Content-based Document Classification using Soft Cosine Measure

By

Md. Shohel Rana

181-25-662

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering.

Supervised By

Md. Zahid Hasan

Assistant Professor & Coordinator of MIS

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2018

APPROVAL

This Project titled **"Content-based Document Classification using Soft Cosine Measure",** submitted by Md Shohel Rana (ID:181-25-662) to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Assistant professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor & Coordinator of MIS
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Muhammad Shrift Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I am declaring that, this project has been done by me under the supervision of **Md. Zahid Hasan, Assistant Professor, Department of CSE,** and Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Md. Zahid Hasan

Assistant Professor and Coordinator of MIS

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Submitted by:

_____

Md. Shohel Rana
ID: 181-25-662
Department of CSE,
Daffodil International University

ACKNOWLEDGEMENT

First of all, our heartiest thanks and gratefulness to Almighty Allah for His divine blessing that makes us capable to complete this project successfully.

We would like to thanks to our honorable teacher & project supervisor **Md. Zahid Hasan, Assistant Professor, Department of CSE,** Daffodil International University for his endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Syed Akhter Hossain**, Head, Department of CSE, for his kind help to finish our project and we are also thankful to all the other faculty and staff members of our department for their co-operation and help.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Document classification is a deep-rooted issue in information retrieval, and it assumes an imperative part in an assortment of applications for an effective management of text and substantial volumes of unstructured data. Automatic document classification can be defined as a content-based assignment of some predefined categories to documents which is for sure less demanding to fetch the relevant data at the right time and for filtering and steering documents directly to users. For recovering data effortlessly at the minimum time, scientists around the globe are attempting to make content-based classifiers and an assortment of classification framework has been developed. Regardless, none of the classification methods is enough effective in light of the fact that they used some conventional algorithms. However, this paper proposes the Soft Cosine Measure as a content-based classification method. This classification method considers the similarity of features in a vector space model rather than considering the features as independent or completely different like all the existing traditional frameworks. For example, the proposed method considers ‘emperor’ and ‘king’ as the same word where all the remaining systems consider these as two different words. Besides, both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) algorithms are used to train the system which confirms the classification accuracy up to 98.06%.

TABLE OF CONTENTS

List of Abbreviations

TF - Term Frequency

IDF - Inverse Document Frequency

CS - Cosine Similarity

SCM - Soft Cosine Measure

ROC - Receiver operating characteristic

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1. Introduction

Document classification refers to the way toward keeping the comparative documents together. This is considered as a major challenge for Information Retrial in light of the fact that getting right documents at the right time is absolutely unimaginable if the records are not legitimately classified and sorted out. Document classification can be done intellectually or automatically. Content-based document classification is an automatic document arrangement where the framework experiences the entire content and groups them in light of the extracted features. There are several well-established algorithms for content-based classifications. But none of these algorithms are fit for dealing with the similitude of the two words meaning the same. So, a much complex algorithm is required equipped for considering the closeness of features in a vector space model (VSM) [1]. The framework proposed in this paper utilizes **Soft Cosine Measure** which finds the likeness of features in VSM.

The proposed framework utilizes both TF [2] and TF-IDF [3] to find the most important words in contents with the goal that no vital term is missed. Subsequently, the framework furnishes with the most precise outcomes and it performs 11.2% superior to Cosine Similarity [4]. The proposed system is tested for 103 times and almost every time classified the document correctly.

1.2. Research Aims and Objectives

The aim of this research is to develop a method that will increase the accuracy of cosine similarity method. In order to develop a system, it is necessary to develop a system that will help to classify the document for government document, cv ranking for human resource management. The system is implemented using Java in NetBeans IDE.

1.3. Report Layout

The report is composed as pursues:

Chapter 1 Introduction

Chapter 2 State of The Art

Chapter 3 Relationship Between Soft Cosine Measure and Cosine Similarity

Chapter 4 Feature Extraction

Chapter 5 Experimental Results

1.4. Summary

This chapter introduces the facts including the report. The proposed systems introduction is also viewed here. The following chapter will provide an overview of existing work related to the proposed system.

CHAPTER 2

STATE OF THE ART

This chapter covers the existing works or related works that have been done to accuracy of my proposed system SCM. The importance of the proposed system is explained in this chapter. How the proposed system is more convenient than the existing system is also discussed in this chapter.

2.1. Existing Works

The evolution of document classification has started a long ago but still, it's a far away from getting saturated. Researchers have been applying various mathematical models to boost a sophisticated document classifier and, in that consequence, a number of documents classification frameworks have been established.

C. Goutte, L. Versoud, E. Gaussier, Eybens used Probabilistic Hierarchical Model for text categorization [5] where Evgeniy Gabrilovich and Shaun Markovitch claimed to increase the classification accuracy by using Support Vector Machine (SVM) [6]. Dieter Merkl introduced an Artificial Neural Network for content-based text classification [7]. Y. H. Li and A. K. Jain conducted an experiment over Naive Bayes Classifier, the Nearest Neighbor Classifier, Decision Trees and a Subspace Method [8] to find out the best-fit algorithm for document classification. On the other hand, Mrs. Sanjivani Tushar Deokar suggested the K-means algorithm [9] while Janani Balakumar proposed an improved Bisecting K-means algorithm for Text Document Clustering [10]. Later, in 2017 S. Adinugroho, Y. A. Sari, M. A. Fauzi, and P. P. Adikara proposed semantic indexing and pillar algorithm to optimize K-means document clustering approach [11]. In the meantime, in 2016, P. Bafna, D. Pramod and A. Vaidya investigated a new document clustering method, TF-IDF [12] and found a more satisfactory result. Some researchers believe, Cosine Similarity ensures the maximum classification accuracy [13]-[15].

