**CLUSTERING NEWS ARTICLES USING K-MEANS AND N-GRAMS**

**BY**

**DESMOND BALA BISANDU (A00019335)**



# SCHOOL OF IT & COMPUTING, AMERICAN UNIVERSITY OF NIGERIA, YOLA, ADAMAWA STATE, NIGERIA

**SPRING, 2018**

**CLUSTERING NEWS ARTICLES USING K-MEANS AND N-GRAMS**

**BY**

**DESMOND BALA BISANDU (A00019335)**



# In partial fulfillment of the requirements for the award of degree of Master of Science (M.Sc.) in Computer Science submitted to the School of IT & Computing, American University of Nigeria, Yola

**SPRING, 2018**

# DECLARATION

I, Desmond Bala BISANDU, declare that the work presented in this thesis entitled ***‘Clustering News Articles Using K-means and N-grams’*** submitted to the School of IT & Computing, American University of Nigeria, in partial fulfillment of the requirements for the award of the Master of Science (M.Sc.) in Computer Science. I have neither plagiarized nor submitted the same work for the award of any other degree. In case this undertaking is found incorrect, my degree may be withdrawn unconditionally by the University.

Date: 16th April, 2018 Desmond Bala Bisandu

Place: Yola A00019335

# CERTIFICATION

I certify that the work in this document has not been previously submitted for a degree nor neither has it been submitted as a part of a requirements for a degree except fully acknowledged within this text.

Student Date

(Desmond Bala Bisandu)

Supervisor Date

(Dr. Rajesh Prasad)

Program Chair Date

(Dr. Narasimha Rao Vajjhala)

Graduate Coordinator, SITC Date

(Dr. Rajesh Prasad)

Dean, SITC Date

(Dr. Mathias Fonkam)

# ACKNOWLEDGEMENT

My gratitude goes to the Almighty God for giving me the strength, knowledge, and courage to carry out this thesis successfully. My profound gratitude goes to my supervisor Dr. Rajesh Prasad for his patience, advice and time to guide me throughout the thesis work, Dr. Mathias Fonkam, Dean SITC for his guide on the thesis, Dr. Salu George Thandekkattu and Dr. Charles Nche for their encouragements, and the entire staff of School of IT & Computing, American University of Nigeria (AUN), Yola, Nigeria.

My sincere appreciation goes to the entire Management of AUN especially faculty member and chair of Computer Science Department Dr. Rao Narasimha Vajjhala and library officials for the kind support through making the AUN environment a learning one.

Also, I have to seize this opportunity to acknowledge the endless effort of my parents Mr. & Mrs. Bala G. M. BISANDU, my friend Ms. Priscilla Tumagodi Adabo for always encouraging me to keep pushing on, and the entire family, including my brothers, sisters, aunties, and uncles for their cooperation and tireless concern for me all through.

Finally, I say a big thanks to my friends and course mates. I wish you success in all your future undertakings. I am most sincerely grateful to all that have contributed in one way or the other to the success of this thesis.

# ABSTRACT

Document clustering is an automatic unsupervised machine learning technique that aimed at grouping related set of items into clusters or subsets. The target is creating clusters with high internal coherence, but different from each other substantially. Simply, items within the same cluster should be highly similar, while maintaining high dissimilarity with items within other clusters. Automatic clustering of documents has played a very significant role in many fields including data mining and information retrieval. This thesis aimed to improve the overall efficiency of a document clustering technique using N-grams and efficient similarity measure. The thesis improves the purity and accuracy of the obtained clusters. The preprocessing method is based on N-grams (sequence of N consecutive characters) which do not give consideration to stop-words or other special punctuations but creates and overlap among the content of a document which further gives room to ignore errors thereby increasing the quality of the clusters to a great extent. This approach clusters the news articles based on their N-grams representation, thereby reducing noise and increase the probability of occurrences of the sequences within the articles document. The proposed clustering technique has parameters which can be changed accordingly at the document representation level in order to improve the efficiency and quality of the generated clusters. The results from the experiment using R programming environment were carried out on real datasets of the *Reuters21578* and *20Newsgropus* proved the effectiveness of the proposed clustering technique at different levels of N-grams in terms of the accuracy and purity of the generated clusters. The results also showed that the proposed clustering technique perform averagely better than the baseline technique both in terms of accuracy and purity with a best results when the window of N-grams = 3.

# TABLE OF CONTENTS

Cover Page i

Title Page ii

[Declaration iii](#_bookmark0)

[Certification iv](#_bookmark1)

[Acknowledgement v](#_bookmark2)

[Abstract vi](#_bookmark3)

[Table Of Contents vii](#_bookmark4)

[List Of Tables xi](#_bookmark5)

[List Of Figures xii](#_bookmark6)

[List Of Abbreviations xiii](#_bookmark7)

[CHAPTER ONE 1](#_bookmark8)

[INTRODUCTION 1](#_bookmark9)

* 1. [Background Of The Study 1](#_bookmark10)
     1. [Motivation For The Study 2](#_bookmark11)
  2. [Data Mining 3](#_bookmark12)
  3. [Clustering 3](#_bookmark13)
  4. [Applications Of Clustering 5](#_bookmark14)
  5. [Problem Statement 6](#_bookmark15)
  6. [Aim And Objectives 7](#_bookmark16)
     1. [Aim 7](#_bookmark17)
     2. [Objectives 7](#_bookmark18)
  7. [Significance Of The Study 7](#_bookmark19)
  8. [Scope And Limitation Of The Study 8](#_bookmark20)
  9. [Thesis Organization 8](#_bookmark21)

[CHAPTER TWO 9](#_bookmark22)

[LITERATURE REVIEW 9](#_bookmark23)

* 1. [Basic Terminologies And Concepts 9](#_bookmark24)
  2. [Textual Document Representation Methods 10](#_bookmark25)
     1. [Word-Based Representation 10](#_bookmark26)
     2. [Term-Based Representation 11](#_bookmark27)
     3. [N-Grams-Based Representation 11](#_bookmark28)
  3. [Textual Document Pre-Processing Methods 12](#_bookmark29)
     1. [Tokenization 12](#_bookmark30)
     2. [Lowercase Conversion 12](#_bookmark31)
     3. [Punctuation Removal 13](#_bookmark32)
     4. [Term Filtering (Stop Words Removal) 13](#_bookmark33)
     5. [Alphanumeric And Short Length Words Removal 13](#_bookmark34)
     6. [Lemmatization 13](#_bookmark35)
     7. [Stemming 14](#_bookmark36)
  4. [Textual Document Similarity Measures 14](#_bookmark37)
     1. [Metric Conditions 14](#_bookmark38)
     2. [Euclidean Distance Measure 15](#_bookmark39)
     3. [Cosine Distance Measure 15](#_bookmark40)
     4. [Jaccard Distance Measure 16](#_bookmark41)
     5. [Manhattan Distance Measure 16](#_bookmark42)
     6. [Pearson Correlation Measure 17](#_bookmark43)
  5. [Categories Of Clustering Algorithms 17](#_bookmark44)
     1. [Hierarchical (Representative-Based) Clustering 18](#_bookmark45)
        1. [Agglomerative (Bottom-Up) Clustering 18](#_bookmark46)
        2. [Divisive (Top-Down) Clustering 18](#_bookmark47)
     2. [Partition-Based Clustering 19](#_bookmark48)
        1. [K-Means Algorithm 19](#_bookmark49)
        2. [K-Medoids Algorithm 19](#_bookmark50)
  6. [Review Of Literature 20](#_bookmark51)

[CHAPTER THREE 24](#_bookmark52)

[METHODOLOGY 24](#_bookmark53)

* 1. [Introduction 24](#_bookmark54)
  2. [Proposed Methodology (Clustering Technique) 24](#_bookmark55)
     1. [Example 3.1 25](#_bookmark56)
  3. [News Articles Pre-Processing 28](#_bookmark57)
  4. [News Articles N-Grams Representation 29](#_bookmark58)
     1. [Representation With 1-Gram 29](#_bookmark59)
     2. [Representation With N-Grams 30](#_bookmark60)
     3. [Weight Normalization 30](#_bookmark61)
  5. [Vector Space Model 31](#_bookmark62)
  6. [Dimensionality Reduction On Vector Features 31](#_bookmark63)
  7. [Improved Sqrt-Cosine Similarity Measure 31](#_bookmark64)
  8. [K-Means Clustering 32](#_bookmark65)
  9. [Evaluation Methods 33](#_bookmark66)

[CHAPTER FOUR 35](#_bookmark67)

[RESULTS AND DISCUSSIONS 35](#_bookmark68)

* 1. [Description Of Datasets 35](#_bookmark69)
  2. [Evaluation Of Experiments 35](#_bookmark70)
     1. [Experimental Set-Up 36](#_bookmark71)
     2. [Results 36](#_bookmark72)
        1. [Result Of Technique With N-Gram Equal To 1 (N=1) 37](#_bookmark73)
        2. [Result Of Technique With N-Grams Equal To 2 (N=2) 38](#_bookmark74)
        3. [Result Of Technique With N-Grams Equal To 3 (N=3) 38](#_bookmark75)
        4. [Result Of Technique With N-Grams Equal To 4 (N=4) 39](#_bookmark76)
        5. [Results Comparison Of Proposed Technique And Baseline Technique 42](#_bookmark77)
  3. [Our Contributions 43](#_bookmark78)

[CHAPTER FIVE 44](#_bookmark79)

[SUMMARY, CONCLUSION, AND FUTURE WORK 44](#_bookmark80)

* 1. [Summary 44](#_bookmark81)
  2. [Conclusion 45](#_bookmark82)
  3. [Future Work 45](#_bookmark83)

[List Of Publications 57](#_bookmark84)

[Appendix 58](#_bookmark85)

[Coding 58](#_bookmark86)

# LIST OF TABLES

Table 3.1: Normalized Weighted Frequencies of N-grams 26

Table 4.1: Real-world datasets summary ....................................**Error! Bookmark not defined.**4 Table 4.2: Average performance of proposed technique with N-gram=1 on the datasets **Error!**

### Bookmark not defined.

Table 4.3: Average performance of proposed technique with N-grams=2 on the datasets **Error!**

### Bookmark not defined.

Table 4.4: Average performance of proposed technique with N-grams=3 on the datasets **Error!**

### Bookmark not defined.

Table 4.5: Average performance of proposed technique with N-grams=4 on the datasets **Error!**

### Bookmark not defined.

Table 4.6: Average performance of the proposed technique as compared to baseline technique on the datasets **Error! Bookmark not defined.**

# LIST OF FIGURES

Figure 1.1: An Example of a Clear Clusters Structure Dataset … **Error! Bookmark not defined.**

Figure 2.1: Euclidean Distance Measure **Error! Bookmark not defined.**

Figure 2.2: Cosine Distance Measure **Error! Bookmark not defined.**

Figure 2.3: An Example of a Dendrogram (Hierarchical Clustering) 18

Figure 2.4: An Example of a Partition Clustering 19

Figure 3.1: Proposed Methodology (Clustering Technique) **Error! Bookmark not defined.**

Figure 3.2: Example Generated Clusters **Error! Bookmark not defined.**

Figure 4.1: Graph between within sum of squares and values of k to determine the best number of clusters **Error! Bookmark not defined.**

Figure 4.2a: Nature of clusters with N-gram=1 on Reuters21578 **Error! Bookmark not defined.**

|  |  |  |
| --- | --- | --- |
| Figure 4.2b: Nature of clusters with N-grams=1 on 20newsgroups **Error!** | **Bookmark** | **not** |
| **defined.**  Figure 4.3a: Nature of clusters with N-grams=2 on Reuters21578 **Error!** | **Bookmark** | **not** |
| **defined.**  Figure 4.3b: Nature of clusters with N-grams=2 on 20newsgroups **Error!** | **Bookmark** | **not** |
| **defined.**  Figure 4.4a: Nature of clusters with N-grams=3 on Reuters21578 **Error!** | **Bookmark** | **not** |
| **defined.**  Figure 4.4b: Nature of clusters with N-grams=3 on 20newsgroups **Error!** | **Bookmark** | **not** |
| **defined.**  Figure 4.5a: Nature of clusters with N-grams=4 on Reuters21578 **Error!** | **Bookmark** | **not** |
| **defined.** |  |  |

Figure 4.5b: Nature of clusters with N-grams=4 on 20newsgroups **Error! Bookmark not defined.**

Figure 4.6a: Bar chart showing Accuracy of proposed technique on different N-grams **Error!**

### Bookmark not defined.

Figure 4.6b: Bar chart showing Purity of proposed technique on different N-grams **Error!**

### Bookmark not defined.

Figure 4.7a: Line chart showing Accuracy of proposed technique on different N-grams **Error!**

### Bookmark not defined.

Figure 4.7b: Line chart showing Purity of proposed technique on different N-grams **Error!**

**Bookmark not defined.**

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| BoW | Bag of Words |
| CI | Clustering Index |
| CLARA | Clustering Large Applications |
| CLARANS | Clustering Large Applications based on  Randomized Search |
| CPU | Central Processing Unit |
| DF | Document Frequency |
| HITS | Hyperlink-Induced Topic Search |
| IDF | Inverse Document Frequency |
| ISC | Improved Square Root Cosine |
| LogTF | Logarithm of Term Frequency |
| ML | Machine Learning |
| N-GRAMS | Number of Grams |
| NLP | Natural Language Processing |
| POS | Part of Speech |
| TF | Term Frequency |
| VSM | Vector Space Model |
| WordNet | Word Network |
| WSS | Within Sum of Squares |
| XML | Extensible Markup Language |

**CHAPTER ONE** **INTRODUCTION**

This chapter introduced the background of the study, motivation of the study, introduction to data mining, clustering and its applications, problem statement, aim and objectives of the study, significance of the study, scope and limitation of the study, and finally the organization of the thesis.

# Background of the Study

The world we live in is full of data. Computers have been accepted as the best means of data storage. This is because of the fact that the data is saved very easily in the computer with high convenience, anybody that has access to a computer is able to do it, and more importantly, many users can share stored information, or send to different locations (Kriegel *et al.,* 2007). As the number of text documents stored in large databases increases, this poses a huge challenge of understanding hidden patterns or relationships in the data. Text data, being not in numerical format, can hardly be analyzed directly using statistical methods. Information overload or drowning in data is a common complaint by people as they see the potential value of information, yet are frustrated in their inability to derive benefit from it due to its volume and complexity (Sowjanya & Shashi, 2010; Han *et al*., 2012).

