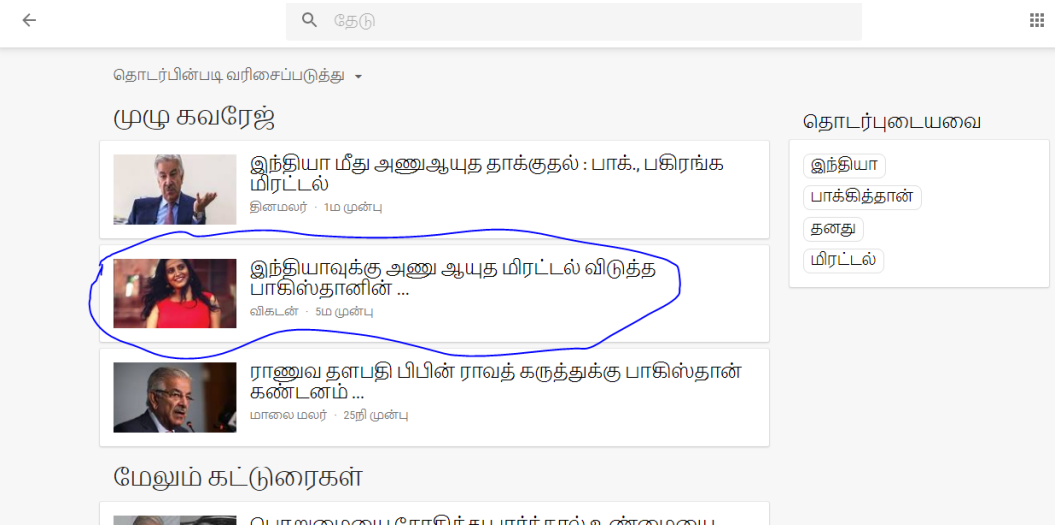

Figure 1.2: Google News Tamil content with unrelated image.

In literature, automatic Tamil document grouping has been done using techniques such as the Vector Space Model (VSM) and Artificial Neural Networks (ANN) [2]. In this approach ANN (93.33%) performs better than VSM (90.33%). The main focus of the classification of Tamil documents and the performance of the system depends on the corpora used to train and the number of corpora used to train.

1.3 Aim and Objectives

The aim of this project is to automatically extract online Tamil news articles and cluster them according to content similarity.

1.4 Methodology

Following are the sub-objectives:

- Build or modify an existing crawler to automatically collect Tamil news from online sources.
- Evaluate the available document/news clustering techniques for other languages and implement them for Tamil news clustering.
- Evaluate the Tamil language processing techniques to preprocess the data to improve the performance of clustering.
- Evaluate the available document representing techniques to find the best approach and apply it to Tamil
- Evaluate the available similarity measurements and find the best approach that works for Tamil

1.5 Organization of the Thesis

The rest of the thesis is organized as follows.

Chapter 2 present the literature work related to this research, specially, web crawling, document preprocessing, document representations, document similarity measurement, clustering methods and cluster evaluations.

Chapter 3 details the work done on building the Tamil news clustering system. It consists of three modules namely, namely the Tamil news article data collection module, the data representation module and the Tamil news article clustering module.

Chapter 4 evaluates the work presented in the thesis. It consists of details of data used for evaluation, experiments carried out and the obtained results along with a discussion on the observed results.

Chapter 5 concludes the thesis with future work.

CHAPTER 2 – LITERATURE REVIEW

2.1 Overview

This section primarily focuses on the available work done on document clustering, text preprocessing techniques, similarity measuring, and clustering text documents. This chapter mainly covers the available approaches in news clustering for other languages, like English. It also focuses on the work available on Tamil document categorization techniques and the stemmer available in Tamil. Moreover, it covers the available techniques in web crawling.

2.2 Clustering

Clustering is an unsupervised technique [2] that groups unlabeled data into meaningful clusters based on similarity [3], where objects in a single cluster are homogeneous and similarity between the objects in the same cluster is high, whereas similarity between different clusters is low. Also clustering does not require any training but little prior information about the data [3].

2.3 Document clustering

Document clustering is the process of grouping similar documents into different clusters. It is a subset of clustering and organizes documents into meaningful clusters based on content similarity. Normally, clustering aims to maximize the similarity within the documents in the clusters and minimize the similarity between the documents in the cluster [3].

Document clustering is mainly involved with the following process [4, 5, 6]:

- Document representation
- Document similarity measure
- Clustering or grouping

2.4 Challenges in document clustering

Document clustering is one of the core concepts studied in information retrieval and text mining. Document clustering has unique challenges compared to clustering in other domains. Some of the identified unique challenges are listed below:

- High dimensionality

Most text representation techniques take each unique term in the document as a dimension for clustering. Typically, documents contain hundreds and thousands of unique terms and a document space contains hundreds and thousands of documents. Since the numbers of unique terms are high, it gets high dimensional space and faces challenges when handling it.

- Ambiguity and synonymy

Sometimes, many terms share a common meaning and a single term can have multiple meanings. So if two documents select two unique terms to express the same context in the document, it's difficult to cluster those documents into a single cluster even though it's talking about the same context.

To overcome this problem, some studies have been conducted. To reduce the dimensionality of terms, researchers used the x^2 test for feature selection [6, 7, 8, 9]. To address the ambiguity of synonymy, researchers used concept representation rather than terms [8].

2.5 Representation of the Textual Documents

Document representation is a basic step in document clustering because currently, there is no method that is able to process the unstructured text data [10]. So as a first step, documents are transformed into a well-defined format to make the clustering process easier. Document representation not only plays an important role in clustering but also plays a major role in information retrieval. Following are some document representation techniques in literature.

2.5.1 Bag of words

Bag of words is mostly used in information retrieval and text mining [11]. Simplest text document representation is introduced in the vector space model framework [6, 7, 12]. In the Bag of words approach, terms appear independently and do not have any details about word order or have grammatical relationship between terms [6, 11]. Simply, it does not capture the relationship between words and loses the context [13]. For example, both “John is quicker than Mary” and “Mary is quicker than John” have the same representation even though it is different in context. But practically, two different documents do not have the exact terms in the exact frequency until it copies the same document [13].

2.5.2 n-grams

N-gram [7, 10] is a language independent text document representation technique [14]. This technique is used in spelling-related applications, string searching, prediction, and speech recognition [14].

N-grams are simply n consecutive characters extracted from the document. To extract n-gram from the documents n size character window, move through the document, moving forward by fixed length (generally one) character at a time [14]. For example: the first two 4-grams in the phrase “general knowledge” are “gene” and “ener”

Usually, n-grams generate more index terms than word-based representation for a document [14] because word-based representation techniques allow the stop word removal, stemming and lemmatization. But n-grams provide the garble tolerance to the system [14].

2.5.3 Bag of phrases

Bag of phrases [8, 10, 15] was introduced to address the data sparseness problem in bag of words [12]. In this approach, instead of words, phrases (sequence of words in the

sentences) are used to represent the document. This approach preserves the semantics qualities but statistical qualities are very reduced.

2.5.4 Vector Space Model

The Vector Space model is one of the most commonly used models to represent text [13]. Document text represented by a numeric vector obtained by a relatively important lexical term in the document is known as a document vector [10].

$$d_i = (w_{1i}, w_{2i}, w_{3i}, \dots \dots w_{|T|i})$$

Here, T: set of terms in the document, |T|: size of vocabulary, w_{ki}: weight frequency of term k in document i

Representing the set of documents in the form of a document vector in a common vector space is called a vector space model. It represents documents as an n-dimensional vector, where n is the number of unique terms in the documents [13]. The commonly used term weighting model is TF-IDF (term frequency–inverse document frequency) [10, 11, 16, 17].

$$\mathrm{TF-IDF_{t,d}} = \mathrm{TF_{t,d}} * \mathrm{IDF_t}$$

Here, $\mathrm{IDF_t} = \frac{\log N}{DF_t}$

N: Number of document in the collection, t: term, d: document

2.5.5 Ontology-based representation

The major challenge of the vector space model is high-dimensional space due to negating the subjectivity and explaining the ability [18]. Ontology based representation [7, 8] mainly aims to consider different views of data and uses an ontology based Concept Selection and Aggregation to build alternative data representation for the text document [18]. Ontology based model also uses the vector space model to represent the

document text but instead of using terms in the document, it represents the concept of terms in the document [7, 10]. In [18] core ontology is defined as follows:

(Core Ontology) A core ontology is a sign system O: = (L,F, C* ,H, ROOT), which consists of

- A lexicon: The lexicon, L, contains a set of natural language terms.
- A set of concepts, C*.
- The reference function f with f :2^{l} 2^{c*}. f links sets of terms {Li } ⊂ L to the set of concepts they refer to. In general, one term may refer to several concepts and one concept may be referred to by several terms. The inverse of f is f-1
- A heterarchy, H: Concepts are taxonomically related by the directed acyclic, transitive, reflexive relation H, (H ⊂ C* × C*).
- A top concept ROOT ∈ C*. For all C ∈ C* it holds: H(C, ROOT).

2.5.6 Continuous Bag-of-Words Model (CBOW)

This is a Word2Vec approach introduced by Mikolov et al. [19]. In this architecture, the order does not affect the projection [19, 20]. CBOW uses a continued distributed representation of the context [19] and predicts the current word using the context.

2.5.7 Skip-gram Model

The Skip-gram model [19, 20, 21] is an efficient approach to creating high quality vector representation for a high volume of unstructured data [20]. The Skip-gram model uses the current word to predict the surrounding word. Mikolov et al. [19] found that an interesting property of the Skip-gram model is that simple vector addition produces meaningful results [20].

e.g: vec(“Russia”) + vec(“river”) is close to vec(“Volga River”) [20]

Mikolov et al [19] use the test data set with a set of questions in the new Semantic-Syntactic Word Relationship to compare the Skip-gram model and CBOW where the CBOW model performs better in the semantic part of the test.

2.6 Preprocessing

2.6.1 Stemming and Lemmatization

Stemming [10, 13, 22] is the process of reducing the given text to their roots. Generally, stemming chops off the ends of words without performing complete morphological analysis [23]. Stemming is a language dependent process. Stemmers improve the recall in information retrieval systems.

e.g., automate(s), automatic, automation all reduced to automat.

Lemmatization [10] is the process of reducing the inflectional/variant form of words to the base form with the use of morphological analysis. Lemmatization aims to do proper reduction and returns the dictionary form of words.

e.g., am, are, and is - be

Car, cars, car's, cars' - car

Both of these techniques help to reduce the dimension of document space by bringing the inflectional form of words to the base form.

2.6.2 Stemmers for Tamil language

Tamil is a morphologically rich language and has high inflection in words. In general, due to the agglutinative nature of the Tamil language, words have more than one morphological suffix. The number of suffixes can vary from 3 to 13 [23].

Rule-based suffix stripping stemmer algorithm

The rule-based suffix stripping stemmer algorithm chops off the suffix of variant Tamil words based on the rules. The algorithms are defined as follows [23]

Step1: Eliminate the entire complex plural

Step2: Eliminate the joint word suffixes.

Step 3: According to the identified suffix, the next possible suffix list is generated using rules.

The problem with the rule-based suffix stripping stemmer is that for some Tamil words it gives an infinite verb.

Light Stemmer

This was introduced to address the infinite verb problem in the rule-based suffix stripping stemmer. Light Stemmer truncates the entire possible suffix with the intention of preserve the meaning of the word [23].

2.6.3 Stop word removal

Stop words are most commonly used words in a language and not descriptive about the document. Generally, stop words have high frequency in documents.

E.g: an, and, by, for, etc.. [English Stop words]

Stop word removal gives a faster and better clustering [13, 22]. By comparing documents with the stop word list, stop words can be removed.

2.7 Document Similarity Measure

Measuring the similarity between documents is a base approach used to identify the resemblance or separation between documents. Text similarity measures play a huge role in document clustering, information retrieval, question and answering and etc. applications.

Terms in the documents can be similar lexically or semantically [17]. Lexically similar terms have the same character sequence and semantically similar terms share common context. Similarity measure calculations come under the following categories:

String based methods - lexical similarity, measure the resemblance between two string words. Eg: Euclidean, Cosine and Jaccard. Here string vectors used to measure the similarity using Euclidean and Cosine.

Corpus based methods - semantic similarity measure, calculate the similarity between words using information from the corpora. Eg: Latent Semantic Analysis (LSA)

Knowledge based methods - semantic similarity measure, calculate the similarity between words using semantic networks. Eg: similarity measure using WordNet

In general, similarity between the objects depends on the properties of object and measures used for calculation [16]. Following are some algorithms widely used in the literature to calculate the similarity in the clustering process.

2.7.1 Euclidean Distance

$$\text{Euclidean}\left(d_i, d_j\right) = \sqrt{\sum_{i=1}^{n}(w_{ti} - w_{tj})^2}$$

Where, d_i , d_j : documents represent by their term vectors.

Term set T= $\{t_1 \ldots\ldots t_m\}$ w: term weight

Euclidean Distance measures the ordinary distance between two points. Commonly used techniques in clustering and default measure used with K-means algorithm [16]. Euclidean Distance is a metric because it satisfies all the following properties [16]:

1. The distance between any two points must be nonnegative, that is, $d(x, y) \geq 0$.

2. The distance between two objects must be zero if and only if the two objects are identical, that is, $d(x, y) = 0$ if and only if $x = y$.

3. Distance must be symmetric, that is, distance from x to y is the same as the distance from y to x, ie.d(x, y) = d(y, x).

4. The measure must satisfy the triangle inequality, which is d(x, z) ≤ d(x, y) + d(y, z).

The Huang [16] experiment shows that on average, the Cosine, Jaccard and Pearson measures perform comparably better than Euclidean distance in text document clustering.

2.7.2 Cosine Measure

Cosine measure defines as: $\frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}}$

Where, A_i, B_i : weighted term in the documents.

When documents are represented in the vector space, similarities between documents correspond with correlation between the vectors. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [17, 22, 24].

Cosine Measure is mostly used in a high-dimensional space and is a popular measure used in document clustering [16, 17]. An interesting thing about Cosine Measure is that it is very efficient even in sparse vectors, and only the non-zero dimensions are taken in to consideration [22].

When experimenting with web page documents, [22] it was observed that the cosine gives better performance than Euclidean and Euclidean-Jaccard. Also, Cosine runs much faster than the other [22].

2.7.3 Jaccard Coefficient

Jaccard Coefficient defined as: $\frac{|A \cap B|}{|A \cup B|}$

Jaccard Coefficient [7, 16, 25] is calculated by dividing the number of intersection terms in the document by the number of union terms in the document. Jaccard Similarity measures the range between 0 to 1, where 0 is when two documents are completely different and 1 is where two documents are similar.

Friburger and Maurel [25] use the proper name for similarity measure. They noticed that the proper name outperforms when the text is small and also that the Jaccard Coefficient performs better than TF-IDF.

2.7.4 Inner Product Measure

A basic similarity measure, also known as dot product. Inner product [4, 7, 26] is calculated as the weighted sum of terms in documents. Inner product between two documents is defined as: $\sum_{i=1}^{n} = A_i * B_i$

Where, n: number of terms in the document vector.

The challenge with the inner product measure is the probability of single component influence [26].

