Bangla Document Categorization using Term Graph

Enamul Hassan
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet - 3114, Bangladesh

Md Nazim Uddin
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet - 3114, Bangladesh

Moudud Ahmed Khan
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet - 3114, Bangladesh

ABSTRACT

Bangla document categorization is an emergent topic now-a-days. Every document has some keywords that reflect its category. Document categorization refers to an automatic categorization of a document based on the keywords it contains. An expedient keyword selection method is necessary to correctly classify a document. TF-IDF [1], Naive Bayes [2][3][4], KNN [5] are some of the trending methods used in Document Categorization. Some of models are also used in Bangla Document Categorization. In this research, Term Graph concept was mainly focused. TGM [5] is never used before for Bangla document categorization. So, the concentration was Term Graph concept mixing with other existing models for categorizing Bangla documents. Experiments are also performed by changing and tuning feature selection method. Maximum 3-size subsets are used in experiment. Features were selected by changing selecting formula. Sometime all features were selected and sometime less important features were removed for increasing accuracy and reducing space complexity.

General Terms

Bangla Document, Categorization, Natural Language Processing.

Keywords

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

1. INTRODUCTION

Document indicates to the written or electronic data which provides information. Categorization refers to assigning predefined classes. Predefined class stands for categories. Each document has its own pattern which indicates its category. Document Categorization means identifying the category of a document by analyzing the keywords present in it. At the beginning of the document categorization process, important keywords are collected from the document. And behavior of impact of keywords in categorizing document differs from one language to another. So, document categorization is language dependent. That means a document categorization model designed for a specific language may not expose expected performance in another language.

Amount of electronic data/documents increasing rapidly. So, grouping of data is very hard now. Lots of works were done by the researchers to make the categorization system efficient and accurate. Naive Bayes Classifier [2][3][4], TF-IDF [1], Chi-Square [6], SVM [7][8], KNN[5][9], SVM with TF-IDF [10], Decision Tree [11], TGM [5], Neural Network [12] are some techniques to categorize an unlabeled document. But most of the researches were done focusing on English language.

Number of researches to categorize a Bangla document is not so high. TF-IDF algorithm along with SVM is used to categorize Bengali documents in a research in 2017. Supervised Learning is used by most of them to learn categorizing a Bengali document. Bangla is one of the most prestigious language all over the world. It is the 7th most spoken language in the world. With the increase of Bangla native speaker in amount, Bangla electronic data/documents are increasing swiftly. So, efficient and proper categorization of Bangla document is desirable.

In this paper, Bangla documents were chosen to categorize them. Supervised learning was used to learn categorizing a Bangla document. The amount of electronic data was increased in last few decades vastly. So, manual classification of these data were almost impossible. So, an automatic system to categorize a document is highly desirable. Lots of documents are producing by newspapers, blogs etc. per day. To efficiently and effectively find a document, documents should also be organized in a nice manner. Document categorization is highly preferable to organize documents in a newspaper or in a blog.

With the increase of electronic data, quantity of false data are also increasing. Manually checking the validity of a document is very time consuming. A good document categorization model designed for false data detection can solve this problem. However, Naive Bayes, TF-IDF method are used in this research to categorize Bangla documents. These techniques are also used by other researchers. Term Graph Model is also used in this research to categorize Bangla documents.

2. LITERATURE REVIEW

At the beginning of this research, some of the works are audited which were performed on Document Categorization. English language is used by most of them. Some of them also used other languages. Some focused techniques were selected and later some of these techniques were applied in this experiment. These techniques are TF/IDF Classifier [1], Naive Bayes Classifier [2][3][4], SVM Based Classifier [7][8], Decision Tree [11], Chi Square Technique [6] etc. Some of them are described in the following section.

2.1 KNN

KNN is one of the mostly used text categorization method. In this method it selects a subset of K element and calculate the confidence for each class. Every test document votes for their own class. From sum of confidence, a class of query document is predicted.

In Pascal Saucy and Guy W. Mineau’s paper [13] on KNN algorithm for text categorization, they used µ-coocurrence for feature selection. In this experiment, 2000 best features are used. They have used CosSim function for calculating similarity between two documents. The summary of the result of their work is given below,

Table 1: Results summary

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	

From the table 1, it is clear that the accuracy is 95.5% of this experiment and total number of features are 267.