Due to the rapid growth of online news articles, journals, books, research papers, and web pages every day; the need on how to quickly find the most important, interesting, valuable, or entertaining items has arisen. This is because we are overwhelmed by the increasing volume of information made available online (Bouras & Tsogkas, 2016; Rupnik *et al.,* 2016). Humans throughout history have used information to achieve lots of great things such as predicting the future to avoid disaster and to make some vital decisions (Butler & Keselj, 2009; Jatowt & Au Yeung, 2011; Bouras & Tsogkas, 2016). The problem of overloading the Internet with this huge amount of information makes searching very tedious to the users, the enormous demands for techniques that will efficiently and effectively derive profitable knowledge from these diverse, unstructured information are highly required (Bouras & Tsogkas, 2013; Popovici *et al.,* 2014; Lwin & Aye, 2017).

One of the most important means to deal with data is classifying or grouping it into clusters or categories. Classification have played an important and an indispensable role throughout human history (Wu *et al.,* 2008). There exist two types of classification, the supervised and unsupervised. In the supervised classification, available predefined knowledge is needed, whereas in the unsupervised classification sometimes referred to as clustering or exploratory data analysis, no predefined labeled data is needed (Agrawal *et al.,* 1998; Tao *et al.,* 2004).

Grouping similar data such as news article based on their characteristics is an important issue. Grouping can be done on the basis of some similarity measures. Several similarity measures (such as Gauging, Jaccard, Euclidean, Edit, and Cosine) have been proposed and applied in computing the similarity between two different textual documents based on character matching, word semantics, and word sense (Damashek, 1995; Huang, 2008; Qiujun, 2010; Svadasa & Jhab, 2014; Akinwale & Niewiadomski, 2015; Sonia, 2016; Huang *et al.,* 2017). The rationale behind every given method of measuring the similarity between two textual documents is based on the increasing quest to improve the quality and the effectiveness of the existing clustering or filtering techniques (Shah & Mahajan, 2012; Sonia, 2016; Singh *et al.,* 2017).

# Motivation for the Study

This research is motivated by the fact that going through an online news portal, we observed the following challenges that needed efficient clustering technique:

* + - 1. The available news articles were large in number.
      2. The news articles were added online each and every day in the large number.
      3. Different sources used to contribute in adding news articles corresponding to the same news.
      4. Real-time update of recommendation had to be generated.

By using an efficient clustering technique the domain of search for recommendation could be reduced because most users are interested in the news that belong to some certain number of clusters. Time efficiency will be improved to an extent.

This research is motivated mainly by investigating the possibilities of improving the effectiveness of text document clustering techniques by pointing out reasons why the already existing techniques (algorithms) are ineffective and getting their solutions.

# Data Mining

Data mining is a field that deals with structured and unstructured type of data to derive knowledge or meaningful information which are previously unknown by using the machine learning algorithms (Shah & Mahajan, 2012; Svadasa & Jhab, 2014; Allahyari *et al.,* 2017). It has been applied in textual documents to predict and group related items in order to create a better and clearer understanding of such items (Grineva *et al.*, 2009; Rothe & Schütze, 2017; Wei *et al.,* 2017). The process of discovering, extracting nontrivial and interesting knowledge or patterns which are previously unknown from unstructured text document is referred to as *text mining* (Agrawal *et al.,* 1998; Singh *et al.,* 2017; Lin *et al.,* 2017). Text mining also known as Knowledge Discovery from Text (KDT) is the process of extracting information of high quality from text (i.e. structured such as RDBMS data, semi-structured such as XML, and unstructured text such as document containing words, videos, and images). This covers a large set of related topics and algorithms used in analyzing text, spanning various computer science bodies of knowledge, which include information retrieval, natural language processing, data mining, machine learning, many application domains web and biomedical sciences (Allahyari *et al.,* 2017).

# Clustering

Clustering has been regarded as one of the most popular algorithms in data mining and have been extensively studied in relation to text (Wu *et al.,* 2008; Reddy, 2017). Clustering has wide- ranging applications in classification (Wu *et al.,* 2008; Allahyari *et al.,* 2017), data visualization (Ferreira *et al.,* 2013), and organizing documents (Slonim & Tishby, 2000; Issal & Ebbesson, 2010). Clustering, in general, is an unsupervised data mining technique that groups highly related set of objects into the same class while maintaining high dissimilarity with other class(es) (Miao *et al.,* 2005; Parapar & Barreiro, 2009; Shah & Mahajan, 2012; Bouras & Tsogkas, 2013). Computation of the similarity is done using a particular similarity measure. Text clustering has different levels of granularities where clusters can be any of this document segments: paragraphs, sentences, or even terms. One of the main techniques used for organizing documents in the quest to enhance retrieval and support browsing, for example, Cutting *et al.,* (2017) applied clustering to produce table of content for a large collection of documents. Text clustering is basically applying the functionality of data mining, clustering analysis, to textual documents (Svadasa &

Jhab, 2014). Clustering analysis have also been applied in numerous areas of life to solve problems such as financial forecasting (Butler & Keselj, 2009), predicting the future expectations (Jatowt & Au Yeung, 2011), and also in the grouping of related news articles or textual documents (Bouras & Tsogkas, 2013; Rupnik *et al.,* 2016).

Clustering has been sometimes referred to as automatic classification (Jajoo, 2008; Bharti & Babu, 2017); however; it is inaccurate because the clusters are not known prior to processing whereas in the case of classification there are pre-defined classes. In clustering, it is the nature of the data and the distribution that will determine membership of a cluster, in contrast to classification where the classifier learns the association among classes and objects from a so- called training set, i.e. correctly labeled dataset beforehand, and then transfer the learned behavior on the unlabeled data set. Figure 1.1 below shows an example of a dataset with a clear cluster structure C1, C2, C3, and C4 respectively:

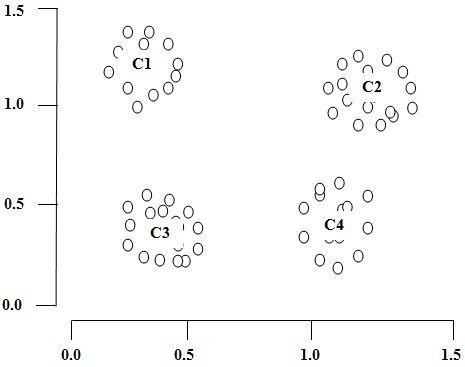


Figure 1.1: An Example of a Clear Cluster Structure Dataset

# Applications of Clustering

Clustering has been regarded as the most common type of unsupervised learning and a major tool applicable in many fields of sciences and businesses. There are many areas where it has been applied to solve problems. The basic directions in which clustering is used have been summarize as follows (Witten, 2004; Ayadi *et al.,* 2016; Allahyari *et al.,* 2017):

1. Finding Similar Documents: This is the feature that is often used when a user spotted a single “good” document in result of search and wants more of such document. The most interesting characteristics here is the discovery of documents that are alike in contrast to search based approaches conceptually that are only able to discover whether the documents share many words that are same using clustering.
2. Organizing Large Document Collection: Finding document relevant to a particular query is the focus of document retrieval, but this has failed to solve the problem of deriving sense from large number of documents that are uncategorized. This is having the challenge of organizing these documents in an identical taxonomy to the one humans would create given enough time and make use of it as a browsing interface to the documents collected originally.
3. Detection of Duplicate Content: There is need to find duplicates or near-duplicates in a large number of documents in many applications. Clustering has been applied in plagiarism detection, grouping related news article and reordering rank of search results (assuring higher diversity among the topmost documents). The description of clusters is rarely needed in such applications.
4. Recommendation Systems: These applications recommend articles to a user based on the articles that the user read already. Clustering of articles makes recommendations in real time possible and improves the quality of the system a lot because only related articles are recommended to the user of the system.
5. Search Optimization: Clustering helps a lot in improving the efficiency and quality of search engines as the query from the user can be first compared to other clusters instead of directly comparing to documents and search results can be arranged easily too.
6. Natural Language Processing (NLP): This is a sub-field in computer science, linguistics, and artificial intelligence that aim at using computers to understand natural language. Many

text mining algorithms extensively apply the NLP techniques, such as part of speech tagging (POS tagging), syntactic parsing, and other analysis in linguistics.

1. Text Summarization: Text mining applications in many fields need to summarize the documents in order to have a precise overview of the large document or a collection of documents on a topic. Summarization techniques are of two categories in general: *extractive* this is a summarization where information units extracted from the original text is comprised in the summary, and contrary *abstractive* this is a summarization where a “synthesized” information that may not be in the original text may be contained in the summary.

# Problem Statement

The degree of purity and accuracy of a clustering technique results is one of the major challenges in text clustering. This is attributed to the basic tasks performed on the text such as: document feature selection, document similarity measure selection, selecting an appropriate algorithm for clustering, clustering algorithms efficiency in terms of Central Processing Unit (CPU) resource, and associating useful conclusion to the final clusters (Shah & Mahajan, 2012; Bouras & Tsogkas, 2013; Svadasa & Jhab, 2014; Khabia & Chandak, 2015; Rupnik *et al.,* 2016; Lwin & Aye, 2017; Huang *et al.,* 2017). The explanation is as follows:

1. Document Feature Selection: This is the process that identify the terms that positively impact on the clustering process.
2. Document Similarity Measure Selection: This is the distance calculation method between two different documents which contribute greatly to the quality of a cluster.
3. Selecting an Appropriate Algorithm for Clustering: There are several algorithms for clustering such as k-means and density-based. Choosing the best algorithm to apply is a challenging task.
4. Clustering Algorithm Efficiency in Terms of CPU resource: This measures how well the intended technique utilizes the available resources.
5. Associating useful Conclusion to the Final Clusters: This is checking the level of information that can be derived from the final clusters.

Many researches have been conducted via text mining approaches in the quest to improve the accuracy and purity of clustering technique and have been applied in areas like; search engine

optimization, detection of plagiarism (Shah & Mahajan, 2012). These researches comprise; a phrasal-based clustering news article (Pera & Ng, 2007); enhancing news articles clustering with word-based N-grams (Bouras & Tsogkas, 2013); a multi-lingual news document similarity and event tracking system using latent semantic indexing technique (Rupnik *et al.,* 2016). Most techniques for textual document clustering have been designed with either traditional similarity measures such as Jaccard, and Cosine, or word-based document representation which have proven to be less effective in terms of clustering accuracy and purity (Bouras & Tsogkas, 2016; Rupnik *et al.,* 2016; Lwin & Aye, 2017; Huang *et al.,* 2017; Sohangir & Wang, 2017a, 2017b).

This thesis proposed a technique for clustering news articles and other related textual documents using an efficient similarity measure known as “improved sqrt-cosine similarity measure” and word-based N-grams. This thesis is based on the enhancement of the weaknesses of the clustering techniques designed by Bouras & Tsogkas, (2016) which uses traditional similarity measure with N-gram and Sohangir & Wang, (2017a, 2017b) which proposed an efficient similarity measure but did not test its suitability using N-grams (sequence of N consecutive characters) data representation technique.

# Aim and Objectives

# Aim

The aim of this research is to enhance an existing clustering technique for news articles and other related textual documents.

# Objectives

The major objectives of this research can be summarised as follows:

* + - 1. To improve the accuracy and purity of a technique for clustering news articles and other textual documents using an efficient similarity measure, N-grams, and k-means clustering algorithm.
      2. To compare the effectiveness of the enhanced clustering technique results with extant technique(s).

# Significance of the Study

Clustering is one of the major machine learning algorithms used in many areas of life which has contributed positively to almost all areas of human endeavors either directly or indirectly. The findings in this study will redound to the benefit of the society most especially in the field of

data mining; specifically the area of unsupervised machine learning algorithms (clustering techniques) design; considering the important role document categorization plays in the effective use of the internet today. The greater demand of techniques that will efficiently optimize grouping related documents such as news articles in the internet justifies the need for effective, clustering approaches. Thus, news articles management systems that apply the recommended clustering approach derived from the result of this study will be able to perform better. The user of such system will experience a more accurate and satisfying service. For the researchers, the study will help them uncover critical areas in the clustering process that many researchers were not able to fully explore. Thus, an enhanced technique of clustering is attained.

# Scope and Limitation of the Study

This research is limited to coming up with an enhanced clustering technique for news articles using an efficient similarity measure known as “Improved sqrt-cosine similarity measure” with N-gram based data representation. An unsupervised machine learning algorithm called k-means only will be used in clustering. This technique can be applied in news articles management systems; information retrieval system in order to improve the quality of grouping related news articles or information for easy access by the users of such systems.

# Thesis Organization

This thesis is organized into five (5) chapters. A brief of the concepts of the remaining chapters follows:

Chapter 2 provides a literature review in relation to news articles clustering techniques, reviews of several algorithms, approaches, and methodologies that have been developed for news articles clustering. The chapter indicates how literature has contributed to this area of research and the approaches that have been used.

Chapter 3 explains the methodology that has been applied in this thesis. It gives the theoretical overview of the methodology and the evaluation measures that are applied to compare the results of the methodology.

Chapter 4 contains the results of the implemented proposed clustering technique in R programming environment, the comparisons, and discussions of the various results.

Chapter 5 contains the final summary, conclusion, and future work of the thesis that can be done using the results and findings of this thesis.

# CHAPTER TWO LITERATURE REVIEW

This chapter explains some basic terminologies and concepts: textual document representation methods, textual document pre-processing methods, textual document similarity measures, categories of clustering algorithms, and finally review of literature of the various existing extant news articles (textual documents) clustering techniques.

# Basic Terminologies and Concepts

Machine Learning (ML) is referred to as systems changes which performs tasks that are associated with artificial intelligence, (Hinton, 1989; Mohri *et al.,* 2012; Bouras & Tsogkas, 2012). Some of these tasks involve diagnosis, recognition, robot control, planning, prediction, descriptive categorization, etc. The ‘changes’ can be either enhancing a system that is already performing or designing an entirely new system (Mitchell, 1997; Villanes *et al.,* 2018; Wang *et al.,* 2018). The field of machine learning has sub-branches dealing with different types of learning tasks viz: Supervised versus Unsupervised learning, Active versus Passive learning, and Online versus Batch learning (Shalev-Shwartz & Ben-David, 2014).

The supervised learning depends on a labeled classes to build a model for prediction/classification while unsupervised learning does not rely on any labeled classes; grouping are done based on some certain similarities (Mohri *et al.,* 2012; Hmeidi *et al.,* 2015). The output of the supervised learning can either be continuous (e.g. linear regression) or discrete (e.g. Classification); but the unsupervised output is mostly descriptive (Manning *et al.,* 2008; Hmeidi *et al.,* 2015).