2.7.5 Pearson Correlation Coefficient

Pearson correlation is defined as: $\frac{n \sum_{i=1}^{n} W_{t,a} * W_{t,b} - TF_a * TF_b}{\sqrt{[n \sum_{i=1}^{n} W_{t,a}^2 - TF_a^2][n \sum_{i=1}^{n} W_{t,b}^2 - TF_b^2]}}$

Where, $F_a = \sum_{i=1}^{n} W_{t,a}$, $TF_b = \sum_{i=1}^{n} W_{t,b}$

Pearson correlation [16] similarity range from -1 to 1 where 1 is the total positive linear correlation, 0 is no linear correlation, and −1 is the total negative linear correlation.

2.7.6 Averaged Kullback-Leibler Divergence

In this approach, the document is treated as the probability distribution of terms [16]. It's an information theory based clustering. The similarity between two documents is calculated by taking the difference between the corresponding probabilities distributions.

Given two distributions, P and Q, the KL divergence is defined as $D_{K,L}(P||Q) = P\log(\frac{P}{Q})$

For documents, the averaged KL divergence:

$$D_{Avg\ K,L}\,(t_a||t_b) = \sum_{t=1}^{m}(\pi_1 \times D(W_{t,a}||W_t) + \pi_2 \times D(W_{t,b}||W_t))$$

where, $\pi_1 = \frac{W_{t,a}}{W_{t,a}+W_{t,b}}$, $\pi_2 = \frac{W_{t,b}}{W_{t,a}+W_{t,b}}$, $W_t = \pi_1 \times W_{t,a} + \pi_2 \times W_{t,b}$

2.7.7 Latent Semantic Analysis (LSA)

Most popular corpus based techniques. Latent semantic analysis (LSA) [5, 15, 17] does not use any knowledge base; it takes raw text as input and parses it into words defined as unique character strings, and separates them into meaningful sentences or paragraphs.

In LSA, as the first step, text is represented as matrices where each row represents a unique word, and each column represents sentences or other context. After that, LSA applies singular value decomposition (SVD) to the matrix. SVD constructs a low-rank approximation to the original term-document matrix, for a value of k that is much smaller than the original rank of the term-document matrix. Then the new k-dimensional representation is used to compute similarities between vectors.

Discussion

Related literature provides contradicting observations on the performance of these similarity measurement methods.

Nalawade et al. [22] noticed that for web page documents, the cosine measure performs better than Euclidean distance and Euclidean-Jaccard and that cosine runs faster than other approaches.

Huang [16] evaluated Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient and averaged Kullback-Leibler divergence with seven different datasets covering newspaper articles, newsgroup posts, and research papers, where Euclidean distance performs the worst, and the average KL divergence and Pearson coefficient tend to outperform the cosine similarity and the Jaccard coefficient.

2.8 Clustering of Textual Documents

2.8.1 Hierarchical methods

Hierarchical clustering [4,10,12,27,28] builds hierarchy of clusters gradually. Hierarchical clustering takes multiple steps to partition the data. Generally, hierarchical clustering is divided into two categories:

1. Agglomerative method:

"Bottom up" approach, where initially, each object is considered as a single cluster and then similar objects are continuously combined and moved up to create a single cluster. The bottom up approach follows these steps [28]:

1. Initially, each document is considered as a single cluster.

2. From all clusters, two are picked with the smallest distance.

3. The selected cluster is merged into a single cluster and replaced with the new merged cluster.

4. The process is repeated until all the documents are in a single cluster.

2. Divisive method :

"Top down" approach. Initially, all the objects are considered as a single cluster and recursively divided into two clusters until single objects get into a single cluster. The divisive method follows these steps:

1. All the documents are considered as a single cluster

2. Repeated until all clusters are singletons

a) A chosen cluster is split

b) The chosen cluster is replaced with the sub-cluster

The similarity of clusters is calculated based on the linkage criterion [12]. Single-linkage, complete-linkage, average-linkage and centroid-linkage are the most commonly used measures. For inter-cluster distance in single-linkage, the shortest distance from any member of one cluster to any member of the other cluster is taken, whereas in complete-linkage, the maximum distance from any member of one cluster to any member of the other cluster is taken, and the average distance is taken in average-linkage. In centroid-linkage, each cluster is represented by its centroid and the inter-cluster distance is taken as the distance between those centroids.

Main challenges of the hierarchical methods are:

- Scale - time complexity $O(n^2)$
- Errors made in the early stage continue because most hierarchical algorithms do not revisit once constructed

The pros of hierarchical methods are:

- Flexible in the level of granularity
- Can handle any attribute type

2.8.2 Partitioning methods

Partitioning algorithms [3, 4,10,12,28] do a single-level division and results in a set of clusters. Here, a single object belongs to a particular cluster. Clusters are represented by a centroid, which is a summary representative of all the data points in the cluster. Centroid may be a real data point in the cluster or may not be a member of the data set. A document is put into a particular cluster if the distance between the centroid of the cluster and the document vector is the smallest when compared to other cluster centers.

K-means Clustering Algorithm

The most simple and popular partition algorithm [4, 27]. K-means algorithms work well with numerical attributes but not with categorical attributes [28]. The K-means algorithm represents a cluster by its centroid, which is the mean of the data points. So it can be affected only by a single outlier [28]. K-means clustering follows the steps shown below:

- The algorithm selects K elements in the dataset as representative of clusters (centroid).
- The rest of the objects in the dataset are assigned to the closest centroid according to similarity.
- The cluster centers are recomputed until cluster centers are not changed

The performance of the algorithm depends on initial centroid selection and number of clusters. The time complexity of the algorithm is linear to the data points.

Variations of K-means Algorithm

1. Bisecting K-means:

Bisecting K-means [4, 27, 28] splits all the data points into two clusters, then from these two clusters, one is selected and split. The approach continues until K clusters are produced. [20] shows that bisecting K-means performs better than standard K-means.

2. K-medoids:

In K-medoids [28] the algorithm selects a centroid from the dataset. The algorithm selects a centroid as medoids so dissimilarity to other data points is minimal. Also, mediod are influenced less by the outliers. Moreover, it presents no limitations on attribute types. K-medoids has two types:

PAM (Partitioning Around Medoids):

CLARA (Clustering LARge Applications)

PAM is the most commonly used K-medoids algorithm.

3. K-Means++:

K-Means++ augments the traditional K-means algorithm with a simple randomized seeding technique showing high accuracy and speed than K-means [4].

4. Adaptive K-Means:

Adaptive K-Means does the partitioning without depending on the initial centroid selection. The algorithm re-arranges the clusters when new elements are added to it. Also, it sometimes merges the clusters and creates new clusters to reflect better partitioning [24].

5. X-means:

Address short comes of K means. X-means [29] decides the appropriate K value so users do not need to define K values. After each iteration of K-means, X-means decides the

subset of the current centroid needed to split the base on BIC (Bayesian Information Criterion) computation.

6. W-Kmeans:

W-Kmeans enhances the standard K-means algorithm by enriching the articles using hypernyms [12]

2.8.3 Model based methods

This is a conceptual approach. Here, data is considered as a sample drawn from independent models [28]. So it generates a model to represent the data. Data assignment to clusters is also handled by the generated model [4, 10].

The Gaussian mixture model is a general probability based model-based method. Kohonen's Self-Organizing Maps (SOM) is another widely used model-based clustering approach [10]. Ant based method is another interesting approach stimulated form real ants [10].

2.8.4 Density based methods

This approach distinguishes the clusters using dense areas separated by sparsely dense areas. This approach is protected from outliers [4, 10, 28] and is capable in handling arbitrary shape clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points to Identify the Clustering Structure) are general density based clustering [4]. This approach is divided into two types [28]:

- Density-Based Connectivity Clustering
- Density Functions Clustering

2.8.5 Grid based methods

This approach divides the data space into cells to form a grid and continues the operations with objects belonging to cells [4, 10]. This approach groups the objects at a hierarchical level. This approach is mainly used for spatial data [28].

2.8.6 Affinity propagation

This approach use message passing between the data points. Each data point receives massage about the availability from exemplars and send the responsibility message to other exemplars. By summing up the availabilities and responsibility data points find the exemplars [30, 31].

2.9 Document/News Clustering Discussion

According to the experimental evaluation carried out by Nanayakara and Ranathunga [4], when applying distance-weighted cosine as a similarity measure with merged similarity of proper names and weighted article title words, Sinhala News Clustering had a marginal improvement in performance. Nanayakkara and Ranathunga [4] also found that removing stop words improves the performance.

Zhao and Karypis [11] proved that partitioning algorithms always performs better than agglomerative algorithms for large document clustering because early agglomerative errors make agglomerative algorithms performance low [11]. However, partitional clustering algorithms perform better for large document datasets and needs relatively low computational requirements [11]. To prove this, Zhao and Karypis [11] introduced constrained agglomerative algorithms, which have the joint features of both partitional and agglomerative algorithms.

Steinbach et al. [27] compare agglomerative hierarchical clustering with K-means where K-means performs better and efficiently in the document domain. Moreover, bisecting K-means performs better than standard K-means. They mention that multiple runs on standard K-means and incremental updating of centroids may be the reasons for K-

means outperforming agglomerative hierarchical clustering. Bisecting K-means creates clusters relatively of the same size, but K-means creates clusters that differ in size so that may lead to why bisecting K-means outperforms. When the document belongs to different classes nearest neighbors, the probability of agglomerative hierarchical clustering mistakes is high, but K-means and the bisecting K-means using the global approach outperforms.

From the Bouras and Tsogkas [12] W-k means (cluster the news articles using WordNet and K-Means) outperforms the standard K-means and partitional algorithms better than hierarchical clustering techniques. In this research, W-k follows a two-step process. The first step enriches the articles using hypernyms and the second step labels the clusters. They find that stemming and finding the noun in the news improves the performance by 5-15% depending on the clustering algorithm.

2.10 News clustering systems

Nalawade et al. [32] introduce an Event Registry system. Event Registry identifies the actual event and creates the cluster across multiple languages. Also, Event Registry differs from the news aggregators because it creates clusters dynamically rather than having a predefined set like business, sports, politics, etc. They collect the news from RSS feeds. Using the bag-of-word model, they represent the articles and maintain the unordered list of words and frequency of the words. By applying TF-IDF, they compute the similarity of the article with centroids of an existing cluster. Similarity between the article and the cluster centroid is calculated using similarity of text, concept mentioned (by wikification - using Wikipedia as the knowledge base link the entities) and from the date difference. If the similarity is higher than the threshold, the article is added into the cluster or a new micro cluster with a single article is created. If the micro cluster reaches 3-6 articles, it is considered as an event. The threshold for the event depends on the language.

Nanayakkara and Ranathunga [4], introduced news clustering system for Sinhala. Nanayakkara and Ranathunga [4], used Term Frequency -Inverse Document Frequency (TF-IDF) vector structure to represent the articles and used similarity measurement based approach to cluster the news. To improve the performances of the system removed the stop words in the articles, distance weighted cosine measure for similarity and merged similarity based on proper names are used.

2.11 Clustering Results Evaluation

Cluster evaluation is used to test the quality of clusters generated by the clustering algorithms. Goodness or quality is the measure used for clustering. Clustering evaluation methods are divided into two core categories.

- Internal Quality Measure [27]

Compare diverse clusters without reference to external knowledge. Here "overall similarity" and dissimilarity between the elements of different clusters are used to calculate the quality.

- External Quality Measure [27]

Clusters generated by the clustering algorithms compared to known classes and quality of clusters measure. Entropy and F-measure are example of external quality measures. Some of the widely used evaluation measures are as follows:

1. Entropy

Entropy [4, 16, 25, 27] measures the quality of clusters based on the distribution of classes in a clusters. Entropy is calculated as follows:

$$E_j = -\sum P_{ij} \log p_{ij}$$

P_{ij} = Probability that member of j belongs to class i

Total entropy for set of clusters is as follows:

$$E_{cs} = \sum \frac{n_j * E_j}{n}$$

Where n_j is the size of cluster j, is the number of clusters, and *n* is the total number of data points.

If a cluster holds data from a single class, the entropy will be zero and smaller entropy means better clusters.

2. Purity

Purity [16] is defined as $P(C_i) = \frac{1}{n_i} \max_h n_i^h$

Where, C_i - cluster, n_i - cluster size, $\max_h n_i^h$ - number of documents from the dominant category in cluster C_i, n_i^h -number of documents from cluster C_i assigned to category h

It calculates the coherence of the cluster. Simply evaluate how much a cluster contains objects from a single class [16]. If purity is high, cluster quality is also high. Purity is 1 for a cluster containing documents from a single class.

3. F-measure (F-score)

F-measure [4, 10, 11, 27] is based on the precision and recalls concepts from information retrieval [27]. Precision and recall are defined as

Precision: the fraction of retrieved documents that are relevant [4]

$$\text{Precision} = \frac{\text{No of relevant items retrieved}}{\text{No of retrieved items}}$$

Recall: the fraction of relevant documents that are retrieved [4]

$$\text{Recall} = \frac{\text{No of relevant items retrieved}}{\text{No of relevant items}}$$

F-score defined as: $f_\beta = (1 + \beta^2) \frac{Precision * Recall}{\beta^2 * Precision + Recall}$ [4]

The balanced F-score (F_1) defined as $\frac{2 * Precision * Recall}{Precision + Recall}$ [4, 27]

F-measure for the total clustering result can be calculated as $F = \sum \frac{n_i}{n} \max (F_1)$ [27]

4. Pairwise F-measure (F-score)

This approach measures the clusters based on how many member pairs two clustering have in common and dissimilar in the cluster [4].

Discussions

Entropy and F-score are widely used in the literature to evaluate the cluster quality. Steinbach et al. [27] used Entropy and F-score to measure cluster quality both performs well.

2.12 Web Crawling

To cluster the news items, it should first be crawled automatically from the internet. Some open source crawlers are:

Crawler4j

Crawler4j provides a simple interface to crawl that supports multithreading and is a Java open source web crawler. It does not support "robots.txt". [4]

Bixo:

It's an open source mining tool. It's based on cascading APIs run top of Hadoop clusters. It is good to handle large collections [4].

DataparkSearch:

Open source search engine and crawler supporting index.[4]

Apache Nutch

Highly extensible and scalable open source web crawler. Written in java and works under Hadoop. It supports batch processing so configuration is a bit challenging. [4]

Sphider

Sphider is a PHP search engine with MySQL. It is open source. It helps to integrate search facility to existing websites. [4] Since it uses a relational database, it is difficult to scale to millions [4]

CHAPTER 3 – METHODOLOGY

3.1 Overview

Figure 3.1 shows the outline of the Tamil News Clustering system. The system consists of three modules, namely,

1. The Tamil News Article Data Collection module
2. Data representation module
3. The Tamil News Article Clustering module

This chapter consists of three sections. In the first section, the sources used to collect the data, and the methodologies used to collect the data are presented. The next section deals with the preprocessing techniques used to process the data and the techniques used to represent the data collected. Finally, the third section is about the approaches used to cluster.