Performance of KNN in experiment done by us is shown in Figure 1.

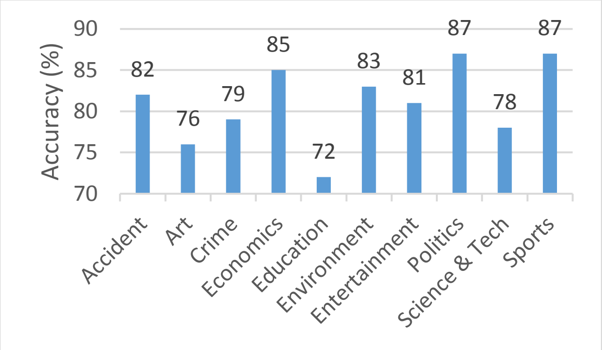


Figure 1: **Performance of KNN**

2.2 SVM

Generally Support Vector Machine or SVM is used for classification problem. Document categorization is very much similar to classification problem. So, SVM is also used 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].

2.3 Entropy based TF/IDF Classifier

Yi-hong Lu and Yan Huang in paper [1] 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 a significant role in their research process.

In the process of feature selection, firstly they have removed all the stop words of a document.

Later, the remaining words were processed by a stemmer.

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

They have chosen 10 categories for their research. They have collected 1000 articles of each category. Among them, 75% were used for training and rest of them were used for testing.

2.4 Naive Bayes Classifier for Arabic Documents

Mohamed EL KOURDI and Tajje-eddine RACHIDI built an automatic system [4] 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 have used 60% of their data for training and rest of them for testing. They achieved 68.78% of average accuracy and 92.8% of maximum accuracy.

3. BASIC COMPONETS OF THE RESEARCH

In this section, the discussion would go on with basic components which is important in this research and experiments. These basic components are essential for all types of experiments on this problem.

3.1 Feature Selection

Feature Selection is major part of any methodology. Based on Feature Selection ac- curacy 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.

3.1.1 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.

3.1.2 Subset Selection

In this type of feature selection sequential term subsets of documents are taken as feature. This type of 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 the accuracy of the methods. In this research, the subset length is fixed maximum four.

3.2 Categorization

Before using data, every dataset goes under preprocessing stage. Then selected features are used by categorization algorithm to categorize document. For training and testing, open source bengali corpus is used. Here, ten categories were chosen firstly. Categories are: Accident, Art, Crime, Economics, Education, Environment, Entertainment, Politics, Science & Tech and 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. Figure 2 describes the process of categorization where predicted category is the result. After predicting query document category, accuracy of each methodology were calculated.

4. EXISTING METHODOLOGIES

A lot of established methods are already exist for document categorization and some of them were also implemented for Bangla document. Though they have a fantastic accuracy, but these accuracy don't meet with expectation. Some methods' accuracy really close to the expectation. In this section, different types of methods will be discussed in details.

4.1 TF-IDF

TF-IDF is one of the well-established method for categorizing document. It is a frequency based statistical method. Here, frequency means number of occurrence of a word in a document. Different kinds of statistical approaches are applied on that frequencies of words.

Table 2: Unique sets of Term in Document Then different types of method works on that indexed subset list.

Index	Features
0	মিনা
1	নিয়মিত
2	স্কুল
3	যায়
4	নিয়মিত, স্কুল
5	স্কুল, যায়
6	নিয়মিত, স্কুল, যায়
7	স্কুল, যায়, স্কুল, পালা
…	…

