Shahjalal University of Science and Technology
Department of Computer Science and Engineering

Bangla Document Categorization Using Term Graph

Md Nazim Uddin
Reg. No.: 2013331038
4th year, 1st Semester

Moudud Ahmed Khan
Reg. No.: 2013331004
4th year, 1st Semester

Department of Computer Science and Engineering

Supervisor
Enamul Hassan
Lecturer
Department of Computer Science and Engineering

March 17, 2018

Bangla Document Categorization Using Term Graph

A Thesis submitted to the

Department of Computer Science and Engineering

Shahjalal University of Science and Technology

Sylhet - 3114, Bangladesh

in partial fulfillment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering

By

Md Nazim Uddin
Reg. No.: 2013331038
4th year, 1st Semester

Moudud Ahmed Khan
Reg. No.: 2013331004
4th year, 1st Semester

Department of Computer Science and Engineering

Supervisor

Enamul Hassan

Lecturer

Department of Computer Science and Engineering

March 17, 2018

Recommendation Letter from Thesis Supervisor

The thesis
entitled " Bangla Document Categorization Using Term Graph"
submitted by the students

1. Md Nazim Uddin
2. Moudud Ahmed Khan

is a record of research work carried out under my supervision and I, hereby, approve that the report be submitted in partial fulfillment of the requirements for the award of their Bachelor Degrees.

Signature of the Supervisor:
Name of the Supervisor: Enamul Hassan
Date: March 17, 2018

Certificate of Acceptance of the Thesis

The thesis
entitled ”Bangla Document Categorization Using Term Graph”
submitted by the students

1. Md Nazim Uddin
2. Moudud Ahmed Khan

on March 17, 2018
as part of the requirements of the course CSE-400, is being approved by the Department of Computer Science and Engineering as a partial fulfillment of the B.Sc.(Engg.) degree of the above students.

Supervisor
Enamul Hassan
Lecturer
Department of Computer
Science and Engineering

Head of the Dept.
Dr Mohammad Reza Selim
Professor
Department of Computer
Science and Engineering

Chairman, Exam.
Committee
M. Jahirul Islam, PhD,
PEng
Professor
Department of Computer
Science and Engineering

Abstract

Automatic categorization of Bangla documents is an exigent topic now-a-days. Every document has some keywords which determine its category. If keywords of a document could be extracted, then a good model can categorize an unknown document efficiently. During keyword selection every document is pre-processed so that only effective keywords remain in the document. Naive Bayes [1][2][3], TF-IDF [4], KNN [5][6] are some existing methods for document categorization which are used for categorizing English documents. Some of them are also used for categorizing Bangla documents by previous researchers. These experiments are performed by changing feature selection method. In the experiments, sometimes top features were selected, sometimes random features were selected and sometimes one-third from top, one-third from middle and one-third from last. These experiments were supervised learning. Term Graph Model [5] is focused model in this thesis. Still it is not used for categorizing Bangla documents. Maximum 4-sized subset would be focused for this experiment in next work.

Keywords: TF-IDF, TGM, KNN, SVM, NB.

Acknowledgements

First, we would like to give thanks to our advisor Enamul Hassan, Lecturer, Department of CSE, SUST, for the constant support to us. His continuous direction to literature study and experiment helps us to conduct our research.

We want to express our gratitude to Asst. Professor Md Saiful Islam for providing his valuable guidance to this research.

Special thanks to Bishwajit Purkaystha (SUST CSE Session 2012-13) for Bangla Stemmer and Abu Shahriar Ratul (SUST CSE Session 2010-11) for the Bangla corpus that is used for the experiment.

We also want to express our gratitude to previous authors and researchers for their direct and indirect contribution in this research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Categorization stands for assignment of class. Here class means a predefined category. Document refers to some written or electronic data which provides information. Document categorization means assigning a specific class to a document from the predefined set of classes.

To categorize a document, first we need to grab all the important words from the document and important words impact in document categorization varies from one language to another language. So, document categorization is language dependent. For that reason, document categorization technique that is developed for a language may not produce the same result while it will be used for another language.

Lots of work are done so far to categorize a document which is written in English. But a few works are done for Bengali document. In this paper we apply our Categorization technique in Bengali documents.

1.1 Background

Day by day quantity of electronic data is increasing exponentially. As data increased rapidly, grouping data to some categorizes is now very hard task.

Lots of research has been done in last few decades to categorize an unlabeled document. Naive Bayes Classifier[1][2][3], Chi-Square[7], TF-IDF[4], SVM[8][9], KNN[5][6], Neural Network[10], Decision Tree[11], SVM with TF-IDF[12], TGM [5] are some techniques that were used in those researches.

Resources, pattern of a language varies from one language to another language. So that a technique that is designed for categorizing document of a language cannot be used for another language. And the research that are done so far, most of them used English document to categorize it.

But number of research on Bengali document is very few. TF-IDF algorithm along with SVM is used to categorize Bengali documents in a research [12] in 2017. Most of them used Supervised Learning to learn categorizing a Bengali document.

Bengali is one of the most popular language in the world. It is in 7th position in the list of most spoken languages. As the number of Bengali native speaker increased, electronic documents written in Bengali language also increased. So categorization of Bengali documents will be desirable.

In this paper, we choose Bengali documents to categorize it. We use supervised learning to learn categorizing a Bengali document.

1.2 Motivation

Vast amount of unlabeled electronic data is increased rapidly in last few decades. Now, classification of this data to some predefined classes is a hard nut to crack. So, an automotive system which can categorize documents to some pre-selected categorizes is highly desirable.

In real world, large amount of documents is being produced in newspapers, blogs etc. Finding one's desired document from this vast amount unlabeled documents will be very time consuming. But if a system can organize all the documents to some predetermined categorize, we can easily find our desired documents from it.

With the vast increase of data in real world, the amount of false data among them also increased. As amount of data is huge, manually tagging true or false to a data is almost an impossible task. So an automotive system to categorizing true and false data is also highly appreciable.

Lots of research has been done to categorize a document which is written in English. But not so much research is conducted to categorize a Bengali document. So we choose Bengali documents in this research.

We use Naive Bayes classifier, TF-IDF, Term Graph technique to categorize a Bengali document.

Chapter 2

Background Study

2.1 Literature Review

Before starting the research we review some papers that work on Document Categorization. Most of them has used English language documents for their research purpose. Few researches are also performed for other languages.

Throughout our literature review we marked some techniques that are used in the most of the researches. They are TF/IDF Classifier, Naive Bayes Classifier, SVM Based Classifier, Decision Tree, Chi Square Technique etc. Some of them are described in the following section.

2.1.1 KNN

KNN is one of the popular method for text categorization. In this method it selects a subset of K element and calculate confidence for each class. Every test document vote for their own class. From sum of confidence we predict a class of query document.