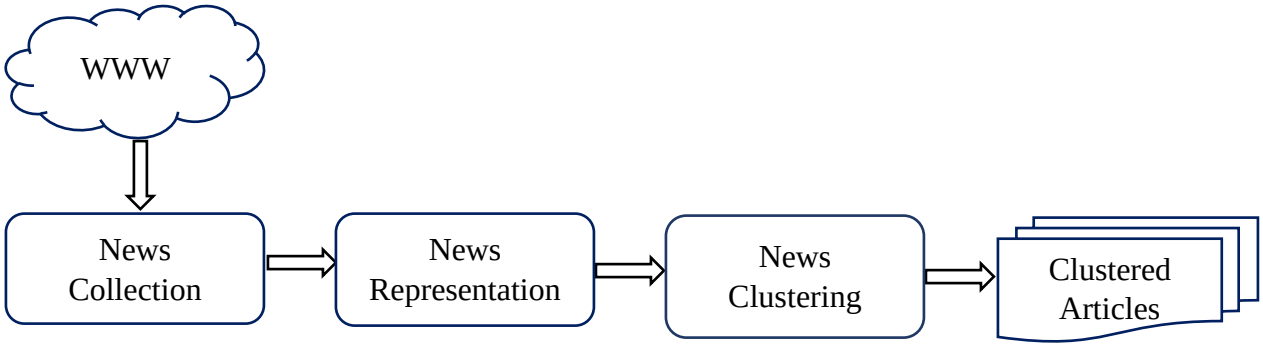

Figure 3.1: Outline of the Tamil News Clustering system

3.2 Tamil News Article Data Collection

This section describes the details of the Tamil news data collection across the web and extracting relevant data from them. This module consists of the details of the Tamil news sources used, crawling methodology used to collect the essential news articles, approaches used to extract the needed content from crawled news article pages and the

common format used to store the collected data. Figure 3.2 shows the outline of the Tamil News Article Data Collection module.

This is an independent module responsible for Tamil news collection. Therefore, this module can be used for different natural language processing tasks like creating the Tamil news corpus.

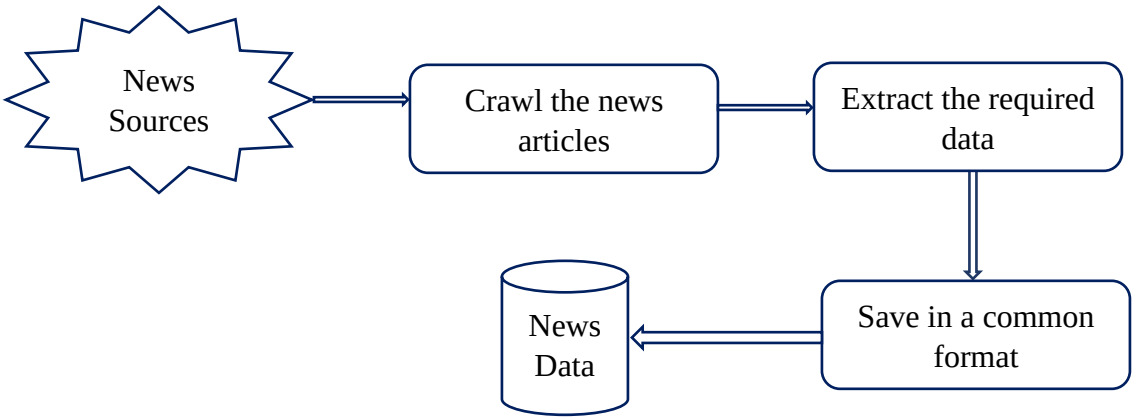

Figure 3.2: Outline of the Tamil News Article Data Collection module

3.2.1 Tamil news sources

There are many online news providers for the Tamil language. Some of them do not update their pages frequently (eg: "Srilanka Mirror"[6]). Further some authors publish the same content to two different websites (Eg: 'Hiru News'[7] and 'Sooriyan News'[8]). Moreover, some are publishing newspapers as digital images (Eg: "Tamli Murasu"[9]). Additionally, when selecting the news sources UTF-8 character encoding support

[6] https://tamil.srilankamirror.com

[7] http://www.hirunews.lk/tamil/

[8] http://www.hirunews.lk/sooriyanfmnews/

[9] http://tm.dinakaran.com/

requirement is also considered. Table 3.1 lists the Tamil news sources used for this research.

Table 3.1: Tamil news sources

News Sources	Source Type	URL
Ada Derana	Online News Site	http://tamil.adaderana.lk/
BBC	Online News Site	https://www.bbc.com/tamil
ITN	Online News Site	http://www.itnnews.lk/ta/
News First	Online News Site	https://www.newsfirst.lk/tamil/
Sooriyan FM News	Online News Site	http://www.hirunews.lk/sooriyanfmnews/
Tamilmirror	Online News Site	https://www.tamilmirror.lk/
Thinakaran	Online Newspaper	http://www.thinakaran.lk/
Thinakkural	Online Newspaper	http://thinakkural.lk/
Virakesari	Online Newspaper	http://www.virakesari.lk/

3.2.2 Web crawling

As the initial step, it is necessary to collect the data from the identified news sources listed above. An open source Python base 'Scrapy'[10] crawler was used for this purpose. Scrapy is a free and open source framework written in Python. Scrapy has good documentation and very good support resources over the web. Scrapy was selected as it is easily configurable and easy to handle using Python. Moreover, the process of extracting the related content and saving the data is made easier by Scrapy.

News Crawler

The main functionality of the news crawler is to decide whether the URL should be visited or not, and then download the pages and extract structured data from the downloaded web pages.

[10] https://scrapy.org

In this research, a crawler is used to crawl the data from news sources. For each individual news source, a unique instant of crawler was used. Crawlers have been defined to decide whether a given link must be visited or not. To visit a URL, the following conditions must be satisfied:

URL must belong to allowed_domains

A static field is used to get rid of irrelevant data. Allowed_domains lists the domains that the crawler is able to crawl. This helps to avoid non relevant web pages such as advertisements getting collected.

Moreover, crawler implementation has strat_urls as seeds and some crawling parameters. Seeds were used to avoid the unwanted data and duplicate data. For example 'http://www.virakesari.lk/category/local' was selected as the seed instead of 'http://www.virakesari.lk' to avoid unrelated news like gossip news. Table 3.2 lists the sample seed URL for the news sources.

Table 3.2: Seed URL list

News Sources	Seed URL
Ada Derana	http://tamil.adaderana.lk/news_archive.php
BBC	https://www.bbc.com/tamil/sri_lanka
ITN	http://www.itnnews.lk/ta/local/
News First	https://www.newsfirst.lk/tamil/category/local/
Sooriyan FM News	http://www.hirunews.lk/sooriyanfmnews/local-news.php
Tamilmirror	http://www.tamilmirror.lk/news/175
Thinakaran	http://www.thinakaran.lk/news
Thinakkural	http://thinakkural.lk/article/category/local
Virakesari	http://www.virakesari.lk/category/local

In addition, the following parameters were used in the crawler:

- FEED_URI (full path the file to be save) = */ %(name)s.xml
- FEED_FORMAT = 'xml'
- FEED_EXPORT_FIELDS = ['URL', 'TITLE', 'BODY', 'DATE']

The crawler starts the execution by making a request to the URLs defined in the start_urls. After that, the crawler calls the callback method and passes the response object as an argument.

3.2.3 News article content collection

Once the content of the URL is downloaded, the URL, title of the news, news content and the date the news was published, are extracted and collected. There is no constant pattern for the design of news web pages, and it varies from one provider to another. So, it become challenging to extract the title, body and date of the news. Moreover, news web sites include advertisement, links to other websites, etc. in the news page in addition to the actual news content.

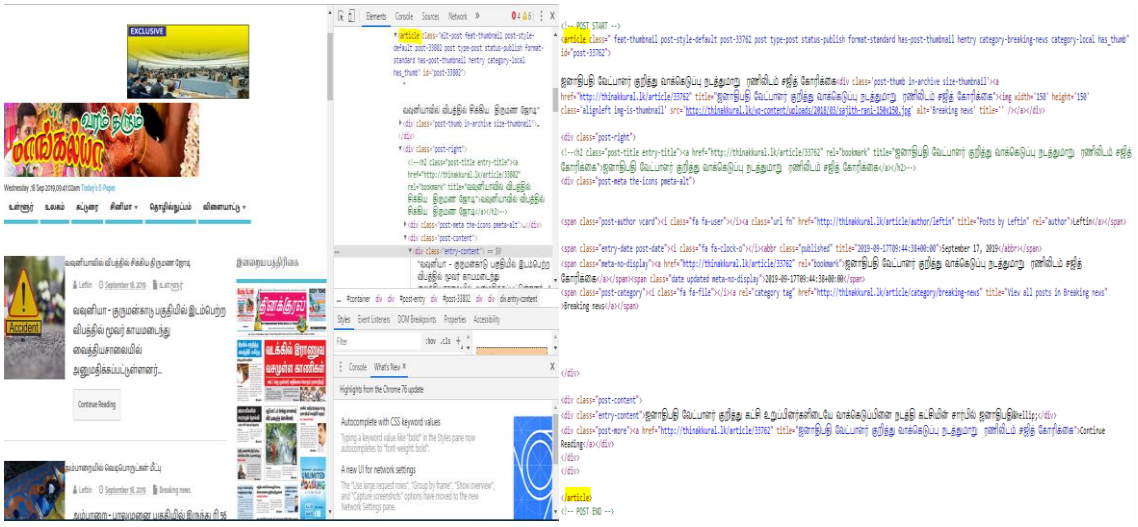

Figure 3.3: News website

To overcome these challenges, the page format of each news source was thoroughly analyzed, and common patterns were identified. After that, the pattern was searched inside the HTML to extract the title, body and date. For example, most websites have article tags and inside that, the title, body and date of the news are included. Figure 3.3 shows the sample news source. This approach is vastly based on heuristic techniques and may need modifications when the news providers update the design of their website. Figure 3.4 shows sample extracted news content.

Figure 3.4: Sample extracted news content

3.2.4 News data storing

Collected data was saved in the .xml file using UTF-8 encoding. The following structure is used to save news data.

```
<item>
        <URL>URL of the page</URL>
        <TITLE>Tile of the news</TITLE>
        <BODY>News description</BODY>
        <DATE>News published date</DATE>
</item>
```

Figure 3.5: News data file structure

Figure 3.6 shows a sample news data file.

<item>
<URL>https://www.newsfirst.lk/tamil/category/sports/</URL>
<TITLE> இலங்கை வீரர்களின் துடுப்பாட்ட சராசரி வீழ்ச்சியடைந்தது ஏன்? </TITLE>
<BODY> இலங்கை கிரிக்கெட் அணியின் பயிற்றுநராக சந்திக்க ஹத்துருசிங்கவும் துடுப்பாட்டப் பயிற்றுநராக திலான் சமரவீரவும் நியமிக்கப்பட்டு ஒரு வருடம் கடந்துள்ளது.</BODY>
<DATE> 01 Dec, 2018 </DATE>
</item>
<item>
<URL>https://www.newsfirst.lk/tamil/category/world/</URL>
<TITLE> நிகரகுவா துணை ஜனாதிபதிக்கு எதிராக தடை விதித்தது அமெரிக்கா </TITLE>
<BODY>மத்திய அமெரிக்க நாடான நிகரகுவாவின் துணை ஜனாதிபதி ரொசாரியோ முரில்லோவிற்கு (Rosario Murillo) எதிராக, அமெரிக்கா தடைகளை விதித்துள்ளது.</BODY>
<DATE> 28 Nov, 2018 </DATE>
</item>
<item>
<URL>https://www.newsfirst.lk/tamil/category/world/</URL>
<TITLE> செவ்வாயில் தரையிறங்கிய இன்சைட் ரோபோ விண்கலம் நாசாவிற்கு புகைப்படம் அனுப்பியது </TITLE>
<BODY> நாசா விண்வெளி மையத்தால் வடிவமைக்கப்பட்ட முதல் ரோபோ விண்கலம் இன்சைட், 7 மாதங்களுக்குப் பிறகு செவ்வாய் கிரகத்தில் வெற்றிகரமாக தரையிறங்கிய அடுத்த சில நிமிடங்களிலேயே தான் எடுத்த புகைப்படங்களை நாசாவிற்கு அனுப்பியுள்ளது.</BODY>
<DATE> 27 Nov, 2018 </DATE>
</item>
<item>
<URL>https://www.newsfirst.lk/tamil/category/world/</URL>
<TITLE> தமிழகத்திற்காக மோடியிடம் நிதியுதவி கோரும் தமிழக முதல்வர் </TITLE>

Figure 3.6: Sample news data file

3.3 Data representation

This section describes the details of the data preprocessing and data representation technologies used. The data saved in the .xml file cannot be directly used for similar calculation and clustering. Data should be converted into a well-defined format. In this research, Term Frequency - Inverse Document Frequency (TF-IDF) and word embedding techniques were used to represent the data.

3.3.1 Pre-processing

The pre-processing approach consists of a few steps. As the first step, an index number was added to every news article for easy handling. After that, data was formatted into the following structure.

<NEWS>

<INDEX>Index number for news</INDEX>

<URL>URL of the page</URL>

<TITLE>Title of the news</TITLE>

<BODY>News description</BODY>

<DATE>News published date</DATE>

</NEWS>

As the next step, individual news articles were identified using the <NEWS> token, and the title and body of the news were identified using the <TITLE> and <BODY> tokens. After that, title and body of the news were tokenized into individual words using the following delimiters:

- White space characters - Space, Tab, Newline/Carriage return
- Punctuation marks

Following the tokenizing process subsequently, the punctuations marks are removed.

3.3.2 Term Frequency-Inverse Document Frequency (TF-IDF) vector

The Term Frequency-Inverse Document Frequency (TF-IDF) for each word is calculated and the TF-IDF vector is constructed for every word of all articles.

3.3.3 Word Embedding model

A skip-gram [19, 20, 21] word embedding model was created using 1 728 158 documents with 21 077 843 words. The model was created with a vector dimension of 300.

In addition to the data collected from the above approach, we have used the data collected by Farhath [40] to create the embedding models.

News articles vector is calculated by averaging the word vector of each and every word in the news article.

3.4 Tamil News Article Clustering

This section explains the process carried out for news clustering. The news article clustering module consists of similarity calculation and clustering.

3.4.1 Similarity calculation

Whether two articles are associated with respect to content or not is decided by using the similarity between the articles. In this research, cosine similarity [4, 17] was used to measure the similarity between news articles, where a high score indicates that the compared two articles are similar in their content and a low score indicates no relation between them.

3.4.2 Tamil news Clustering

News articles need to be clustered with respect to their content similarity. Using a labeled data set to group the news articles is very challenging since the news changes dynamically. It is very hard to predict the number of clusters in advance since the number of clusters heavily depends on the collected news article content. These make it impossible to use popular clustering algorithms like K-means, agglomerative clustering algorithms etc.

Bearing in mind the above factors, algorithms that can define the number of clusters dynamically according to the content were selected. It eliminates the need to define the number of clusters beforehand.

In this study, two clustering algorithms were used.

- One pass algorithm

  This approach was introduced by Leban et al. [32] and used by Nanayakkara and Ranathunga [4] to cluster the articles. This algorithm was able to define the number of clusters according to the content, no need to define the number of clusters earlier. A THRESHOLD value is used to define the clusters. Optimal THRESHOLD value was obtained by carrying out a number of experiments with test data sets.

  The algorithm works as follows,

  1 Allocate the first news article to the first cluster and make it the centroid of the cluster

2 For all remaining articles,

- Calculate the similarity between the news article and each of the existing cluster centroids
- Determine the maximum similarity (Similarity Max) value for a particular news article
- If Similarity Max >= THRESHOLD add the current article to the cluster with the maximum similarity value
- Update the cluster centroid by getting the mean value of all the news articles
- If Similarity Max < THRESHOLD add the current news article to a new cluster and update the clusters list

- Affinity propagation

As discussed in Section 2.8.6, affinity propagation was selected as the clustering algorithm since it is capable of clustering data and finding the number of clusters simultaneously.