Unfortunately, none of these classifiers could overcome some common drawbacks which lead the scientist to develop a content-based document classification framework, a well more intelligent classification model, equipped for classifying any text documents just by experiencing its content.

2.2. Summary

The propose system overcomes the lacking of existing systems. This chapter discusses about all the existing systems and shows how the proposed system is convenient over the existing systems. In the next chapter explanation of SCM is given in details.

CHAPTER 3

AN INVESTIGATION INTO THE RELATIONSHIP BETWEEN SOFT COSINE MEASURE AND COSINE SIMILARITY

3.1. Introduction

Soft Cosine Measure, a new concept in classification considers the pairs of features [16] to discover the similitude between two vectors in a vector space model (VSM) [17]. Although Soft Cosine Measure Soft Cosine has derived from Cosine Similarity, there is a major distinction between these two concepts. Cosine comparability ordinarily considers the cosine of the angle between two non-zero vectors to discover the similitude between them [18] where Soft Cosine Measure calculates comparability between features for computation of similarity of objects in a Vector Space Model (VSM) [16]. For, two N-dimension vectors α and β the Soft Cosine Similarity can be calculated as follows:

Soft Cosine (α,β) = $\frac{\sum_{i,j}^{N} s_{ij}\alpha_i\beta_j}{\sqrt{\sum_{i,j}^{N} s_{ij}\alpha_i\alpha_j}\sqrt{\sum_{i,j}^{N} s_{ij}\beta_i\beta_j}}$; Where, $s_{i,j}$= similarity ($feature_i$, $feature_j$)

If,$s_{i,i}$=1 and $s_{i,j}$= 0 for i ≠ j then, Soft Cosine (α,β) = $\frac{\sum_{i,j}^{N} \alpha_i\beta_i}{\sqrt{\sum_{i,j}^{N} \alpha_i\alpha_i}\sqrt{\sum_{i,j}^{N} \beta_i\beta_i}}$

$= \frac{\sum_{i=1}^{N} \alpha_i\beta_i}{\sqrt{\sum_{i=1}^{N} {\alpha_i}^2}\sqrt{\sum_{i=1}^{N} {\beta_i}^2}}$

$= \frac{\alpha \,.\, \beta}{\|\alpha\|\|\beta\|}$ = Cosine Similarity.

So, when there is no similarity between the features of the objects, Soft Cosine Measure winds up proportional to the regular Cosine Similarity formula.

3.2. Data Processing

To classify any document into a given number of categories, the framework should be trained with a legitimate set of data. To ensure the maximum accurate feedback from any system, the training data should properly be preprocessed. Data processing in the proposed system is done in a couple of steps introduced in the following diagram.

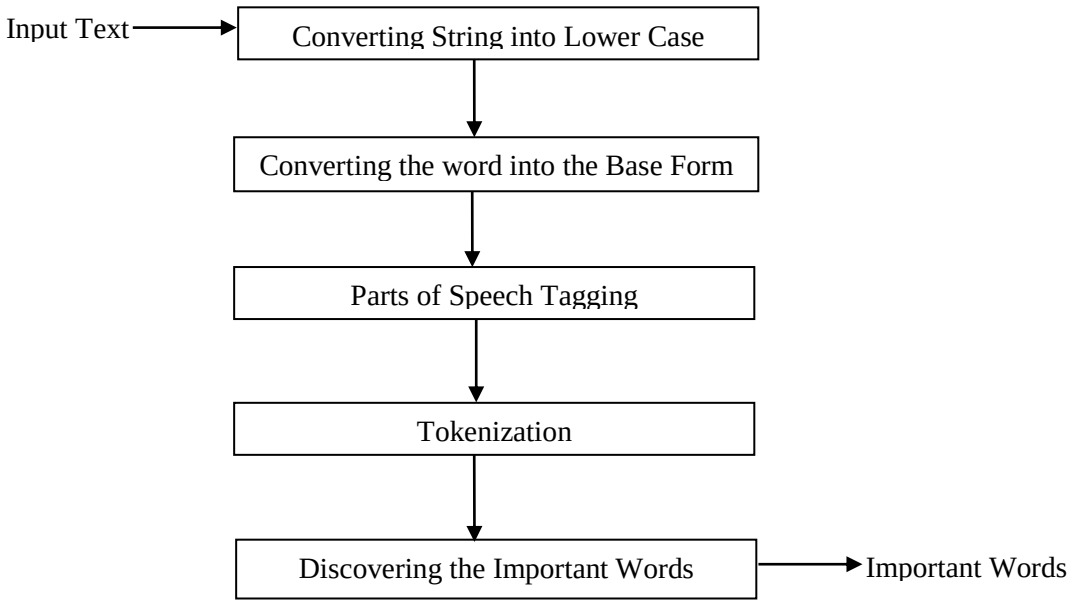


Fig 1: Data Processing Diagram

3.3 Removing Punctuation

Punctuation moving is a vital task in natural language processing. There are quantities of approach to expel punctuation from a text. In the proposed system a regular expression is used to dispose of all the punctuations.