The term ‘Bag of Words’ (BoW) is a concept used for document representation which considers the number of times each term (word/phrase) occurs but ignoring the order (Šilić *et al.,* 2007; Allahyari *et al.,* 2017). The vector representation is derived from the ‘bag of words’ and can be analyzed using some dimensionality reduction algorithms from statistics and machine learning (Witten & Frank, 2005; Allahyari *et al.,* 2017). Another term ‘N-grams’ is used to refer to a sequence of ‘n’ consecutive words or characters appearing in a collection of textual document (Chomsky & Jelinek, 2006; Šilić *et al.,* 2007; Al-Shalabi, 2017).

# Textual Document Representation Methods

Clustering algorithms do not interpret text documents directly. An indexing procedure is needed to map text document dk to a compactable representation of its content (Shafiei *et al.,* 2006; Reddy & Kumar, 2012). The selection format representation of textual document depends on what an individual accept as the most important text unit (lexical semantic problem) and the rules for natural language for combining these units (compositional semantics problem) where the problem in the latter is usually neglected (Shafiei *et al.,* 2006; Bartosiak & Nitoń, 2015). Vector-Space Model (VSM) based idea is one of the most widely used textual document representation methods; that represents each document by a vector of weight of *m* “features” extracted from the textual document, as shown in the equation 2.1:

## 𝑑𝑘 = (𝑤1𝑘, 𝑤2𝑘, … , 𝑤𝑛𝑘) (2.1)

Where *n* represents the number of features and wj is the *ith* feature weight. The value of the weight of a feature is the measure of how much the feature contributes to the meaning of the document *dk*. The different approaches for document representations differ due to the following:

* + 1. Understanding what a feature is?
    2. Computing feature weights differences.

Generally, the three major ways of representing textual documents are word-based, term- based, and N-grams-based representations (Shafiei *et al.,* 2006; Khabia & Chandak, 2015; Singh *et al.,* 2017).

# Word-Based Representation

The choice for representing a feature in this method is by identifying features with words (García-Hernández & Ledeneva, 2009; McCarthy *et al.,* 2016). This is called often either the ‘words sets’ or the ‘bag of words’ method of representing textual documents and depends on whether it is a binary weight or not (Shafiei *et al.,* 2006). Every element in the representation indicates the absence or presence of a word within a textual documents, information about sequence is lost as the words get much and is the major limitation for this particular method (Graovac, 2014; Garz *et al.,* 2018).

# Term-Based Representation

Terms are multi-words or sometimes referred to as phrases are also used as features of document vectors. They represent the total number of candidate phrase occurrences (Fejer & Omar, 2015). Term document representation has the ability of significant dimensionality reduction. Thus, some researchers alluded that it gives a better result than the word-based representation in text corpora that are special (Shafiei *et al.,* 2006). But in experiments, results of this method have not been very encouraging uniformly (Witten & Frank, 2005; Shafiei *et al.,* 2006; Bouras & Tsogkas, 2014; Huang *et al.,* 2017).

# N-grams-Based Representation

N-grams-Based representation is an independent language method of textual document representation. Documents are transformed into features with high dimensional vectors which correspond to a contiguous sub-string (Shafiei *et al.,* 2006). This is a gram of adjacent *n* characters from an alphabet say *M* (Lebret & Collobert, 2014)*.* Hence, distinct N-grams number in a given text is less than or equal to |*M*|*n*. Showing that the dimension of the N-grams vector of the feature is very high even for a value of *n* that is moderate. However, the N-grams are not present in a textual document, thus the dimension is reduced substantially. For instance, if there are 8929 unique bigrams (without stops words) in a dataset. Generally, when performing the N- gram formation of feature vectors all characters in uppercase are converted to lowercase characters and all spaces are assumed to be punctuation and normalization of the vectors are done.

To extract character N-gram form a textual document is like having a wide window of n characters across the textual document character by character or word by word. At each window n characters, positions are covered, which define a single N-gram. A single-space or multi-space are treated as one space and a space is used to be replaced by any non-letter character (Buck *et al.,* 2014). The N-grams in byte are N-grams that are generated from the sequence of bytes as appeared in the files, without applying any type of preprocessing (Buck *et al.,* 2014). To compare with the stemming and stop words removal applied in word-based or term-based, N- gram-based representation is having an advantage of robustness and less sensitive to typographical or grammatical errors and preparation of linguistic is not needed making it language independent. However, the representation in N-grams most especially word-based-N-

gram is less effective in dimensionality reduction before passing it to the text mining algorithms (Khabia & Chandak, 2015). The benefit of this is better achieved using stemming and stop word removal instead of word-based-N-gram representation (Shafiei *et al.,* 2006; Khabia & Chandak, 2015).

# Textual Document Pre-processing Methods

Pre-processing is one of the components of text mining algorithms that is very key. For instance, the framework of the traditional text document categorization comprises of pre- processing, extraction of features, selection of features, and the categorization stage (Witten, 2004; Svadasa & Jhab, 2014; Allahyari *et al.,* 2017). Even though it has been confirmed that extraction of features (Li *et al.,* 2007; Kanya & Geetha, 2013), selection of features (Hasan & Ng, 2014), and classification algorithms (Wu *et al.,* 2008), are having important impact on the process of classification, the stage of pre-processing have a noticeable effect on this success. The impact of pre-processing tasks in relation to text document classification has been widely studied (Allahyari *et al.,* 2017). The traditional pre-processing step consists of the following tasks: tokenization, term filtering or stop word removal, lemmatization, and stemming (Qian & Zhai, 2014).

# Tokenization

Tokenization is the process where text document or text strings are broken into tokens having only individual pieces (Kaur, 2015). This task breaks sequence of character to pieces (words/phrases) called tokens and throws away some characters such as special characters or punctuation marks. Then the generated list of tokens is used for processing further (Allahyari *et al.,* 2017). For instance given the text: {“When text error exist in a text, it is bad”} it is tokenized as: {“When” “text” “error” “exist” “in” “a” “text” “,” “it” “is” “bad”}.

# Lowercase Conversion

This is the conversion of all capital letters into small letters in order to have a unified formatted text before performing the next step of processing (Kaur, 2015). For instance for the text {“*When”*} and {“*when”*} will be treated as different feature. Therefore, the text is converted from : {“When” “text” “error” “exist” “in” “a” “text” “,” “it” “is” “bad”} to {“when” “text” “error” “exist” “in” “a” “text” “,” “it” “is” “bad”}.

# Punctuation Removal

This is the step of pre-processing that removed all the punctuations because it makes up the word count in any textual document but it is not having importance in data mining (Kaur, 2015). Therefore, removing the punctuations from the text {“when” “text” “error” “exist” “in” “a” “text” “,” “it” “is” “bad”} becomes {“when” “text” “error” “exist” “in” “a” “text” “it” “is” “bad”}.

# Term Filtering (Stop words Removal)

Stop words are words that are common and have little or no lexical content. Thus, they do not have any significance and need to be removed to save space and search time. This is usually performed to remove some words. The common form of filtering is the removal of stop-words. The stop-words are those words that appear frequently in a textual document without having content information (Example conjunctions, prepositions, etc). Words that are similar occurring quite often in a text are said to have little information to help differentiate documents and also word that occurs rarely are of no significant relevance and can be removed also from the textual document (Kaur, 2015; Allahyari *et al.,* 2017). For instance, applying this to the text {“when” “text” “error” “exist” “in” “a” “text” “it” “is” “bad”} becomes {“text” “error” “exist” “text” “bad”}

# Alphanumeric and Short Length Words Removal

This step of the pre-processing is performed in order to remove words with characters less than three but in some cases, even words with less than three characters are also important (Kaur, 2015). Therefore, this pre-processing is not applied to every kind of dataset because it can be ignored for example in the text {“text” “error” “exist” “text” “bad”} all the words with characters less than three are also important and there are no alphanumeric word example of such word is{“abcde1234”}.

# Lemmatization

This process is considered also as a feature extraction process, it converts different form of words inflated into a base form that can be processed by the same object. This is also considered as words morphology analysis i.e. grouping all inflated words together so that it can be analyzed as a single entity (Allahyari *et al.,* 2017). In other words, these methods always try mapping verbs to infinitive tense and nouns to single form. Before performing lemmatization to a textual

documents the part of speech (POS) of each word should be specified but because part of speech identification (POS) is error-prone and tedious, in practice researchers preferred to use stemming (Allahyari *et al.,* 2017). For instance, lemmatizing the word “nations” becomes “nation”.

# Stemming

This method is performed to obtain the stem (root) of the words derived. These algorithms are highly language dependent (Moral *et al.,* 2014; Allahyari *et al.,* 2017). The first of this algorithm was introduced in the year 1968 by Julie B. Lovins (Allahyari *et al.,* 2017), but the most widely used stemming method in English is the one introduced in the year 1980 by Martin

F. Porter (Kar, 2016; Allahyari *et al.,* 2017).

# Textual Document Similarity Measures

The distance between textual documents is defined by their similarity. Similarity measure function is applied to clustering algorithms as a condition to getting exact clusters from huge datasets (Allahyari *et al.,* 2017; Jeya & Bala, 2018). In determining how similar or dis-similar two textual documents are these functions help. First of all the documents must be converted to a vector format because similarity measures techniques and clustering algorithms cannot interpret textual documents in their original forms. The most widely used format of representation applied on any of the textual documents representation method is the vector space model and it is same as in equation 2.1 above (Allahyari *et al.,* 2017).

# Metric Conditions

The applicability of a similarity or distance measure on textual documents is subject to some metric conditions (Huang, 2008; Gillot & Cerisara, 2011). These conditions are listed below: Given a set with objects *p, q,* and *D(p, q)* is their distance. Then:

* + - 1. Nonnegativity: Their distance must be non-negative, that is, *D(p, q)*  0.
      2. Nil: Their distance is zero if and only if they are identical, that is, *D(p, q)* = 0, *iff p = q*.
      3. Symmetry: They must have symmetric distance, that is, distance from any object to the other must be the same, i.e. *D(p, q)* = *D(q, p).*
      4. Triangular inequality: Their measure must satisfy the triangle inequality property, that is,

*D(p, r)*  *D(p, q) + D(q, r).*

# Euclidean Distance Measure

The Euclidean distance is the sum of square of the difference between coordinates of two objects (Witten, 2004; Han *et al.,* 2012; Aggarwal, 2015). The Euclidean measure *D* between two vectors *a* and *b* can be calculated by the following equation 2.2:

## 𝐷 = √∑𝑛

(𝑎

− 𝑏 )2

(2.2)

𝑘=1

i𝑘

j𝑘

For example, given two documents vectors X and Y in Figure 2.1:

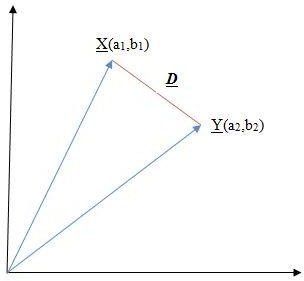


Figure 2.1: Euclidean Distance Measure

Where *D* is the Euclidean distance of the vector and can be calculated as in Equation 2.3:

𝐷 = √(𝑎1 − 𝑎2)2 + (𝑏1 − 𝑏2)2 (2.3)

# Cosine Distance Measure

The Cosine distance of two document vectors X and Y is the measure of the angle of Cosine between the document vectors (Witten, 2004; Han *et al.,* 2012; Aggarwal, 2015). It can be calculated as in Equation 2.3:

𝑐𝑜𝑠 𝜃 = K̅̅→̅̅→.̅̅̅F̅→

̅→ ̅→

|K||F|

(2.4)

For example, given two document vectors X and Y in Figure 2.2:

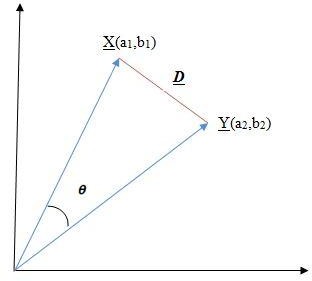


Figure 2.2: Cosine Distance Measure

# Jaccard Distance Measure

The Jaccard measure is also known as the Tanimoto Coefficient. It is the measure of the intersection divided by the union of the two textual documents (Pandit & Gupta, 2011; Huang *et al.,* 2017). This calculates terms sum and compares it with weighted sum of the terms occurring in either of the two documents but not shared among the two documents (Akinwale & Niewiadomski, 2015; Kaur, 2015). It can be calculated as in equation 2.5 or equation 2.6:

## 𝐽𝑎𝑐𝑐𝑎𝑟𝑑 (X, 𝑌) = KF

| |2 | |2

X + Y −KF

(2.5)

or

## 𝐽𝑎𝑐𝑐𝑎𝑟𝑑 (X, 𝑌) = |K ∩F|

|K 𝖴F|

# Manhattan Distance Measure

(2.6)

This calculates the absolute value of the difference between two data points (Spina *et al.,* 2014; Kaur, 2015): Given two data points *i,j* of n-dimensions, their Manhattan distance can be calculated as in equation 2.7:

## 𝐷ij = ∑n

|xi𝑘 − xj𝑘|

(2.7)

k=1

where 𝐷ij is the distance and |xi𝑘 − xj𝑘| is the difference between the two documents data points

# Pearson Correlation Measure

This is a correlation coefficient that measures how two sets of data fit on a straight line precisely. It has a value that is between -1 and 1. The correlation coefficient can be concluded by the following conditions.

* + - 1. If two documents have a correlation coefficient of 1, it indicates that there is a positive linear relationship.
      2. If two documents have a correlation coefficient of -1, it indicates that there is a negative linear relationship.
      3. If two documents have a correlation coefficient of 0, it indicates that there is no linear relationship (Gillot & Cerisara, 2011; Kaur, 2015).

The Pearson Correlation measure between two documents a and b Sim(a, b), can be calculated by the equation 2.8:

∑ ab− ∑ a ∑ b

## 𝑆i𝑚(𝑎, 𝑏) = n

√(∑ a2(∑ a)2)(∑ b2(∑ b)2)

(2.8)

n n

where n is the dimension of the text document vector, ∑ ab is the sum of the product of the two documents.

Clustering techniques uses similarity measures in calculating the similarity between documents before grouping them into their appropriate categories.