3.4.3 Baseline system setup

The baseline system was configured with TF-IDF as the document representing technology and one pass algorithm was configured as the clustering algorithm. This approach was used as the benchmark to compare the other approaches.

3.5 Alternative approach

By introducing different document representation techniques and different clustering approaches, multiple alternative approaches were tried out to get the best result for Tamil news clustering. Those are discussed below.

TF-IDF with Affinity propagation Algorithm

In this approach TF-IDF was used as the document representing technology and Affinity propagation was used for the clustering.

Word Embeddings with One pass Algorithm

In this case, word embedding was used to represent the document and One pass algorithm was used for the clustering.

Word Embeddings with Affinity propagation Algorithm

In this case, word embedding was used to represent the document and Affinity propagation was used for clustering.

3.6 Clustering based on title

Meedeniya and Perera [39] have demonstrated that text document grouping based on the article titles gives high performance than groping based on the article body for English documents. To verify the validity for the Tamil news article titles, the following approaches were applied.

1. TF-IDF with One pass propagation algorithm
2. TF-IDF with Affinity propagation algorithm
3. Word Embeddings with One pass algorithm
4. Word Embeddings with Affinity propagation algorithm

3.7 Summary

This chapter discussed the background details of data sources, technical details of automatic Tamil news collection and the clustering the collected articles according to their content similarity. We used TF-IDF and word embedding techniques to represent Tamil news articles, cosine measure used for the similarity calculation and we used one pass clustering algorithm and affinity propagation for the clustering process. Further we were able to create a word embedding model for Tamil news.

CHAPTER 4 – EVALUATION AND ANALYSIS

4.1 Overview

This chapter presents the evaluation results and analysis of the techniques presented in the thesis so far. To complete this study, how the objectives of this research have been achieved is described in this chapter with evidence on contributions made by the thesis. The results obtained were carried out through statistical analysis and are presented in this chapter. The set of experiments done under each topic were discussed here under the respective topics.

4.2 Data Collection

As the initial step, dissimilar sets of Tamil news articles were collected from multiple news sources. News articles were collected on non-adjacent days to get the variation in the data set. Each data set includes news articles collected from nine different news sources listed in Table 3.1.

As the next step, baseline article groups were identified in order to use when comparing the system generated clusters. Baseline news clusters were obtained by manual clustering of articles. Table 4.1 shows the summary of collected article sets and manual clustering output. For example, data set 3 contains 211 news articles and of that, 106 news articles were clustered into 25 groups and the remaining 105 news articles were not put into any clusters and were treated as single articles.

Table 4.1: Details of the Datasets

Data Set	**Number of Articles**	**Number of Clusters**	**Average Cluster Size**	**Largest Cluster Size**	**Grouped Articles Percentage**
Data Set 1	235	33	4.30	14	60.42%
Data Set 2	260	34	3.94	8	51.53%

Data Set 3	211	25	4.2	18	49.76%
Data Set 4	236	30	4.13	15	52.54%
Data Set 5	270	27	5	25	50%
Data Set 6	193	27	4.4	8	61.13%
Data Set 7	168	28	3.25	7	54.16%
Data Set 8	258	32	3.75	13	46.51%
Data Set 9	254	33	3.61	17	46.85%
Data Set 10	255	32	3.91	20	49.01%

In order to evaluate the manual clustering, three data sets were given to two more people and their clustering were compared with the above manual clusters. Fleiss' Kappa statistic [4, 42] was calculated to evaluate the similarity between the manual clusters. Table 4.2 show Kappa statistic values of agreement between manual clustering. The evaluation scores of three manual clustering show good agreement. For instance, in the dataset 5, all three evaluators agreed on 40 clusters while they had different opinion on 3 cluster. Out of 270 news articles only 6 news articles differently clustered.

Table 4.2: Kappa statistic values of agreement between manual groupings

Dataset	**Fleiss' kappa value**
Dataset 5	0.956
Dataset 6	0.940
Dataset 7	0.867
Mean (Average)	0.921

4.3 Evaluation Method

The performance of the system was evaluated based on the Pairwise F-score. As mentioned in Section 2.11, pairwise F-score is appropriate for evaluating the cluttering outputs with different numbers of clusters.

For all cases to be evaluated, the pairwise F-score was calculated as below, with respect to the above described manual clustering result.

For all the unique data sets [4],

1 The text file holding the data set was fed into the system and system-generated clusters were saved.

2 The F-score was calculated as follows:

Calculate the number of news articles both manual cluttering and system generated clustering share (a)

Calculate the news articles, that only in manual clustering contains (b)

Calculate the news articles, that only in system clustering contains (c)

F-score= $\frac{2a}{(2a+b+c)}$

4.4 Baseline (TF-IDF with One pass algorithm)

The baseline system was configured as per the description given in Chapter 3. In order to evaluate the effectiveness of the baseline system, the following steps were carried out.

1. Test data sets were fed into the system and system generated groups were recorded.
2. For each test data set, the F-score with respect to the manual grouping were calculated.
3. F-score obtained for each data set was recoded and analyzed.

Figure 4.1 shows the obtained pairwise F-score values and Table 4.3 statistically analyzes the obtained F-score values.

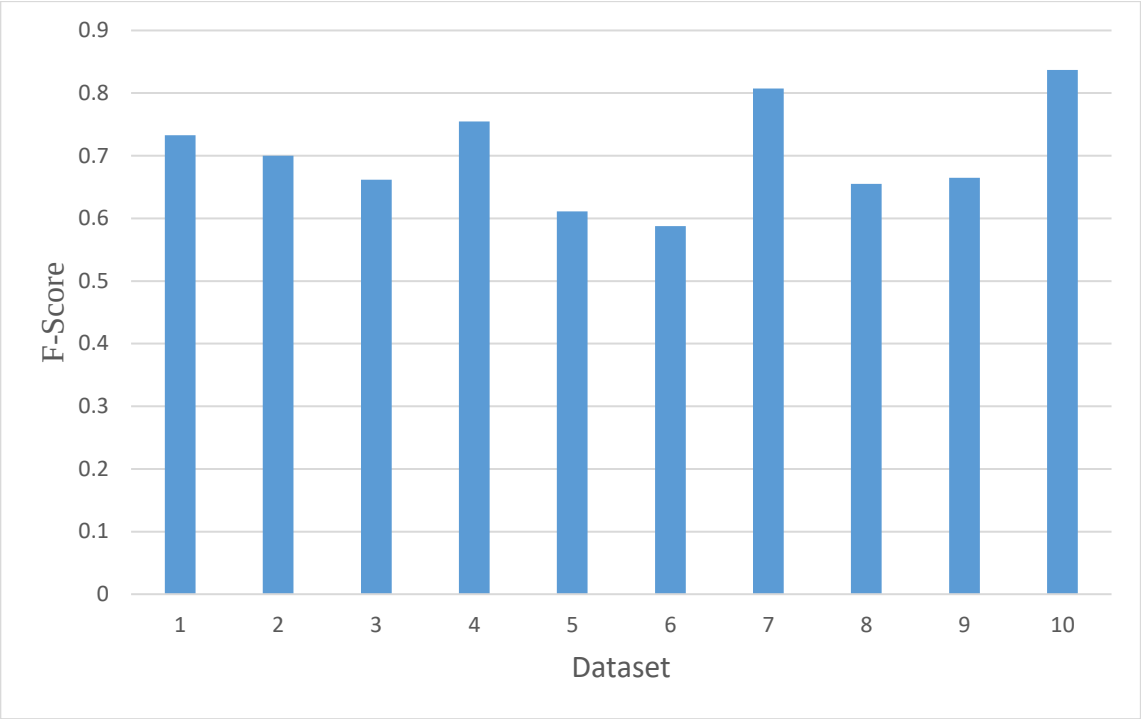

Figure 4.1: Effect of Baseline

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

Mean (Average)	0.701
Median	0.682
Minimum	0.588
Maximum	0.837
Standard Deviation	0.077

4.5 Effectiveness of clustering based on TF-IDF with Affinity propagation Algorithm

In order to evaluate the effectiveness of TF-IDF with Affinity propagation algorithm in Tamil news clustering, the F-score with respect to the manual grouping was calculated for each dataset. Figure 4.2 shows the obtained pairwise F-score values with the baseline result and Table 4.4 statistically analyzes the obtained F-score values.

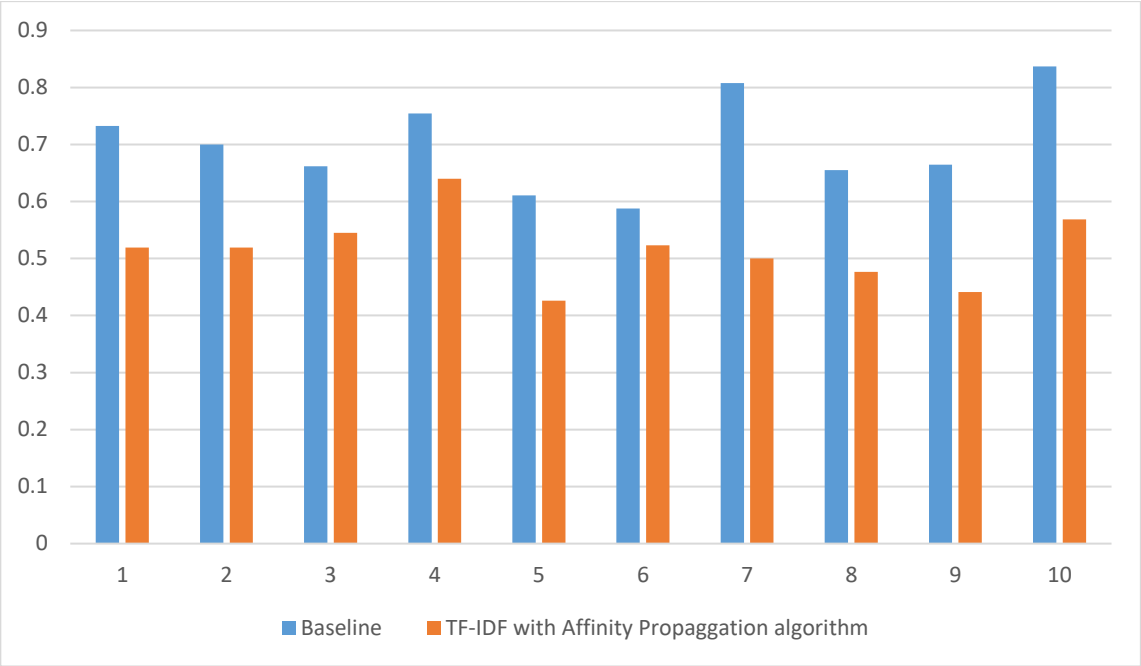

Figure 4.2: Effectiveness of TF-IDF with Affinity propagation Algorithm

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

	Baseline (TF-IDF with One pass algorithm)	TF-IDF with Affinity propagation Algorithm
Mean (Average)	0.701	0.516
Median	0.682	0.519
Minimum	0.588	0.426
Maximum	0.837	0.639
Standard Deviation	0.077	0.059

According to the obtained result, the baseline performs better than the TF-IDF with Affinity propagation approach in Tamil news clustering in all the aspects like mean, median, minimum, maximum and Standard Deviation of obtained F-score values. This shows that the One pass algorithm performs better than Affinity propagation with the TF-IDF document representing technique in Tamil news clustering.

4.6 Effectiveness of clustering based on Word Embeddings with One pass Algorithm

The evaluation scores of using word embedding along with one pass algorithm to Tamil news clustering is described in this section. The evaluation of clustering quality is calculated by measuring the F-score with respect to manual clustering. The statistical analysis of the obtained F-score values is shown in Table 4.5 and Figure 4.3 shows the obtained pairwise F-score.

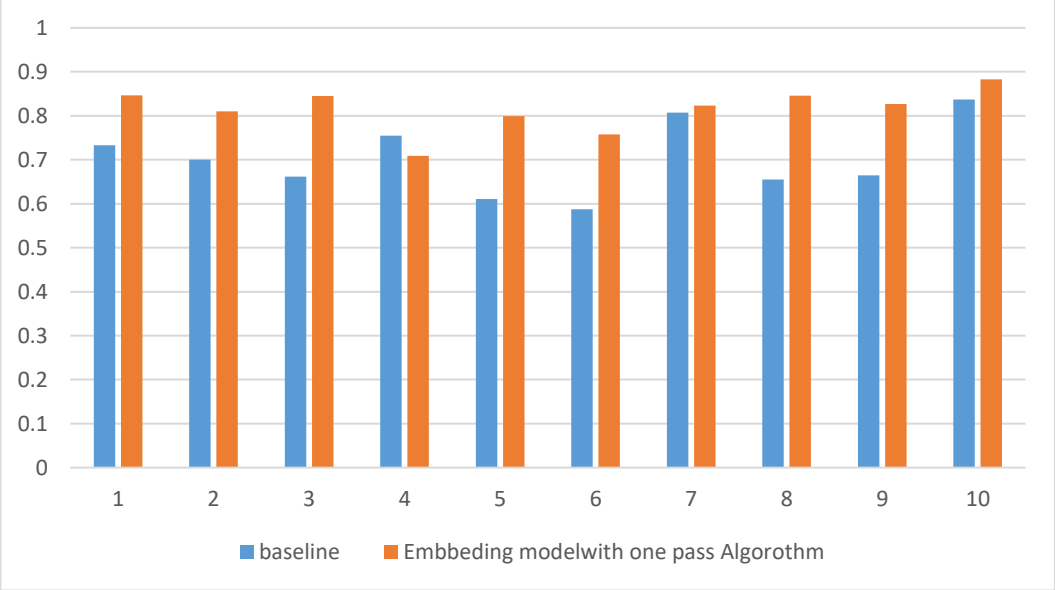

Figure 4.3: Effectiveness of Word Embeddings with One pass Algorithm

Table 4.5: Statistical analysis of Word Embeddings with One pass Algorithm

	Baseline (TF-IDF with One pass algorithm)	Embedding with One pass Algorithm
Mean (Average)	0.701	0.815
Median	0.682	0.825
Minimum	0.588	0.709
Maximum	0.837	0.883
Standard Deviation	0.077	0.047

In Figure 4.3 and Table 4.5, it is clearly shown that using word embedding with one pass algorithm performs better than the baseline in all the aspects like mean, median, minimum, maximum and Standard Deviation of obtained F-score values. Also, it gave promising quality improvements in clustering the Tamil news articles. Moreover, it clearly show that word embedding for document repartition performs far better than TF-IDF in Tamil news clustering.

4.7 Effectiveness of clustering based on Word Embeddings with Affinity propagation Algorithm

The experimental results reported herewith are for the techniques explained under section 3.5.3. The evaluation results are reported in Figure 4.4 and Table 4.6.