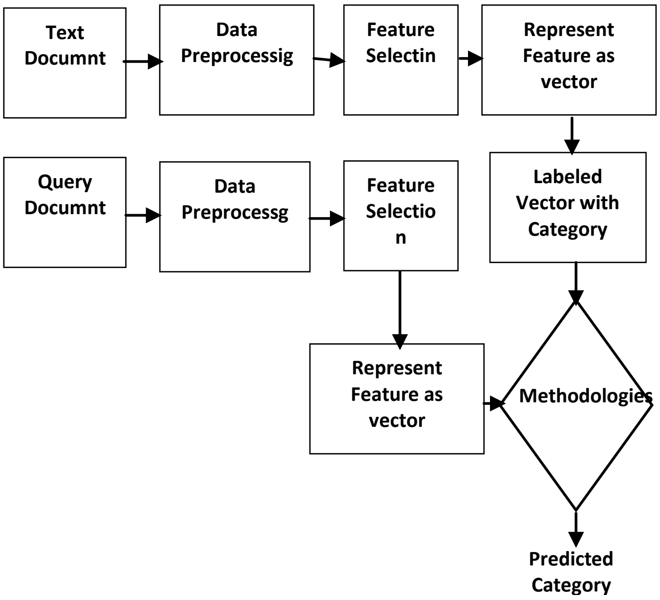


Figure 2: Work-flow of Document Categorization

A weight is given to all the features. Weight of each word is calculated using TF/IDF algorithm.

Weight of a word in j category is calculated using the following equation:

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

Here, W_j = Weight of word w in category j
$TF(w, D)$ = Term Frequency of word w in all the documents of category 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

For each training document, feature vector is calculated (labeled documents).

A prototype vector is created for each category based on training data sets. Prototype vector for a category is build using each document prototype vector of that category. All documents prototype vector is being added for creating that categories prototype vector. Now for categorizing an unknown document, at first weighted feature vector of that document is calculated. After that cosine similarity is measured with each category and target document prototype vector. Maximum similar category is declared as expected category.

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

Here, H(dd) is the category of document dd

Best experiment is conducted by Yi-hong Lu and Yan Huang in 2009 of this method. They have used 10 newsgroup dataset for their experiment. This groups are misc, atheism, hardware, autos, graphics, forsale, baseball, motorcycles and windows.

Performance of TF-IDF in this experiment is shown in figure 3 and 4.

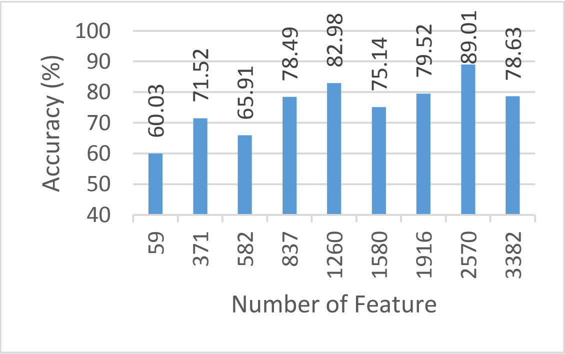


Figure 3: Performance of TF-IDF

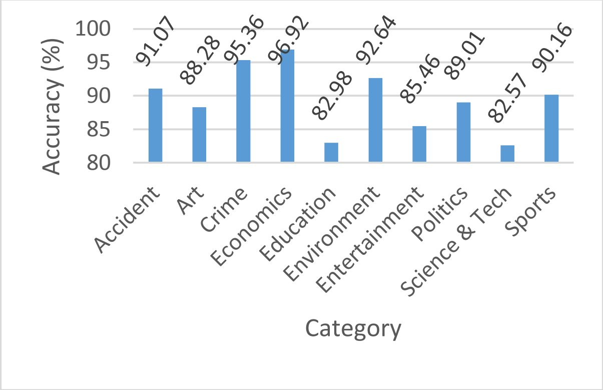


Figure 4: Accuracy by category in TF-IDF

4.2 VSM

Vector Space Model is a way of representing a text document into a vector form. This is a popular form for representation a document among most of the studied methodologies. Term Vector Model is another known name of it. Main purpose of this model is information retrieving and formatting.

Document is nothing but a bag of words. Each unique words information is considered as a dimension in vector space. 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: Binary VSM and Weighted VSM.

In Binary VSM, each unique term value is either 1 or 0. If one term i exists in a document D then value of W_i will be 1, otherwise 0. For calculation purpose total number of unique words will be the dimension of a vector. Let look to an example. Let, 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 its value. Value of a term can be calculated in many ways. In some methodology term frequency is used as value, some use TF-IDF value. 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.

$$cos\theta = \frac{\mathrm{Vd\,.\,Vq}}{||Vd||\ ||Vq||}$$

Here, Vd = Vector of document d, Vq = Vector of query q.