In Pascal Saucy and Guy W. Mineau's paper [13] on KNN algorithm for text categorization, they use μ-coocurrence for feature selection. In this experiment 2000 best features is used. They use *CosSim* function for calculating similarity between two document. In below their results summary is given,

Task Name	μ-cooccu+	
	#f	%
Course	35	96.9
Reuters1	8	98.3
Spam	77	95
Prisoner	9	90
Beethoven	8	85
News	130	85.2
McrAvrg	95.5	
#feature	267	

Figure 2.1: Results summary

From above figure we see accuracy 95.5% of this experiment. Total number of features were 267.

2.1.2 SVM

Support Vector Machine or SVM is used for classification problem. As document categorization is one type of classification problem, so SVM is a popular approach for document categorization. For each document one n-dimension point is created and SVM create a hyper-line for separate them.

Still now 95.53% accuracy is gained for SVM using each document as a vector and 21578 dataset is used for this experiment. This experiment is done by Mubaid and Umair, 2006[14] .

2.1.3 Entropy based TF/IDF Classifier

Yi-hong Lu and Yan Huang in paper [4] used TF/IDF classifier to categorize English documents. They use bag-of-words of a document to represent the document as a feature vector. So sequence of words in a document is not matter in their research. Rocchio Relevance Feedback machine learning classifier has an significant role in their research process.

In the process of feature selection firstly they remove all the stop words of a document. Later, the remaining words are processed by a stemmer.

In their research they use Porter Stemmer to get the stems of a document. This stems are used to create the feature vector of a document.

They choose 10 categories for their research. They collect 1000 articles of each category. 75% among them are used for training and rest of them are used for testing.

2.1.4 Naive Bayes Classifier for Arabic Documents

Mohamed EL KOURDI and Tajje-eddine RACHIDI built an automatic system [3] to classify Arabic documents. They use Naive Bayes classification technique.

They choose five predefined categorizes and perform their experiment considering this five categorizes. They perform their experiment using 300 documents of each class. They also execute their experiment after cross validating their data set to view how the accuracy is impacted after the cross validation of their dataset.

They use 60% of their data for training and rest of them for testing.

They achieve 68.78% of average accuracy and 92.8% of maximum accuracy.

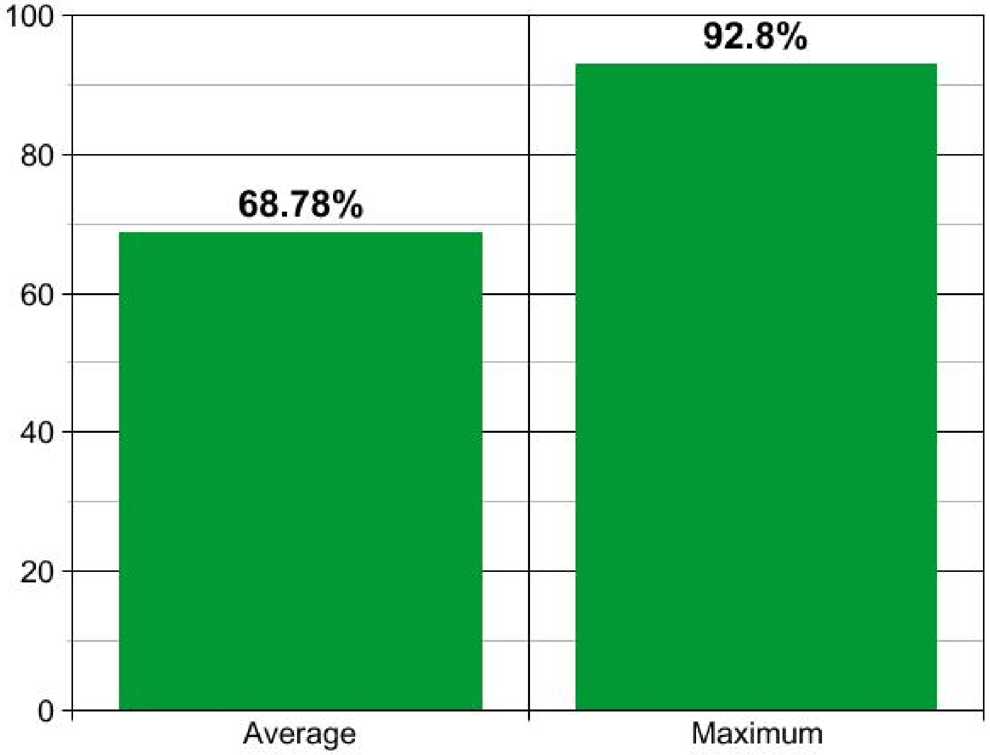


Figure 2.2: Naive Bayes Accuracy Graph

2.2 Basics

In this section we will discuss about basic components which is important in our research and experiments. These basic component is essential for all types of experiment on this problem.

2.2.1 Feature Selection

Feature Selection is major part of any methodology. Based on Feature Selection accuracy of may be boost up or reduced. In document categorization different method use different types of feature selection. Here two types of feature selection will be discussed. These are,

1. Vector Model

2. Subset selection

2.2.1.1 Vector Model

In this type of feature selection, documents are represented as vector. Each dimension represents as a term. Dimension of vector is number of unique terms in collection. In some methodology each dimension value is 1(if term is present in document) or 0(if not). Some other methods use different strategy for calculating each dimension value. Term Frequency and TF-IDF are two popular strategy for determining dimension value. Term Frequency means number of times a term occur in a document. Let's look to an example,

আমি ভাত খেয়েছি, সে কাচ্ছি খেয়েছে

After preprocessing above documents looks like,

ভাত খাই কাচ্ছি খাই

At the end of preprocessing, document divide into two subsegment and has only three unique terms in total. Now each unique term should be indexed uniquely. Following Table shows that,

Index	Feature
0	ভাত
1	কাচ্ছি
2	খাই

Table 2.1: Unique Terms of Document

Binary vector of sub-segments is,

$$V_{s1} = \{1, 0, 1\}$$

$$V_{s2} = \{0, 1, 1\}$$

Term Frequency vector of document looks like,

$$V_d = \{1, 1, 2\}$$

2.2.1.2 Subset Selection

In this type of feature selection sequential term subsets of documents are taken as feature. This type feature selection is used for Term Graph Model. In subset selection maximum subset length is very important. Because small increase of subset length causes huge number of feature, which need large memory to handle and take long execution time. But too small length of subset also reduce our methods accuracy. In our research we fixed our subset length maximum four. Let's look to an example,

মিনা নিয়মিত স্কুলে যায় ।
রাজু ও নিয়মিত স্কুলে যায় , তবে সে স্কুল পালায় ।

After preprocessing documents looks like,

মিনা নিয়মিত স্কুল যায়
রাজু নিয়মিত স্কুল যায় স্কুল পালা

Here, unique terms are মিনা, নিয়মিত, স্কুল, যায়, রাজু, পালা . Subsets are,

{ মিনা } , { নিয়মিত } , { স্কুল } , { যায় } , ... , { নিয়মিত, স্কুল } , { স্কুল, যায় } , ... , { নিয়মিত, স্কুল, যায় }, ... , { স্কুল, যায়, স্কুল, পালা }

After creating subsets, every subset must indexed uniquely,

Index	Feature
0	মিনা
1	নিয়মিত
2	স্কুল
3	যায়
4	নিয়মিত, স্কুল
5	স্কুল, যায়
6	নিয়মিত, স্কুল, যায়
7	স্কুল, যায়, স্কুল, পালা
⋮	⋮

Table 2.2: Unique sets of Term in Document

Then different types of method works on that indexed subset list.