3.4 Converting String into Lowe Case Letter

For processing the data in the most convenient way all the letters in the textual content has been converted into the lower-case form.

3.5 Converting the word into the Base Form

A critical errand in natural language processing is to convert all the words into their base form. This causes a framework to comprehend words regardless of whether they are in various structures. For instance,

Word in different Form	Base Word
ran, run, running, runs, runner	run
good, better, best	good

Table 1: Words in various forms and their base form

To convert the words into their base form, Streamer Porter Algorithm [19], [20] is utilized in this system.

3.6 Parts of Speech Tagging

The system proposed in this paper has used Latent Analogy [21] for tagging parts of speech. Only noun and verbs are used to train the system.

3.7 Tokenization

The process of breaking up the content into distinct meaningful units is recognized as tokenization.

Tokenization is an important task for this system as the system makes the vector with some particular words or tokens.

3.8 Discovering the Important Words

All the words in a text file are not equally important for a specific purpose. So, researchers have taken this task to as a challenge to find the important words out from a text document. As a result, various algorithms have been developed to discover the vital words from content. However, the proposed framework utilizes two most proficient algorithms to choose the critical word: (TF) and TF-IDF.

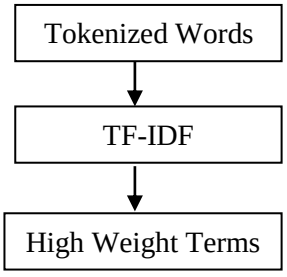


Fig 2: Important Words Acquisition

3.8.1 Term Frequency (TF) and Tern Frequency Inverse Document Frequency (TF-IDF)

Term Frequency [2] is the calculation of how many times each word appears in a text document.

Term Frequency (TF) in documents can be calculated by using the logarithmic scale.

$tf_{t,d} = \begin{cases} \log(1 + f_{t,d}), & if\ f_{t,d} > 0 \\ 0, & otherwise \end{cases}$; where, t defines a term, d is document and $tf_{t,d}$ is the frequency of the term in the documents.

Inverse Document Frequency [22], [23] is the calculation that determines whether a term is common or rare across all the documents. Inverse Document Frequency can be easily calculated by the following formula.

$idf_{t,D} = \log \frac{|D|}{|d \in D : t \in d|}$; where, d is a document and D is the set of documents.

Tern Frequency Inverse Document Frequency [3] calculates the high weighted terms in a set of documents.

$tf - idf_{t,d,D} = tf_{t,d} * idf_{t,D}$; where, t is a term, d is a document and D is a set of documents.

The implementation of TF-IDF mimics the following algorithm:

1. Set d ←Text Document, t← a specific term; t € d
2. Set n ←Number of times term t appears in d, m ←Total number of terms in d;
3. Compute X ← n ÷ m;
4. Set N← Total number of documents, M← Number of documents with t; M € N
5. Compute Y← N ÷ M;
6. Set K ← Term Frequency Inverse Document Frequency
7. Compute K← X × Y;
8. Print K;

CHAPTER 4

FEATURE EXTRACTION

4.1. Introduction

The basic convenience of Soft Cosine Measure over Cosine Similarity is its capability of computing similarity between two documents by using their inner features regardless of whether they don't have any physical comparability. For doing that, Soft Cosine Measure discovers features for each vital term in all the documents in an archive set. For that purpose, it uses a dictionary that profits all the possible words with the same meaning for a given word. This process is utilized at both training time and system handling time. The entire procedure is graphically spoken to in the accompanying segment.

4.2 Feature Extraction during System Training:

For preparing the framework, vital terms are gathered from a substantial amount of data set which indicates a solid plausibility of having duplicate terms. So, it becomes essential to extract feature at data preparing time to remove redundant High Weigh Terms. The imperative terms are principally put away in a cluster and the accompanying algorithm is utilized to evacuate all the redundant data.

1. Start
2. set i to 0
3. if A[i+1]==A[i]
   3.1 remove A[i+1]
   3.2 else Look for A[i] in Lexicon and put all possible features of A[i] into B[j]
      3.2.1 set j to 0
      3.2.2 if A[i+1]==B[j]
         3.2.2.1 remove A[i+1]
      3.2.3 else set j to j+1;
   3.3 set i to i+1;
4. Stop

4.3 Feature Extraction During Similarity Scoring

Feature extraction during classification is the key errand that enables the proposed framework performs better than all other existing frameworks. This task is performed in accordance with the following algorithm.