"." denotes dot product of two vector.

||Vd|| denotes length of vector Vd.

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

$$Accuracy = \frac{\text{Accurate Prediction}}{\text{Total Prediction}}$$

4.3 Knn + Cosine Similarity + Tf-Idf

This is a hybrid model. Basically this is a KNN based model, in which TF-IDF is used for vector generating and for calculating similarity cosine similarity is used. This model process diagram is given in figure 5. This hybrid model proposed and experimented by Fazla Elahi Md. Jubayer and Syed Ikhtiar Ahmed in 2015. Accuracy of this model comparing with other existing model is too low. Only 68.07% is achieved by this model. 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.

4.4 Naïve Bayes

This method is one of the best method based on statistical approach. It is developed based on famous probability bayes theorem. Many researcher used this approach for different language. it is used by Mohamed EL KOURDI and Tajje-eddine RACHIDI for arabic language. Naive bayes classification technique is used by them for this purpose.

Before applying Naive Bayes Classification technique they preprocess each document. Stop words and all unnecessary words are removed at the first step of preprocessing. All vowels also must be removed from document. After that roots of each remaining words are extracted. This preprocessed document is used later to do the experiment. The system is trained using labeled documents. 60% of all the datasets is used to train the system.

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

Following equation is used for posteriori probability:

$$P(C_i|D) = \frac{\mathrm{P}(C_i) * P(D|C_i)}{\mathrm{P(D)}}$$

Accuracy of this method is calculated using naive approach. For arabic language experiment researchers use only 5 category and for each category they use more than 300 document. These document are crawled from internet.

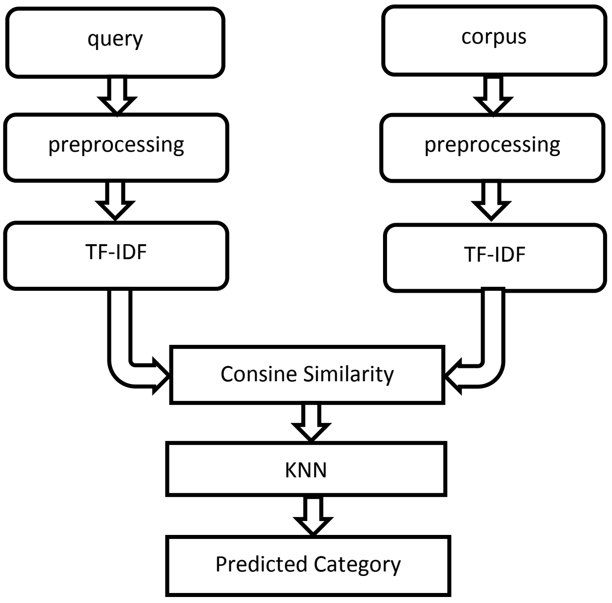
In this experiment maximum accuracy 90% has been gained at sports category and minimum 40% has been gained at culture category. The result of the experiment is given in table 3.

Table 3: Accuracy of different category

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

4.5 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 documents.



yy

Figure 5: KNN+TF-IDF Process

Accuracy of this system for each of the 5 classes is displayed below using bar diagram in figure 6.

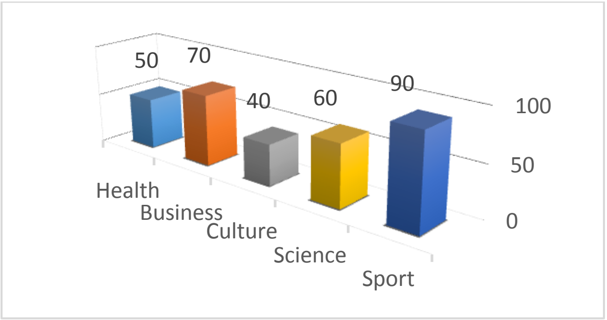


Figure 6: Accuracy of different categories in Naïve Bayes.

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, weight in edges are added. 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. Here, the discussion is done 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.

There is a problem in this model. In this method, huge amount of data must be handled. The size of processed data could be reduced. Some items of a document never play any role in categorization. But they reduced accuracy of the method. So these items can be removed. These items are called stop words. All stop words of a document has been removed. Stop words are all types of pronoun (আমি, তুমি, সে, আমরা, তারা, তিনি etc.), conjunction (ও, এবং 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 ? বাসায় 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 বাসায় কেমন আছে আন্টি কেমন আছে কোথায় যাইতেছ

Now each words stemmed to its 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.