2.2.2 Categorization

We use our proposed methodology for classification. Before using data every dataset will go under preprocessing stage. Then selected feature will be used by categorization algorithm to categorize document. For training and testing open source bengali corpus is used. Here we work on ten category. Categories are,

1. Accident
2. Art
3. Crime
4. Economics
5. Education
6. Environment
7. Entertainment

8. Politics

9. Science & Tech

10. Sports

Most important thing about categorization problem is that every category is linearly separable. In SVM method every document is plotted as a point in n-dimensional(if vector has n feature) space and create a hyper-line for separating categories. Following diagram describes our process of categorization,

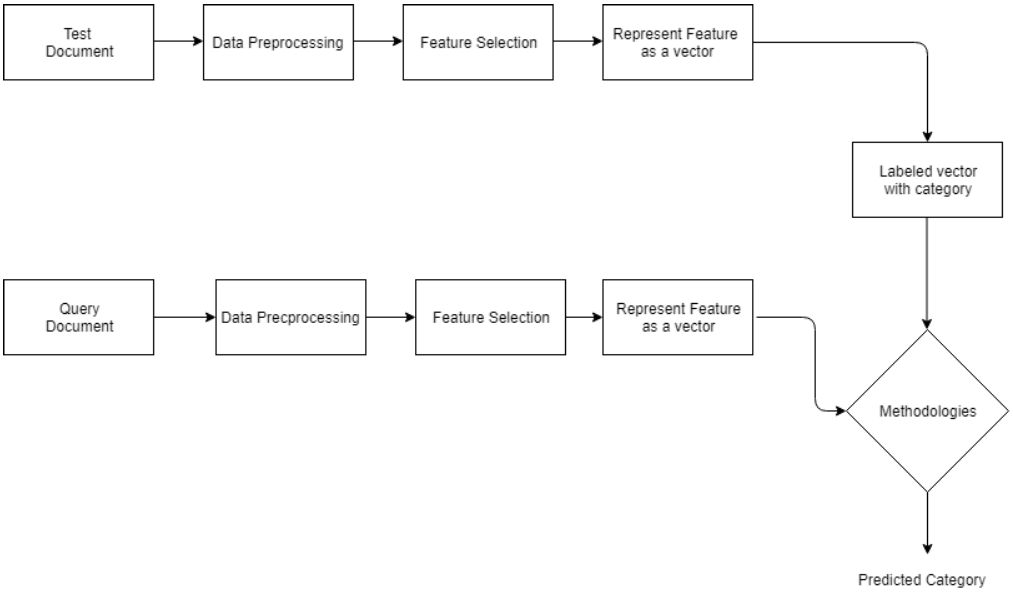


Figure 2.3: Work-flow of Document Categorization

Here predicted category is our result. After predicting query document category we calculate accuracy of each methodology.

Chapter 3

Existing Methodology and Result

For document categorization there are so many methodology exist. Some of them are impressive and some of them aren't enough good. In this chapter we will discuss few of them.

3.1 TF-IDF

A document categorization research has been conducted by Yi-hong Lu and Yan Huang in 2009. They use English documents for their research purpose. First they find the features of a document, then using TF/IDF algorithm they give weightage to this features. After this they use Rochio Relevance Feedback algorithm to categorize a unlabeled document.

Throughout their research they use feature vector to represent a document. Construction of feature vector has been done by grabbing the bag-of-words of a document. This words are used as features of the document.

They maintain the following steps to select the features of a document:

- Removing the stop words, such as a, an, the etc. This are the most frequent words in a document. Removal of this kind of words produce efficient result.

- Then, a stemmer is used to find all the stem of a document. They use porter stemmer for their stemming process.
- All the remaining unique words along with frequency in the document are used as the features of this document.

A weight a given to all the features. Weightage is calculated using TF/IDF algorithm. Weight of a word in class j is calculated using the following equation:

$$W_j = TF(w, D) \times IDF(k)$$

Here,

$W_j \ = \ Weight\ of\ word\ w\ in\ class\ j$

$TF(w,\ D) \ = \ Term\ Frequency\ of\ word\ w\ in\ all\ the\ documents\ of\ class\ j$

$IDF(k) \ = \ Inverse\ Document\ Frequency$

Inverse Document Frequency of a word is calculated using the following equation.

$$IDF(k) \ = \ log(\ |D| DF_k\ )$$

Here,

$|D| \ = \ Total\ number\ of\ documents$

$DF_k \ = \ Number\ of\ documents\ where\ this\ word\ appeared$

Feature vector is calculated for each training documents(labeled documents).

Then, a prototype vector is calculated for each categories that are found in test data. Prototype vector of a class c is calculated by adding all the document feature vectors of the corresponding class.

Now for a given document, weighted feature vector for this document is calculated first. Then, cosine measure of new document vector and each categories prototype vector is computed. Class of prototype vector that produce maximum cosine measure is assumed as the new document class.

$$H(dd) = max_{c\varepsilon C}\ cos(c,\ dd)$$

Here,

$$H(dd)\ is\ the\ category\ of\ document\ dd$$

They use 10 newsgroup dataset for their experiment. This groups are represented in the following table,

misc
atheism
hardware
autos
graphics
forsale
comp.sys.ibm.pc.hardware
baseball
motorcycles
comp.windows.x

Table 3.1: Newsgroups used by Yi-hong Lu

3.2 VSM

Vector Space Model is a model which is used to represent a text document as a vector(an algebraic form). VSM is used in most of the methodology of document categorization which we study. It is also known as term vector model. Information retrieval and formating is its main purpose.

Document is a collection of different terms. Information about each unique term of a document is a vector dimension. Let, D is a document then vector of document D will be represented as

$$V_D = \{W_1, W_2, W_3, ..., W_n\}$$

Here, $W_1, W_2, W_3, ..., W_n$ represents each unique term information of document D. Two types of vector representation is used in different methodology. They are

1. Binary VSM
2. Weighted VSM

In Binary VSM, each unique term value is either 1 or 0. If one term i is present in document D then value of W_i will be 1, otherwise 0. For scaling purpose total number of unique words will be the dimension of a vector. Let look to an example. In dataset have only three documents. Documents are,

আমি ভালো আছি

তুমি কেমন আছো ?

