Text Documents Classification in Uzbek Language

O.J. Babomuradov Ozod, N. S. Mamatov, L. B. Boboev, B.Otaxonova

Abstract: ***This article deals with intellectual analyzing technologies, which classify texts in Uzbek language, in which the Bernoulli and multi-nominal models are considered. The textual documents used in this research are from the authentic sources of The State National Information Agency of Uzbekistan. To compare the probability methods of classification, 600 documents of 6 types of categories, with 169205 words, have been used.***

Keywords: Uzbek Language, classification

I. INTRODUCTION

Text classification is done manually, with the help of expert instructions and machine learning methods [4-6]. Automatic classification of texts is mostly based on the concept of "similarity." Normally, such texts store similar words and word phrases in them.

One of the widespread methods of pre-processing of texts is Bag of Words [1]. In this model, firstly we create vocabulary V out of the words from the pre-set of texts. A histogram vector is created based on the number of repetitions of the words in the texts that match the vocabulary. Some methods look to shorten the vocabulary [2], and some improve the histogram by using the weight scheme. For example: TF-IDF (term frequency – inverse document frequency) method [1, 3].

In some cases of text classification, based on intellectual information technologies, naive Bayes classifier could be helpful, but it will be problematic when we try to classify a natural language automatically. In order to solve these issues, parameters are normalized.

II. STATEMENT OF A PROBLEM AND THE CONCEPT OF THE PROBLEM DECISION

Assume we have a V set of words of a language. Usually V set is called vocabulary. The validity of the V N $(N=|V|)$ is equal to the number of words in it. Based on the V set, a vector of $S=(S_1,S_2,...,S_N)$ words is formed. The $K=\bigcup_{i=1}^{m}K_i$ set of texts, say, is categories.

Say, we have a D_j text of K_i category: $(i=\overline{1,m};\ j=\overline{1,p})$. The probability of the D_j text lying in the K_i set, according to the Bayes theorem $P(K_i \mid D_j)$, is equal to:

$$P(K_i \mid D_j)=\frac{P(D_j \mid K_i)P(K_i)}{P(D_j)} \Rightarrow P(D_j \mid K_i)P(K_i) \quad (1)$$

With the given D_j text, G_j set of words is formed and $W_j=(w_{j1},w_{j2},...,w_{jr})$ vector of words is created, that matches the G_j set.

Based on the vector of S words, $X_j^{\,i}$ Boolean vector, with N dimension, is formed:

$$x_{jt}=\begin{cases}1, & if \quad s_t=w_{je} \qquad j=\overline{1,p};\\ 0, & otherwise \qquad t=\overline{1,N}; e=\overline{1,r};\end{cases}$$

If the probability of the s_t word is in the K_i is $P(s_t \mid K_i)$, then the probability of s_t is not in K_i equals to $(1-P(s_t \mid K_i))$. Then according to (1), the probability of D_j text belongs to K_i will be determined thus:

$$P(D_j \mid K_i)=P(S \mid K_i)=\prod_{t=1}^{|V|}[x_t P(s_t \mid K_i)+(1-x_t)(1-P(s_t \mid K_i))] \quad (2)$$

Say the number of documents (that have s_t words from K_i) is $\eta_{K_i}(s_t)$ and the number of documents that belong to that category is N_{K_i}; Then the probability of s_t word is equal to:

$$\hat{P}(s_t \mid K_i)=\frac{\eta_{K_i}(s_t)}{N_{K_i}} \quad (3)$$

If the total number of the learning documents is N, then the probability of documents belonging to K_i is:

$$\hat{P}(K_i)=\frac{N_{K_i}}{N} \quad (4)$$

The Bernoulli model of classification of the set of learning documents and the texts of K_i category is carried out through the following steps:

1. V vocabulary is created.
2. Learning.
3. Classification

To determine a category of a non-classified document D, the combinations (1) and (2) are used:

$$P(K_i \mid S)\Rightarrow P(S \mid K_i)P(K_i)\Rightarrow$$
$$P(K_i)\prod_{t=1}^{|V|}[x_t P(s_t \mid K_i)+(1-x_t)(1-P(s_t \mid K_i))] \quad (5)$$

In order to classify texts of greater magnitude, usually multi-nominal model is used, which is more effective than the Bernoulli model. Below is a detailed explanation of it.