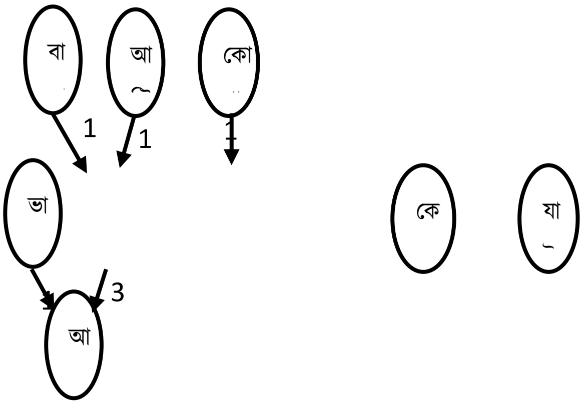


Figure 7: Graph represent of above document

Let’s see the graphical representation of above document in graph in figure 7. The adjacency list are given below.

11

{ভাল} ⇒ 1, {আছ} ⇒ 1, {কেমন} ⇒ 1, {বাসা} ⇒ 1, {আন্টি} ⇒ 1, {কোথা} ⇒ 1, {যাই} ⇒ 1, {ভাল, আছ} ⇒ 1, {কেমন, আছ} ⇒ 3,

{বাসা, কেমন} ⇒ 1, {আন্টি, কেমন} ⇒ 1, {কোথা, যাই} ⇒ 1,

{কোথা, যাই} ⇒ 1, {বাসা, কেমন, আছ} ⇒ 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.

5. EXPERIMENTS AND RESULTS

The main target of this research is enhancing its accuracy as some research already has been done on this topic. From previous research which already has been done by Fazla Elahi Md. Jubayer and Syed Ikhtiar Ahmed on this topic, it is reported that the maximum accuracy 92% was gained. Different hybrid method had been applied for acquiring this accuracy. TF-IDF + KNN, CHISQUARE+SVM, TF-DF + SVM are some hybrid which is used in previous research.

Term Graph Model is a trending concept now-a-days for document categorization. This model is so much popular because of its high accuracy. But unfortunately this concept hasn’t been applied yet for Bangla language. So, some methods based on this model for categorization are proposed.

5.1 Weighted TGM

Weighted TGM is the first proposed method based on term graph model for Bangla language by us. In this method term graph concept is used with a weight assigning method. Different sized subsets are used in this method.

At first of this method preprocessed steps are done as usually. All unnecessary words and stop words must be removed. All characters other than alphabets must be removed at preprocessing steps. An efficient stammer is needed for this method. Using stammer rest of the words of document will be converted into its root word. After that root words are being grouped into different sizes. This group size is an important parameter of this method. For assigning weight different approaches will be followed. Suppose one is just keep frequency as weight and for more than one word group weight will be summation of subset words frequency. Now for measuring all group in same scale let multiply two size subset with 10, three size with 100 and so on. Another way of assigning weight is, measure them in scale one. So, scaling schema is done as follows,

1. *size_subset_cost = weight × 0.2*
2. *size_subset_cost = weight × 0.3*
3. *size_subset_cost = weight × 0.5*

This all will be done with training data set. For categorizing an unknown document, it will also go under preprocessing and other necessary prerequisite steps. After that all remaining root words will be grouped into same size as training documents. At next step total weight is calculated of the query document for each category. Maximum weighted category will be labeled as query document category.

First experiment was done with maximum 2 size subset of term. In this experiment total 30,000 document has been used where for each category 2500 documents has been used. From 2500 documents 1800 documents for training purpose and 700 documents for testing purpose. For calculating weight of each term following rules were followed,

1_size_subset_weight = frequency × 1

1_size_subset_weight = frequency × 10

For predicting category of an unlabeled document as usual approach has been followed. Maximum weighted category is final category. Following chart (Figure 8) shows us accuracy against each category in this experiment.

In this experiment, the maximum achieved accuracy is 93.61% in politics category. The average accuracy of this experiment is 72.05 %.