চল ঘুরে আসি

In datasets, unique words are আমি , কেমন , আছ ,ভালো ,তুমি , চল , ঘুর , আস. There are total eight(8) unique words. So, dimension for each document will be 8 and binary vector of each document is

$$V_{first} = \{1,0,1,1,0,0,0,0\}$$

$$V_{second} = \{0,1,1,0,1,0,0,0\}$$

$$V_{third} = \{0,0,0,0,0,1,1,1\}$$

On the other hand, in Weighted VSM each unique term value represent it's weight. Weight of a term can be calculated in many ways. In some methodology term frequency is used as weight, some use TF-IDF value as weight. Term Frequency means how many times that term occur in certain document. If document is

আমি ভালো আছি, তুমি কি ভালো আছো ?

Unique words are আমি , ভাল , আছ , তুমি , কি . Total five unique words. So Weighted VSM will be

$$V_D = \{1,2,2,1,1\}$$

For predicting a document category, query document is converted into a vector. Let's call it query vector(V_q). Each document of test data set will also be converted into vector. Query vector will be represented as same kind of vector as test documents vector.

Similarity is calculated by comparing angles deviation between query document and test documents vector. For each test document calculate cosine of angle between vectors.

$$cos\theta = \frac{V_d.V_q}{||V_d||\ ||V_q||}$$

Here,

V_d = Vector of document d.

V_q = Vector of query q.

"." denotes dot product of two vector.

$||V_d||$ denotes length of vector V_d.

Query document is labeled by document category which deviation is minimum with query document. In this method accuracy is calculated by the ratio of number of accurate prediction and total prediction.

$$Accuracy = \frac{AccuratePrediction}{TotalPrediction}$$

3.3 KNN + Cosine Similarity + TF-IDF

This model is developed on KNN where TF-IDF is used for vector generating and cosine similarity is used for calculating similarity. This model process diagram is given below,

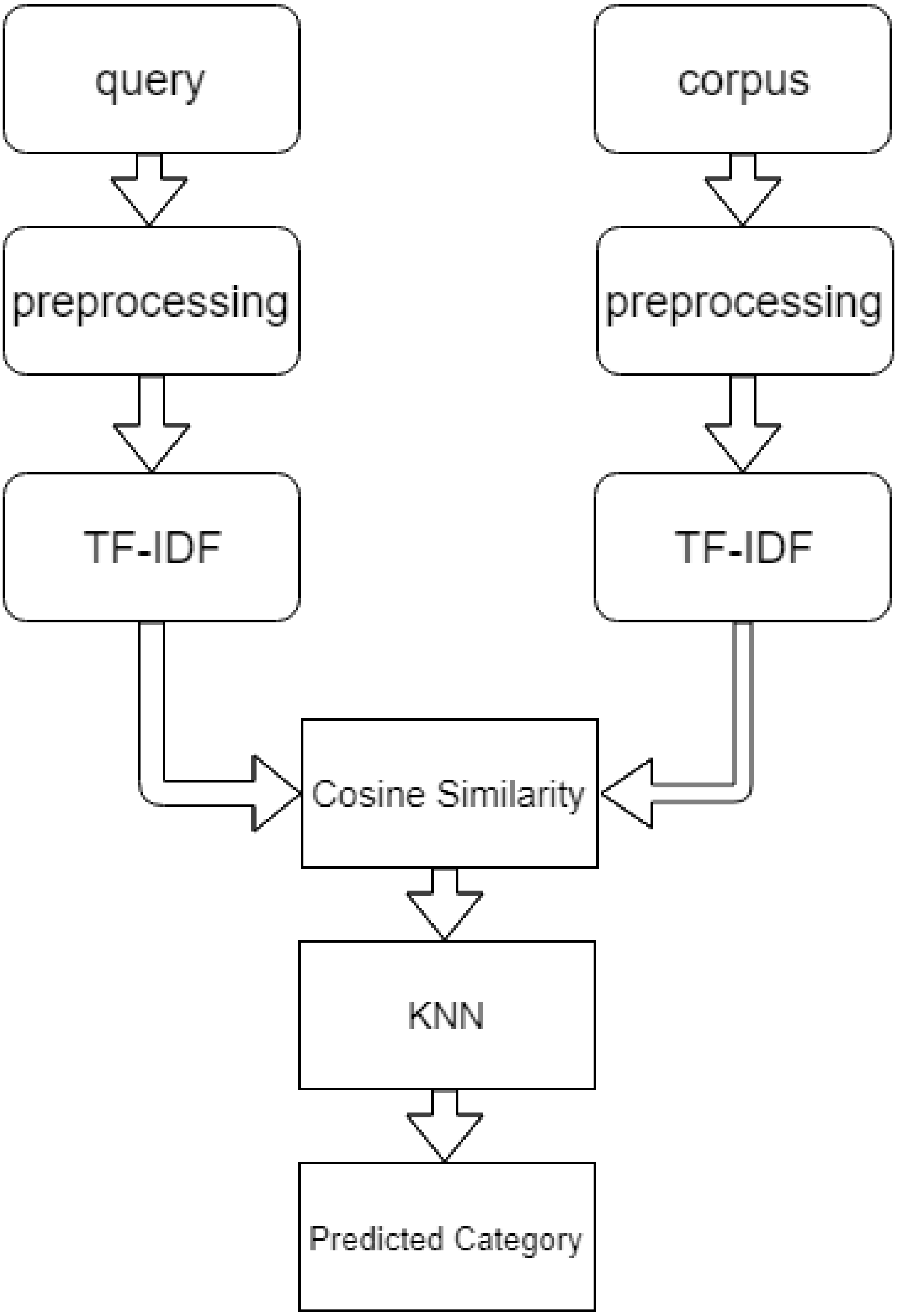


Figure 3.1: KNN+TF-IDF Process

This model accuracy is too low by comparing with other model. This experiment is done by Fazla Elahi Md. Jubayer and Syed Ikhtiar Ahmed in 2015. In this model accuracy was 68.07%. Here 10% of total data was test data and rest of them was in train set. This model is one of the lowest accuracy model in document categorization.

3.4 Naive Bayes

Mohamed EL KOURDI and Tajje-eddine RACHIDI has completed a research in document categorization using Naive Bayes Classification technique. They use Arabic documents in their research.

Before applying Naive Bayes Classification technique they preprocess each document.

In this method preprocessing of a document done by removing stopwords from the document. Vowels are discarded from the document and then roots of each remaining words are extracted. This preprocessed document is used later to do the experiment.

First the system is trained using labeled documents. 60% of all the datasets is used to train the system.