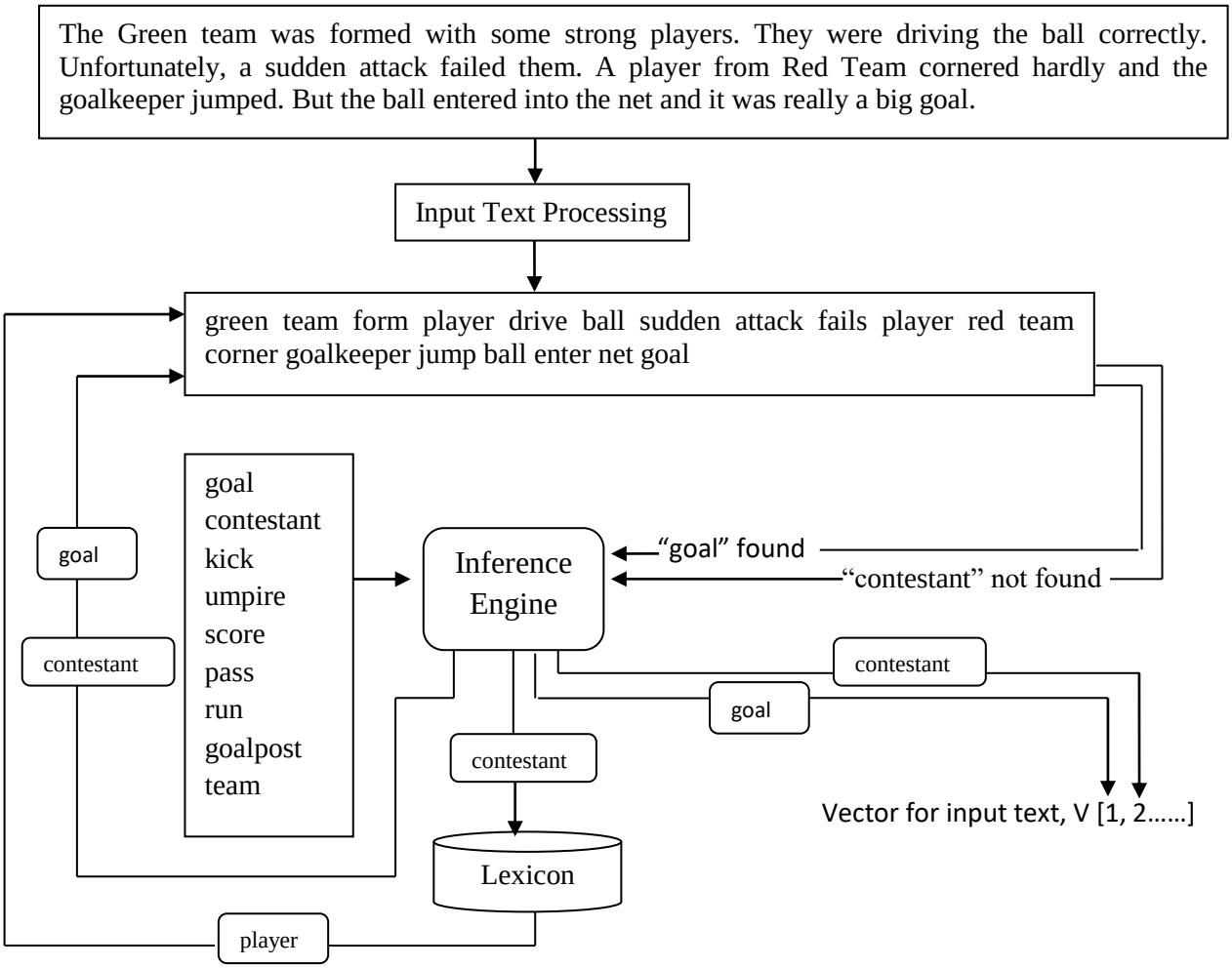


Fig 3: Feature Extraction Procedure

4.4 Feature Vectors Construction

In the wake of finding the essential words, the framework figures the Term Frequency for those words and builds the final vectors with the resultant numeric qualities. The following algorithm demonstrates the vector construction process.

1. Start
2. Declare High Weight Terms as $T_{i\,;\,1\leq i\leq n}$
3. Declare The document as $d_{i\,;\,1\leq i\leq n}$
4. Compute Term Frequency, TF forT_1 in Document, d_1
5. Extract Features for T_1 in d_1
   5.1 Add all TF for T_1
   5.2 Rerun the Sum
6. Repeat Step 4 for n times
7. Store the result in a vector, V
8. Repeat step 3 and 4 for n times
9. End

4.5 Implementation of the System

Soft Similarity or Soft Cosine Measure classifies text documents in view of the contents it bears. For that, only a cosine angle between the features of the trained data and the input data is estimated. An architectural design of how Soft Cosine Measure classifies the text documents is presented in bellow.