As the result of previous experiment was not good enough, so another experiment was done by changing its perimeters. In this experiment 27600 documents has been used. For each category 2300 documents has been used where 2000 documents has been used for training purpose and 300 documents has been used for testing purpose. Parameters in calculating weight also has been changed in this experiment. In this experiment following rules were followed,

1_size_subset_weight = frequency × 2

1_size_subset_weight = frequency × 4

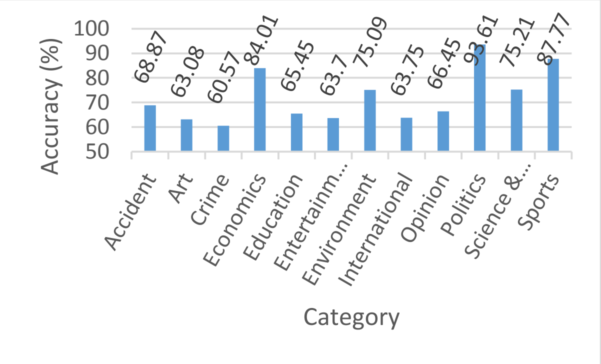


Figure 8: Weighted TGM with 1:10 weight

Querying for a document label was as usual. Following (Figure 9) chart shows us details about this experiment result.

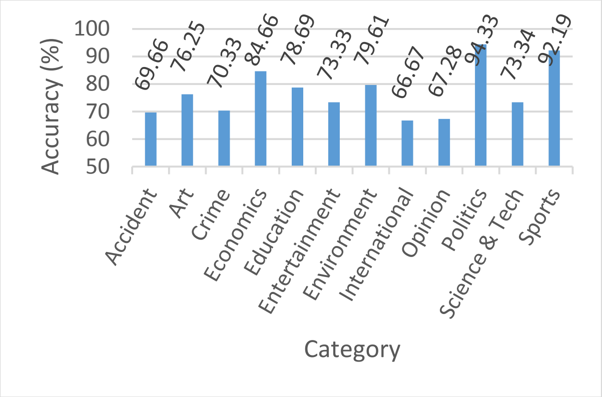


Figure 9: Weighted TGM with 2:4 weight

In this experiment maximum 94.33% has been gained and average accuracy was 77.90%. Accuracy has been improved by this experiment from previous one. For each individual category except one accuracy also has been increased. Following chart (Figure 10) shows us comparison between two experiments result category by category.

Though accuracy has been increased but average accuracy didn't match with the expectation. So after a lot analysis on details result of previous experiment, some point has been figured out which may reduce accuracy.

Some document belongs to multiple category. Suppose one document about international politics, it belongs to both international and politics category. Another category is opinion which can bias any other category. Because opinion can be about any topic. So another experiment was done by removing following three topics: Art, International, Opinion.

In this experiment with 9 category accuracy has been increased. Following table 4 shows us result of this experiment.

In this experiment maximum 95.28% accuracy has been gained and average accuracy was 83.42%.

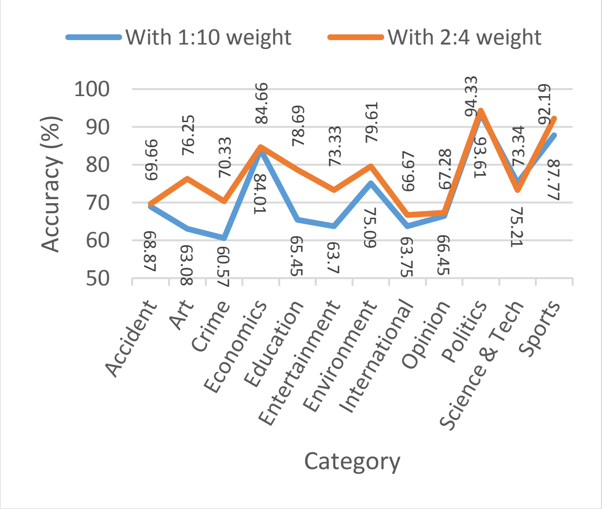


Figure 10: Comparison between two experiments result

Table 4: Accuracy with 9 category

Category	Accuracy
Accident	79.01%
Crime	88.63%
Economics	86.67%
Education	79.34%
Entertainment	80.34%
Environment	68.67%
Politics	94.34%
Science & Technology	78.53%
Sports	95.28%

5.2 TGM+VSM

This method is developed using term graph concept with combining vector space model. Here weighted VSM will be used for representing documents.