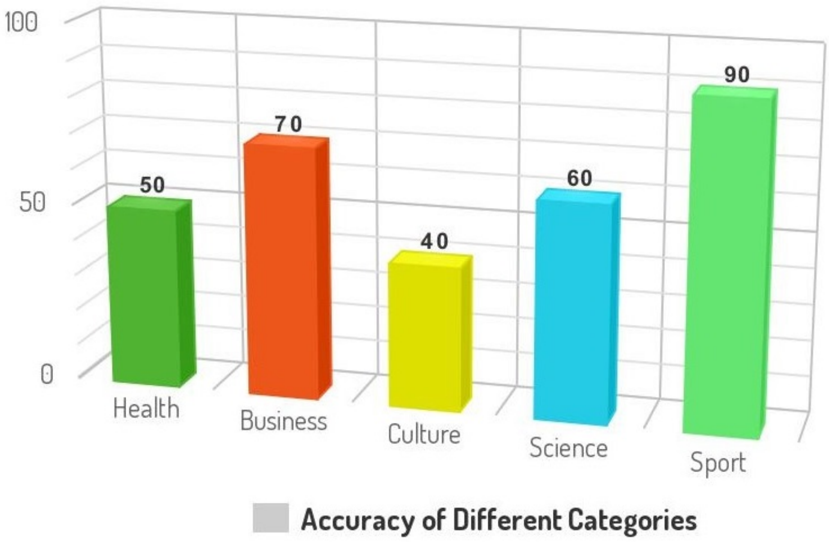
In the time of a unlabeled document, posteriori probability of the document for each class is calculated using the Naive Bayes classifier. Class which provides maximum posteriori probability is selected for this document.

Following equation is used for posteriori probability:

$$P\left(C_i|D\right) = \frac{P\left(C_i\right) \times P\left(D|C_i\right)}{P\left(D\right)}$$

Performance of this method is calculated based on the percentage of error. Author of this method use 5 classes for the experiment. They collect at least 300 web documents for each of the classes.

Accuracy of this system for each of the 5 classes is displayed below using bar diagram:


100
50
0
50
70
40
60
90
Health
Business
Culture
Science
Sport
Accuracy of Different Categories

Chapter 4

Proposed Methodology

Enhance accuracy of Bangla document categorization is our main concern in this research. From previous work we saw maximum accuracy was 92.54% and they use different mix model such that, TF-IDF + KNN, CHISQUARE+SVM, TF-DF + SVM etc. Term Graph Model is a trending model now-a-days for document categorization which accuracy is quite high. But still it is not implemented for Bangla document categorization. So we implement it for Bangla document categorization.

4.1 Term Graph Model

Term Graph Model or TGM is a known method which is used to categorize document. It is slightly different from other methods. In other methods of categorization relevant position of terms never comes in consideration. But in Term Graph Model relevant position of terms play a vital role in calculating similarity between two document. Here documents represents as a graph model.

Node or Vertex is a fundamental unit of which graphs are formed and edges are a line which connect two nodes.

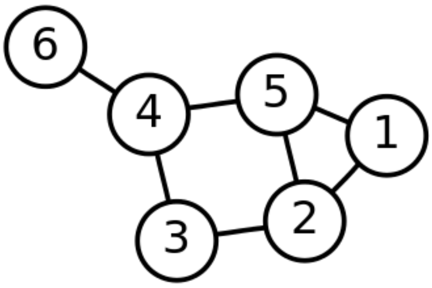


Figure 4.1: an example of graph

Here circle denotes node and joining line between nodes denote edge. This is an undirected graph.

In TGM each unique terms of a document is considered as a node. There is an edge between two nodes if and only if two terms are stay sequentially. But it is less effective if weight of edges are unit, because some sequence of terms occur more than other which effects documents categorization. So we add weight in edges. Weight of an edge is number of times this pair occur sequentially.

Now, in this model consideration of length of sequential word vary experiment to experiment. We discuss here considering at most 4 length sequence. For every document an adjacency list is created of single, double, triple and 4 members sequence following their occurrence which is retrieved from graph. Every set of words of test documents is labeled by exactly one of the category. Category is determined by most occurrences category.

For a query document, it is also represented as graph and adjacency list. Category of query documents is determined by voting of its adjacency list items. Each item vote for its labeled category. Maximum voted category is desired category of query document.

There's a problem in this model. In this method we must handle huge amount of data. We can reduce size of processed data. Some items of a document never play any role in categorization. But they reduce accuracy of method. So these items can be removed. These items are called stop words. We remove all stop words of a document. Stop words are all types of pronoun (আমি , তুমি , সে , আমরা , তারা , তিনি etc.) , conjunction (ও , এবং , কিন্তু etc.), preposition , exclamation. All types of terms which count has small deviation between categories also removed.

Let's look to an example,

আমি ভালো আছি , তুমি কেমন আছ ? বাসার সবাই কেমন আছে? আন্টি কেমন আছে?

তুমি কোথায় যাইতেছ?

At first all stop words are remove. After removing stop words documents looks like,

ভালো আছি কেমন আছি বাসার কেমন আছে আন্টি কেমন আছে

কোথায় যাইতেছ

Now each words stemmed to it's root. So after stemmed words looks like,

ভালো ⇒ ভাল

আছি,আছে ⇒ আছ

কোথায় ⇒ কোথা

যাইতেছ ⇒ যাই

বাসার ⇒ বাসা

After all types of pre-process now documents look like,

ভাল আছ কেমন আছ বাসা কেমন আছ আন্টি কেমন আছ

কোথা যাই

Unique terms of above document are ভাল , আছ , কেমন , বাসা , আন্টি , কোথা , যাই . Each unique term represent a node. There must be an edge between two consecutive terms node. Weight of an edge is number of occurrence of that pair in document.

Let's see the graphical representation of above document in graph,

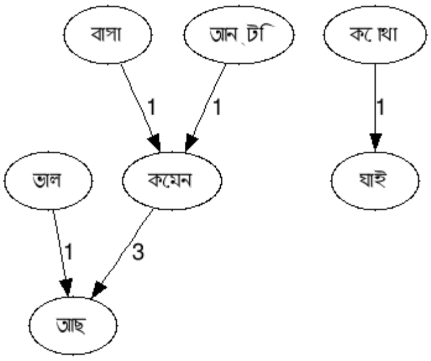


Figure 4.2: Graph represent of above document

Here, adjacency list are,

$\{$ ভাল $\} \Rightarrow 1$

$\{$ আছ $\} \Rightarrow 1$

$\{$ কেমন $\} \Rightarrow 1$

$\{$ বাসা $\} \Rightarrow 1$

$\{$ আন্টি $\} \Rightarrow 1$

$\{$ কোথা $\} \Rightarrow 1$

$\{$ যাই $\} \Rightarrow 1$

$\{$ ভাল, আছ $\} \Rightarrow 1$

$\{$ কেমন, আছ $\} \Rightarrow 3$

$\{$ বাসা, কেমন $\} \Rightarrow 1$

$\{$ আন্টি, কেমন $\} \Rightarrow 1$

$\{$ কোথা, যাই $\} \Rightarrow 1$

$\{$ কোথা, যাই $\} \Rightarrow 1$

$\{$ বাসা, কেমন, আছ $\} \Rightarrow 1$

Term Graph methodology is space and time consuming. But its accuracy is quite high, almost 98%. This is one of the best method for document categorization.

Chapter 5

Data Sets

Data is one of the most essential component of any research. More data you will use, chances of getting expected accuracy will be higher.

Our research is completely data driven. We use Bengali documents for our experiment. So we put a great deal of time to collect Bengali documents. Sources of Bengali documents and challenges we face during data collection and preprocessing will be discussed in the later sections.