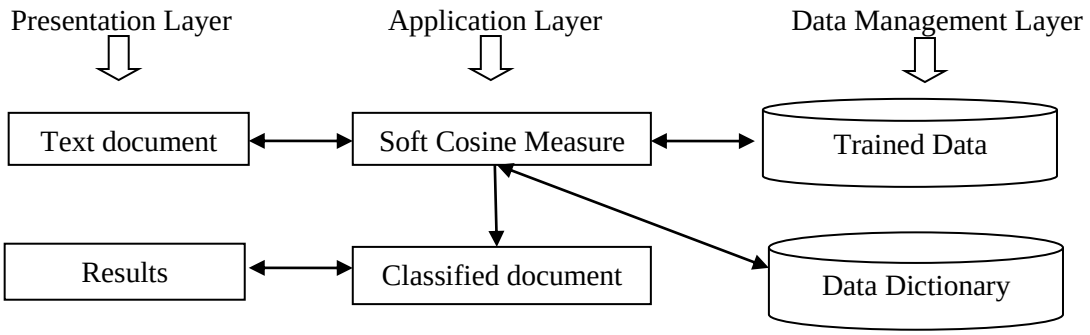
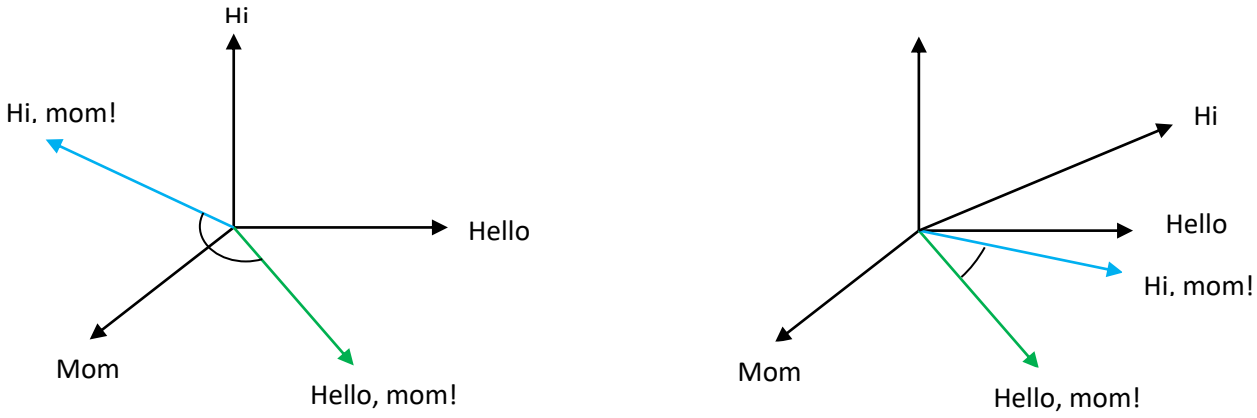


Fig 4: System Architecture for Document Classification

Soft Cosine Measure considers the features in VSM which makes it equipped for figuring the likenesses between two documents regardless of whether they do not have any word in

common. It uses a lexicon to extricate the features that it actually takes into account to quantify the similarity between the meaning of two words rather than the words themselves.



Cosine Similarity Soft Cosine Measure

Fig 5: Performance Comparison between CS and SCM

Based on the similarity scores, performed by different text documents, the system classifies the documents into some the predefined classes. A very straightforward algorithm for the proposed system is given below.

1. Start
2. Scan the input text, T
3. Process T
4. Extract the feature of T from the lexicon, D
5. Make a feature vector, V_i
6. Find the similarity score between V_i and V_t; $\{V_t \mid V_t \in V;\ |V| = n\}$ wh V= trained data vector and n= number of classes
7. Store the score in a list, L
8. Repeat Step 6 and 7 for n times
9. Find the biggest score from L
10. Make the final decision to put T into S_t; $\{S_t \mid S_t \in S;\ |S| = |V|\}$ S= Set of classes.
11. End

4.6. Summary

The proposed system is implemented in java. Different parts of the implementation are described in this chapter. The next chapter will discuss about the results of the system.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Introduction

A number of experiments have been conducted over the proposed system and each of these experiments confirms that Soft Cosine Measure performs better than Cosine Similarity. The similarity scores of Cosine Similarity and Soft Cosine Measure between two sample documents are presented in the next table.

Document 1	Document 2	Cosine Similarity Score	Soft Cosine Measure Score
Every Mom is the most amazing person for her children - she's their heroine. Mom always knows how her child feels and can help with any problem. Mom can make the most complicated braid and explain fractions; Mom can help to wake her child up in the morning and hug her tightly when she's sad.	Each Mom is the most stunning individual for her kids - she's their champion. Mother dependably knows how her kid feels and can help with any issue. Mother can make the most entangled mesh and clarify divisions; Mom can get her kid up toward the beginning of the day and embrace her firmly when she's pitiful.	0.5041	0.9997

Table 2: Similarity Score Comparison between CS and SCM

The above table shows that Soft Cosine Measure performs 49.56% better than the Cosine Similarity. This experiment has been done for 50 distinct documents and each time Soft Cosine Measure performs around 45% better than Cosine Similarity.

