ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/314666043

Semi supervised keyword based bengali document categorization

Conference Paper · September 2016

DOI: 10.1109/CEEICT.2016.7873040

CITATIONS
5

READS
99

4 authors, including:

Abdullah Al Maruf
East West University (Bangladesh)
73 PUBLICATIONS **366** CITATIONS
SEE PROFILE

Md Saiful Islam
Shahjalal University of Science and Technology
70 PUBLICATIONS **237** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Sentiment Analysis and Opinion Mining in Bangla View project

Artificial Intelligence Lab View project

All content following this page was uploaded by Md Saiful Islam on 25 December 2018.

The user has requested enhancement of the downloaded file.

Semi Supervised Keyword Based Bengali Document Categorization

Fahim Quadery*, Abdullah Al Maruf†, Tamjid Ahmed‡, Md. Saiful Islam§
Department of Computer Science and Engineering, Shahjalal University of Science and Technology
Kumargaon, Sylhet-3114, Bangladesh
Email: *fahim@student.sust.edu, †maruf@student.sust.edu, ‡tamjid@student.sust.edu,§saiful-cse@sust.edu

***Abstract*—Document Categorization is an area of important research over the last couple of decades. The basic task in document categorization is classifying a given document in some predefined classes. Bengali is among the top ten most spoken languages in the world and is spoken by more than 200 million people, but the candid truth is, it still lacks significant research efforts in the area of Bengali Document Categorization. In the first phase of this paper a model has been designed that extracts keywords from a Bengali document. We crawled over 35000 news documents form popular Bengali newspapers and journals. Those documents have been stemmed and less significant words are removed using stemmer and Parts-of-Speech(POS) tagger.Statistical approach is used to extract keywords form the documents. Then probabilistic distribution and semi supervised learning with Naïve Bayes algorithm is used to approximate the category of a given Bengali document.Result and statistical data show the effectiveness of this model.**

***Index Terms*—Chi-square, Probability distribution, Bengali Document, Word frequency, Naïve Bayes, supervised learning, semi supervised learning, Training Dataset, Pipilika, Chorki, Co-occurrence matrix, Classifier, Stemmer, POS tagger, Corpus.**

I. Introduction

Document categorization has a vast application in this computer dominated time. In this paper we will discuss what document categorization is, why it is important and step by step how a model has been built that classifies a Bengali document in some predefined classes.

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. Document categorization has versatile applications. Spam filtering, sentiment analysis, language identification, finding similar articles, automatically determining the genre of a text.

II. Bengali Document Categorization and Related Works

Document categorization is now an important topic of library science. There is a healthy number of veteran researchers working on this topic all over the globe. But all these research are mostly focused on English documents categorization. Despite of being one of the most widely spoken language, for Bengali document categorization, there is a little work and resource available sporadically on the World Wide Web.

Most of the online news portals categorize their news manually. However few research works have been done on Bengali document categorization. [1] Proposed a n-gram based technique for document categorization. In [2] Naive-Bayes classification is used to predict category for news articles. Over 7000 documents were crawled from newspapers in [2]. Documents were tokenized, stemmed and stop words were removed. Then documents were treated as a vector and using Naive Bayes Classifier documents were categorized. There are two Bengali search engines called "Pipilika" and "Chorki" that use Bengali document categorization. This paper tested and learned about [1], [2] and Pipilika search engine's approach, robustness of their models, i.e. pros and cons. And also compared their work with this paper.

III. Background Studies

There are different approaches available for English and Bengali language for the keyword extraction. There is statistical approach which uses statistical models like chi-squared [3] , N-gram [4] [1] etc. There is linguistic approach that mainly analyze the linguistic feature of the language to find out the important keyword of the documents [5]. Machine learning approach which is supervised or semi-supervised learning approach [6]. Noun phrase extraction approach extracts noun phrase from the document and tries to find out the important keyword from the document using statistical analysis known as Conditional Random Field (CRF) model [7]. Graph base approach is to build a graph using the linguistic analysis of the document. Then give a leap word and propagate the value to the other node of the graph [8]. Then finds out the most weighted keyword for the documents. In this paper a semi supervised machine learning based approach on top of statistical analysis is introduced which required studying several machine learning models. And finally found a good model that suits Bengali language perfectly. Study results on machine learning models and statistics is described in the paper.