# Categories of Clustering Algorithms

Clustering algorithms are broadly classified into Hierarchical (representative-based) algorithms, partition-based algorithms, density-based algorithms, term clustering algorithms, core topic-based algorithms, and tweets clustering using hashtags (Shi *et al.,* 2018). Using any of the category of the algorithms depend on individuals interested output (Wu *et al.,* 2008;

Hosseinimotlagh & Papalexakis, 2018). For the scope of this thesis, only the Hierarchical and partition-based are discussed briefly. Hierarchical is sometimes referred to as hard clustering while the partition is also refers to as soft clustering due to their characteristics (Pande & Khandelwal, 2014).

# Hierarchical (Representative-Based) Clustering

This is one of the most popularly used clustering algorithms and it is generally of two types: Agglomerative Hierarchical Clustering (Bottom-Up) and Divisive Hierarchical Clustering (Top- Down). This clustering algorithm method treats documents like tree structures known as Dendrograms (Lwin & Aye, 2017).

# Agglomerative (Bottom-Up) Clustering

This is a hierarchical clustering that starts considering end nodes or leaves as individual clusters and keeps moving in a hierarchy until a final cluster or root nodes is reached and having all the most similar documents merged together (Shah & Mahajan, 2012; Han *et al.,* 2012; Kaur, 2015).

# Divisive (Top-Down) Clustering

This is a hierarchical clustering that is applied when smaller clusters are divided from one set of documents (single cluster). At each step, clusters are divided hierarchically so that each cluster represents set of information that is unique (Shah & Mahajan, 2012; Han *et al.,* 2012; Kaur, 2015). The Figure 2.3 shows an example of dendrogram with hierarchical clustering.

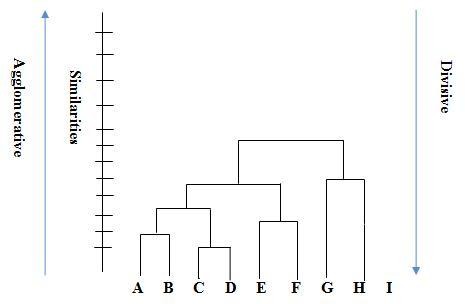


Figure 2.3: An Example of a Dendrogram (Hierarchical Clustering)

# Partition-Based Clustering

These algorithms work in an iterative manner by separating data into different clusters or groups by reducing some objective function. This is also referred to as centroid clustering (Boomija & Phil, 2008). For proper partitioning of a data that contains documents vectors, a textual document is compared with the mean of the documents so that at a minimum an objective function for the document to be associated with that clusters is formed (Kaur, 2015; Houthuys *et al.,* 2018). So many types of partition-based clustering exist such as k-means, k-medoids, CLARA, and CLARANS but for the scope of this thesis, only the k-means and k-medoids are discussed.

# K-means Algorithm

This is one of the simple unsupervised learning algorithms that is applied to solve a well- known clustering problem. The idea is finding k centroids, for individual cluster. When the choice of the centroid is good it provides a better cluster. The main aim of this algorithm is to minimize the function of object, which is the square error function in this case (Boomija & Phil, 2008; Brbić & Kopriva, 2018).

# K-medoids Algorithm

This is an unsupervised clustering algorithm that uses object representation as point of reference instead of using the mean value of objects in each cluster, unlike the k-means that is sensitive to outliers because objects with extreme large value may affect the distribution of the data, therefore, affecting the quality of the clusters (Boomija & Phil, 2008). Figure 2.4 shows an example of partition clustering.

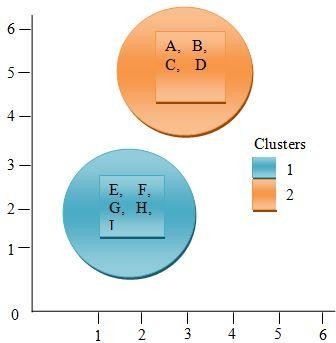


Figure 2.4: An Example of a Partition Clustering

# Review of Literature

News articles clustering technique is a wide area of research that has been on for a very long time in history which includes several tasks that range from segmenting events of news streams to tracking and detecting events (Damashek, 1995; Kyle *et al.,* 2012). Clustering techniques or methods are proposed based on some form of documents presentation, similarity, and machine learning algorithms (Mihalcea & Tarau, 2005; Saini, 2018). Research works on news article clustering techniques can be broadly categorized in two: *Word-based clustering techniques* and *N-grams-based clustering techniques* (Shafiei *et al.,* 2006; Qiujun, 2010; Ifrim *et al.,* 2014; Rupnik *et al.,* 2016). Formally, the focus was on clustering techniques for related news articles or documents and the clusters are the basis for extracting information needed (Toda & Kataoka, 2005; Nyman *et al.,* 2018). Finding hidden features and clustering these features in trying to identify some events from a news article is the latter of clustering techniques designs (Miao *et al.,* 2005; Shah & Mahajan, 2012; Mele & Crestani, 2017).

In the last few decades, research works focused on improving the efficiency of news clustering techniques. The challenge is that the traditional approaches are designed with language dependencies using traditional similarity measure and cannot perform very efficiently as the number of news reports increased in different languages, which some are considered to be short, noisy, and published at a speed that is very high (Bouras & Tsogkas, 2013; Tao *et al.,* 2018). One of the main problems in news article clustering techniques is the ability to perform efficiently on any kind of news articles without distinguishing the language which the news articles or documents are presented (Bouras & Tsogkas, 2013).

Miao *et al.,* (2005) proposed a documents-clustering method using character N-grams and compared with the terms and words based clusters results, the technique applied character N- grams to build a feature document frequency (DF × IDF) scheme. The result of the experiment of their technique shows that using character N-grams give the best clustering result. Toda & Kataoka, (2005) proposed a method for clustering news article which addressed the problem of retrieving information from information retrieval systems using a named entity extraction, terms, and finally labeling the classes of the document from the term list. Their technique finds the maximum set of terms as features representing the categories of the news within some specific time window. They identified the most frequent term features and list them within a window.

Then, the terms are grouped and analysis technique is applied to determine the most frequent terms. The extraction of these terms may result to a very large number of terms especially if the preprocessing methods are not applied. Moreover, describing the detected terms in the news using a single word set may not be intuitive and can be very difficult to interpret by humans. In Newman *et al.,* (2006) the authors present an approach to analyze entities and topics from a news article using statistical topic models this approach only consider how topics can be generated from a news article but does not consider categorizing the news articles into some clusters or to describe them. In Ikeda *et al.,* (2006) a technique that can automatically link blog entries with news articles related to such blogs was proposed using vector space model and cosine similarity to calculate the distance between the blog and the news without knowing the category of the news article. However, this method does not apply any clustering technique to know the number of news that belong to what type of news and also the method of document representation used is the word-based which have been proven to be less efficient though they tried to improve the effectiveness of their techniques using an intuitive weighing method (Naughton *et al.,* 2006).

Huang, (2008) conduct an evaluation of clustering techniques based on text similarity measures and confirmed that Euclidean distance performed worst in the clustering thereby making any clustering technique that used it less effective. Parapar & Barreiro, (2009) concluded from their experiment on the various clustering algorithms that using N-grams on any clustering algorithm helps in increasing the effectiveness in the performance of such algorithm and proposes an approach to reduce computational load on the existing clustering algorithms by using fingerprint method in trimming the size of the documents before applying the clustering algorithm and their approach perform very well with respect to saving memory and time in computation, while Karol & Mangat, (2013) uses particle swarm optimization to evaluate their proposed clustering technique that was designed using cosine similarity measure. They affirmed that the method of document representation also affect the quality of a clustering technique. A similar problem was addressed by Bouras & Tsogkas, (2010) who investigated the application of a great spectrum of clustering algorithms, as well as the measure of distance calculated by designing a news article clustering technique that make use of three different similarity measure. From the experiment, the results show that despite the simplicity of the k-means algorithm if the right preprocessing methods are applied, it increases the efficiency of the clustering technique.

Analogously Qiujun, (2010) proposed an approach for extracting the news content which is based on twin pages with the same features (specifically noisy similarity). The similarity measure applied is based on edit distance because of its simplicity which gave a fairly high complexity. This technique was designed just to check the appropriateness of applying text cleaning techniques on unstructured data from different web pages before clustering.

Park *et al.,* (2011) proposed a news articles clustering technique for contrasting contentious issues in news article from oppositions that is based of words feature by using the disputant relations similarity with the issues at hand. This technique used the word-based representation with the HITS algorithm to calculate the similarities between different discourses. Li *et al.,* (2011) proposed a two-stage scalable personalized news recommendation clustering technique which is based on intrinsic user interest, however, this technique uses hierarchy and topic detection which uses cosine similarity to calculate the distances between the interest of users before categorizing. Bouras & Tsogkas, (2012) proposed a news articles clustering technique using keywords and WordNet by applying cosine similarity to calculate the distances between the keywords then finally clustering using the weighted k-means clustering. Consequently, Bouras & Tsogkas, (2013) proposed a method based on word-based N-grams techniques, “Bag of Word”, WordNet which helps to enrich the N-grams word list by clustering the k-means core processes and the Wk-means that extend k-means. This method was implemented on N-grams based clustering system without given consideration to improving the similarity measure. The performance of the technique was measured using the clustering index (*CI*) with k-means and a previously proposed Wk-means algorithm. Though the research validated the improvement on the use of N-grams based data representation but did not check if improvement on the similarity measure used on the news articles can also improve the coherence of the news articles clusters. Qian & Zhai, (2014) proposed a multi-view clustering technique for selecting features in an unsupervised way for text-images web news data where images learn from a local orthogonal non-negative matrix factorization for labeling. However, this technique was designed on the bases of views on a particular image. Analogously, Xia *et al.,* (2015) proposed a clustering technique for social news using a topic model known as discriminative bi-term which exclude bi- terms that are less indicative by topical terms discrimination from general and specific documents. This technique, however, is language dependent because of the discrimination

attached to the specific document which makes it not flexible.

Other recent and popular techniques for clustering news articles and textual documents are: Bouras & Tsogkas, (2016) designed a document clustering system that help to solve new user problem based on WordNet database and minimal user ratings. This system is implemented using word-based N-grams which fetched articles from the database and make recommendation to new users. The result of the experiment shows that changing the value of the “n” has great impact on the clustering; Rupnik *et al.,* (2016) designed a method that can track events written in different languages and can also conduct articles comparison from different languages for making predictions of events. This was implemented using document similarity measures which is based on the cross-languages Wikipedia. The method was implemented on a multi-language system with semantic-based feature selection using probabilistic cosine similarity measure; Lwin & Aye, (2017) proposed a method for document clustering systems using hierarchical clustering based on the number of occurrences of word representations in the dataset not on the frequency of the items. Jaccard similarity measure was used for calculating the similarity between the documents; Sohangir & Wang, (2017a, 2017b) proposed a similarity measure based on the Hellinger distance known as “Improved sqrt-cosine similarity measurement”. This similarity measure was tested on different datasets and compared with other existing similarity measures on clustering textual document that contains data with high dimensionality and was proven to be more robust in contributing to the accuracy and the purity of the clusters. This measure was tested only on the “Bag of Words” (BoW), but was not tested on sequence of character or words representation of documents such as N-grams; and Santhiya & Bhuvaneswari, (2018) designed a clustering technique and implemented it on a system applying MapReduce framework for classification of crime in news articles using MongoDB. However, the authors indicated that there is need to design a technique that can automatically categorize crime from different sources irrespective of the language used to present the news.

Therefore, this thesis, focus on unsupervised ML techniques, we rather propose a technique for clustering news articles and other related textual documents using different data mining techniques such as the “Improved sqrt-cosine similarity measure” and the N-grams based data representation to enhance a clustering technique with the aim of achieving an optimal accuracy and purity of the proposed technique.

# CHAPTER THREE METHODOLOGY

This chapter explains the methodology of the proposed clustering technique used in this thesis. It gives the theoretical overview of the methodology with respect to each step involved viz: news articles pre-processing, N-grams, normalization method, vector space model, dimensionality reduction, improved sqrt-cosine similarity measurement, k-means clustering algorithm, and finally the evaluation method used in comparing the results of the proposed technique.

# Introduction

From the literature review in the previous chapter, the following points were indicated and the proposed technique is intended to concentrate on the points and then evaluate the results:

* + 1. The dimensionality of features is reduced by the pre-processing and it is important because it affects the results of the clustering technique by improving the accuracy and purity.
    2. The accuracy and type of clusters generated are affected by the input parameters to clustering algorithms.
    3. Similar news articles or documents in same clusters can also be divided iteratively into sub- clusters, which can provide more efficient insight of knowledge discovery.
    4. The method of data representation and similarity measure used affects the accuracy and purity of a clustering technique.

# Proposed Methodology (Clustering Technique)

The proposed clustering technique consist of the following steps: news articles pre- processing, news article representation using N-grams, vector space model of the news articles, dimensionality reduction using threshold on the feature vector, improved sqrt-cosine similarity measurement. Finally, k-means clustering is applied on the obtain vector and clusters of news articles are obtain. In the end, these clusters of news articles are evaluated with a view to discovering knowledge. Figure 3.1 shows all the steps of the proposed technique and the implemented steps are explained in the sub-sections that follow. Example 3.1 explains the proposed methodology in details.

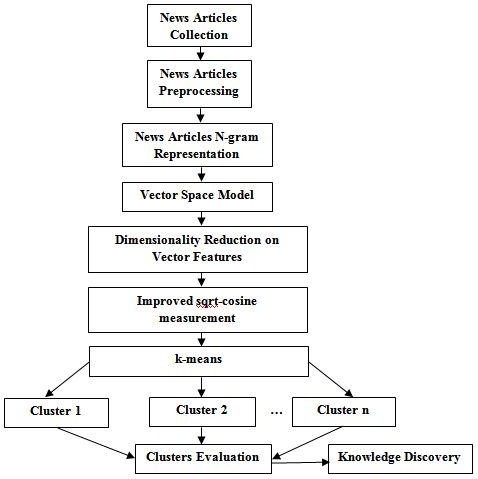


Figure 3.1: Proposed Methodology (Clustering Technique)

# 3.2.1 Example 3.1

Suppose given articles are A1, A2, and A3. A1: I am here

A2: I wont

A3: I am a boy

### STEP 1: Pre-processing the given articles

A1: I\_am\_here A2: I\_wont

A3: I\_am\_a\_boy

### STEP 2: N-gram representation using (Using character N-gram, N=4)

A1: {I\_am, \_am\_, am\_h, m\_he, \_her, here} A2: {I\_wo, \_won, wont}

A3: {I\_am, \_am\_, am\_a, m\_a\_, \_a\_b, a\_bo, \_boy}

### STEP 3: Weight Normalization

Sf*i*: Sequences frequency

W*s,* Ai: Log weight of sequences

Normalized weight: W*s,* Ai = 1 + log sfs,Ai, if sf,Ai > 0 otherwise, W*s,* Ai = 0. The Table 3.1 represent the normalized weight frequencies of the N-grams.