Revised Version Manuscript Received on 16 September, 2019.

* Correspondence Author

O.J.Babomuradov Ozod1, Tashkent university of information technologies named after Muhammad al-Khwarizmi

N. S. Mamatov2, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan

L. B. Boboev3, Scientific and Innovation Center of Information and Communication Technologies at TUIT named after Al-Kharezmi, Tashkent, Uzbekistan

B.Otaxonova4, Tashkent university of information technologies named after Muhammad al-Khwarizmi

In the multi-nominal model, a vector of signs is created based on the repetition of a word in a vocabulary-based text.

Multi-nominal division comprises the basis of multi-nominal model. Multi-nominal coefficient for N words of m type is calculated with the below formula:

$$M_k = \frac{N!}{n_1! n_2! \ldots n_2!}$$

Here, n_i is the amount of repetition i word from the given vocabulary.

Multi-nominal division of words based on the multi-nominal coefficient is calculated with the following formula:

$$P(N) = \frac{N!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N} = \frac{N!}{\prod_{t=1}^{N} n_t!} \prod_{t=1}^{N} p_t^{n_t} \quad (6)$$

Here, the probability of n_i words' sequence is divided by the $\prod_{t=1}^{N} p_t^{n_t}$ multiplication, and classify the target.

Say, n_t is the frequency of s_t word in a D_j document. In that case, the probability of s_t is in the K_i equals to: $P(s_t | K_i)$. Then, the probability of D_j text belongs to K_i, i.e. the probability of S words belong to K_i is:

$$P(D_j | K_i) = P(S | K_i) = \frac{N!}{\prod_{t=1}^{|V|} n_t!} \prod_{t=1}^{|V|} P(s_t | K_i)^{n_t} \Rightarrow \prod_{t=1}^{|V|} P(s_t | K_i)^{n_t} \quad (7)$$

Due to the fact that the normalization doesn't concern whether the s_t word is the property of any class, it is not necessary to conduct a normalization.

In the multi-nominal model, the probability of the $P(s_t | K_i)$ category document and $P(K_i)$ category will develop parameters for the model. Whether D_j document belongs to K_i category, is created by evaluating the parameters of a set of learning documents, and valued with 1 or 0. When the total number of documents is N, $P(s_t | K_i)$ probability is determined through the below formula:

$$\hat{P}(s_t | K_i) = \frac{\sum_{j=1}^{N} n_{jt} z_{fi}}{\sum_{f=1}^{|V|} \sum_{j=1}^{N} n_{jf} z_{ji}} = \frac{n_i(s_t)}{\sum_{f=1}^{|V|} n_i(s_f)} \quad (8)$$

$\{Y_1, Y_2, ..., Y_l\}$ is formed based on the set of learning documents, that is, if Y_t belongs to K_i category, z_{ti} variable is 1, otherwise it is 0.

Say, Y set of learning documents and K set of categories are given, the algorithm of text classification based on multi-nominal model would be as follows:

1. V vocabulary is developed;
2. The followings will be calculated:
   - N – total number of documents
   - N_k – the number of documents, that belong to category k, is determined $k = \overline{1, K}$
   - n_{it} the frequency of the word s_t in D_i document, for each word in V, is calculated; simultaneously, the $n_i(s_t)$ frequency of s_t words in K_i category documents is determined;
3. Using (8), $P(s_t | K_i)$ probability is calculated.
4. Using (4), $P(K_i)$ probability is calculated.
5. Whether a text belongs to K_i category is found out thus.

When classifying the D_j document, the category probability is calculated through the combinations of (1) and (7):

$$P(K_i | D_j) = P(K_i | S) \Rightarrow P(S | K_i) P(K_i) \Rightarrow P(K_i) \prod_{t=1}^{|V|} P(s_t | K_i)^{n_t} \quad (9)$$

Unlike the Bernoulli model, in the multi-nominal model, words that don't exist ($s_t = 0$) in a document don't affect the probability ($p^0 = 1$).

If the words in a document are symbolized as u, the probability is calculated as follows:

$$P(K_i | D_j) \Rightarrow P(K_i) \prod_{t=1}^{len(D)} P(u_t | K_i) \quad (10)$$

Here, u_t is the t-nth word in document D_j.