A. Chi-square Distribution

Chi-Square distribution can be used to measure the difference between the actual count and the expected count of distinct elements [3]. In this paper Chi-Square was used to determine important words in a document. Taking degree of freedom one the equation of Chi-Square distribution is

$$\chi^2 = \sum_{i=0}^{n} \frac{\{o_i - \varepsilon_i\}^2}{\varepsilon_i} \tag{1}$$

Where,
χ^2 = Chi-Square Value
o_i = Observed frequency
ε_i = Expected frequency

B. Naïve Bayes Classifier

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between features [9]. Naive Bayes probalilistic classification model was first used for text classification by D. Lewis [10].

Let U a Universe with all the possible outcomes (of an experiment for instance), and we are interested in some subset of them, namely some event. Say we are studying event A and B.Probability of event A and B happening individually is

$$P(A) = \frac{|A|}{|U|} \tag{2}$$

and,

$$P(B) = \frac{|B|}{|U|} \tag{3}$$

now, given B probability of event AB is

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{4}$$

and vice versa, given event A, probability of event AB is

$$P(B|A) = \frac{P(AB)}{P(A)} \tag{5}$$

Putting this two equations together we get,

$$P(A|B)P(B) = P(B|A)P(A) \tag{6}$$

and finally,

$$P(A\ B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

Which is Bayes theorem. What actually the insight here is to train this algorithm with some features against some classes, and when we feed the model some features, it returns us classes.

IV. Methodology

In this paper, classifier was trained with more than 35,000 Bengali documents mostly collected from newspapers and journals. First each of these were Stemmed and less important words(i.e. অব্যয়) were omitted from the document. then the most frequent words (top 30%) were filtered out. After finding out the frequent words, the Chi-Squared distribution value of each terms with the frequent words was calculated. Those apparently sporadic terms having higher Chi-Squared value are tend to be more important in that document.

Once the important words for all documents under a category are found, then all those important words are treated like a single document for that particular category. Another Chi-square was used on important words to filter out the keywords for that particular category. This second Chi-square trimmed our training data set, streamlined the model and had a big impact on category approximation accuracy and pre-processing CPU time.

When a new document is to be classified, keywords of that documents are extracted and for that document we used Naïve Bayes classifier on those keywords with each category. And this approach captivated document categorization.

A. Extracting Keywords

According to [3] Chi-Squared value for each of the terms of the document with the most frequent terms,

$$\chi^2 = \sum_{g \epsilon G} \frac{\{freq(w,g) - p_g n_w\}^2}{P_g n_w} \tag{8}$$

A document contains sentences of different lengths. If a term appears in a long sentence, it is more likely to co-occur with many terms. If a term appears in a short sentence, it is less likely to co-occur with other terms [3]. Here,

P_g As (the sum of the total number of terms in sentences where g appears) divided by (the total number of terms in the document).

n_w As the total number of terms in sentences where w appears.

Again $n_w P_g$ represents the expected frequency of co-occurrence. However, it's value becomes more sophisticated. A term co-occurring with a particular term g ϵ G has a high χ^2 value. However, these terms are sometimes adjuncts of term g and not important terms. [3]

Here is the value of Chi-square on sample data-set.

Table I
Chi-Square values of words

No	Word	Chi-Square value
1.	পুলিশ	46901.248
2.	উপজেলা	33105.5258
3.	গতকাল	29368.5810
4.	রাত	28220.190
5.	জানা	28218.628

After the calculation of first Chi-Square on the input set, top 20% words were filtered out according to higher value. On those set of words, Chi-Square was used second time to filter out once again. And hence end up with the key words.

Table II
Chi-Square values after 2nd pass

No	Word	Chi-Square value
1.	পুলিশ	1324.22
2.	উপজেলা	966.98
3.	গতকাল	914.41
4.	রাত	619.34
5.	জানা	469.90

B. Classifying Documents

As mentioned previously, once the key words are extracted, Naïve Bayes classifier is used on those key words to train the model. To serve this purpose, the classifier is trained with over 35,000 documents of 12 categories.

1) Training the Naïve Bayes Model: The occurrence of each keyword under each class is counted. From this a table that contains the count of features against all classes is made. A sample table is shown III. This table shows some frequent key words and their corresponding occurrences under different classes [3]. From this table the Prior Probability, Probability of features and Probability of Likelihood are calculated. All these terms are essential for calculating Naïve Bayes and are described here in gist.