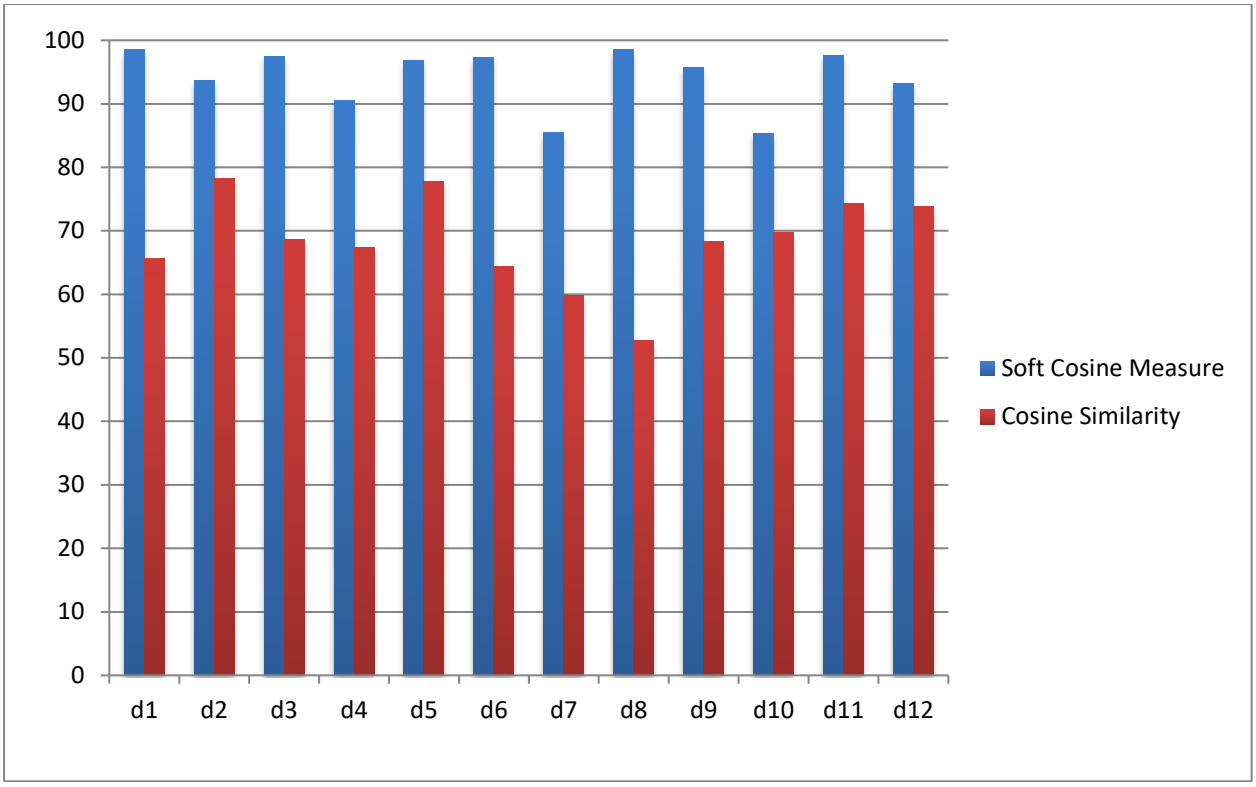


Fig 6: Similarity Scores of Soft Cosine Measure and Cosine Similarity for 12 different documents

However, the proposed system has been tested for 114 times with 103 different documents to classify into 5 categories. Each time the system has classified 101 documents correctly which secures its accuracy rate up to 98.06%.

Documents	System Selected Category	Actual Category	Remark
Doc1	History	History	✓
Doc 2	History	History	✓
Doc3	Literature	Literature	✓
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
Doc48	Comics	Comics	✓
Doc49	Politics	History	X
Doc50	Science	Science	✓
Doc51	Science	Science	✓
Doc52	Literature	Literature	✓
Doc53	History	Politics	X
.	.	.	.
.	.	.	.
.	.	.	.
.			.
Doc100	Comics	Comics	✓

Doc101	History	History	✓
Doc102	Literature	Literature	✓
Doc103	Politics	Politics	✓

Table 3: Document Classification Results of Soft Cosine Measure

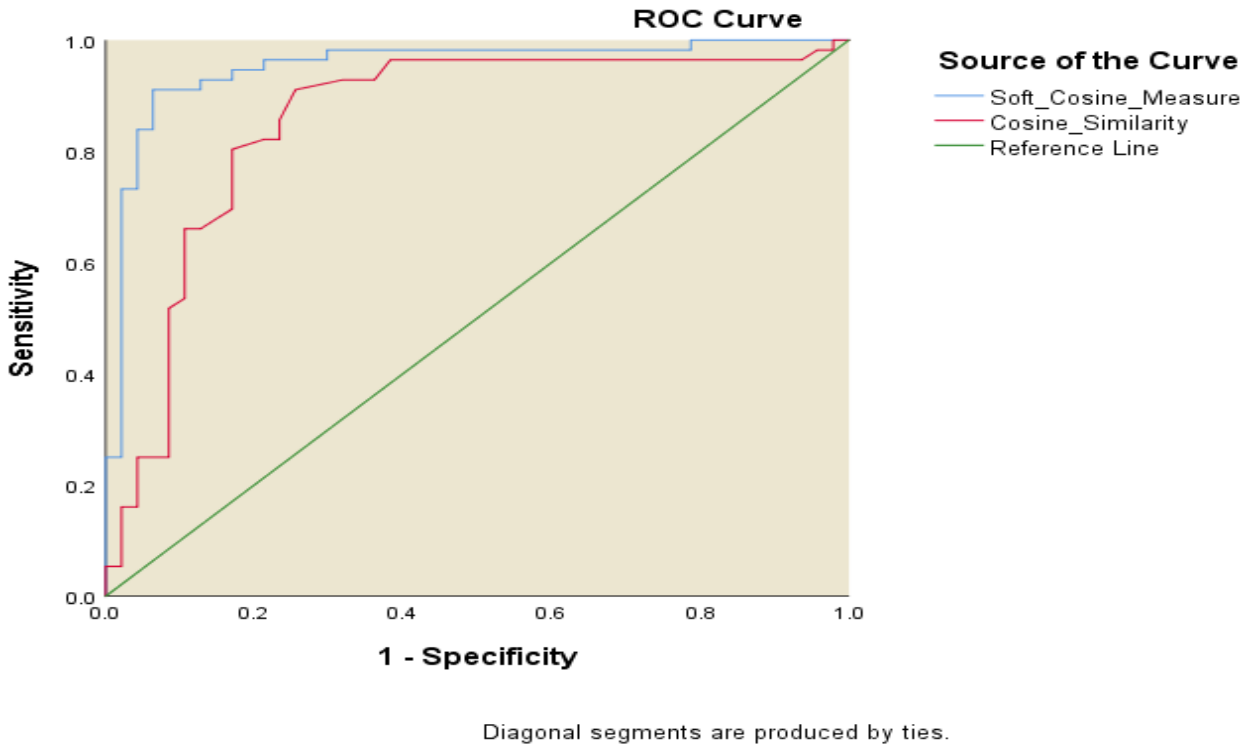


Fig 7. ROC curves illustrate the classification accuracy of Soft Cosine Measure and Cosine Similarity