We need data for two phases. They are Training Phase and Testing Phase.

All of our collected documents are labeled documents. But we separate this documents in two portion. One for Training, another for Testing.

For training, we use 60% of collected labeled Bengali documents.

For testing, we use the rest of collected labeled Bengali documents. Then we calculate the accuracy from the result given by our system and our labeled tag.

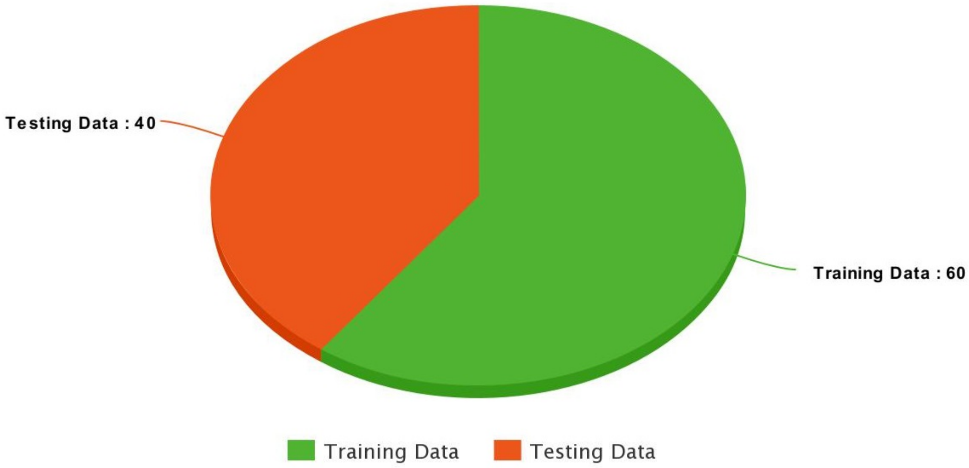


Figure 5.1: Partition of Data

In this research we work with ten categories. We collect various amount of data of this categories. Quantity of data of all categories is given below.

Category	Amount
Accident	5500
Education	10000
Art	2500
Science and Technology	2500
Sports	11000
Entertainment	9400
Economics	5100
Environment	6200
Politics	18000
Crime	7500

Table 5.1: Amount of data in different category

5.1 Data Sources

Currently there are some Bengali blogs and newspaper from where we can collect Bengali documents. Most of the blogs and newspapers manually categorize all of their documents. So, we can use this labeled documents for our training phase. And to measure the accuracy of our system this labeled documents can also be used.

We collect corpus from the previous group that work on this topic. We use this corpus in our experiments that are performed so far.

To increase the efficiency, we need to train our system with huge amount of data. So we will collect more Bengali documents from various sources and will add it to our corpus.

We will use following sources for collecting Bengali documents:

- Prothom-Alo Online Newspaper
- BDNews24
- Somewhereinblog.net
- RoarBangla Media

5.2 Data Challenges

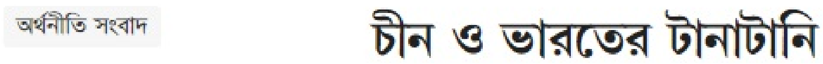
During data collection we face various challenges. Some of them are mentioned below.

5.2.1 Unstructured Data

Documents stored in the Blogs and Online Newspapers are not well structured. We have to face extensive difficulties to grab documents from Bengali blogs.

5.2.2 Multiple Categories

Some of the documents have multiple categories. But our research concern with 1 category for each document. So, in case of a document containing multiple categories we face the problem of assigning a single category.



অর্থনীতি সংবাদ

চীন ও ভারতের টানাটানি

Above article is found in the Economics section of Prothom Alo, same article can also be found in the International section. In that case we face the problem.

5.2.3 Stemming Problem

After applying stemmer in a Bengali document many of the important words are removed or essential portion of some words are removed. This type of documents won't produce proper result.

5.3 Data Preprocess

Data Preprocess is a major part in document categorization. If raw data is used in categorization methods then accuracy of methods will be low. So, for use data in methods we must preprocess raw data. In preprocessing steps we must be concern on some matter,

First, Information lose.

Second, Unnecessary things must be removed.

Third, Which terms effects all documents equally must be removed.

At the beginning of preprocessing document is segmented into sub-segments. Sub-segments are created based on punctuation marks. Then each sub-segments process individually. Then each sub-segments is processed in following way,

- Remove all types of special symbol(Exp. <, >, (,), ?, {, }, etc). These symbol hasn't any impact in document categorization. On the other hand if we keep these symbol it may reduce our accuracy.

- Collect all terms. Terms means words. These are separated by whitespace. Terms must be collected by keeping their sequence information. Because it may be needed in some methodology.

- Remove all types of pronoun(আমি, তুমি, সে, আমরা, তোমারা, তারা etc) and conjunction(ও, এবং, কিন্তু etc). These words also have no impact in document category. We collect a list of words which are unrelated with classification.

- In last step we find base form of each word. This is called stemming. One word is used differently based on situation. Such as, যাই used as যাইতেছি when task is running. So for getting better performance each word must be stemmed. This process also known as root finding.

Chapter 6

Experiments and Results

6.1 KNN

KNN or K-Nearest Neighbor is another excellent approach of document categorization. This method is used in a large range in document categorization. In this method a document is classified on voting.

At first, we convert each document to a vector. Then we select a set of K neighbors of test document. For each class *c* we calculate sum of confidence of all documents which belong to our selected set. From these sum of confidence we get our predicted class of query document. For comparing query document Q with test document T we use ***cosSim*** formula.

For binary vector model it looks like,

$$CosSim(Q,D) = \frac{C}{\sqrt{A * B}}$$

Here, A is total number of terms in vector Q, B is total number of terms in vector D and C is total number of terms which is common in both Q and D.

We use total 2017 document for our experiment. We run our algorithm on different value of K for getting a better result. In below figure we give our statistics of accuracy of different category,