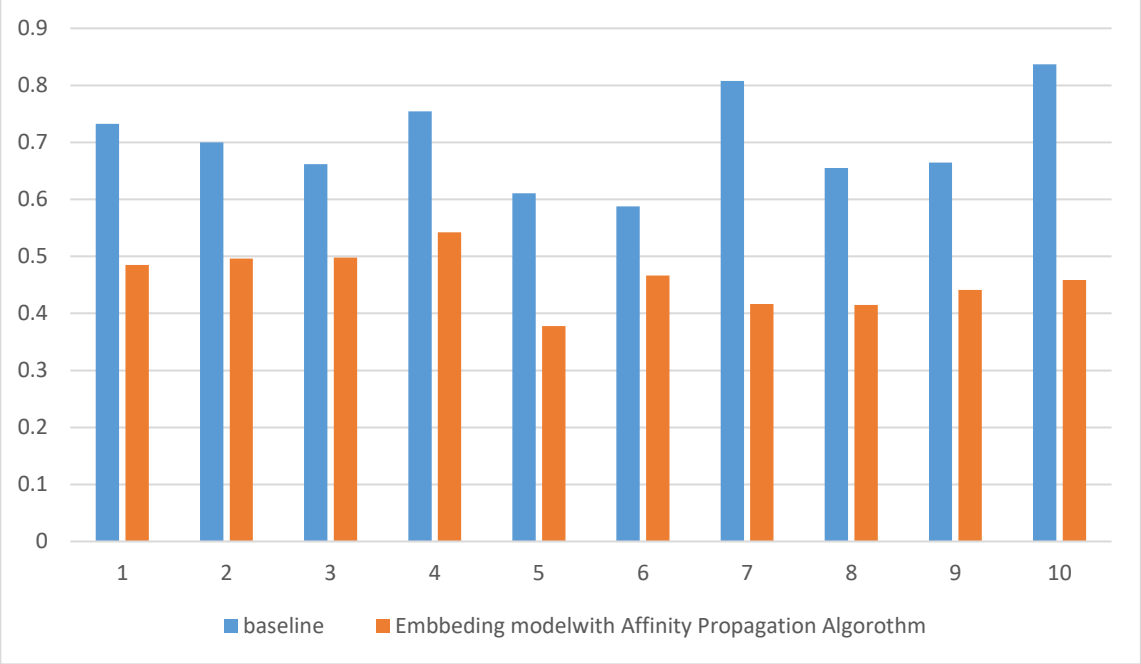

Figure 4.4: Effectiveness of Word Embeddings with Affinity propagation Algorithm

Table 4.6: Statistical analysis of Word Embeddings with Affinity propagation Algorithm

	Baseline (TF-IDF with One pass algorithm)	Embeddings with Affinity propagation Algorithm
Mean (Average)	0.701	0.459
Median	0.682	0.463
Minimum	0.588	0.378
Maximum	0.837	0.542
Standard Deviation	0.077	0.046

As reported in Figure 4.4 and Table 4.6, the baseline system performs better than the Word Embeddings with Affinity propagation Algorithm in Tamil news clustering in all the aspects like mean, median, minimum, maximum and Standard Deviation of obtained F-score values.

4.8 Effectiveness of clustering based on article title

To measure the performance of the clustering base on the title, the following steps were carried out.

1. Each data set text file was fed into the system and the generated results were recorded here; both news title and body content are taken to account
2. Each data set text file was fed into the system and the generated results were recorded here with only the news title
3. The F-score for above cases with respect to manual clustering was calculated

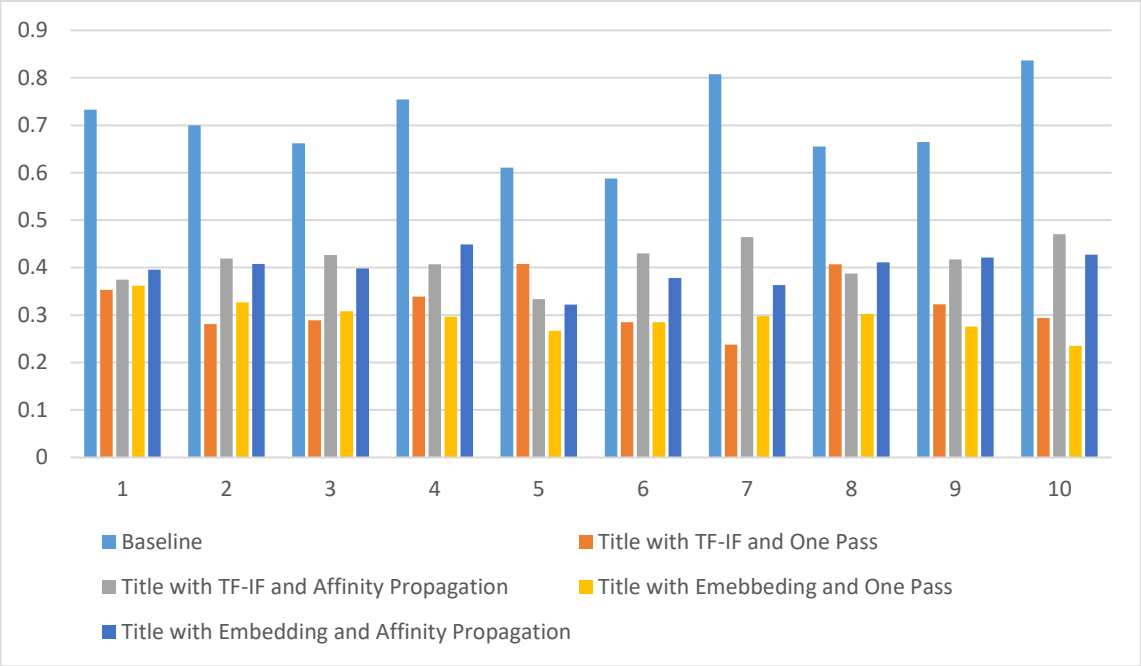

Figure 4.5: Tamil news clustering based on the news title

Table 4.7: Statistical analysis of Tamil news clustering based on the news title

	Baseline	TF-IDF with one Pass	TF-IDF with affinity propagation	Word embedding with one pass	Word embedding with affinity propagation
Mean (Average)	0.701	0.322	0.413	0.296	0.397
Median	0.682	0.308	0.418	0.297	0.403
Minimum	0.588	0.238	0.333	0.235	0.322
Maximum	0.837	0.407	0.471	0.362	0.449
Standard Deviation	0.077	0.053	0.039	0.032	0.034

The result gained from the experiments clearly shows that clustering just using news titles is not successful. This observation agrees with the observation made by Nanayakkara and Ranathunga [4] on clustering Sinhala news articles and observation contrasts with the observation made by Meedeniya and Perera [39] on a set of English documents. It may be due to,

- Some news titles being very short with only two or three words providing very little information. For examples, விசேட கலந்துரையாடல்...!!(special discussion), தீர்ப்பு வெளியானது (Judgment Released), ஏற்க தயார் (Ready to accept), 60 பேர் பலி (60 killed).
- Some news titles report the event indirectly. For example:
  In the following sample, both news bodies are talking about "YouTube service" but in the first news title, it is directly pointing to the event whereas in the second title, it is indirectly pointing.

<NEWS>
<TITLE> UPDATE: YOUTUBE சேவை வழமைக்குத் திரும்பியது (Youtube services back to normal) </TITLE>
<BODY> உலகின் முதல்நிலை இணையத்தளமான யூடியூப்பின் சேவை மீண்டும் வழமைக்குத் திரும்பியுள்ளது. </BODY>
</NEWS>

<NEWS>
<TITLE> உலகையே பரபரப்பாக்கியுள்ள விடயம்..!! (The thing which makes world excited)
</TITLE>
<BODY> உலகின் பிரபல சமூக வலைத்தளமான youtube தற்போது செயலிழந்துள்ளதாக தெரிவிக்கப்பட்டுள்ளது. </BODY>
</NEWS>

The following subsections discuss how title-based clustering and title and body-based clustering works for different approaches:

4.8.1 Effectiveness of clustering based on article title using TF-IDF with One pass algorithm

As shown in Figure 4.6 and Table 4.8, title and body-based clustering performs better than only title based clustering in the TF-IDF with the one pass algorithm approach.

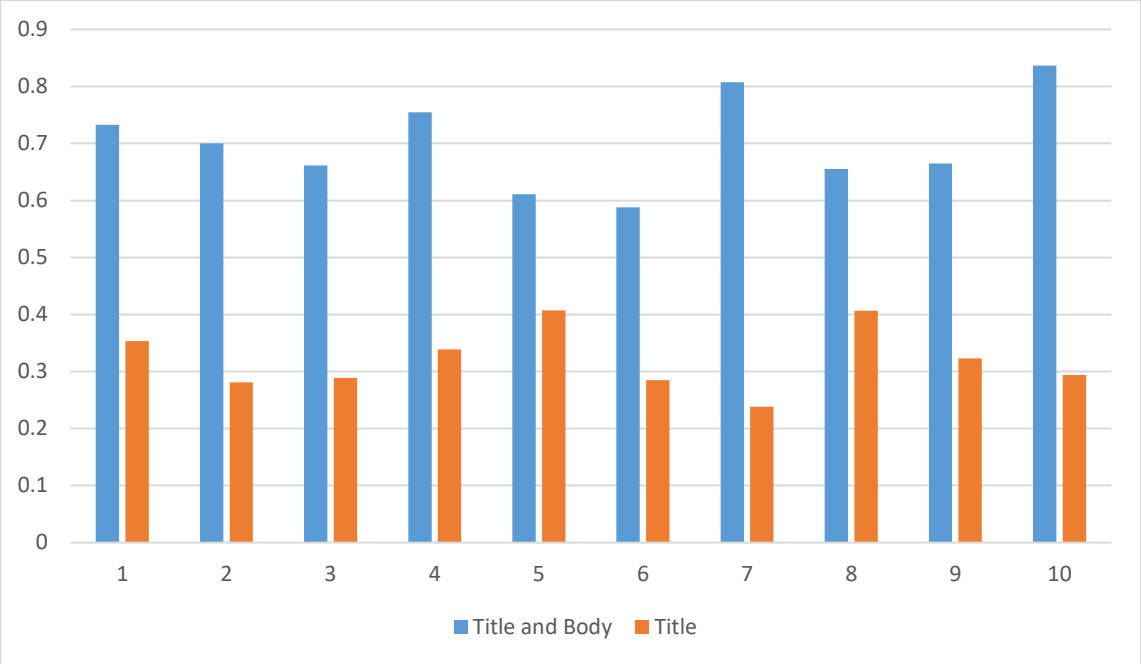

Figure 4.6: Tamil news clustering based on TF-IDF with One pass algorithm

Table 4.8: Statistical analysis of Tamil news clustering based on TF-IDF with One pass algorithm

	Title and Body	Title
Mean (Average)	0.701	0.322
Median	0.682	0.308
Minimum	0.588	0.238
Maximum	0.837	0.407
Standard Deviation	0.077	0.053

4.8.2 Effectiveness of clustering based on article title using the TF-IDF with Affinity Propagation Algorithm

As shown in Figure 4.7 and Table 4.9 title and body-based clustering performs better than only title based clustering in the TF-IDF with affinity propagation algorithm approach.

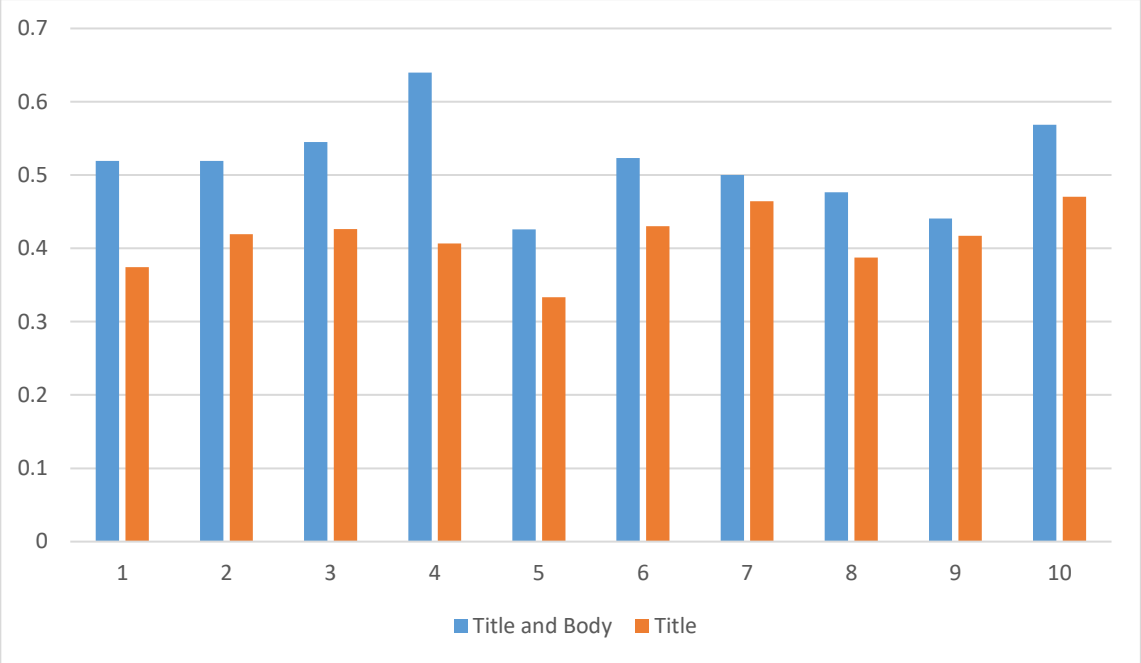

Figure 4.7: Tamil news clustering based on TF-IDF with Affinity propagation algorithm

Table 4.9: Statistical analysis of Tamil news clustering based on the TF-IDF with Affinity propagation algorithm

	Title and Body	Title
Mean (Average)	0.516	0.413
Median	0.519	0.418
Minimum	0.426	0.333
Maximum	0.639	0.471
Standard Deviation	0.059	0.039

4.8.3 Effectiveness of clustering based on article title using Word Embedding with One pass algorithm

As shown in Figure 4.8 and Table 4.10, title and body-based clustering performs better than only title based clustering in word embedding with one pass algorithm approach.

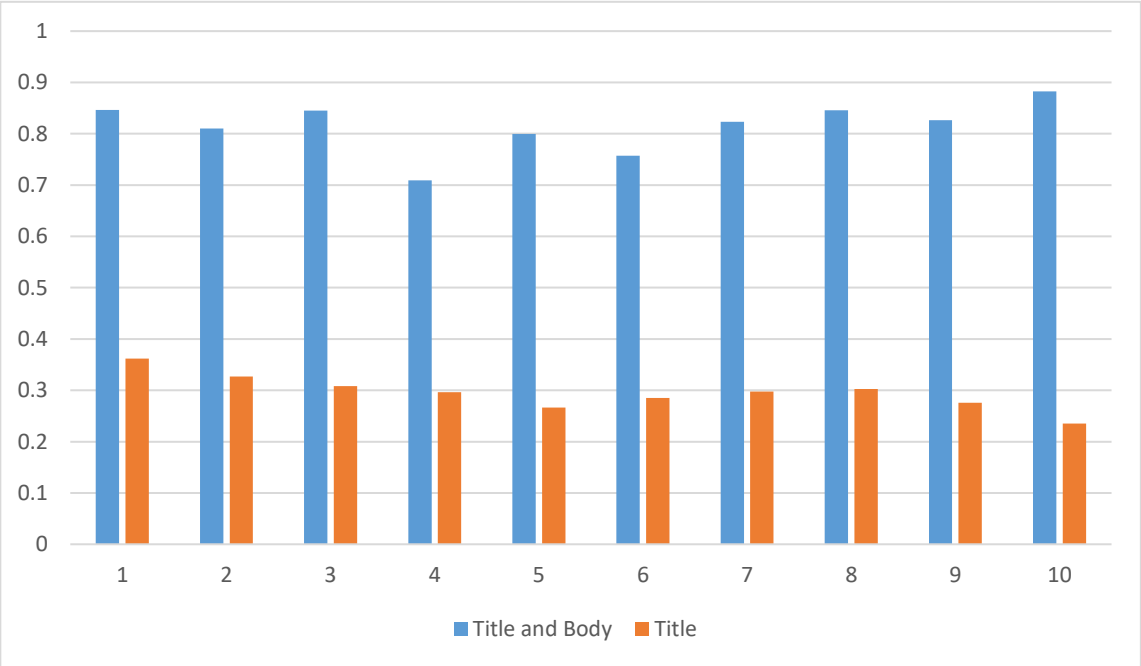

Figure 4.8: Tamil news clustering based on Word Embedding with the One pass algorithm

Table 4.10: Statistical analysis of Tamil news clustering based on Word Embedding with the One pass algorithm

	Title and Body	Title
Mean (Average)	0.815	0.296
Median	0.825	0.297
Minimum	0.709	0.235
Maximum	0.883	0.362
Standard Deviation	0.047	0.032

4.8.4 Effectiveness of clustering based on article title using Word Embedding with Affinity Propagation Algorithm

As shown in Figure 4.9 and Table 4.11, title and body-based clustering performs better than only title based clustering in word embedding with the affinity propagation algorithm approach.