The above figure shows two ROC (Receiver Operating Characteristics Curve) curves. The blue curve represents the classification accuracy of Soft Cosine Measure where the red curve depicts the accuracy of Cosine Similarity. The AUC (area under the curve) of Soft Cosine Measure is 0.925 where AUC of Cosine Similarity is 0.840. So, it turns out to be evident that, the classification accuracy of Soft Cosine Measure is superior to the Cosine Similarity.

5.2 Results Analysis

Though document classification is a very important task in natural language processing for its extended use case, it is yet a big challenge to come across the most intense precision in document classification. Researchers have tried numerous strategies to locate the most extreme precision in content-based document classification. According to a contemporary research outcome [24], it has been clear to researchers that Support Vector Machine (SVM)

provides the maximum accurate result than any other methods in document classification which is estimated 90.26%. But another recent research demonstrates that cosine similarity classifies content-based document more efficiently that SVM and its accuracy reaches up to 93.9% [4]. However, this research with adequate evidence clarifies that Soft Cosine Measure performs better than Cosine Similarity. A very straightforward comparison among different classification accuracy is given in the bellow table.

Methodology	Accuracy (%)
SVM	90.26
Decision Tree	76.99
K Nearest Neighbor	84.60
Naive Bayes	84.70
Cosine Similarity	93.90
Soft Cosine Measure	**98.30**

Table 4: Classification accuracy of different Methods

For better understating of the system accuracy, a graphical presentation of is also provided in the next section.

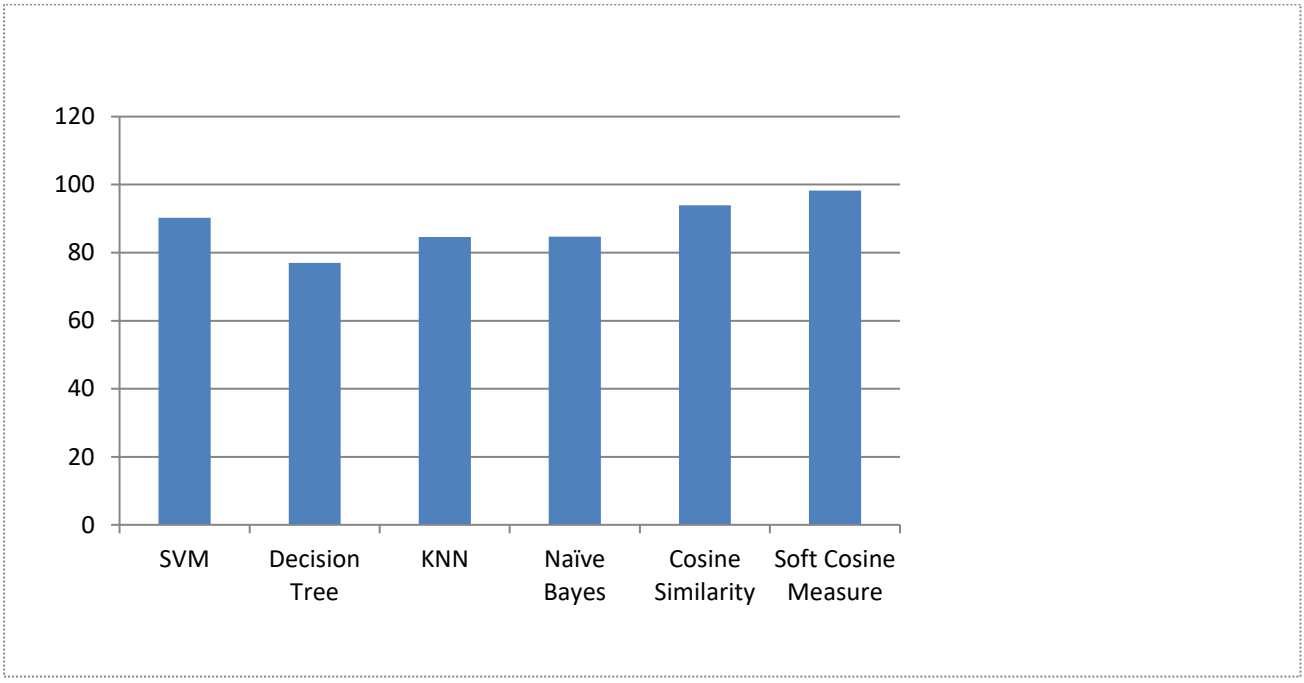


Figure 8: Classification Accuracy of Different Methodologies

5.3 Conclusion

Soft Cosine Measure is a state-of-the-art mathematical model that considers the features in a vector space model to quantify the comparability between two text documents. The proposed system uses this mathematical model to construct a content-based document classification framework. To classify any document the system considers the edge between the component vectors of the given documents and the readied data. The system secures its precision rate up to 98.3% which is vastly improved than some other existing framework.