***Table 3.1: Normalized weighted frequencies of N-grams***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sequences Frequency of sequences in article Log Frequency weight of sequences in article**  (N-grams = 4) A1 A2 A3 A1 A2 A3 | | | | | | |
| I\_am | 1 | 0 | 1 | 1 | 0 | 1 |
| \_am\_ | 1 | 0 | 1 | 1 | 0 | 1 |
| am\_h | 1 | 0 | 0 | 1 | 0 | 0 |
| m\_he | 1 | 0 | 0 | 1 | 0 | 0 |
| \_her | 1 | 0 | 0 | 1 | 0 | 0 |
| Here | 1 | 0 | 0 | 1 | 0 | 0 |
| I\_wo | 0 | 1 | 0 | 0 | 1 | 0 |
| \_won | 0 | 1 | 0 | 0 | 1 | 0 |
| wont | 0 | 1 | 0 | 0 | 1 | 0 |
| m\_a\_ | 0 | 0 | 1 | 0 | 0 | 1 |
| \_a\_b | 0 | 0 | 1 | 0 | 0 | 1 |
| a\_bo | 0 | 0 | 1 | 0 | 0 | 1 |
| \_boy | 0 | 0 | 1 | 0 | 0 | 1 |

### STEP 4: Vector Space Model

A1: {1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0}

A2: {0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0}

A3: {1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1}

### STEP 5: Feature Extraction

A1: {1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0}

A2: {0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0}

A3: {1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1}

### STEP 6: Distances Calculation using improved sqrt-cosine similarity measure

Using equation 3.3 the distance is calculated as follows:

𝑛

∑ √𝐴1i𝐴2i

𝐼𝑆𝐶(𝐴1, 𝐴2) =

i=1 **=**

𝑛 i=1

√(∑

i=1

𝐴1i) √(∑𝑛

𝐴2i)

√1\*0+ √1\*0+√1\*0+ √1\*0+ √1\*0+ √1\*0 +√0\*1+√0\*1+√1\*0+√0\*0+√0\*0+√0\*0 **=** 0

= **0.0000**

√(1+1+1+1+1+1+0+0+0+0+0+0+0) \* √(0+0+0+0+0+0+1+1+1+0+0+0+0)

Using the same equation 3.3 as above give the following results: ISC (A1, A2) = 0.0000

ISC (A1, A3) = 0.3333 ISC (A2, A3) = 0.0000

√(6) √(3)

**STEP 7: Apply clustering algorithm (K-means).** Based on the distances above the articles are clustered as follows: Cluster 1 = {A1, A3} and Cluster 2 = {A2}. As shown below.

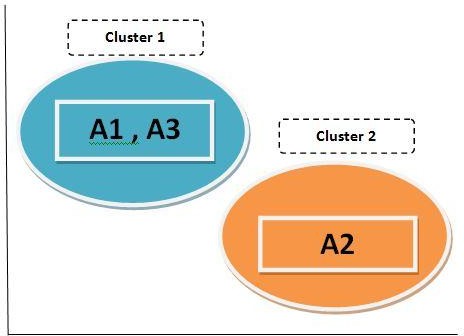


Figure 3.2: Example Generated Clusters

**Step 8: Evaluation**. The clusters generated in Figure 3.2 can be evaluated in terms of accuracy and purity as follows below:

### Accuracy of the clustering results

F-Measure is a measure of a test's accuracy. It considers both the precision and the recall. Higher F-measure means better clustering technique.

Class M = 3 Cluster 1 (C1) = 2

Cluster 2 (C2) = 1

Number of common pairs in M and C (*tp*) = 2 Number of common pairs in class M (*fn + tp*) = 3 Number of pairs in clustering C (*fp + tp*) = 2 Precision = *tp* / (*fp + tp*) = 2/2 = 1

Recall = *tp* / (*fn + tp*) = 2/3 = 0.66

F-Measure = (2\* Precision \* Recall) / (Precision + Recall) = (2 \* 1 \* 0.66) / (1 + 0.66) = 1.32 / 1.66 = **0.7952**

### Purity of the clustering results

This is the measures the quality of a single cluster Ci referring to pij. The higher purity means better clustering technique.

Class M = 3 Cluster 1 (C1) = 2

Cluster 2 (C2) = 1

Number of common pairs in M and C (tp) = 2 Number of common pairs in class M (fn + tp) = 3 Number of pairs in clustering C (fp + tp) = 1 Purity = (1/ | Ci |) \* maxj {pij}

maxj {pij} = largest number of objects common in M and Ci Purity = (1/ |2|) \* 2 = **1** or Purity = (1/ |1|) \* 1 = **1**

# News Articles Pre-processing

Clustering textual document such as news articles is one of the various tasks in text mining because the data you are dealing with is unstructured in nature, a huge amount of unstructured or semi-structured data are stored in text format like blogs, research papers, books, web pages, email messages, XML documents, news articles etc (Khabia & Chandak, 2015). To obtain a structured data format form the unstructured data format, some sequence of long operations are required to convert this text to the desired format before performing any further processing. Some of these operations are: stop words removal, lemmatization, stemming, and white space

removal which are mostly applied for the traditional “bag of words” methods of document representation. The pre-processing technique applied in this thesis is different from the traditional methods; this is because using N-grams has a special pre-processing method that helps in converting any non-character letter or any multiple spaces due to error at any part of the document into a single space and stop words don’t have any effect in the final result of the clustering like it is having in the “bag of words” documents representation method. For instance, given the articles {“A1: I am here, A2: I won't, A3: I am a boy”} after pre-processing it, the result will be {“A1: I\_am\_here, A2: I\_won\_t, A3: I\_am\_a\_boy”}. The advantage of using this method is that it does not perform all the other stages of pre-processing like in the traditional “bag of words” (BOW) documents representation methods thereby saving computational time (Lebret & Collobert, 2014).

# News Articles N-grams Representation

N-grams news articles representation is a sequence of ‘textual units’ in ‘n’ adjacencies (sequence of text of length ‘n’), which are extracted from a particular news article. The context of interest determined the level which a ‘textual units’ can be identified either as a character, word or byte. In this thesis, we are using the word-based N-grams which is conceptualized by having a small sliding window placed in a given text of a sequence, where the only visible words are the ones within the ‘n’ window placed at a given time (Bouras & Tsogkas, 2013; Glasserman & Mamaysky, 2017). At each window position, only the words within it are recorded. Some schemes allow window sliding of more than a single word record for each N- gram.

# Representation with 1-gram

This is the simplest N-gram called the unigram with n = 1, which falls back to the traditional ‘bag of words’. This practice is typically for news articles representation as ‘bag of words’ (Khabia & Chandak, 2015). Documents and ‘bag of words’ dimensions are comprised in the vector space model dimensions. For the purpose of word extraction from the textual document, the text processing in the traditional ‘bag of words’ are applied. For instance, given the articles {“A1: I am here, A2: I won't, A3: I am a boy”} the 1-Gram representation will be

{“A1: ‘I’ ‘am’ ‘here’, A2: ‘I’ ‘won't’, A3: ‘I’ ‘am’ ‘a’ ‘boy’ ”}.

# Representation with N-grams

A long string of symbols are extracted from the documents known as N-grams. This sequence is in the form of character, word or byte. If the account of word is taken as the sequence, text semantics is captured better (Chomsky & Jelinek, 2006; Khabia & Chandak, 2015; Bouras & Tsogkas, 2016). Thus, we consider the word-based N-grams, N-grams are collection of adjacent words. This means that bi-gram, tri-grams, etc. are obtained. For the N- gram method of representing the news articles, the removal of stop words are not needed and also other pre-processing such as stemming is not needed. Thus, the uses of N-grams help in ignoring any grammatical or typographical errors in the articles. For instance, given the articles

{“A1: I am here, A2: I won't, A3: I am a boy”} the character quad-grams representation will be

{A1: I\_am, \_am\_, am\_h, m\_he, \_her, here, A2: I\_wo, \_won, won't, A3: I\_am, \_am\_, am\_a, m\_a\_,

\_a\_b, a\_bo, \_boy}.

# Weight Normalization

Weighting generated N-grams is needed in order to normalize the frequency of the respective N-grams. The most used weighting method in text mining and information retrieval is the Inverse Document Frequency and Term Frequency, this gives the measure of the number of relevant terms in the collected document (Sohangir & Wang, 2017a, 2017b). Term frequency denotes the total number of sequence of N-grams repeated in a given document. While the total number of times an N-grams sequence occurs in a given document is the Inverse Document Frequency i.e. the count of documents that contains the significant sequence. The term frequency is calculated using equation 3.1.

𝑡ƒi,j

= 𝑚i,j

∑j 𝑚𝑘,j

𝑘

(3.1)

Where 𝑚i,j represent the number of times relevant terms sequence ti appears in the document dj and sum of occurrence of terms in the whole document dj.

In this thesis, we used the weighting formula in the equation 3.2. This is because of its

simplicity and yet having a more accurate result (Sohangir & Wang, 2017a, 2017b).

𝑤𝑡,𝑑 = {1+ log10 𝑡ƒ𝑡,𝑑 if 𝑡ƒ𝑡,𝑑 >0 𝑜𝑡ℎ𝑒𝑟wi𝑠𝑒 (3.2)

0

Where 𝑤𝑡,𝑑 the normalized log frequency weight and 𝑡ƒ𝑡,𝑑 is the term frequency of the sequence in the document dj

# Vector Space Model

An appropriate numerical model is needed to represent the text document dataset in order for the clustering algorithm to be able to process it. The vector space model is the most appropriate numerical model presented with term weighting technique that is suitable (Shafiei *et al.,* 2006; Khabia & Chandak, 2015; Singh *et al.,* 2017). The log term frequency (LogTF) weighting scheme is used in this thesis as sequence of terms weight. Also, the log frequency weighting is normalized to unity i.e. 𝑤𝑡,𝑑<=1. The LogTF takes the log of the term frequency based on the condition if the term frequency is greater than zero.

# Dimensionality Reduction on Vector Features

This is needed because text documents with short length form several N-grams. Therefore, a huge number of features are generated from the N-gram based representation model. The dimensionality reduction task is necessary for the sake of time computational complexity (Shafiei *et al.,* 2006; Khabia & Chandak, 2015; Singh *et al.,* 2017). During the news articles clustering, only the dimensions of the feature vector are reduced, which means that only the number of features to be used for the clustering we are interested in is reduced. A threshold is applied to the 𝑤𝑡,𝑑 values of vector space model in order to select the features. In N-grams, the highest total number of 𝑤𝑡,𝑑 weight in the news articles text collection is selected; N-grams are selected as feature for the news articles clustering from the news articles collection. Therefore, 50 per cent (50%) of the dimensions are successfully reduced.

# Improved Sqrt-Cosine Similarity Measure

High-dimensional data information retrieval is very common, but there is a problem of space when working with the Euclidean distances (Sohangir & Wang, 2017a, 2017b). In machine learning, when dealing with higher dimensional data, Euclidean distances are rarely considered to be effective measurements. Aggarwal, (2012) proves Euclidean norm measurement from both empirical and theoretical perspective not to be an effective metric for data mining applications on high-dimensional data. Because in high dimensional spaces there is concentration on the

distance, the distance ratio from far and near neighbors is almost one to a given target. Due to this, distances from different data points do not have variations between them. Also, Aggarwal, (2012) investigated the behavior of the Euclidean norm measures in high-dimensional space. From the results of this investigation, it was discovered that the lower value of dimensions tends to perform better. In other words, application on high-dimensional data, distances such as the Hellinger, is favored than the Euclidean distances. Therefore, in this thesis, equation 3.3 represents the similarity measurement used to calculate the distances between the normalized vectors of the news articles.

𝑛

∑ √𝑝i𝑞i

## 𝐼𝑆𝐶(𝑝, 𝑞) =

i=1

i=1

(3.3)

𝑛 i=1

√(∑

𝑝i) √(∑𝑛

𝑞i)

where each document is normalized and the square root of their normalized form, that is,

𝑛 i=1

√(∑

𝑝i) is used. 𝑝i is document one normalized form, qi is document two normalized form,

and 𝐼𝑆𝐶(𝑝, q) is their similarity measure. equation 3.3 above is known as the efficient similarity measure. It has been chosen because it has been proven as an effective measure compared to other “state of the art” similarity measures for textual document clustering on word-based document representation but has not been tested with the N-grams-based document representation (Sohangir & Wang, 2017a, 2017b).

# K-means Clustering

This is considered one of the simplest unsupervised machine learning algorithms used in grouping objects that are similar. This was designed by J. MacQueen (1967) and thereafter J. A. Hartigan and M. A. Wong (1975) (Han *et al.,* 2012). The news articles clustering groups the news collection into different groups such that news in the same group share features in common. k-means clustering is used for clustering the news articles. This algorithm partitions news automatically into k different clusters. To determining the value of k to be used while clustering, the sum of square error method is used. In this thesis, we used the k value in which the Within-group distance Sum of Squares (WSS) is smallest, from the result of the estimated Within Sum of Square (WSS), this is because we want to adhere to the main goal of clustering which is reducing the within sum of square distance of the clusters.

# Evaluation Methods

In order to evaluate the extent to which the proposed clustering algorithm has been enhanced and to know the effect of extracting N-grams within the context of clustering news articles through the offline experiment, we compared our technique with other techniques that are recent using metrics like the purity and the accuracy (Karol & Mangat, 2013; Sohangir & Wang, 2017a, 2017b). Equations 3.4, 3.5, 3.6, 3.7, 3.8, and 3.9 explain the evaluation metrics respectively. Let the clustering of document *D* be represented by k = {k1, k2, …, kc}, k\* = {k1\*, k2\*, …, kl\*} represents the “correct” class set of *D.*

The Entropy Ej of each cluster can be represented by the equation 3.4:

## 𝐸j = − ∑ 𝑝ij log(𝑝ij) (3.4)

where 𝑝ij is the probability that cluster j member belong to the category i.

The total entropy can be calculated by the addition of the entropies of all the weighted size of each cluster as represented by the equation 3.5:

𝑐

## 𝐸𝑛𝑡𝑟𝑜𝑝𝑦(𝐶) = ∑

j =1

𝑛j \* 𝐸j

𝐷

(3.5)

where, c is the total number of clusters, nj the size of cluster *jth* and ***D*** is total number of documents.