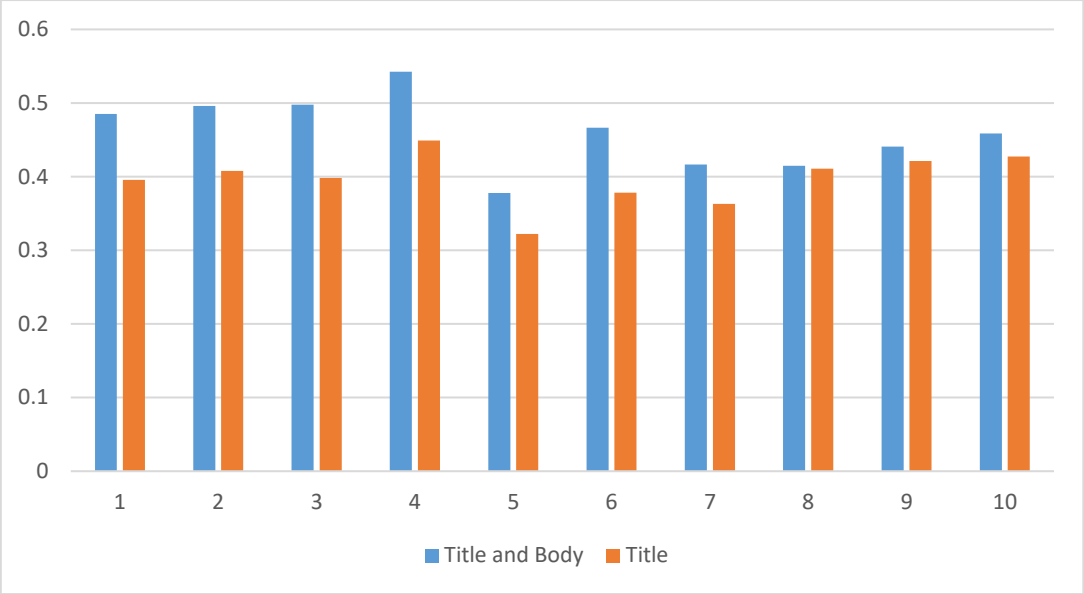

Figure 4.9: Tamil news clustering based on Word Embedding with Affinity Propagation Algorithm

Table 4.11: Statistical analysis of Tamil news clustering based on Word Embedding with Affinity Propagation Algorithm

	Title and Body	Title
Mean (Average)	0.459	0.397
Median	0.463	0.403
Minimum	0.378	0.322
Maximum	0.542	0.449
Standard Deviation	0.046	0.034

4.9 Effectiveness of document representing techniques and clustering techniques.

With the motive to achieve more fluency in the output, multiple experiments were carried out with different clustering algorithms and different document representing approaches. Results acquired for those experiments are discussed below.

4.9.1 TF-IDF Vs Word embedding

To identify the best document representing approach for Tamil, we have compared the results obtained from different clustering approaches in Table 4.12.

Table 4.12: Statistical analysis of documents representing approaches

	One pass Algorithm		Affinity propagation Algorithm	
	TF-IDF	Word Embedding	TF-IDF	Word Embedding
Mean (Average)	0.701	0.815	0.413	0.459
Median	0.682	0.825	0.418	0.463
Minimum	0.588	0.709	0.333	0.378
Maximum	0.837	0.883	0.471	0.542
Standard Deviation	0.077	0.047	0.039	0.046

The table clearly shows that word embedding outperforms than TF-IDF in both cases in all the aspects like mean, median, minimum and maximum of obtained F-score values.

4.9.2 One pass clustering algorithm Vs Affinity propagation

To identify the best clustering approach for Tamil, we have compared the results obtained from different word representing approaches, as shown in Table 4.14.

Table 4.14: Statistical analysis of clustering approaches for Tamil news

	Word Embedding		TF-IDF	
	One pass Algorithm	Affinity propagation Algorithm	One pass Algorithm	Affinity propagation Algorithm
Mean (Average)	0.815	0.459	0.701	0.516
Median	0.825	0.463	0.682	0.519
Minimum	0.709	0.378	0.588	0.426
Maximum	0.883	0.542	0.837	0.639
Standard Deviation	0.047	0.046	0.077	0.059

The table 4.14 clearly shows that one pass algorithm outperforms affinity propagation under both document representing approaches in all the aspects like mean, median, minimum and maximum of obtained F-score values.

4.10 Overall Performance

As shown above, clustering news articles with title and body using word embedding and one pass clustering algorithm gives better results. Therefore this combination is considered as overall system performance. Table 4.15 shows pairwise F-score values obtained for each data set and Table 4.16 show the summary of it. Figure 4.10 shows the graphical representation of pairwise F-score values obtained for each dataset.

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

Dataset	**Pairwise F-Score**
Dataset 1	0.846
Dataset 2	0.81

Dataset 3	0.845
Dataset 4	0.709
Dataset 5	0.799
Dataset 6	0.757
Dataset 7	0.823
Dataset 8	0.846
Dataset 9	0.827
Dataset 10	0.883

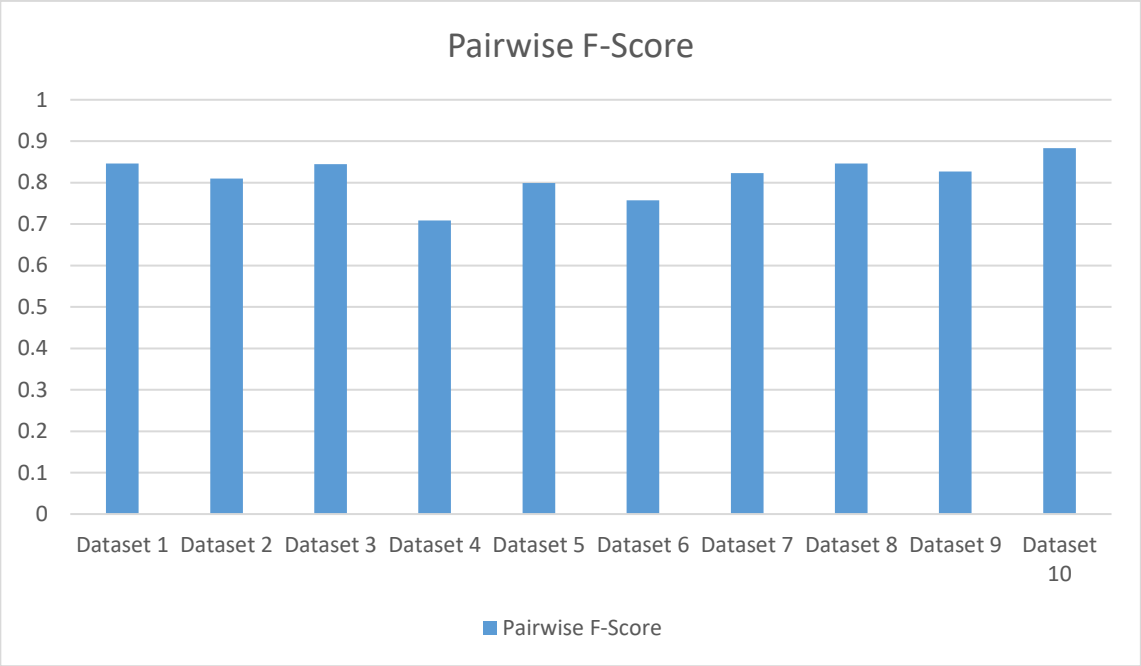

Figure 4.10: Pairwise F-score values obtained for each data set

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

	Embedding with One pass Algorithm
Mean (Average)	0.815
Median	0.825
Minimum	0.709

Maximum	0.883
Standard Deviation	0.047

As described in Table 4.16, we were able to attain an average of 0.815 with a standard deviation of 0.047 Maximum was achieved for dataset 10 and minimum results were obtained for dataset 4.

4.11 Discussion

We were able to develop a news clustering system for Tamil news with average accuracy of 0.81 (measured with pairwise F-score) with respect to the manual clustering. The best performance was achieved by applying word embedding and one pass algorithm. Even though the system was able to cluster similar news in most of the cases, it has failed in some cases.

For example, in the following case, the system has clustered four news articles taking three different instances related to corpse into one cluster. In the example, first and fourth articles talk about similar instances and the second and third are talking of two different instances.

<NEWS>
<CLUSTER> 24 </CLUSTER>
<TITLE> ஆற்றிலிருந்து ஆண் ஒருவரின் சடலம் மீட்பு (A male corpse recovery from a river) </TITLE>
<BODY>திம்புள்ள, பத்தனை பொலிஸ் பிரிவிற்குட்பட்ட குயின்ஸ்பெரி கீழ்பிரிவில் உள்ள கிளை ஆற்றிலிருந்து ஆண் ஒருவரின் சடலம் இன்று (07) காலை மீட்கப்பட்டுள்ளது. </BODY>
</NEWS>

<NEWS>
<CLUSTER> 24 </CLUSTER>
<TITLE> சடலம் ஒன்று மீட்பு (A corpse recovery) </TITLE>

<BODY>பலாங்கொடை - பின்னவல - அமுபிட்டிய பிரதேசத்தில் உள்ள வனப்பகுதி ஒன்றில் இருந்து அடையாளம் காணப்படாத சடலம் ஒன்று மீட்கப்பட்டுள்ளது. </BODY>
</NEWS>

<NEWS>
<CLUSTER> 24 </CLUSTER>
<TITLE>மர்மமான முறையில் உயிரிழந்த ஆணின் சடலம் மீட்பு (Recovery of the corpse of a mysteriously dead man) </TITLE>
<BODY>புத்தளம் பொலிஸ் பிரிவுக்குட்பட்ட கல்லடி தம்மன்னா கிராமத்தில் மர்மமான முறையில் உயிரிழந்த ஒருவரின் சடலத்தை இன்று</BODY>
</NEWS>

<NEWS>
<CLUSTER> 24 </CLUSTER>
<TITLE>இரு பிள்ளைகளின் தந்தை சடலமாக மீட்பு (A father of two children recover as corpse) </TITLE>
<BODY>பத்தனை பொலிஸ் பிரிவுக்குட்பட்ட குயின்ஸ் பெரி தோட்டத்திலுள்ள நீரோடையிலிருந்து ஆண் ஒருவரின் சடலம் ஒன்று இன்று </BODY>
</NEWS>

System failed to differentiate the following four new items related to arrest. First and third news talking about arrest because of keeping guns and second and fourth are talking about a foreign citizen's arrest

<NEWS>
<CLUSTER> 23 </CLUSTER>
<TITLE>துப்பாக்கி ஒன்றுடன் நபர் ஒருவர் கைது (Person arrested with a gun) </TITLE>
<BODY>கொஸ்கொட, தூவமோதர பகுதியில் உள்ள வீடொன்றில் இருந்து வெளிநாட்டில் உற்பத்தி செய்யப்பட்ட துப்பாக்கி ஒன்றுடன் நபர் ஒருவர் கைது செய்யப்பட்டுள்ளார். </BODY>
</NEWS>

<NEWS>
<CLUSTER> 23 </CLUSTER>
<TITLE> போலி நாணயத் தாள்களுடன் வெளிநாட்டு பிரஜை கைது (Foreign citizen arrested with fake currency notes) </TITLE>
<BODY> போலி நாணயத் தாள்களை வைத்திருந்த வெளிநாட்டு பிரஜை ஒருவர் பிலியந்தலை, மடபாத்த பிரதேசத்தில் வைத்து கைது செய்யப்பட்டுள்ளார். </BODY>
</NEWS>

<NEWS>
<CLUSTER> 23 </CLUSTER>
<TITLE> துப்பாக்கி மற்றும் ஹெரோயினுடன் இருவர் கைது (Two arrested with gun and heroin) </TITLE>
<BODY> கொஸ்கொட - துவேமோதர பகுதியில் துப்பாக்கியுடன் இரண்டு பேர் கைது செய்யப்பட்டுள்ளனர். </BODY>
</NEWS>

<NEWS>
<CLUSTER> 23 </CLUSTER>
<TITLE> வெளிநாட்டவர் ஒருவர் கைது. (A foreign citizen arrested) </TITLE>
<BODY> 5 ஆயிராம் ரூபா போலி நாணயத்தாள்களை கைவசம் வைத்திருந்த வெளிநாட்டவர் ஒருவர் கைது செய்யப்பட்டுள்ளார். </BODY>
</NEWS>

Further, the system was not able to cluster the few related articles. For example, it was unable to identify the similarity of the following news items since they are written in two different ways. All these news articles are talking about a criminal who was hanged for sexually abusing a young girl.

- பாகிஸ்தானில் சிறுமி பாலியல் துஷ்பிரயோகம்: குற்றவாளி தூக்கிலிடப்பட்டார் (young girl Sexual Abuse in Pakistan: Criminal Hanged)
- பாகிஸ்தான் சிறுமி வல்லுறவு - கொலை: தூக்கிலிடப்பட்டார் குற்றம்சாட்டப்பட்டவர் (Pakistan young Girl Rape - Murder: offender hanged)

- 9 சிறுமிகளை கற்பழித்து கொன்ற இம்ரான் அலி தூக்கிலிடப்பட்டான் (Imran Ali hanged for raping 9 young girls)
- பாக்கிஸ்தானில் சிறுமி படுகொலை குற்றவாளிக்கு தூக்குதண்டனை நிறைவேற்றம் (Execution of the criminal of a minor murder in Pakistan)

Further, an interesting thing we can observe is when we apply word embedding to group some news articles which were written in different styles by different news providers. For example, in the first case இப்ராஹிம் மொஹம்மட் சொலி and இப்ராகிம் மொகமது சாலிக் refer to Ibrahim Mohamed Solih, which can be grouped together when we use word embedding and it failed to group together when we apply TF-IDF.