As like other methods in training phase each document goes under several preprocessing steps. Finally only root words are exist in document. After that root words are grouped into different sizes. In this experiment maximum 3 sized group will be used. Each document will be represented as a vector where each dimension is a subset of terms. One problem was arisen in this approach. When at most 3 sized subsets are taken then it needs a huge memory for processing. So only top 10,000 words according to frequency were used in this method for building vector for each category. Here many options exist for determining how value will be calculated. Some of them are just only frequency summation or TF-IDF value in a modified way.

5.2.1 *Experiment with Frequency*

In this experiment frequency of each subset was used as value in vector space model. For each category one vector was built. Total 30,000 documents was used in this experiment. 80% of them was used as training document and 20% of them was used as testing document.

At first each document is being processed. Then 1-size subset, 2-size subset and 3-size subset frequency is calculated. After creating vector for each document, all vectors of a category was merged. This merge is done by sum up frequency of each subset. Now, only top 10,000 frequent subsets are kept for making vector. For all category this procedure is followed for creating vector of that category.

For an unknown document at first a vector is made for every category. This is done by calculating frequency of each subset of that categories frequency vector. After that similarity is calculated between unknown document vector and that certain category vector. Similarity is calculated using cosine similarity. In this experiment 67.63% accuracy was gained. This accuracy is the worst accuracy among all of the experiment results. For individual category maximum 84.26% has been gained in this experiment. After the research, cause of this low accuracy is figured out. If VSM model is used in experiment then only frequency use as weight reduce accuracy in a worst way. Because certain subsets frequency difference doesn't make an effect on determining category.

Following figure 11 shows the comparison of accuracy between previous methods with it.

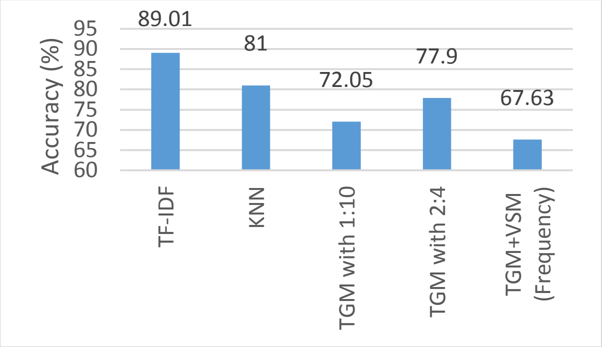


Figure 11: Comparison of Accuracy between Different Methods

5.3 TF-IDF+KNN

This is a hybrid method. In this method, two concepts TGM and VSM are used. At first, for each document maximum 2-size subset is calculated and their frequency is counted. Using the frequency of vector TF-IDF value is calculated for each document. With most high TF-IDF value of 1000 subsets of each document is used for creating vector of that document.

For query document, vector also calculated using TF-IDF value. This TF-IDF value is calculated based on testing documents. After calculating vector for query document KNN method is applied for finding category of that unlabeled document. Cosine similarity method is used for calculating similarity. This experiment is a time consuming experiment because for each query whole data set need to be processed. For each query two time whole data set is being processed, one for calculating TF-IDF value and another for calculating similarity. 2500 documents of each category is used for this experiment. Among them 2100 documents is used for training purpose and 400 document is used for testing purpose. Following figure 12 shows us accuracy against each category what is gained in this experiment.

In this experiment maximum 95.29% accuracy was gained and minimum accuracy was 85.39%. Average accuracy 90.995% was gained in this hybrid experiment.

This experiment result also doesn't meet with expectation because of biased category. So, another experiment is done by removing some category which biases the categorization. Following category is removed from the experiment: Art, Environment, International, Opinion. Other parameter of this experiment was same as before experiment. Following figure 13 shows us result of this experiment with 8 category.