Recall of cluster *j* with respect to i class, recall (j, i) can be represented by the equation 3.6:

𝑟𝑒𝑐𝑎𝑙𝑙 (i, j) = |𝑘j ∩𝑘i\*|

\* (3.6)

|𝑘i |

where, |𝑘j ∩ 𝑘i\*| represents the number of members of class i in cluster j, |𝑘i\*| represents the number of members of class i.

The precision cluster j with respect to i class, precision (i, j) can also be represented by the

equation 3.7:

𝑝𝑟𝑒𝑐i𝑠i𝑜𝑛 (i, j) = |𝑘j ∩𝑘i\*|

(3.7)

|𝑘j |

where, |𝑘j ∩ 𝑘i\*| represents the number of members of class i in cluster j, |𝑘i | represents the number of members of cluster j

Then, F-measure is the combination of both precision and recall represented by the equation

3.8

𝐹(i, j) = 2\*precision (i,j)\*recall (i,j)

precision (i,j)+recall (i,j)

(3.8)

The F-measure of the overall quality of the cluster set k can be represented by the equation 3.9:

## 𝐹 =

𝑙

## ∑

i=1

|𝑘i\*|

𝐷

## 𝑚𝑎𝑥{𝐹(i, j)}, j = 1, 2, …, c

(3.9)

where the max is taking from all clusters, ***D*** is total number of documents.

Generally, higher purity and higher accuracy means better clustering solution (Shah & Mahajan, 2012; Lwin & Aye, 2017). The results of the analysis of the proposed clustering technique is discussed in chapter four.

# CHAPTER FOUR RESULTS AND DISCUSSIONS

The proposed news articles clustering technique is implemented using R programming environment version 3.4.3. This chapter describes the nature of the datasets, experiment evaluation, experimental set-up, results discussions, and finally our contributions.

# Description of Datasets

The experiment in this thesis was conducted on two different datasets from different domains of application. Table 4.1 presents the list of the datasets. We used these sets because they are used commonly as benchmark for text clustering techniques, classification, and other machine learning algorithm’s validation and test. A more elaborate description of the used sets is as follows:

* + 1. The *Reuters-21578* is a document collections appearing at the Reuters newswire since 1987. *R1* and *R21* are Reuters-21578 data subsets categorization text. The Reuters personnel labeled the contents after collecting the document set.
    2. *20Newsgroups* is a dataset collection of different newsgroups of about 20Newsgroups, around 20,000 newsgroup documents are contained; it is one of the most used dataset in text mining (Sohangir & Wang, 2017a, 2017b).

***Table 4.1: Real-world datasets summary***

|  |
| --- |
| **Datasets No. of Samples No. of Dimensions No. of Class** |
| Reuters 2900 1000 1-21  20Newsgroups 11293 1000 20 |

# Evaluation of Experiments

In evaluating clustering techniques, the general goal is to evaluate the techniques based on some criteria’s such as accuracy and purity of the generated clusters results. Clustering news articles is a novel task, therefore no test collection exist since the articles are in the form of unstructured data. The number of items in each cluster represents the level relationships that exist between them. Any clustering technique that generates clusters with high accuracy level and high purity level is considered to be the best. The goal of every clustering technique is to achieve better performance than the existing ones.

# Experimental Set-Up

The experiment in this thesis has been implemented in R version 3.4.3. Before running the experiment, we needed to determine the best value of k (number of clusters) between 2 to 10 that will be suitable and give best results based on our data using “within sum of square” distance analysis (Elbow method) (Khabia & Chandak, 2015). The Figure 4.1 shows the plot of total number of “within sum of square” against number of clusters k. This was done in order to check the number of clusters that will help in given high-quality result. As indicated in the Figure for the value k = 3 the ratio of change of the within sum of squares tends to have the first elbow like appearance at k = 3 compared to the other k’s, this shows that k = 3 is the best choice of the number of clusters. Therefore, the number of cluster used in this thesis is 3 based on the analysis conducted.

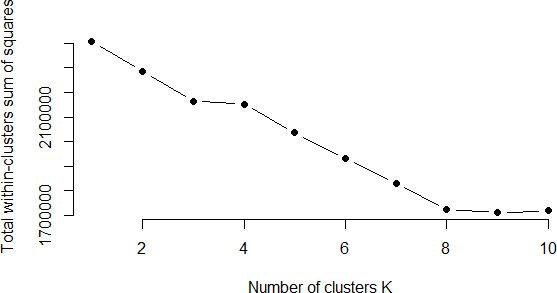


Figure 4.1: Graph between within sum of squares and different values of k to determine the

best number of clusters

# Results

The results of the experiment is recorded in this section, the various values of N-gram results purity and accuracy are compared in the experiment. The performance metrics across all the two datasets with only the k-means learner are the focus in the first instance. The experiment is run ten (10) times and the average of the accuracy and the purity are recorded. At the end, the performance of the best N-grams results of our proposed clustering technique is compared with the baseline clustering technique (Sohangir & Wang, 2017a, 2017b) which was enhanced in

order to check effectiveness and the level of improvement by the proposed clustering technique in this thesis.

# Result of Technique with N-gram Equal to 1 (N=1)

Figure 4.2a, Figure 4.2b and Table 4.2 show the nature of the cluster, average purity and accuracy of the proposed technique with the value of N-gram = 1 on both the reuters21578 and 20newsgroup datasets respectively.



Figure 4.2a: Nature of clusters with N-gram=1 on Reuters21578

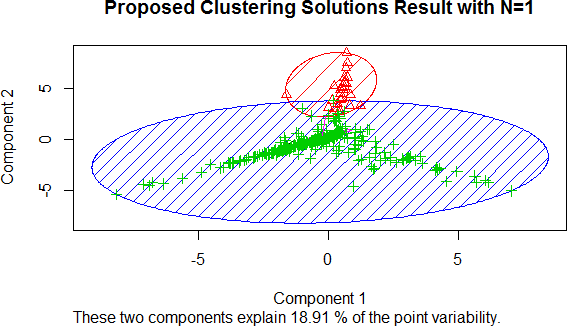


Figure 4.2b: Nature of clusters with N-gram=1 on 20newsgroups

***Table 4.2: Average performance of proposed technique with N-gram=1 on the datasets***

|  |
| --- |
| **Datasets Accuracy Purity** |
| Reuters 0.3135 0.5521  20Newsgroups 0.3001 0.5600 |

# Result of Technique with N-grams Equal to 2 (N=2)

Figure 4.3a, Figure 4.3b and Table 4.3 show the nature of the cluster, average purity and accuracy of the proposed technique with the value of N-grams = 2 on both the reuters21578 and 20newsgroup datasets respectively.

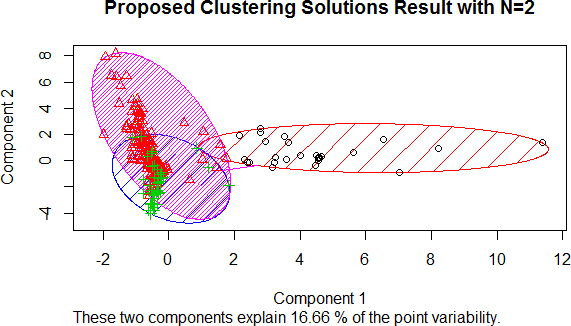


Figure 4.3a: Nature of clusters with N-grams=2 on Reuters21578

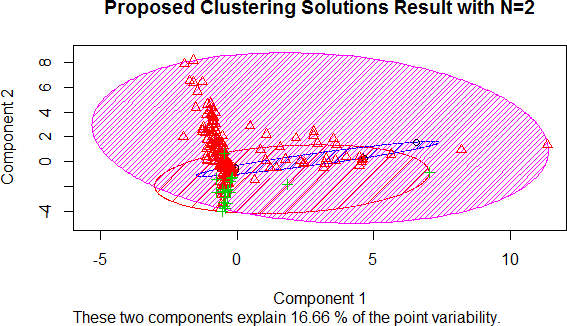


Figure 4.3b: Nature of clusters with N-grams=2 on 20newsgroups

***Table 4.3: Average performance of proposed technique with N-grams=2 on the datasets***

|  |
| --- |
| **Datasets Accuracy Purity** |
| Reuters 0.3435 0.6021  20Newsgroups 0.3201 0.5900 |

# Result of Technique with N-grams Equal to 3 (N=3)

Figure 4.4a, Figure 4.4b and Table 4.4 show the nature of the cluster, average purity and accuracy of the proposed technique with the value of N-grams = 3 on both the reuters21578 and 20newsgroup datasets respectively.

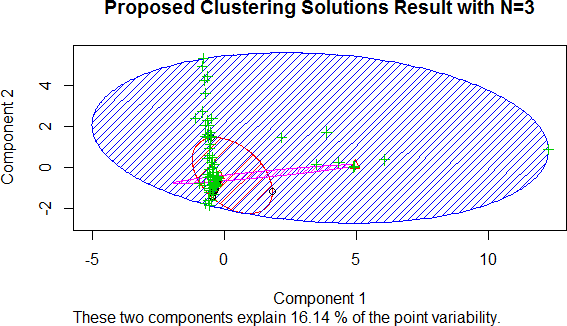


Figure 4.4a: Nature of clusters with N-grams=3 on Reuters21578

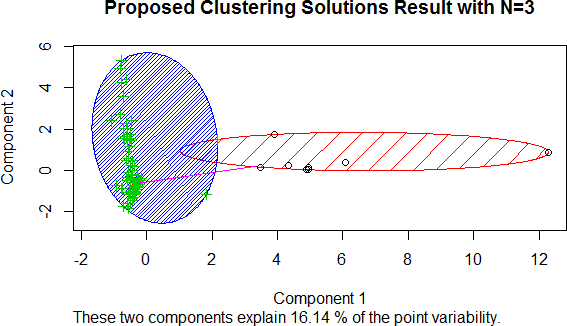


Figure 4.4b: Nature of clusters with N-grams=3 on 20newsgroups

***Table 4.4: Average performance of proposed technique with N-grams=3 on the datasets***

|  |
| --- |
| **Datasets Accuracy Purity** |
| Reuters 0.3950 0.9418  20Newsgroups 0.3801 0.9200 |

# Result of Technique with N-grams Equal to 4 (N=4)

Figure 4.5a, Figure 4.5b and Table 4.5 show the nature of the cluster, average purity and accuracy of the proposed technique with the value of N-grams = 4 on both the reuters21578 and 20newsgroup datasets respectively.

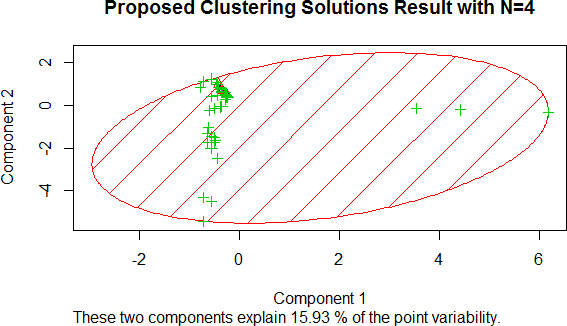


Figure 4.5a: Nature of clusters with N-grams=4 on Reuters21578

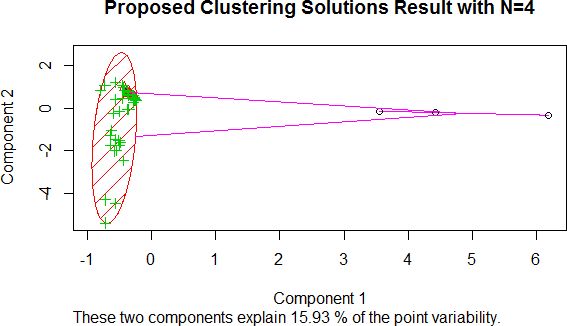


Figure 4.5b: Nature of clusters with N-grams=4 on 20newsgroups

***Table 4.5: Average performance of proposed technique with N-grams=4 on the datasets***

|  |
| --- |
| **Datasets Accuracy Purity** |
| Reuters 0.2950 0.4418  20Newsgroups 0.2701 0.4100 |

Tables 4.2, 4.3, 4.4 and 4.5 show the performance of the proposed technique on the two datasets base on one-gram, bi-grams, tri-grams, and quad-grams respectively, as indicated in the accuracy and purity, the proposed clustering technique is better on reuters21578 as compared to the result on the 20newsgroups but on average the technique perform averagely good on purity with a value of 44% to 94% which is an indication of better performance as compare to the baseline technique results (Sohangir & Wang, 2017a). Though the performance of this technique is measured without given consideration to the time and spaces complexities because the primary

aim is to achieve better purity and accuracy level from the proposed technique regardless of execution time. Looking at the accuracy level of the proposed technique the performance is not too encouraging having the values ranging from 29% to 39% average but compared to the baseline technique the proposed technique have been able to show improvement in accuracy with 4% which is an enhancement. Figure 4.6a, Figure 4.6b, Figure 4.7a, and Figure 4.7b are charts showing the plot of accuracy and purity against one-gram, bi-grams, tri-grams, and quad-grams on the datasets respectively, the line chart charts plot on the various values of N-grams clearly shows the drop in both accuracy and purity on the two datasets as the values of N-grams is equal or greater than 4 (N-grams=>4) which indicates poor clustering result as the number of N-grams grows higher than three. It is clear from the chart that when value of N-grams=3 the technique tend to achieve better results both in accuracy and purity of the clusters generated.