Table III
Keyword co-occurrence table

Feature / Class	গতকাল	দেশ	পুলিশ	জানান	বাংলাদেশ	মোট
Accident	1064	42	761	877	71	1938
Opinion	291	3948	1049	284	2885	8457
Crime	1611	197	1679	1571	211	5269
Total	2966	4187	3489	2732	3057	15664

Now the Prior probability is guessing a category blind folded. For instance, base rate or prior probability of the above category are calculated below.

$$P(accident) = \frac{1938}{15664}$$
$$P(opinion) = \frac{8457}{15664}$$
$$P(crime) = \frac{5269}{15664}$$

Probability of Evidence or Feature Probability is the probability of getting an individual feature from all the features. Feature Probability of first 3 features is calculated.

$$P(\text{গতকাল}) = 2966/15664$$
$$P(\text{পুলিশ}) = 3489/15664$$
$$P(\text{বাংলাদেশ}) = 3057/15664$$

2) Calculating the Probability Distribution: Probability of Likelihood of a feature is nothing but the conditional probability of finding a feature, while a class is given. Let's calculate the likelihood of first few features on Opinion category

$$P(\text{পুলিশ}/\text{Opinion}) = 1049/8457$$
$$P(\text{জানান}/\text{Opinion}) = 284/8457$$
$$P(\text{বাংলাদেশ}/\text{Opinion}) = 2885/8457$$

Let's calculate the probability distribution of a set of features F. Which is the multiplication of likelihood of each feature from the feature set over the probability of evidence. For example, after calculating the probability of the feature set for all the classes, classes are sorted in descending order of probability value. Category with highest probability in the classifier is more likely being the category of that document.

V. Training Data and Statistics

A. Classes or Categories

This paper used 12 different categories. These categories cover a major portion of Bengali news categories. The categories that were worked with are given.

Table IV
Category List

Accident	Art	Crime	Economics
Environment	International	Education	Entertainment
Opinion	Science-tech	Politics	Sports

The other existing Bangla news categorization model from "Pipilika" search engine worked on four categories. Categories that Pipilika used are shown in V

Table V
Pipilika Search Engine's Category List

Economics	Entertainment	International	Sports

From the Naïve Bayes equation it is trivial to show that increasing in number of classes will reduce the performance of the classifier. Thus the number of classes and the performance is reversely related.

B. Training Model

Naïve Bayes model is trained with a total of 35,580 unique text files. Twelve categories of Bengali documents are used.

This dataset is used to train the classification model by calculating the prior probability and evidence probability. Then for every category the conditional probability is calculated using the key words of that category. Once all the conditional probabilities are calculated, then it is very straightforward to calculate the probability of a certain class for a given feature set. The number of training set has an almost linear relation with performance.

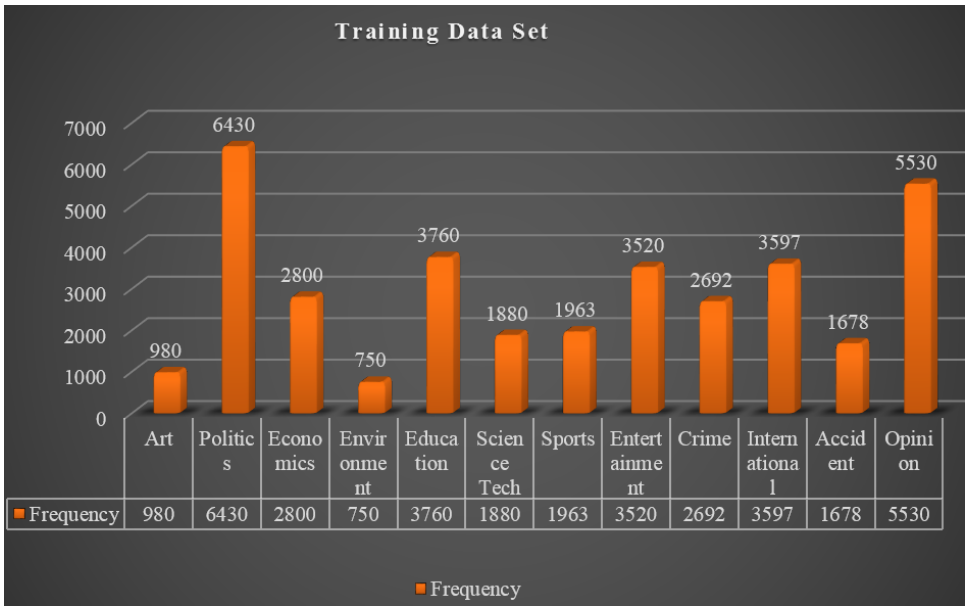


Figure 1. Training Data-set.