5.4 Future Work

In future, the limitations that come out from this system will be removed and a better system will be developed. A rich knowledge base will be developed with a huge number of rule bases in the next version of the system. In next version, the system will be more accurate, robust and powerful. The system errors of this research will be removed in future system by a training module. An online ranking system will be developed in future.

References

[1] D.L. Lee, HueiChuang,K.Seamons,"Document ranking and the vector-space model", IEEE Software ,Volume: 14, Issue: 2, Mar/Apr 1997 ,DOI: 10.1109/52.582976

[2] Mikio Yamamoto, Kenneth W. Church, "Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus",Computational Linguistics archiveVolume 27 Issue 1, March 2001,Pages 1-30, doi:10.1162/089120101300346787

[3] Rung-Ching Chen, Jui-Yuan Liang, Ren-Hao Pan, "Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency" Expert Systems with Applications, Volume 34, Issue 1, January 2008, Pages 488-501

[4] Radha mothukuri, Nagaraju.M, DivyaChilukuri, "SIMILARITY MEASURE FOR TEXT CLASSIFICATION ", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 5, Issue 6, November - December 2016 ,ISSN 2278-6856

[5] C. Goutte, L. Versoud, E. Gaussier, Eybens, "Method forMulti-class, multi-label categorization using probabilistic hierarchical modeling", U.S Patent 7 139 754 B2, Nov 21, 2006

[6] Evgeniy Gabrilovich and Shaun Markovitch, "Text Categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5", Proceedings of 21st International Conference on Machine Learning, 2004.

[7] Dieter Merkl,"CONTENT-BASED DOCUMENT CLASSIFICATION WITH HIGHLY COMPRESSED INPUT DATA",Proc. 5th Int'l Conference on Artificial Neural Networks (ICANN'95). Paris. Oct 9-13. 1995. pp Vol 2: 239-244

[8] Y. H. Li A. K. Jain,"Classification of Text Documents", The Computer Journal, Volume 41, Issue 8, 1 January 1998, Pages 537–546, https://doi.org/10.1093/comjnl/41.8.537

[9] Sanjivani Tushar Deokar., International Journal of Technology and Engineering Science [IJTES] TM Vol 1 (4), pp 282 – 286, July 2013

[10] Janani Balakumar, "An Improved Bisecting K-means Algorithm for Text Document Clustering"International Journal of Knowledge Based Computer Systems,Volume 4 Issue 2,2016

[11] S. Adinugroho, Y. A. Sari, M. A. Fauzi and P. P. Adikara, "Optimizing K-means text document clustering using latent semantic indexing and pillar algorithm," 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), Dubai, 2017, pp. 81-85.,doi: 10.1109/ISCBI.2017.8053549

[12] P. Bafna, D. Pramod and A. Vaidya, "Document clustering: TF-IDF approach," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 61-66. doi: 10.1109/ICEEOT.2016.7754750

[13] L. Muflikhah and B. Baharudin, "Document Clustering Using Concept Space and Cosine Similarity Measurement," 2009 International Conference on Computer Technology and Development, Kota Kinabalu, 2009, pp. 58-62. doi: 10.1109/ICCTD.2009.206.

[14] [Li B., Han L. (2013) Distance Weighted Cosine Similarity Measure for Text Classification. In: Yin H. et al. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer, Berlin, Heidelberg

[15] M. L. Aishwarya and K. Selvi, "An intelligent similarity measure for effective text document clustering," 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-5.doi: 10.1109/ICCTIDE.2016.7725342

[16] *Sidorov, Grigori; Gelbukh, Alexander; Gómez-Adorno, Helena; Pinto, David.* "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model". *Computación y Sistemas.* ***18*** *(3): 491–504. doi:10.13053/CyS-18-3-2043*. Retrieved 7 October 2014.

[17] *J*ThomasMikolov et al. Efficient Estimation of Word Representations in Vector Space ,arXiv:1301.3781v3 [cs.CL] 7 Sep 2013

[18] Li B., Han L. (2013) Distance Weighted Cosine Similarity Measure for Text Classification. In: Yin H. et al. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2013. IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer, Berlin, Heidelberg

[19] M.F. Porter, (1980) "An algorithm for suffix stripping", Program, Vol. 14 Issue: 3, pp.130-137, https://doi.org/10.1108/eb046814

[20] Atharva Joshi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 266-26

[21] Jerome R. Bellegarda, Part-of-Speech Tagging by Latent Analogy IEEE Journal of Selected Topics in Signal Processing; Volume: 4, Issue: 6, Dec. 2010 ;DOI: 10.1109/JSTSP.2010.2075970

[22] Church K., Gale W. (1999) Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. In: Armstrong S., Church K., Isabelle P., Manzi S., Tzoukermann E., Yarowsky D. (eds) Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology, vol 11. Springer, Dordrecht,https://doi.org/10.1007/978-94-017-2390-9_18

[23] Stephen Robertson, (2004) "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of Documentation, Vol. 60 Issue: 5, pp.503-520, https://doi.org/10.1108/00220410410560582

[24] Choudhury, S., Batra, T., Hughes, C., & LEMMATIZER, L. (2016). Content-based and link-based methods for categorical webpage classification.