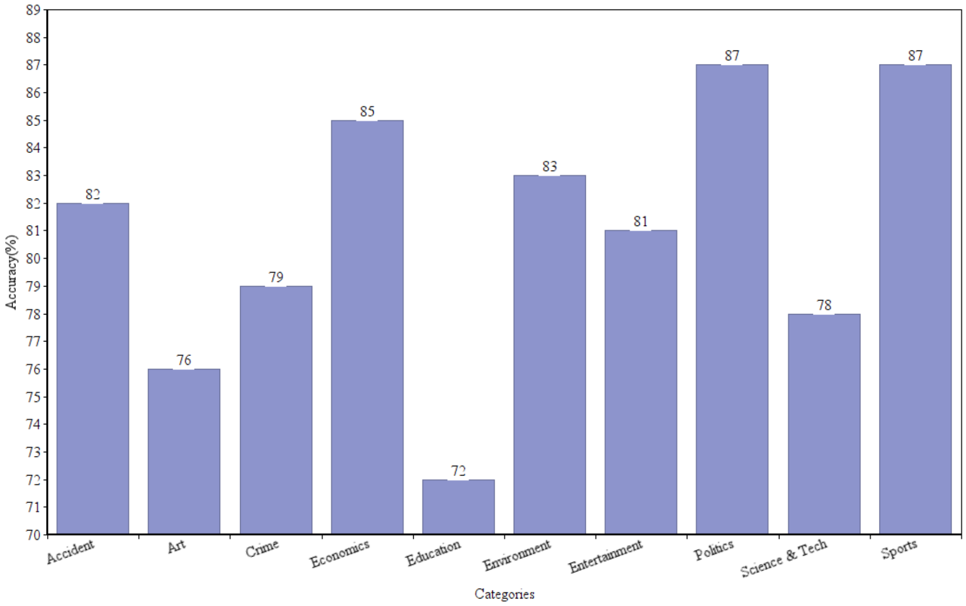


Figure 6.1: Performance of KNN

From above figure we see our maximum accuracy hits 87% in two categories(politics and sports) and minimum accuracy is 72% which is in education category. Average accuracy of our approach is 81.0%. We will try to increase accuracy of this method by changing feature selection in future. Main disadvantage of KNN is it takes more time for large number of test dataset.

6.2 TF-IDF

Term Frequency - Inverse Document Frequency(TF-IDF) is an effective way for calculating value of each dimension in vector model. TF-IDF calculates each term relative information in a document. We used at most 3382 feature for categorization. At first we train our system with a small number of feature then we increase it slowly. But result's behave in an unexpected way.

We get our maximum accuracy when we use 2570 feature. Our maximum accuracy is 89.01% in average. Here accuracy means for same number of feature mean of accuracy of different category.

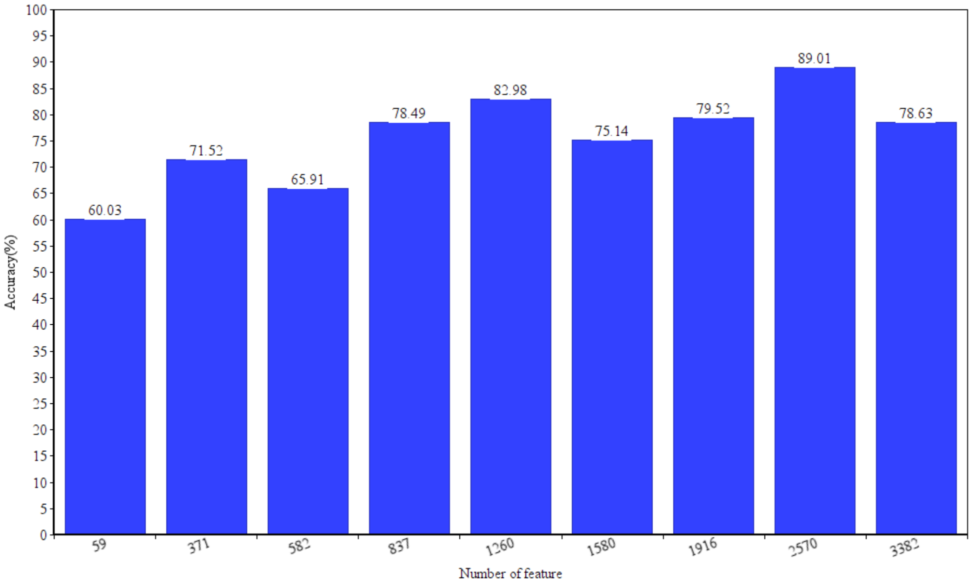


Figure 6.2: Performance of TF-IDF

Above figure shows graphical view of mean accuracy against different number of feature. We use 60% data for training purpose and 40% for testing purpose. We change it to 80% training and 20% testing but it decreased accuracy because of less testing dataset. For each individual category accuracy is quite good.

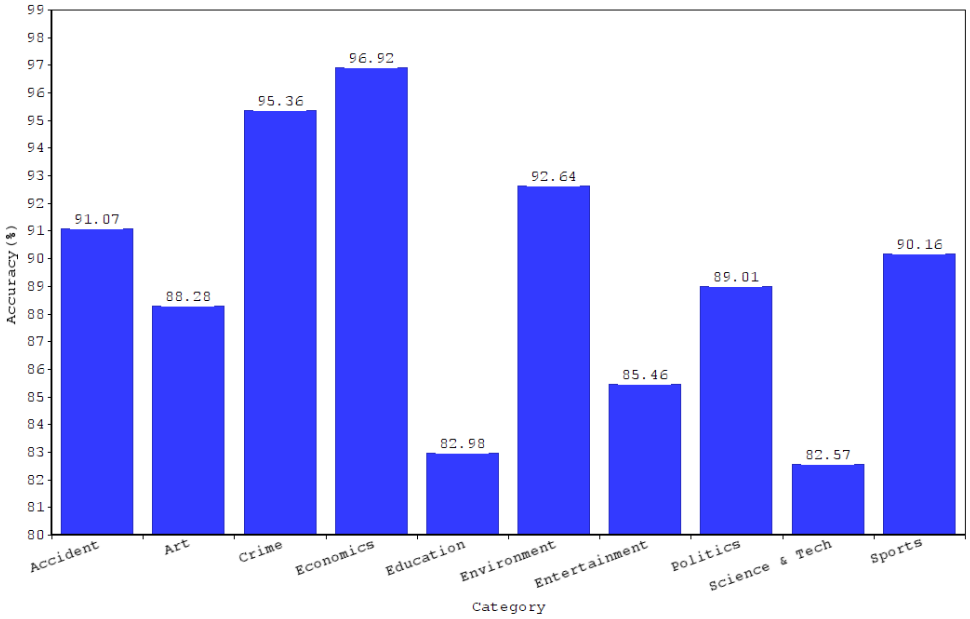


Figure 6.3: Accuracy by category in TF-IDF

Above figure shows maximum accuracy against category when we use TF-IDF. Here we see, we get maximum accuracy in economics category and it's 96.92%.

When we try this method we faced some problem. We will try another approach of implementing TF-IDF in future to get a better result.

6.3 Naive Bayes

Naive Bayes classifier is a probabilistic document categorization model. Different probabilistic equations are applied in this experiment.

This experiment is done for five categories. We collect 1000 documents of each category. A list of categories that are considered in this experiment is given below:

- Education
- Politics
- Entertainment

- Science
- Sport

In this experiment 85% documents of each category is used for training and rest of them is used for testing.