C. Statistical Analysis

Here some statistics of the training data is presented. For the lack of working space, the category names in English alphabets are mapped. The mapping is shown in Table VI.

Table VI
Class Mapped to Alphabet

A	Accident	B	Art	C	Crime
D	Economics	E	Education	F	Entertainment
G	Environment	H	International	I	Opinion
J	Science-Tech	K	Politics	L	Sports

Now here in Table VIII are the most frequent words of each category. Table VII shows the top 5 words with highest Chi-Square value for each class.

Table VII
Top Five Words With Maximum Chi-Square Value for Five Category

A	Word	উপজেলা	নিহত	গতকাল	পুলিশ	জানান
	χ^2 Value	20	15	14	13	12
B	Word	কথা	মানুষ	শেষ	দেখা	মন
	χ^2 Value	175	163	155	154	150
C	Word	পুলিশ	উপজেলা	গতকাল	রাত	জানা
	χ^2 Value	469	331	293	282	282
D	Word	বাংলাদেশ	টাকা	ব্যাংক	দেশ	হাজার
	χ^2 Value	50	45	42	39	34
L	Word	ম্যাচ	দল	বিশ্বকাপ	রান	বাংলাদেশ
	χ^2 Value	35	32	28	27	25

VI. Performance Analysis

A. Accuracy of The Model

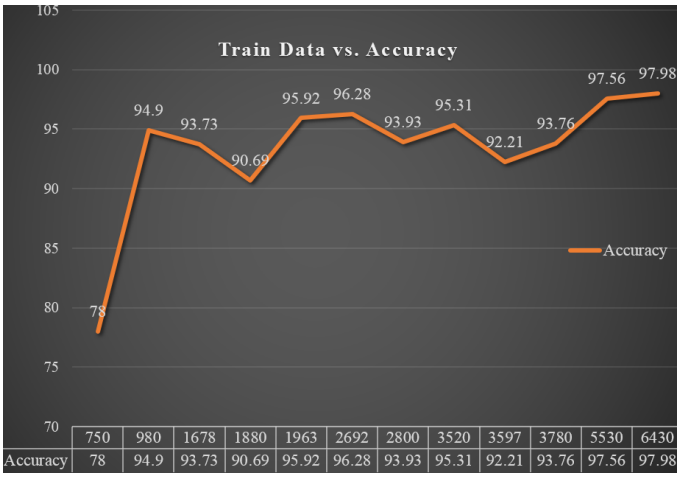
From observation here it is clear that there is almost a linear relation between the size of training data-set and output efficiency. Higher number of training data for a given category tends to give more accurate approximation. This data-set vs. accuracy graph is shown in figure 2.

From the linear relationship in this graph it shows that this model with a larger dataset can provide even more accurate category approximation. So a way of improving this model is increasing the training dataset.

It is easy to observe the relation between training set

Table VIII
Most Frequent Words

A	word	দিক	উপজেলা	গতকাল	জানান	নিহত
	count	1220	1069	1064	877	871
B	word	কথা	মন	মধ্যে	মানুষ	পা
	Count	600	582	509	497	488
C	Word	পুলিশ	গতকাল	জানান	থানা	উপজেলা
	Count	1679	1611	1571	1510	1477
D	Word	বাংলাদেশ	মধ্যে	দেশ	হবে	টাকা
	Count	1552	1483	1390	1300	1264
E	Word	শিক্ষার্থী	হবে	আজ	ঢাকা	অংশ
	Count	2976	2560	2509	2507	2326
F	Word	ছবি	শুরু	হবে	করেছেন	মধ্যে
	Count	1547	1369	1352	1202	1187
G	Word	গেছে	দেখা	মধ্যে	এলাকা	পা
	Count	418	399	389	370	357
H	Word	গত	খবর	গতকাল	দেশ	মধ্যে
	Count	1963	1920	1683	1677	1592
I	Word	হবে	দেশ	মধ্যে	হল	কথা
	Count	4104	3948	3648	3554	3444
J	Word	কথা	করেন	গতকাল	হবে	সরকার
	Count	3795	3648	3433	3354	3253
K	Word	তথ্য	কাজ	ব্যাবহার	তৈরি	নতুন
	Count	885	825	823	798	796
L	Word	ম্যাচ	দল	বিশ্বকাপ	শেষ	পঞ্চম
	Count	1317	1300	1000	976	880



	750	980	1678	1880	1963	2692	2800	3520	3597	3780	5530	6430
Accuracy	78	94.9	93.73	90.69	95.92	96.28	93.93	95.31	92.21	93.76	97.56	97.98

Figure 2. Relation Between Training Data-set and Efficiency.

and category approximation is not strictly linear. From observation, performance depends on availability of keywords in training set under each category. This concludes, a better stemming of dataset will increase the performance of this model.