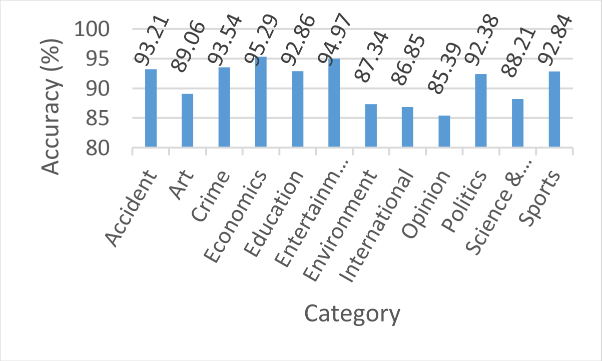


Figure 12: Accuracy of TGM+VSM+TF-IDF+KNN Experiment

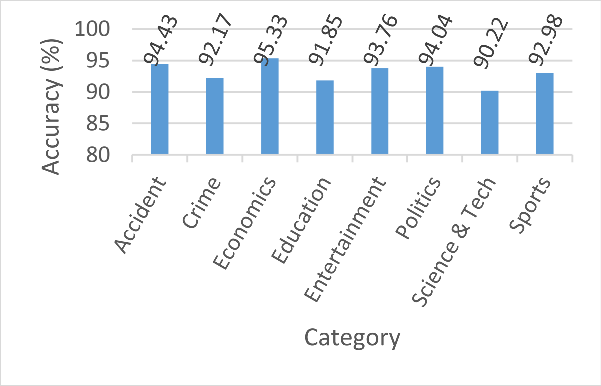


Figure 13: Accuracy of TGM+VSM+TF-IDF+KNN Experiment with 8 Category

In this experiment, average accuracy 93.10% has been gained where maximum accuracy was 95.33% and minimum accuracy was 90.22%.

In this research maximum 93.10% accuracy is gained. This accuracy is gained in the experiment where TGM and VSM concept is used with TF-IDF value calculating approach and KNN method for determining unknown documents category. This experiment is done with 8 category. Following figure 14 shows the comparison between different methods accuracy.

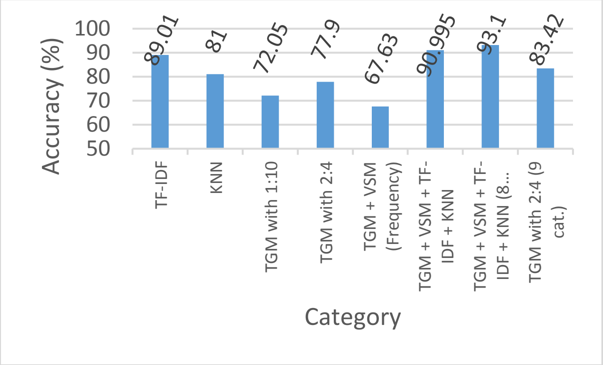


Figure 14: Comparison of Different Methods Accuracy

6. DATA SET

In this research, work has been done with twelve categories. Table 5 shows the quantity of the collected data.

The following sources are used for collecting Bengali documents: Prothom-Alo Online Newspaper, BDNews24, Somewhereinblog.net, RoarBangla Media, Bangla Corpus.

Table 5: **Amount of data in different category**

Category	Amount	Category	Amount
Accident	6350	Economics	5351
Education	12389	Environment	6852
Art	1669	Politics	20439
Science and Technology	2906	Crime	8840
Sports	12086	Opinion	2530
Entertainment	10131	International	5922

7. ACKNOWLEDGMENTS

Without some peoples' involvement, this work must be void. So, they must be mentioned in this paper. First gratitude goes to Md Saiful Islam, Assistant Professor, Department of CSE, SUST, Sylhet, BD. Secondly, the big thanks goes to Bishwajit Purkayastha for Bangla Stemmer and Abu Shahriar Ratul for Bangla Corpus. Lastly, the department of CSE, SUST must be thanked for giving the opportunity to do this work.

8. REFERENCES

[1] Y. h. Lu and Y. Huang, "Document categorization with entropy based tf/idf classi- fier," vol. 4, pp. 269–273, May 2009.

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

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

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

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

[7] C.-h. Chan, A. Sun, and E.-P. Lim, "Automated online news classification with per- sonalization," 03 2002.

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

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

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

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

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

[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 distribu- tional clustering and learning logic," 2006