0.5

0.4

0.3

0.2

0.1

0

Reuters

20Newsgrouops

1 2 3 4

**N-Grams**

**Accuracy**

Figure 4.6a: Bar chart showing accuracy of proposed technique on different N-grams

1

0.8

0.6

0.4

0.2

0

Reuters

20Newsgrouops

1 2 3 4

**N-Grams**

**Purity**

Figure 4.6b: Bar chart showing purity of proposed technique on different N-grams

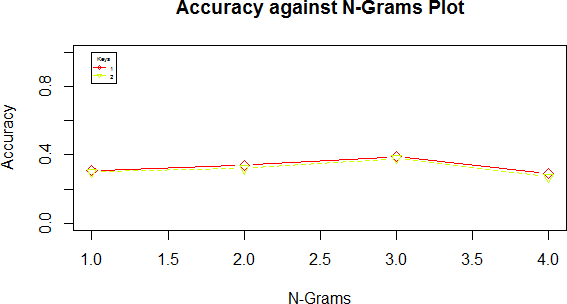


Figure 4.7a: Line chart showing accuracy of proposed technique on different N-grams

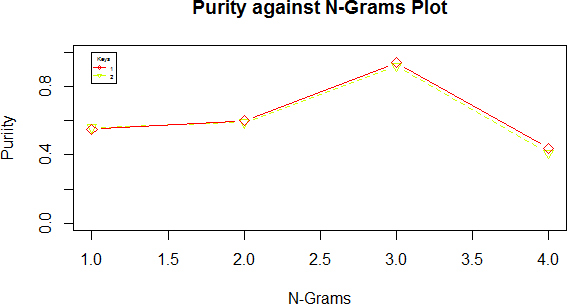


Figure 4.7b: Line chart showing purity of proposed technique on different N-grams

# Results Comparison of Proposed Technique and Baseline Technique

Table 4.6 shows the best average purity and accuracy of the proposed technique as compared to the baseline technique (Sohangir & Wang, 2017a) on both the reuters21578 and 20newsgroup datasets respectively.

***Table 4.6: Average performance of the proposed technique as compared to baseline technique on the datasets***

|  |  |
| --- | --- |
| **Techniques Datasets Accuracy Purity** | |
| Proposed Clustering Technique | Reuters 0.3950 0.9418 |
| 20Newsgroups 0.3801 0.9200 |
| Baseline Clustering Technique (N=1) | Reuters 0.3135 0.5521  20Newsgroups 0.3001 0.5600 |

The average performance of the results in Table 4.6 across the all the datasets shows that our proposed technique has a better performance than the baseline technique with the highest accuracy of **0.3950** and purity of **0.9418** which is an improvement of the baseline line technique. This means that on both datasets our proposed technique has been able to increase accuracy and purity with **15%** and **36%** as compared to best results achieved by the baseline techniques respectively. But when the value of N-grams equal or greater than four (N-grams=>4) the accuracy and purity tends to be poor which is in line with what the literature has proven in past (Khabia & Chandak, 2015; Bouras & Tsogkas, 2016).

# Our Contributions

The following are the contributions made in this thesis:

* + 1. Applied various N-grams methods of preprocessing textual data.
    2. Applying N-grams based representation of data (text) with an efficient similarity measure known as “improved sqrt-cosine similarity”.
    3. Applying an unsupervised machine learning algorithm known as k-means for clustering the dataset.
    4. The results of our technique compared with an extant technique in the literature and also shows how better our technique is from the existing one (Sohangir & Wang, 2017a, 2017b).

# CHAPTER FIVE

# SUMMARY, CONCLUSION, AND FUTURE WORK

This chapter contains the final summary, conclusion, and future work of the thesis that can be done using the results and findings of this thesis.

# Summary

News articles clustering technique is an area of research that is promising and challenging. It has been used in different fields of data mining to discover important and meaningful patterns from data. Datasets are divided into groups containing similar items. In this thesis, clustering techniques has been studied in order to perform news articles clustering using N-grams and the improved sqrt-cosine similarity so as to improve the accuracy and purity of an existing technique. Generally, clustering algorithms are of two types: hierarchical and partition based clustering but this thesis only looked at a clustering technique that involves the partition based (k-means).

This thesis has explored methodologies for news articles clustering task. The problem with word-based document representation has been indicated through the literature. The inconsistency level of word-based document representation due to its language dependant has been overcome using the N-grams document representations. Besides proposing the technique for clustering news articles and other related textual documents, the technique can be successfully applied in other disciplines as well, since it is a tool for clustering and comparison of objects other the news articles.

Clustering methodologies, based on partitioning, defines number of clusters to represent the input parameter, a centroid is randomly selected for a cluster and produce clusters of flat shape. The accuracy of such algorithms as such, depends on the input parameter primarily selected. In addition to this, literature have shown that the method of representing the document to be clustered and the type of similarity measure used also contribute immensely to the accuracy and the purity of such technique. This is most especially applicable when we are dealing with large datasets or data that has high dimensions such as textual document.

# Conclusion

This thesis rationale is to find the importance of document representation and similarity measurement in enhancing the accuracy and purity of a clustering technique using k-means algorithm. Numerous experiments were conducted on two datasets: Reuters21578 and 20Newsgroups using N-grams document representation with different values of the N-grams such as one-gram, bi-gram, tri-gram, and quad-gram, to check their resultant effect and level of improvement on the performance of the clustering technique with the improved sqrt-cosine similarity measure and k-means algorithm. The experiment was conducted on R version 3.4.3 programming environment. The experiment indicated good performance of the proposed clustering technique on various values of N-grams. Moreover, compared to the word-based clustering technique it also has better results. Based on the analysis conducted the following conclusions were reached:

* + 1. The accuracy and purity of the clustering technique is best at N-grams = 3 with the value of **39%**, **94%** on reuters21578 and **38%**, **92%** on 20Newsgroups respectively. But when the value of N-grams tends to equal or greater than 4 (i.e. N-grams  4), the result become poor which is also in line with what the literature supports. Our proposed technique performs better than the previous technique (N-gram = 1) studied in the literature in terms of accuracy and purity.
    2. The method of document representation and similarity measures contributes a lot to the accuracy and purity, also in the overall effectiveness of clustering technique.
    3. Having the best number of clusters in k-means provide clusters with a very low variability between **18.91%** and **16.14%** as shown from the results of our study. When N-grams = 3, in our proposed clustering technique, a better clustering result with low inconsistency between elements of the same clusters is achieved.
    4. Clustering techniques designed based on N-grams increased more positive probabilities in items search (as N-grams increases the alphabet size), since it allows errors and expands the overlapping of the contents of a particular news article or textual document.

# Future Work

The news articles clustering technique proposed in this thesis is at its very rudimentary stage and many possible improvements can be added and implemented. In addition to the results

evidence in this thesis some of the following are the possible implementation and extensions that can be made:

* + 1. A system can be implemented based on the proposed clustering technique in order to test the technique on an online live streaming news data.
    2. Since clustering only indicates the presence of articles N-grams in a group but does not give a specific label to the elements in a particular group, therefore it will be an interesting work to apply two or more classifiers so as to have clear labels of the document classes. This is to further test which classifier gives the best result.
    3. The proposed technique can be implemented with different developed similarity measure in the future to further check the validity of the proposed technique.
    4. Implementation of the technique on different unsupervised machine learning algorithms in order to measure the accuracy, purity, and complexities both in time and space.

# REFERENCES

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. New York, NY: Springer Science & Business Media.

Aggarwal, C. C. (2015). *Data mining*: *The textbook*. Cham: Springer International Publishing. Agrawal, R., Dimitrios, G., & Frank, L. (1998). Mining process models from workflow logs. In

*International Conference on Extending Database Technology.* (pp. 467–483). Springer Berlin Heidelberg.

Akinwale, A., & Niewiadomski, A. (2015). Efficient similarity measures for texts matching.

*Journal of Applied Computer Science*, *23*(1), 7–28.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief survey of text mining: Classification, clustering, and extraction techniques. *ArXiv Preprint ArXiv: 1707.02919, 21*(5)*,* 23–33.

Al-Shalabi, R. E. F. (2017). The effect of N-gram indexing on Arabic documents retrieval.

*International Journal of Computer Science Issues (IJCSI)*, *14*(1), 115–125.

Ayadi, R., Maraoui, M., & Zrigui, M. (2016). A survey of Arabic text representation and classification methods. *Research in Computing Science*, *117*(12), 51–62.

Bartosiak, T., & Nitoń, B. (2015). Walenty–extending valence dictionary with semantics.

*Selected Problems in Information Technologies*, *5*(3), 34–48.

Bharti, S. K., & Babu, K. S. (2017). Automatic keyword extraction for text summarization: A survey. *ArXiv Preprint ArXiv: 1704.03242, 20*(4), 12–20.

Boomija, M. D., & Phil, M. (2008). Comparison of partition based clustering algorithms. *Journal of Computer Applications*, *1*(4), 18–21.

Bouras, C., & Tsogkas, V. (2010). Assigning web news to clusters. In *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on* (pp. 1–6). IEEE.

Bouras, C., & Tsogkas, V. (2012). A clustering technique for news articles using WordNet.

*Knowledge-Based Systems*, *36*(8), 115–128.

Bouras, C., & Tsogkas, V. (2013). Enhancing news articles clustering using word N-grams. In

*DATA* (pp. 53–60). ACM Press.

Bouras, C., & Tsogkas, V. (2014). Evaluating the unification of multiple information retrieval techniques into a news indexing service. In *DATA* (pp. 33–40). ACM Press.

Bouras, C., & Tsogkas, V. (2016). Assisting cluster coherency via N-grams and clustering as a tool to deal with the new user problem. *International Journal of Machine Learning and Cybernetics*, *7*(2), 171–184.

Brbić, M., & Kopriva, I. (2018). Multi-view low-rank sparse subspace clustering. *Pattern Recognition*, *73*(3), 247–258.

Buck, C., Heafield, K., & Van-Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *LREC* (Vol. 2, pp. 4). IEEE.

Butler, M., & Keselj, V. (2009). Financial forecasting using character N-gram analysis and readability scores of annual reports. In *Canadian Conference on AI* (pp. 39–51). Springer, Berlin, Heidelberg.

Chomsky, N., & Jelinek, F. (2006). *N-grams*. Cambridge University Press.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (2017). Scatter/gather: A cluster- based approach to browsing large document collections. In *ACM SIGIR Forum* (Vol. 51, No. 2, pp. 148–159). ACM Press.

Damashek, M. (1995). Gauging similarity with N-grams: Language-independent categorization of text. *Science, New Series*, *267*(5199), 843–848.

Fejer, H. N., & Omar, N. (2015). Automatic multi-document Arabic text summarization using clustering and keyphrase extraction. *Journal of Artificial Intelligence*, *8*(1), 1–9.

Ferreira, R., Freitas, F., Cabral, L. de S., Lins, R. D., Lima, R., Franca, G., … & Favaro, L. (2013). A four dimension graph model for automatic text summarization. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT*) (pp. 389–396). IEEE Computer Society.

García-Hernández, R. A., & Ledeneva, Y. (2009). Word sequence models for single text summarization. In *Advances in Computer-Human Interactions, 2009. ACHI'09. Second International Conferences on* (pp. 44–48). IEEE.

Garz, M., Sorensen, J., & Stone, D. F. (2018). Congeniality and news demand: Evidence from Facebook. *Hamburg Media School* (pp.43). Hamburg Press.

Gillot, C., & Cerisara, C. (2011). Similarity language model. In *Twelfth Annual Conference of the International Speech Communication Association* (pp. 23–30). ACM Press.

Graovac, J. (2014). A variant of N-gram based language-independent text categorization.

*Intelligent Data Analysis*, *18*(4), 677–695.

Grineva, M., Grinev, M., & Lizorkin, D. (2009). Extracting key terms from noisy and multi- theme documents. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 661–670). ACM Press.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques, data management system* (3. ed). Amsterdam: Elsevier ; Morgan Kaufmann.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1262–1273).

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight- space. *Neural Computation*, *1*(1), 143–150.

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic Arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, *41*(1), 114–124.

Hosseinimotlagh, S., & Papalexakis, E. E. (2018). Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. *University of California Riverside* (pp. 1–8). University of California Press.

Houthuys, L., Langone, R., & Suykens, J. A. K. (2018). Multi-view kernel spectral clustering.

*Information Fusion*, *44*(1), 46–56.

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand* (pp. 49–56).

Huang, A., Yi-Fei, W. E. I., & Mei, S. O. N. G. (2017). Visual word based similar image retrieval/optimization by hamming distance. *Sciencepaper Online*, *2*(12), 1–12.

Ifrim, G., Shi, B., & Brigadir, I. (2014). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014* (pp. 12–25). ACM Press.

Ikeda, D., Fujiki, T., & Okumura, M. (2006). Automatically linking news articles to blog entries. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 78– 82). AAAI Press.

Issal, C., & Ebbesson, M. (2010). *Document clustering*. Chalmers University of Technology, Thesis, Sweden.

Jajoo, P. (2008). *Document clustering.* India Institute of Technology Kharagpur, Thesis. India.

Jatowt, A., & Au Yeung, C. M. (2011). Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 1259–1264). ACM Press.

Jeya, T., & Bala, V. M. (2018). Socirank: Identifying and ranking prevalent news topics using social media factors. In *Internal Conference on Next Gen. Tech.* (pp. 13–18). Sengutar Eng. College.

Kanya, N., & Geetha, S. (2013). Information extraction-a text mining approach. In *Proceedings of the 4th International MIDP Colloquium, Bloemfontein, South Africa, Daelemans* (pp. 141–160). Multilingualism and Electronic Language Management*.*

Kar, M. (2016). *Summarization of changes in dynamic text collections*. (Doctoral dissertation, Universidade do Porto, Portugal.

Karol, S., & Mangat, V. (2013). Evaluation of text document clustering approach based on particle swarm optimization. *Open Computer Science*, *3*(2), 69–90.

Kaur, N. (2015). *A combinatorial tweet clustering methodology utilizing inter and intra cosine similarity* (Doctoral dissertation, Faculty of Graduate Studies and Research, University of Regina), Canada.

Khabia, A., & Chandak, M. B. (2015). A cluster-based approach with N-grams at word level for document classification. *International Journal of Computer Applications*, *117*(23), 12– 22.

Kriegel, H. P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., & Zimek, A. (2007).

Future trends in data mining. *Data Mining and Knowledge Discovery*, *15*(1), 87–97.

Kyle, A., Obizhaeva, A., Sinha, N., & Tuzun, T. (2012). News articles and the invariance hypothesis. *CEFRN: Center for Economic and Financial Research at New Economic School*, *34*(3), 1–44.

Lebret, R., & Collobert, R. (2014). N-gram-based low-dimensional representation for document classification. *ArXiv Preprint ArXiv*, *1412.6277*, *10*(12), 1–8.

Li, J., Fan, Q., & Zhang, K. (2007). Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences*, *12*(5), 917–921.

Li, L., Wang, D., Li, T., Knox, D., & Padmanabhan, B. (2011). SCENE: A scalable two-stage personalized news recommendation system. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 125–134). ACM Press.

Lin, Z. H. U., Xiang, Z. H. A. N. G., Wenju, Z. H. A. N. G., Xuhui, H. U. A. N. G., Xuhui, G. U.