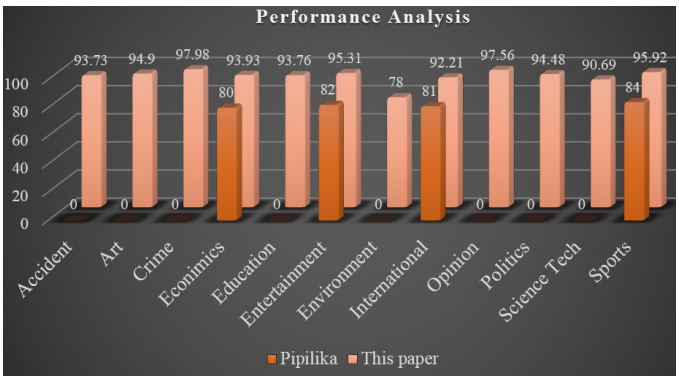


Figure 3. Performance Comparison Between this Paper and Pipilika Search Engine's model.

B. Result of Other Related Works

[1] Used one year "Daily Prothom-Alo" news corpus. It tested the model with 25 randomly selected documents. Accuracy ranges from 20% to 100% (for 3 categories using tri-gram) over 6 categories. [2] Used more than seven thousand documents extracted form "Daily Prothom-Alo" and manually tagged six hundred twenty eight news where two hundred eight items were used as training set. It achieved a highest precision rate of 0.8.

Figure 3 shows the performance analysis between this model and Pipilika search engine's model which is developed under "SUST NLP research lab". Pipilika search engine achieved a average accuracy of 82% over four category. It shows this paper outperforms Pipilika's model in every aspects.

VII. Conclusion

Document categorization is a dabbling topic these days. However for Bengali, a significant approach is yet to emerge. Here we developed a prolific model that approximates adequately. We achieved a average accuracy of almost 93% with maximum of 97.98% accuracy. This paper discussed the problem, what is Bengali Document categorization application, what already have done by others in this topic. Then the gist of proposed approach is given. It described out derived approach which is dedicated to describe the in-depth of this classification model and step by step described Chi-Square value of probability distribution and Naïve Bayes classifier. Next section was dedicated to depict different properties of the training data and statistical analysis of data-set. Then it showed the performance analysis of the model. Overall this paper proposed the importance of document categorization for Bengali language and developed an efficient approach for Bengali document categorization.

References

[1] M. Mansur, N. UzZaman, and M. Khan, "Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus," in *9th International Conference on Computer and Information Technology*, Dhaka, Bangladesh, 2006.

[2] A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla News Classification using Naive Bayes classifier," in *16th International Conference on Computer and Information Technology (ICCIT)*, Khulna, Bangladesh, March 2014, pp. 336–371.

[3] Y. Matsuo and M. Ishizuka, "Keyword Extraction from a Single Document using Word Co-ocurrence Statistical Information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 1, pp. 157–169, 2004.

[4] K. Sarkar, "An N-Gram Base Method for Bengali Key phrase Extraction," in *International Conference of Information Systems for Indian Languages*, Patiala, India, March 2011, pp. 36–41.

[5] J. Kaur and V. Gupta, "Effective Approaches For Extraction Of Keywords," *International Journal of Computer Science Issues*, vol. 7, no. 6, November 2010.

[6] P. D. Turney, "Learning algorithm for keyphrase extraction," *Journal of Information Retrieval*, vol. 2, no. 4, pp. 303–336, 2000.

[7] D. B. Bracewell, F. Ren, and S. Kuriowa, "Multilingual single document keyword extraction for information retrieval," in *International Conference on Natural Language Processing and Knowledge Engineering*, Wuhan, China, 2005, pp. 517–522.

[8] Y. Ohsawa, N. E. Benson, and M. Yachida, "Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor," in *Advances in Digital Libraries Conference*, Washington DC, USA, 98, pp. 12–18.

[9] S. Ting, W. Ip, and A. H. Tsang, "Is Na ve Bayes a Good Classifier for Document Classification," *International Journal of Software Engineering and its Applications*, vol. 5, no. 3, July 2011.

[10] D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European Conference on Machine Learning(ECML)*, Chemnitz, Germany, April 1998, pp. 4–15.

View publication stats