<NEWS>
<CLUSTER> 37 </CLUSTER>
<TITLE> மாலைத்தீவின் 7ஆவது ஜனாதிபதியாக இப்ராஹிம் மொஹம்மட் சொலி பதவியேற்பு </TITLE>
<BODY>மாலைத்தீவின் 7ஆவது ஜனாதிபதியாக இப்ராஹிம் மொஹம்மட் சொலி (Ibrahim Mohamed Solih) இன்று பதவிப்பிரமாணம் செய்து கொண்டார் . </BODY>
</NEWS>

<NEWS>
<CLUSTER> 37 </CLUSTER>
<TITLE>மாலைதீவின் புதிய ஜனாதிபதியானார் இப்ராகிம் மொகமது சாலிக் </TITLE>
<BODY> மாலைதீவில் அன்மையில் இடம்பெற்ற ஜனாதிபதி தேரிதலில். எதிர்க்கட்சிகள் சார்பில் போட்டியிட்ட இப்ராகிம் முகமது சாலிக் வெற்றி </BODY>
</NEWS>

In the below example, Basil Rajapaksa is written in different ways like பசில், பெஸில் by different news providers. This is addressed successfully by our approach.

<NEWS>
<CLUSTER> 92 </CLUSTER>
<TITLE> ஜனாதிபதியும் பசிலும் ஒருமணிநேரம் பேச்சு </TITLE>

<BODY> ஜனாதிபதி மைத்திரிபால சிறிசேனவும் , முன்னாள் அமைச்சர் பசில் ராஜபக்ஷவும் , நேற்றிரவு </BODY>
</NEWS>

<NEWS>
<CLUSTER> 92 </CLUSTER>
<TITLE> ஜனாதிபதி - பஷிலுக்கிடையில் முக்கிய பேச்சு </TITLE>
<BODY> ஜனாதிபதி மைத்திரிபால சிறிசேன மற்றும் முன்னாள் பொருளாதார அமைச்சர் பசில் ராஜபக்ஷவுக்கிடையில் முக்கிய சந்திப்பொன்று </BODY>
</NEWS>

<NEWS>
<CLUSTER> 92 </CLUSTER>
<TITLE>ஜனாதிபதி - பசில் ராஜபக்ஸ இடையில் பேச்சுவார்த்தை</TITLE>
<BODY> ஜனாதிபதி மைத்திரிபால சிறிசேனவிற்கும் பசில் ராஜபக்ஸவிற்கு இடையில் பேச்சுவார்த்தை நடைபெற்றுள்ளது. </BODY>
</NEWS>

<NEWS>
<CLUSTER> 92 </CLUSTER>
<TITLE>ஜனாதிபதி மற்றும் முன்னாள் அமைச்சர் பெசிலும் சந்திப்பு </TITLE>
<BODY>ஜனாதிபதி மைத்ரிபால சிறிசேனவுக்கும் முன்னாள் அமைச்சர் பெஸில் ராஜபக்ஷவுக்கும் இடையிலான சந்திப்பு ஒன்று நேற்றிரவு இடம்பெற்றது. </BODY>
</NEWS>

Another interesting thing we observed when applying word embedding is the ability to cluster news with inflated words. Unfortunately, when applying the TF-IDF it failed to group the news with inflated words. For example, in the following case,

the word நாடு is written in two different forms like நாடு, and நாட்டில்.

The word மழை can be written in three different form such as மழையுடன், மழை and மழையுடனான

நாடு முழுவதும் மழையுடன் கூடிய வானிலை நிலைமை அதிகரிக்கும்
இன்று மழை அல்லது இடியுடன் கூடிய மழை
நாட்டில் மழையுடனான வானிலை நீடிப்பதற்கான சாத்தியம்

Problems:

There is no hard and fast rule to categorize similar articles and it is not possible to define such a rule as well. There are disagreements even between the manual evaluators in defining the clusters. Therefore, defining a golden stand for a data set also became challenging and difficult.

For example, two human evaluators have grouped the three news articles below into one cluster, while the third one has grouped the first two into one cluster and the third one into a different cluster. So for the golden stand set we agree to put all three into one cluster.

- பொட்டு அம்மான் உயிருடன் இருக்கிறாரா? (Is Pottu Amman alive?)
- பொட்டு அம்மான் உயிருடன் இருக்கிறாரா? என்ன சொல்கிறார்கள் முன்னாள் போராளிகள்? (Is Pottu Amman alive? What are the ex- fighters saying?)
- கருணா அம்மானின் கருத்து உண்மைக்கு புறம்பானது (Karuna Amman's comment is untrue)

Sometimes news providers report the news in an indirect way. For example, in this case news is related to weather change in the country but it has mention as எச்சரிக்கை..!! இன்று பிற்பகல் நாட்டில் ஏற்படவுள்ள மாற்றம்!! (Warning .. !! This afternoon the change going to happen in the country!!).

News reporting include poetical reporting styles in some cases. For example " விலை கொடு, செவி மடு ": டெல்லியை உலுக்கிய இந்திய விவசாயிகளின் போராட்டம் (Pay the price, listen: Indian farmers struggle that rocked Delhi)

Some news providers use English words in the middle of the news. Following are some examples,

Body Building Physique and Junior Championship 2018 நிகழ்வுக்கு RS Steel இணை அனுசரணை (RS Steel sponsoring the event Body Building Physique and Junior Championship 2018)

“Anchor Students with Talent” வட மாகாண மாணவர்களின் திறமைக்கு மகுடம் சூட்டிய பிரம்மாண்ட பயணம் (“Anchor Students with Talent”)

Another challenge is handling the synonyms. Tamil is a language that has a large volume of synonyms. This makes it difficult to identify the same news with different alternative words. To overcome the synonyms issue, some research based on English and some other languages used external resources like WordNet [12]. For example, the following four news articles use three different words (பாலியல் துஷ்பிரயோகம், வல்லுறவு, கற்பழித்து) to say abuse

- பாகிஸ்தானில் சிறுமி பாலியல் துஷ்பிரயோகம்: குற்றவாளி தூக்கிலிடப்பட்டார் (young girl Sexual Abuse in Pakistan: Criminal Hanged)
- பாகிஸ்தான் சிறுமி வல்லுறவு - கொலை: தூக்கிலிடப்பட்டார் குற்றம்சாட்டப்பட்டவர் (Pakistan young Girl Rape - Murder: offender hanged)
- 9 சிறுமிகளை கற்பழித்து கொன்ற இம்ரான் அலி தூக்கிலிடப்பட்டான் (Imran Ali hanged for raping 9 young girls)

Sometime a news article covers multiple instances. For example, the following case reporting two different firing instances. First incident happened in kuda Oya and second incident happened in the Gevithupura. So, this news can be clustered into two different groups in the ideal situation if assignment to more than one cluster was allowed.

<TITLE>இருவேறு துப்பாக்கிச்சூட்டு சம்பவங்களில் இருவர் பலி </TITLE>
<BODY> குடாஓய மற்றும் கொவிதுபுர பிரதேசத்தில் இடம்பெற்ற இருவேறு துப்பாக்கிச்சூட்டு சம்பவங்களில் இருவர் உயிரிழந்துள்ளனர். கொவிதுபுர - உனானயாய போவல பிரதேசத்திலுள்ள வீட்டுக்கு முன்னால் ஒருவர் சுட்டுக்கொலைசெய்யப்பட்டுள்ளார். இது தொடர்பில் கிடைக்கப்பெற்ற தகவலுக்கமைய விசாரணைகள் முன்னெடுக்கப்பட்டுள்ளதாக பொலிஸார் தெரிவித்துள்ளனர். இவ்வாறு உயிரிழந்தவர் சியம்பலாண்டுவ பகுதியை சேர்ந்த நபரென தெரியவந்துள்ளத. </BODY>

CHAPTER 5 – CONCLUSION AND FUTURE WORK

5.1 Conclusion

News aggregators support the readers to view multiple news providers' news in a single point. At the moment, the only news aggregator supporting Tamil news is Google news, and it has some noticeable shortages. Therefore, this research focused on creating a news aggregator for the Tamil language.

In this study, TF-IDF and word embedding document-representing techniques were experimented with the one pass clustering algorithm and the affinity clustering algorithm. TF-IDF with one pass algorithm was considered as the baseline system. Word embedding with one pass clustering algorithm gave the best result. Further, we applied all these approaches only to the news title to group the news. That did not show any positive impact on the result. From this study we were able to build a system to collect news from online sources and group them according to the content similarity. Further we have created a word embedding model with 21 077 843 words. From our study, we have observed that word embedding outperforms the TF-IDF for both clustering algorithms and that word embedding is able to identify the word written in different styles by different publishers. Also, word embedding can handle the inflected words, whereas TF-IDF fails. One pass clustering algorithm outperforms the affinity propagation algorithm in both document representing approaches. It may be due to having a high number of single article clusters in the dataset.

5.2 Future Work

There are number of possible ways to improve the current system performance.

- Evaluate the effectiveness by experimenting with stop words removal
- Use document-level embeddings such as Doc2Vec to represent the news articles and can evaluate the effectiveness of the system

- Use of different weights for title and body

REFERENCES

[1] E. Ulken ‘A Question of Balance: Are Google News Search Results Politically Biased?’, *USC Annenberg School for Communication*. Available: http://ulken.com/thesis/googlenews-bias-study.pdf, 2005.

[2] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, B. Palaniappan “Automatic classification of Tamil documents using vector space model and artificial neural network” *Expert Systems with Applications: Elsevier*, Vol. 36, pp.10914–10918, 2009.

[3] V.G.S. Sharma, "Recent Developments in Text Clustering Techniques," *International Journal of Computer Applications*, vol. 37, no. 6, pp. 14-19, 2012.

[4] P. Nanayakkara and S. Ranathunga, “Clustering sinhala news articles using corpus-based similarity measures,” in *2018 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2018, pp. 437– 442.

[5] T. K. Landauer, P. W. Foltz and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259-284, 1998.

[6] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *lcml*, vol. 97, pp. 412-420, 1997.

[7] A. Karima, E. Zakaria, T. G. Yamina, "Arabic text categorization: A comparative study of different representation modes", *Journal of Theoretical and Applied Information Technology*, vol. 38, no. 1, pp. 1-5, 2012.

[8] Z Elberrichi. ,A. Rahmoun, and M. Bentaalah, "Using WordNet for Text Categorization", *International Arab Journal of Information Technology(IAJIT)*, 5(1): 16-24, 2008.

[9] H. Schütze, D. Hull, and A Pedersen, “A Comparison of Classifiers and Document Representations for the Routing Problem,” in *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, New York, pp. 229-237, 1995.

[10] A. Amine, Z. Elberrichi and M. Simonet, "Evaluation of Text Clustering Methods Using WordNet," *The International Arab Journal of Information Technology*, vol. 7, no. 4, pp. 349-357, 2010.

[11] J Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proc..of the eleventh international conference on Information and knowledge management*, McLean, Virginia, USA, 2002, pp. 515-524

[12] C. Bouras and V. Tsogkas, "A clustering technique for news articles using WordNet," *Knowledge-Based Systems*, vol. 36, pp. 115-128, 2012.

[13] Marcus Lönnberg, Love Yregård, "Large scale news article clustering", *M.S. thesis, Dept. of Comput. Sci. and Eng., Chalmers University of Technology, Gothenburg*, Sweden, 2013

[14] D. Shen, J. Liu J, C. Nicholas and E. Miller, "Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System," *Computer Journal of Digital Information*, vol. 1, no. 5, pp. 257-265, 1999.

[15] T Hofmann, "Probmap: A Probabilistic Approach for Mapping Large Document Collections", *Journal for Intelligent Data Analysis*, vol. 4, pp. 149-164, 2000.

[16] A. Huang, "Similarity measures for text document clustering," in *Proc. New Zealand Computer Science Research Student Conf. (NZCSRSC 2008)*, Christchurch, New Zealand, pp. 49-56, 2008.

[17] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, 2013.

[18] A. Hotho, A. Maedche and S. Staab, "Ontology-based text document clustering," *KI*, vol. 16, no. 4, pp. 48-54, 2002.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *In Proceedings of ICLR Workshops Track,* 2013.

[20] T. Mikolov, I Sutskever, K. Chen, G. Corrado and J. Dean.Distributed representations of words and phrases and their compositionality. *In Proc. Advances in Neural Information Processing Systems* 26 3111–3119 (2013).

[21] Y. Goldberg, O. Levy. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method, 2014.

[22] R. Nalawade, A. Samal and K. Avhad, "Improved Similarity Measure For Text Classification And Clustering," *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, no. 5, pp. 214-219, 2016.

[23] M.Thangarasu and Dr.R.Manavalan, "Stemmers for Tamil Language: Performance Analyses," *International Journal for Computer Science & Engineering Technology*, Vol. 4 No. 07 Jul 2013

[24] S. Bhatia, "Adaptive K-Means Clustering," in *FLAIRS Conference*, 2004. pp. 695-699

[25] N. Friburger and D. Maurel, "Textual similarity based on proper names," *in Proc. of the workshop Mathematical/Formal Methods in Information Retrieval*, 2002, pp.155-167.

[26] G. Furnas and W. Jones, "Pictures of Relevance: A Geometric Analysis of Similarity Measures," *Computer Journal of the American Society for Information Science*, vol. 38, no. 6, pp. 420-442, 1987.

[27] M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques," *KDD workshop on text mining*, vol. 400, no. 1, pp. 525-526, 2000.

[28] Rai, P., and Singh, S. "A Survey of Clustering Techniques," *International Journal of Computer Applications*, 7(12), 1-5, (2010, October).

[29] A. Moore and D. Pelleg, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in *Proc. of the Seventeenth International Conference on Machine Learning*, pp. 727-734, 2000.

[30] Y. He, Q. Chen, X. Wang, R. Xu, X. Bai and X. Meng, "An adaptive affinity propagation document clustering," *The 7th International Conference on Informatics and Systems (INFOS)*, Cairo, 2010, pp. 1-7, 2010.

[31] Shailendra Kumar Shrivastava, J L Rana and R C Jain. Article: "Text Document Clustering based on Phrase Similarity using Affinity Propagation". *International Journal of Computer Applications* 61(18):38-44, January 2013.

[32] G. Leban, B. Fortuna and M. Grobelnik, "Using News Articles for Real-time Cross-Lingual Event Detection and Filtering," *in Proc. of the NewsIR'16 Workshop at ECIR*, Padua, Italy, 2016. pp. 33-38

[33] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651-666, 2010.

[34] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130-137, 1980.

[35] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[36] David D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, p.37-50, June 21-24, 1992, Copenhagen, Denmark

[37] G. Salton , A. Wong , C. S. Yang, "A vector space model for automatic indexing", *Communications of the ACM*, v.18 n.11, p.613-620, Nov. 1975

[38] M.Thangarasu and Dr.R.Manavalan, "Design and Development of Stemmer for Tamil Language: Cluster Analysis," *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3 No. 07, 2013

[39] D. Pfitzner, R. Leibbrandt and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," *Knowledge and Information Systems*, vol. 19, no. 3, pp. 361-394, 2009.

[40] Farook Fathima Farhath, " Sinhala - Tamil Statistical Machine Translation (SMT) for Official Documents," MPhil. thesis, Dept. of Comput. Sci. and Eng., Univ. of Moratuwa, Sri Lanka, 2018.

[41] D. A. Meedeniya and A. S. Perera, "A Comparative Study on Data Representation to Categorize Text Documents," in *Twentieth International Conference on Software Engineering and Knowledge Engineering*,2008.