A. N., Naiyang, & Zhigang, L. U. O. (2017). Unsupervised domain adaptation with joint supervised sparse coding and discriminative regularization term. *Research Fund for the Doctoral Program of Higher Education of China Foundation, PaperEduCn 2*(5), 1–10*.*

Lwin, M. T., & Aye, M. M. (2017). A modified hierarchical agglomerative approach for efficient document clustering system. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *29*(1), 228–238.

Glasserman, P. & Mamaysky, H., (2017). *Does unusual news forecast market stress?.* Working paper, Office of Financial Research in Columbia University (pp. 1–64). Columbia University Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol.

1, No. 1, pp. 496). Cambridge: Cambridge University Press.

McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability.

*Computational Linguistics*, *42*(2), 246–275.

Mele, I., & Crestani, F. (2017). Event detection for heterogeneous news streams. In *International Conference on Applications of Natural Language to Information Systems* (pp. 110–123). Springer, Cham.

Miao, Y., Kešelj, V., & Milios, E. (2005). Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 357–358). ACM Press.

Mihalcea, R., & Tarau, P. (2005). A language-independent algorithm for single and multiple document summarizations. In *Companion Volume to the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts* (pp.12–20). IJCNLP.

Mitchell, T. M. (1997). *Machine learning* (International ed., [Reprint.]). New York, NY: McGraw-Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*.

Cambridge, Mass: Cambridge University Press.

Moral, C., de Antonio, A., Imbert, R., & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research: An International Electronic Journal*, *19*(1), 1-10.

Naughton, M., Kushmerick, N., & Carthy, J. (2006). Clustering sentences for discovering events in news articles. In *European Conference on Information Retrieval* (pp. 535–538). Springer, Berlin, Heidelberg.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In *International Conference on Intelligence and Security Informatics* (pp. 93–104). Springer, Berlin, Heidelberg.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018). *News and narratives in financial systems: exploiting big data for systemic risk assessment* (pp. 1– 58)*.* Bank of England.

Pande, V. C., & Khandelwal, A. S. (2014). A survey of different text mining techniques.

*IBMRD’s Journal of Management & Research*, *3*(1), 125–133.

Pandit, S., & Gupta, S. (2011). A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science*, *2*(1), 29–31.

Parapar, J., & Barreiro, A. (2009). Evaluation of text clustering algorithms with N-gram-based document fingerprints. In *European Conference in Advances in Information Retrieval* (pp. 645–653). Springer, Berlin, Heidelberg.

Park, S., Lee, K., & Song, J. (2011). Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 340–349). Association for Computational Linguistics.

Pera, M., & Ng, Y. K. (2007). Finding similar RSS news articles using correlation-based phrase matching. In *International Conference on Knowledge Science, Engineering and Management* (pp. 336–348). Springer, Berlin, Heidelberg.

Popovici, R., Weiler, A., & Grossniklaus, M. (2014). On-line clustering for real-time topic detection in social media streaming data. In *SNOW 2014 Data Challenge* (Vol. 8, pp. 57– 63).

Qian, M., & Zhai, C. (2014). Unsupervised feature selection for multi-view clustering on text- image web news data. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management* (pp. 1963–1966). ACM Press.

Qiujun, L. A. N. (2010). Extraction of news content for text mining based on edit distance.

*Journal of Computational Information Systems*, *6*(11), 3761–3777.

Villanes, A., Griffiths, E., Rappa, M., & Healey, C. G. (2018). Dengue fever surveillance in India using text mining in public media. *The American Journal of Tropical Medicine and Hygiene*, *98*(1), 181–191.

Reddy, P. D. K. (2017). A survey on short text conceptualization and clustering. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *2*(4), 859–862.

Reddy, Y. S., & Kumar, A. S. (2012). An efficient approach for web document summarization by sentence ranking. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(7), 7–15.

Rothe, S., & Schütze, H. (2017). AutoExtend: Combining word embeddings with semantic resources. *Computational Linguistics*, *43*(3), 593–617.

Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., & Grobelnik, M. (2016). News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, *55*(2), 283–316.

Saini, A. (2018). An approach to data mining. *Anjali Saini, International Journal of Computer Science and Mobile Applications*, *6*(1), 31–37.

Santhiya, K., & Bhuvaneswari, V. (2018). An automated MapReduce framework for crime classification of news articles using MongoDB. *International Journal of Applied Engineering Research*, *13*(1), 131–136.

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J., & Spiteri, R. (2006). A systematic study of document representation and dimension reduction for text clustering. *Technical Report Technical Report CS-2006-05, Faculty of Computer Science, Dalhousie University Halifax, NS B3H 1W5, Canada* (pp.1–45)*.*

Shah, N., & Mahajan, S. (2012). Document clustering: a detailed review. *International Journal of Applied Information Systems*, *4*(5), 30–38.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge New York, NY; Port Melbourne Delhi Singapore: Cambridge University Press.

Shi, B., Poghosyan, G., Ifrim, G., & Hurley, N. (2018). Hashtagger+: efficient high-coverage social tagging of streaming news. *IEEE Transactions on Knowledge and Data Engineering*, *30*(1), 43–58.

Šilić, A., Chauchat, J., Basic, B. D., & Morin, A. (2007). N-grams and morphological normalization in text classification: A comparison on a Croatian-English parallel corpus. In *Portuguese Conference on Artificial Intelligence* (pp. 671–682). Springer, Berlin, Heidelberg.

Singh, K. N., Devi, H. M., & Mahanta, A. K. (2017). Document representation techniques and their effect on the document clustering and classification: A review. *International Journal of Advanced Research in Computer Science*, *8*(5), 12–20.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 208–215). ACM Press.

Sohangir, S., & Wang, D. (2017a). Improved sqrt-cosine similarity measurement. *Journal of Big Data*, *4*(1), 25–38.

Sohangir, S., & Wang, D. (2017b). Document understanding using improved sqrt-cosine similarity. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on* (pp. 278–279). IEEE.

Sonia, F. G. (2016). A novel committee-base clustering method (pp. 70). *Departamento de Informatica: Pontificia Universidade Catolica Do Rio De Janeiro (Master’s Thesis).* Brazil.

Sowjanya, M. A., & Shashi, M. (2010). Cluster feature-based incremental clustering approach (cfica) for numerical data. *International Journal of Computer Science and Network Security*, *10*(9), 73–79.

Spina, D., Gonzalo, J., & Amigó, E. (2014). Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 527–536). ACM Press.

Svadasa, T., & Jhab, J. (2014). A literature survey on text document clustering and ontology- based techniques. *International Journal of Innovative and Emerging Research in Engineering*, *1*(2), 8–11.

Tao, H., Hou, C., Liu, X., Yi, D., & Zhu, J. (2018). Reliable multi-view clustering. In *AAAI* (pp.

1–8). AAAI Press.

Tao, Y., Christos, F., Dimitris, P., & Bin, L. (2004). Prediction and indexing of moving objects with unknown motion patterns. In *Proceedings of the 2004 ACM Sigmod International Conference on Management of Data* (pp. 611–622). ACM Press.

Toda, H., & Kataoka, R. (2005). A clustering method for news articles retrieval system. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web* (pp. 988–989). ACM Press.

Wang, H., Zhang, F., Xie, X., & Guo, M. (2018). DKN: Deep knowledge-aware network for news recommendation. *ACM*, *2*(1), 1–10.

Wei, B., Liu, J., Ma, J., & Zheng, Q. (2017). DOTE: Automatic domain-specific term extraction from Wikipedia. *Sciencepaper Online*, *12*(1), 1–12.

Witten, I. H. (2004). *Text mining.*CRC Press.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*

(2. ed). San Francisco, Calif.: Kaufmann [u.a.].

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.

Xia, Y., Tang, N., Hussain, A., & Cambria, E. (2015). Discriminative bi-term topic model for headline-based social news clustering. In *FLAIRS Conference* (pp. 311–316). ACM Press.

# LIST OF PUBLICATIONS

1. **Desmond Bala Bisandu, Rajesh Prasad, & Musa Muhammad Liman.** “Clustering News Articles using Efficient Similarity Measure and N-grams”. *Int. J. of Knowledge Engineering and Data Mining(IJKEDM).* Status: Communicated.
2. **Desmond Bala Bisandu, Rajesh Prasad, & Musa Muhammad Liman**. “Data Clustering using Efficient Similarity Measures: A Critical Review”. *Journal of Science and Technology (Pertanika Journals)*. Status: Communicated.
3. **Musa Muhammad Liman, Rajesh Prasad, Mathias Fonkam, & Desmond Bala Bisandu**. “A Framework for Text Mining using String Matching and Bit-Parallelism”. In *14th International Conf. on Data Science (ICDATA’18 : July 30th – August 2nd , 2018, LAS VEGAS, USA).* Status: Accepted.

# APPENDIX CODING

###########LoadingLibrariesNeededInRVersion 3.4.3################ library("tm")

library("quanteda") library("RWeka") library("proxy")

library("factoextra")#DeterminingtheBestValueofK library("NbClust")#DeterminingtheBestValueofK library("stream")

library("ngram") library("reshape2")

###################ReadingDataset############################

setwd("C:/Users/DESMOND/Desktop") reuters<-read.csv("newsCorpora.csv") #dim(reuter)

options(header=TRUE, stringsAsFactors = FALSE,FileEncoding="latin1") setwd("C:/Users/DESMOND/Desktop")

#text <- readLines("vocab.enron.txt") reuters <- Corpus(VectorSource(text))

reuters<- Corpus(DirSource("C:/Users/DESMOND/Desktop/XMLdataset"), #readerControl=list(reader=readReut21578XML))

reut21578 <- system.file("texts", "crude", package = "tm") reuters <- VCorpus(DirSource(reut21578, mode = "binary"),

readerControl =list(reader = readReut21578XMLasPlain)) #######################N-gramspreprocessingmethod###################### #reuters<-preprocess(reuters, case = "lower", remove.punct = FALSE,

#remove.numbers = FALSE, fix.spacing = TRUE)

#reuters #inspect(reuters)

#####################NGramsFunction###########################

NGram <- function(x) NGramTokenizer(x, Weka\_control(min = 1, max = 1))

##########PreprocessingStage#TermNormalization################ tdm <-TermDocumentMatrix(reuters,

control = list(tokenize = NGram, weighting = function(x)

weightTfIdf(x, normalize = TRUE), stopwords = FALSE,

fix.spacing = TRUE)) #inspect(tdm)

#######################TransformToMatrix######################

matrx <- as.matrix(tdm) matrx<-scale(matrx) #matrx

##############K=2-10WSSToDeterminedBestValueOfK##########3####

#set.seed(123) #k.max <- 10 #data <-matrx

#wss <- sapply(1:k.max,

#function(k){kmeans(data, k, nstart=10 )$tot.withinss}) #plot(1:k.max, wss,

#type="b", pch = 19, frame = FALSE, #xlab="Number of clusters K",

#ylab="Total within-clusters sum of squares") #wss<-fviz\_nbclust(data, kmeans, method = "wss") + #geom\_vline(xintercept = 1:10, linetype = 2) #plot(wss)

################DistanceMeasurementFunction################### Isc<-function(x,y){

similarity<-(sum(sqrt(x\*y)))/(sqrt(sum(x)) \* sqrt(sum(y))) return((similarity) \* 180 / pi)} ######################PairwiseDistance########################

pr\_DB$set\_entry(FUN = Isc, names = c("Isc") ) d1 <- dist(matrx, method = "Isc") #pr\_DB$delete\_entry("Isc")

#d1<-dist(matrx, method = "Hellinger") #d1

#summary(pr\_DB, "long")

######################K-MeansClustering#######################

clust<-kmeans(d1, 3, nstart=25)

clust

#matrx<-matrx[,-20]

clusplot(matrx, clust$cluster,color=TRUE,main="Proposed Clustering Solutions Result with N=2",shade=TRUE,col.p = clust$cluster)

######################EVALUATIONMETRIC######################## ######################ClustersPurity##########################

clust<- DSD\_Gaussians(k=4, d=2)

matrx <- DSC\_DStream(gridsize=0.05, Cm=1.5) update(matrx, clust, 500)

#plot(dstream, stream, xlab = "Principal Component 1",

#ylab = "Principal Component 2", main="Proposed Clustering Solutions Result") evaluate(matrx, clust, measure=c("purity"), n=100)

evaluate(matrx, clust, measure=c("entropy"), n=100)

############################################################## #########################ClusterAccuracy######################

#evaluate(matrx, clust, measure=c("precision"), n=100) #evaluate(matrx, clust, measure=c("recall"), n=100) evaluate(matrx, clust, measure=c("F1"), n=100)

######################LineGraphPlot###########################

# Create Line Chart

# convert factor to numeric for convenience #Orange$Tree <- (0.31,0.34,0.39,0.29)

#Orange$Tree

#ntrees <- 5 #ntrees

# get the range for the x and y axis #xrange <- range(1:4)

#yrange <- range(0.0:0.9) # set up the plot

#plot(xrange, yrange, type="n", xlab="N-grams", #ylab="Purity" )

#colors <- rainBoW(ntrees) #linetype <- c(1:ntrees) #plotchar <- seq(5,5+ntrees,1) # add lines

#for (i in 5:ntrees-3) {

#tree <- subset(Orange, Tree==i) #lines(c(1,2,3,4),c(0.56,0.59,0.92,0.41), type="b", lwd=1.5,

#lty=linetype[i], col=colors[i], pch=plotchar[i]) #}

# add a title and subtitle

#title("Purity against N-grams Plot", "Number of grams") # add a legend

#legend(xrange[1], yrange[2], 1:2, cex=0.4, col=colors, #pch=plotchar, lty=linetype, title="Keys")

#################################################################### #########################EndOfCode############################

#chart\_data <- melt(matrx, id='x') #names(chart\_data) <- c('x', 'func', 'value') #ggplot() +

#geom\_line(data = chart\_data, aes(x = x, y = value, color = func), size = 1)+ #xlab("N-grams") +

#ylab("Accuracy") #x <- c(1,2,3,4,5,6)

#y <- c(3,5,2,4,1,4)

#ggplot(data=chart\_data, aes(x= "N-grams", y= "Accuracy", group = "Agencia\_ID", colour = as.factor("Agencia\_ID")))

#geom\_line()

#x <- c(1,2,3,4,5,6)

#y <- c(3,5,2,4,1,4)

#z <- c(2,3,4,3,2,1)

#plot(x,y,type="b") #lines(x,z,col="red",type="b") #x <- c(1,2,3,4,5,6)

#y <- c(3,5,2,4,1,4)

#z <- c(2,3,4,3,2,1)

#plot(x,y)