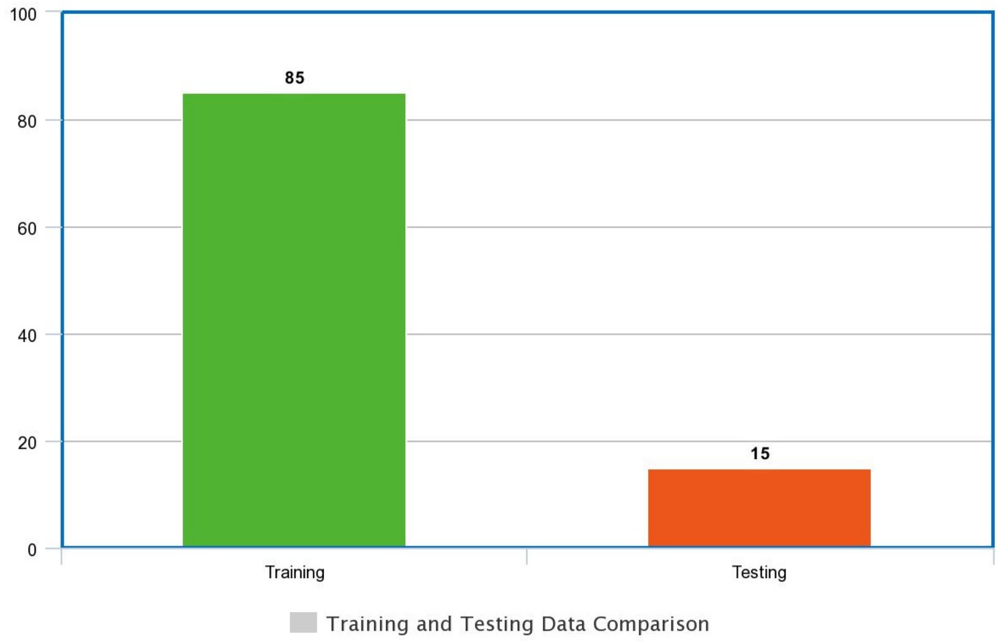


Figure 6.4: Naive Bayes Data Comparison

In the training phase, we execute all the necessary calculation that are needed in the testing phase. We calculate $P(C_i)$ and $P(D|C_i)$ of each category.

Here,

$P(C_i)$ *means the probability of the category.*

$P(D|C_i)$ *means the probability of document D belongs to category* C_i.

Then in the testing phase, for each unlabeled document we calculate its posteriori probability for each category. Category that produces maximum posteriori probability is our assigned category to this document.

We calculate error for each category. For each document we know its actual category. so if the returned category of this document from our system and its actual category mismatched we count it as error.

Then accuracy of a category is calculated using following equation:

$$Accuracy_c = \frac{Total\ Error_c}{Total\ Documents_c}$$

We perform this experiment by cross validating our dataset. Average, maximum and minimum accuracy obtained in the cross validation is displayed in the following table.

	Education	Politics	Entertainment	Science	Sport
Average	67%	62%	67%	68%	64%
Maximum	80%	75%	80%	78%	82%
Minimum	59%	52%	55%	61%	60%

Table 6.1: Accuracy of different category

6.4 TGM

In our research our main focus is on this model. In Proposed Methodology section we described how this model work. In this section we describe how this model will be implemented.

At first we will pre-process whole dataset using pre-processing system. Now comes our feature selection part. We can do this in a few way. It's the most important part of this experiment, because it directly impact our accuracy. We will select our feature in three ways and then apply our algorithm. These ways are,

1. Without any change after pre-processing.

2. Apply TF-IDF function and keep only those words whose TF-IDF value is greater than a certain threshold.

3. Count each term frequency and keep only words whose frequency is greater than a certain threshold.

We iterate every word of a document. For each consecutive word we make a pair if it doesn't exist otherwise increase frequency of this pair. Then we take consecutive next two word of current word and make a tuple if it doesn't exist otherwise increase frequency. We do this sentence wise.

Now we have two types of data. One is pair and another is tuple. We apply different ratio weight between tuple and pair. For each query we also convert document in pair and tuple. We will find match between query document and each category. We will do this in two ways.

Firstly, we will use TF-IDF method on pair and tuple individually and find cosine similarity between query document and each category.

Secondly, we calculate value for each category by matching it with query document for pair and tuple individually. Then multiply pair and tuple value with their ratio and at last summarize them. Maximum point category is our expected category.

Chapter 7

Conclusion

Lots of work have been done so far to categorize English documents. Improvements in Bengali documents categorization is not provoked so far. But in the recent years amount of Bengali Data is increasing exponentially. So we think that construction of an automated system to categorize Bengali documents will be an epoch making work.

In this report, we try to explain the background and emphasize of our research, what types of works have been done so far by others and what we have done so far.

First few months of our research we spend time to read some research papers that are written on this topic so that we can acquire some ideas of currently going trend. In the literature review section of this report we mention some of our studied papers.

Some new methods are proposed based on the previous techniques and would be tried to increase their accuracy.

7.1 Future Work

We already proposed some methods and performed some experiments. But our achieved accuracy is not so good. We will modify our approaches and do different experiments to achieve a better result.

7.1.1 Data Collection

We conduct our experiments using the corpus of previous group. But we think that, if we can enrich our collections with more Bengali documents our system will be more accurate.

We already identify some sources for Bengali documents. We will collect documents from this sources to enrich our collections.

References

[1] Y. Wang, J. Hodges, and B. Tang, “Classification of web documents using a naive bayes method,” pp. 560– 564, 12 2003.

[2] Y. Matsuo and M. Ishizuka, “Keyword extraction from a single document using word co-occurrence statistical information,” vol. 13, 03 2003.

[3] M. EL KOURDI, A. BENSAID, and T.-e. Rachidi, “Automatic arabic document categorization based on the naïve bayes algorithm,” 08 2004.

[4] Y. h. Lu and Y. Huang, “Document categorization with entropy based tf/idf classifier,” vol. 4, pp. 269–273, May 2009.

[5] V. Bijalwan, P. Kumari, J. Espada, and V. Semwal, “Knn based machine learning approach for text and document mining,” vol. 7, 06 2014.

[6] Y. Wang and Z.-O. Wang, “A fast knn algorithm for text categorization,” vol. 6, pp. 3436 – 3441, 09 2007.

[7] M. Alexandrov, A. Gelbukh, and G. Lozovoi, “Chi-square classifier for document categorization,” pp. 457–459, 02 2001.

[8] C.-h. Chan, A. Sun, and E.-P. Lim, “Automated online news classification with personalization,” 03 2002.

[9] A. Mesleh, “Support vector machines based arabic language text classification system: Feature selection comparative study.” pp. 11–16, 01 2007.

[10] Z. Chen, C. Ni, and Y. L. Murphey, “Neural network approaches for text document categorization,” pp. 1054–1060, 2006.

[11] S. Weiss, C. Apte, F. Damerau, and S. Weiss, “Text mining with decision trees and decision rules,” 10 1999.

[12] M. S. Islam, F. Elahi, and S. Ikhtiar Ahmed, “A support vector machinemixed with tf-idf algorithm to categorize bengali document,” 04 2017.

[13] P. Soucy and G. Mineau, “A simple knn algorithm for text categorization,” pp. 647 – 648, 02 2001.

[14] H. Al-Mubaid and S. A. Umair, “A new text categorization technique using distributional clustering and learning logic,” 2006.