[42] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

APPENDIX A: TAMIL SENTENCES WITH TRANSLITERATION AND TRANSLATION

விசேட கலந்துரையாடல் - viceta kalanturaiyatal - Special Discussion

தீர்ப்பு வெளியானது - tirppu veliyanatu - The judgment was released

ஏற்க தயார் - erka tayar - Ready to accept

60 பேர் பலி - 60 per pali - 60 killed

UPDATE: YOUTUBE சேவை வழமைக்குத் திரும்பியது - UPDATE: YOUTUBE cevai valamaikkut tirumpiyatu - YouTube services back to normal

உலகின் முதல்நிலை இணையத்தளமான யூடியூப்பின் சேவை மீண்டும் வழமைக்குத் திரும்பியுள்ளது - ulakin mutalnilai inaiyattalamana yutiyuppin cevai mintum valamaikkut tirumpiyullatu - The world's premier internet site, YouTube’s service back to normal

உலகையே பரபரப்பாக்கியுள்ள விடயம்..!! - ulakaiye paraparappakkiyulla vitayam..!! - The thing which makes world excited

உலகின் பிரபல சமூக வலைத்தளமான youtube தற்போது செயலிழந்துள்ளதாக தெரிவிக்கப்பட்டுள்ளது - ulakin pirapala camuka valaittalamana youtube tarpotu ceyalilantullataka terivikkappattullatu - It is reported that YouTube, the world's most popular social website, is currently defunct

ஆற்றிலிருந்து ஆண் ஒருவரின் சடலம் மீட்பு - arriliruntu an oruvarin catalam mitpu - A male corpse recovery from a river

திம்புள்ள, பத்தனை பொலிஸ் பிரிவிற்குட்பட்ட குயின்ஸ்பெரி கீழ்பிரிவில் உள்ள கிளை ஆற்றிலிருந்து ஆண் ஒருவரின் சடலம் இன்று (07) காலை மீட்கப்பட்டுள்ளது - timpulla, pattanai polis pirivirkutpatta kuyinsperi kilpirivil ulla kilai arriliruntu an oruvarin catalam inru (07) kalai mitkappattullatu. - Body of man rescued from a branch river in the kuinsberry down division which is under coming under Timbulle, Paththana police division

சடலம் ஒன்று மீட்பு - catalam onru mitpu - A corpse recovery

பலாங்கொடை - பின்னவல - அமுபிட்டிய பிரதேசத்தில் உள்ள வனப்பகுதி ஒன்றில் இருந்து அடையாளம் காணப்படாத சடலம் ஒன்று மீட்கப்பட்டுள்ளது - palankotai - pinnavala - amupittiya piratecattil ulla vanappakuti onril iruntu ataiyalam kanappatata catalam onru mitkappattullatu. - Unidentified body recovered from a forest area in Palangoda – Pinnawala – Ammutiya

மர்மமான முறையில் உயிரிழந்த ஆணின் சடலம் மீட்பு - marmamana muraiyil uyirilanta anin catalam mitpu -Recovery of the corpse of a mysteriously dead man

புத்தளம் பொலிஸ் பிரிவுக்குட்பட்ட கல்லடி தம்மன்னா கிராமத்தில் மர்மமான முறையில் உயிரிழந்த ஒருவரின் சடலத்தை இன்று - puttalam polis pirivukkutpatta kallati tam'manna kiramattil marmamana muraiyil uyirilanta oruvarin catalattai inru - Puttalam police division Kallady tammanna village, the corpse of a person who died in mysterious circumstances Today

இரு பிள்ளைகளின் தந்தை சடலமாக மீட்பு - iru pillaikalin tantai catalamaka mitpu - A father of two children recover as corpse

பத்தனை பொலிஸ் பிரிவுக்குட்பட்ட குயின்ஸ் பெரி தோட்டத்திலுள்ள நீரோடையிலிருந்து ஆண் ஒருவரின் சடலம் ஒன்று இன்று - pattanai polis

pirivukkutpatta kuyins peri tottattilulla nirotaiyiliruntu an oruvarin catalam onru inru - One man's body from a stream in the kuin bery garden Paththane police division today

சடலம் ஒன்று மீட்பு - catalam onru mitpu - A corpse recovery

பாகிஸ்தானில் சிறுமி பாலியல் துஷ்பிரயோகம்: குற்றவாளி தூக்கிலிடப்பட்டார் - pakistanil cirumi paliyal tuspirayokam : kurravali tukkilitappattar - young girl Sexual Abuse in Pakistan: Criminal Hanged

பாகிஸ்தானில் இவ்வருடம் ஜனவரி மாதத்தில் 6 வயது சிறுமி பாலியல் துஷ்பிரயோகம் செய்யப்பட்டு கொல்லப்பட்ட வழக்கில் நீதிமன்றத்தால் தூக்குத்தண்டனை விதிக்கப்பட்ட நபர் இன்று (17) தூக்கிலிடப்பட்டார் - pakistanil ivvarutam janavari matattil 6 vayatu cirumi paliyal tuspirayokam ceyyappattu kollappatta valakkil nitimanrattal tukkuttantanai vitikkappatta napar inru (17) tukkilitappattar - Man sentenced to death for rape for murder of a 6-year-old girl this year January 6 in Pakistan

பாகிஸ்தான் சிறுமி வல்லுறவு - கொலை: தூக்கிலிடப்பட்டார் குற்றம்சாட்டப்பட்டவர் - pakistan cirumi valluravu - kolai: tukkilitappattar kurramcattappattavar - Pakistan young Girl Rape - Murder: offender hanged

இம்ரான் அலி தூக்கிலிடப்படும் காட்சியை நேரில் பார்த்ததாகக் கூறிய ஜைனபின் தந்தை அமீன் அன்சாரி , அந்த நிகழ்ச்சியை தொலைக்காட்சியில் ஒளிபரபப்படவில்லை என்று வருத்தம் தெரிவித்தார் - imran ali tukkilitappatum katciyai neril parttatakak kuriya jainapin tantai amin ancari , anta nikalcciyai tolaikkatciyil oliparapappatavillai enru varuttam terivittar . - Zainab's father, Amin Ansari, who witnessed the scene of Imran Ali's execution, expressed regret that the show was not broadcast on television.

9 சிறுமிகளை கற்பழித்து கொன்ற இம்ரான் அலி தூக்கிலிடப்பட்டான் - 9 cirumikalai karpalittu konra imran ali tukkilitappattan - Imran Ali hanged for raping 9 young girls

பாகிஸ்தானில் லாகூர் பகுதியை சேர்ந்தவன் இம்ரான் அலி (30) . இவன் 9 சிறுமிகளை கற்பழித்தான். அவர்களில் 7 வயது சிறுமியை கற்பழித்து கொன்று அவளது உடலை - pakistanil lakur pakutiyai cerntavan imran ali (30) . ivan 9 cirumikalai karpalittan. avarkalil 7 vayatu cirumiyai karpalittu konru avalatu utalai - Imran Ali (30) hails from Lahore, Pakistan. He raped 9 little girls. 7-year-old girl raped and killed her body

பாக்கிஸ்தானில் சிறுமி படுகொலை குற்றவாளிக்கு தூக்குதண்டனை நிறைவேற்றம் - pakkistanil cirumi patukolai kurravalikku tukkutantanai niraiverram - Execution of the criminal of a minor murder in Pakistan

எனது மகளின் மரணத்திற்கு காரணமானவரிற்கு மரணதண்டனை நிறைவேற்றப்பட்டதை தொலைக்காட்சியில் காண்பிக்கவில்லை என்பதே எனது - enatu makalin maranattirku karanamanavarirku maranatantanai niraiverrappattatai tolaikkatciyil kanpikkavillai enpate enatu - The fact that the execution of the person responsible for my daughter's death is not shown on television

மாலைத்தீவின் 7ஆவது ஜனாதிபதியாக இப்ராஹிம் மொஹம்மட் சொலி பதவியேற்பு malaittivin 7avatu janatipatiyaka iprahim moham'mat coli pataviyerpu - Ibrahim Mohammed Soli sworn in as Maldives' 7th president

மாலைத்தீவின் 7ஆவது ஜனாதிபதியாக இப்ராஹிம் மொஹம்மட் சொலி (Ibrahim Mohamed Solih) இன்று பதவிப்பிரமாணம் செய்து கொண்டார் - malaittivin 7avatu janatipatiyaka iprahim moham'mat coli (Ibrahim Mohamed Solih) inru patavippiramanam ceytu kontar - Ibrahim Mohamed Solih takes oath as the 7th President of the Maldives

மாலைதீவின் புதிய ஜனாதிபதியானார் இப்ராகிம் மொகமது சாலிக் - malaitivin putiya janatipatiyanar iprakim mokamatu calik - The new President of the Maldives is Ibrahim Mohammed Saliq

மாலைதீவில் அன்மையில் இடம்பெற்ற ஜனாதிபதி தேரிதலில். எதிர்க்கட்சிகள் சார்பில் போட்டியிட்ட இப்ராகிம் முகமது சாலிக் வெற்றி - malaitivil anmaiyil itamperra janatipati teritalil. etirkkatcikal carpil pottiyitta iprakim mukamatu calik verri - In the recent presidential election in the Maldives. Ibrahim Mohammed Salik wins on behalf of opposition parties

ஜனாதிபதியும் பசிலும் ஒருமணிநேரம் பேச்சு - janatipatiyum pacilum orumanineram peccu - President and Basil talk for one hour

ஜனாதிபதி மைத்திரிபால சிறிசேனவும் , முன்னாள் அமைச்சர் பசில் ராஜபக்ஷவும் , நேற்றிரவு janatipati maittiripala ciricenavum , munnal amaiccar pacil rajapaksavum , nerriravu - President Maithripala Sirisena and former Minister Basil Rajapaksa yesterday night

ஜனாதிபதி - பஷிலுக்கிடையில் முக்கிய பேச்சு - janatipati - pasilukkitaiyil mukkiya peccu -Key talk between the President and Bashil

ஜனாதிபதி மைத்திரிபால சிறிசேன மற்றும் முன்னாள் பொருளாதார அமைச்சர் பசில் ராஜபக்ஷவுக்கிடையில் முக்கிய சந்திப்பொன்று - janatipati maittiripala ciricena marrum munnal porulatara amaiccar pacil rajapaksavukkitaiyil mukkiya cantipponru - A key meeting between President Maithripala Sirisena and former Economic Minister Basil Rajapaksa

ஜனாதிபதி - பசில் ராஜபக்ஸ இடையில் பேச்சுவார்த்தை janatipati - pacil rajapaksa itaiyil peccuvarttai - Talks between President and Basil Rajapaksa

ஜனாதிபதி மைத்திரிபால சிறிசேனவிற்கும் பசில் ராஜபக்ஸவிற்கு இடையில் பேச்சுவார்த்தை நடைபெற்றுள்ளது. - janatipati maittiripala ciricenavirkum pacil rajapaksavirku itaiyil peccuvarttai nataiperrullatu. - Talks between President Maithripala Sirisena and Basil Rajapakse

ஜனாதிபதி மற்றும் முன்னாள் அமைச்சர் பெசிலும் சந்திப்பு - janatipati marrum munnal amaiccar pecilum cantippu - President and Former Minister Basil Meets

ஜனாதிபதி மைத்ரிபால சிறிசேனவுக்கும் முன்னாள் அமைச்சர் பெஸில் ராஜபக்ஷவுக்கும் இடையிலான சந்திப்பு ஒன்று நேற்றிரவு இடம்பெற்றது. - janatipati maitripala ciricenavukkum munnal amaiccar pesil rajapaksavukkum itaiyilana cantippu onru nerriravu itamperratu. - A meeting between President Maithripala Sirisena and former Minister Basil Rajapaksa took place last night.

நாடு முழுவதும் மழையுடன் கூடிய வானிலை நிலைமை அதிகரிக்கும் - natu muluvatum malaiyutan kutiya vanilai nilaimai atikarikkum - The weather conditions will increase with rainfall across the country

இன்று மழை அல்லது இடியுடன் கூடிய மழை - inru malai allatu itiyutan kutiya malai - Rain or thunderstorm today

நாட்டில் மழையுடனான வானிலை நீடிப்பதற்கான சாத்தியம் - nattil malaiyutanana vanilai nitippatarkana cattiyam - Possibility of prolonged rainy weather in the country

பொட்டு அம்மான் உயிருடன் இருக்கிறாரா? - pottu am'man uyirutan irukkirara? -Is Pottu Amman alive?

பொட்டு அம்மான் உயிருடன் இருக்கிறாரா? என்ன சொல்கிறார்கள் முன்னாள் போராளிகள்? -pottu am'man uyirutan irukkirara? enna colkirarkal munnal poralikal? - Is Pottu Amman alive? What saying ex- fighters?

கருணா அம்மானின் கருத்து உண்மைக்கு புறம்பானது - karuna am'manin karuttu unmaikku purampanatu - Karuna Amman's comment is untrue

எச்சரிக்கை..!! இன்று பிற்பகல் நாட்டில் ஏற்படவுள்ள மாற்றம்!! - eccarikkai..!! inru pirpakal nattil erpatavulla marram!! - Warning .. !! This afternoon the change going to happen in the country!!.

விலை கொடு, செவி மடு ": டெல்லியை உலுக்கிய இந்திய விவசாயிகளின் போராட்டம் " vilai kotu, cevi matu ": telliyai ulukkiya intiya vivacayikalin porattam - Pay the price, listen: Indian farmers struggle that rocked Delhi

பாலியல் துஷ்பிரயோகம் - paliyal tuspirayokam - Sexual abuse

வல்லுறவு - valluravu - Rape

கற்பழித்து - karpalittu – Rape

இருவேறு துப்பாக்கிச்சூட்டு சம்பவங்களில் இருவர் பலி - iruveru tuppakkiccuttu campavankalil iruvar pali - Two victims by two different shootings

குடாஓய மற்றும் கொவிதுபுர பிரதேசத்தில் இடம்பெற்ற இருவேறு துப்பாக்கிச்சூட்டு சம்பவங்களில் இருவர் உயிரிழந்துள்ளனர். கொவிதுபுர - உனானயாய போவல பிரதேசத்திலுள்ள வீட்டுக்கு முன்னால் ஒருவர் சுட்டுக்கொலைசெய்யப்பட்டுள்ளார். இது தொடர்பில் கிடைக்கப்பெற்ற தகவலுக்கமைய விசாரணைகள் முன்னெடுக்கப்பட்டுள்ளதாக பொலிஸார் தெரிவித்துள்ளனர். இவ்வாறு உயிரிழந்தவர் சியம்பலாண்டுவ பகுதியை சேர்ந்த நபரென தெரியவந்துள்ளத. - kuta'oya marrum kovitupura piratecattil itamperra iruveru tuppakkiccuttu campavankalil iruvar uyirilantullanar. kovitupura - unanayaya povala piratecattilulla vittukku munnal oruvar

cuttukkolaiceyyappattullar. itu totarpil kitaikkapperra takavalukkamaiya vicaranaikal munnetukkappattullataka polisar terivittullanar. ivvaru uyirilantavar ciyampalantuva pakutiyai cernta naparena teriyavantullata. - Two people were killed in two separate shootings in Kudaoya and Kovitupura. One person shot dead in front of house in Kovitupura unnaya. Police said investigations are being carried out according to information available. The victim has been identified as a resident of Siyambalanduwa.