In experimental procedure, the change in time of transformation was observed. TfidfVectorizer and HashingVectorizer transformation approaches were used to verify the reliability of results, as shown in Figure 1.

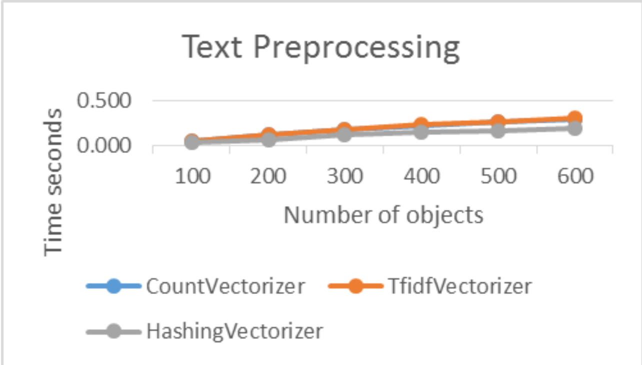

Fig. 1: Time-consuming comparison of different types of transformation

Here are results of algorithm based on Naive Bayes probability with CountVectorizer. Classification accuracy being increased from 78% to 88%.

When applying TfidfVectorizer approach, speed and accuracy were low: 78-88% accuracy was obtained.

The following table shows the result of a comparison models' accuracy and time consuming (Table 1).

Table 1: Comparison of classification models

Model	Precision	Time
BernoulliNB	0.65	0.017
MultinomialNB	**0.86**	**0.009**
LinearSVC	0.82	0.89
Perceptron	0.86	0.019

To assess the effectiveness of the classification models such as the Bernoulli and multi-nominal, 600 documents, with 169205 words of 6 categories in it, have been used and with the set of documents, a 28343-word vocabulary has been created.

Retrieval Number: B14930982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1493.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

When testing the classification of the selected texts with the Bernoulli model, average accuracy was 65% and it took 17.28 milliseconds. As for the multi-nominal model, the accuracy came to about 86% and it took 9.79 milliseconds. Experimental research works have proven the multi-nominal model more accurate and faster than the Bernoulli.

III. CONCLUSION

Pre-processing of texts, with Bernoulli and multi-nominal methods, has been looked through. The space for symbols, which is the most important for the classification of texts in Uzbek language, and mathematical way of classification have been developed. In order to make them recognizable, texts of various themes were formed and classified into categories. The results show the effectiveness of the multi-nominal model, when classifying the texts of bigger size.

REFERENCES

1. C.D. Manning, P. Raghavan, H. Schütze. “Introduction to Information Retrieval”, Cambridge University Press, Cambridge (2008)
2. R.E.Madsen, S.Sigurdsson, L.K.Hansen, J.Larsen. “Pruning the vocabulary for better context recognition”, Proceedings of the International Conference on Pattern Recognition, 2 (2004), pp. 483-488
3. J.H. Paik. “A novel tf-idf weighting scheme for effective ranking”, Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (2013), pp. 343-352
4. S.B. Kim, K.S. Han, H.C. Rim, S.H. Myaeng. “Some Effective Techniques for Naive Bayes Text Classification”, IEEE Transactions on Knowledge and Data Engineering, December 2006
5. Y.Yiming, J.Thorsten “Text categorization”, (2008), Scholarpedia, 3(5):4242
6. P.Philipp, W.Bonnie “Stable classification of text genres” (2011) Association for Computational Linguistics Vol. 37, No 2., pp.385-393,
7. B.S.Harish, D.S.Guru, S.Manjunath “Representation and classification of text documents: A Brief Review”, IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” RTIPPR, 2010., pp.110-119
8. N.Kamal, K.Andrew, T.Sebastian, M.Tom “Text classification from labeled and unlabeled documents using EM”, Machine Learning, 1–34 Kluwer Academic Publishers, Boston.
9. A.A.Alekseev, A.S.Katusyev, A.E.Kirillov, A.P.Kirpichnikov "Classification of text documents based on text mining technology" Computer Science, Computer Science and Management, Bulletin of the Technical University. 2016. T.19, No.18, pp.116-119
10. I.V.Polyakov, T.V.Sokolova, A.A.Chepovskii, A.M.Chepovsky "The problem of classification of texts and differentiating signs" Bulletin of the Novosibirsk State University of shock. Series: Information technology. 2015. Vol. 13, No. 2., pp. 55-63

.