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 \( S = (S_1, S_2, ..., S_s) \) words is formed. The \( K = \bigcup_{i=1}^{n} K_i \) set of texts, say, is categories.

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

\[
P(K_i | D_i) = \frac{P(D_i | K_i) P(K_i)}{P(D_i)} = \frac{P(D_i | K_i) P(K_i)}{P(D_i)} (1)
\]

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

Based on the vector of \( S \) words, \( \bar{x}_j \) Boolean vector, with \( N \) dimension, is formed:

\[
x_j = \begin{cases} 1, & \text{if } s_j = w_{j1}, j = \overline{1,p}; \\ 0, & \text{otherwise } t = \overline{1,N}; e = \overline{1,r}; \\
\end{cases}
\]

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

\[
P(D_i | K_i) = P(S | K_i) = \prod_{j=1}^{n} P(s_j | K_i) + (1 - \bar{x}_j)(1 - P(s_j | K_i)) (2)
\]

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

\[
\hat{P}(s_j | K_i) = \frac{\eta_{s_j}(s_i)}{N_{k_i}} (3)
\]

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

\[
\hat{P}(K_i) = \frac{N_{k_i}}{N} (4)
\]

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:

\[
P(K_i | S) \Rightarrow P(S | K_i) \hat{P}(K_i) \Rightarrow P(K_i \prod_{j=1}^{n} [x_j P(s_j | K_i) + (1 - \bar{x}_j)(1 - P(s_j | K_i)))] (5)
\]

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.
Text Documents Classification in Uzbek Language

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:

\[ M_i = \frac{N!}{n_1!n_2!...n_J!} \]

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:

\[ P(N) = \frac{N!}{n_1!n_2!...n_J!}p_1^{n_1}...p_J^{n_J} = \frac{N!}{\prod_{i=1}^{J}n_i}p_1^{n_1} \]  \(6\)

Here, the probability of \(n\) words’ sequence is divided by the \(\prod_{i=1}^{J}n_i\) multiplication, and classify the target.

Say, \(n_i\) is the frequency of \(s_i\) word in a \(D_i\) document. In that case, the probability of \(s_i\) is in the \(K_i\) equals to: \(P(s_i \mid K_i)\).

Then, the probability of \(D_i\) text belongs to \(K_i\), i.e. the probability of \(S\) words belong to \(K_i\) is:

\[ P(D_i \mid K_i) = \prod_{i=1}^{n} P(s_i \mid K_i) \]  \(7\)

Due to the fact that the normalization doesn’t concern whether the \(s_j\) word is the property of any class, it is not necessary to conduct a normalization.

In the multi-nominal model, the probability of the \(P(s_i \mid K_i)\) category document and \(P(K_i)\) category will develop parameters for the model. Whether \(D_i\) 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 \(N\), \(P(s_i \mid K_i)\) probability is determined through the below formula:

\[ \hat{P}(s_i \mid K_i) = \frac{\sum_{\substack{\gamma \in \{1,2,...,N\}\n \gamma \neq 0}} n_i(s_i)}{\sum_{\substack{\gamma \in \{1,2,...,N\}\n \gamma \neq 0}} n_i(s_i) + \sum_{\substack{\gamma \in \{1,2,...,N\}\n \gamma = 0}} n_i(s_i)} \]  \(8\)

\(Y, Y_1, ..., Y_n\) is formed based on the set of learning documents, that is, if \(Y_i\) belongs to \(K_i\) category, \(z_n\) 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
   - \(N_i\) – the number of documents, that belong to category \(k\), is determined \(k = 1, K\)
   - \(n_i\) the frequency of the word \(s_i\) in \(D_i\) document, for each word in \(V\), is calculated; simultaneously, the \(n_j(s_i)\) frequency of \(s_j\) words in \(K_i\) category documents is determined;
3. Using (8), \(P(s_i \mid K_i)\) 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 \(D\) document, the category probability is calculated through the combinations of (1) and (7):

\[ P(K_i \mid D) = \frac{P(D \mid K_i) \cdot P(K_i)}{\sum_{k=1}^{K} P(D \mid K_k) \cdot P(K_k)} \]  \(9\)

Unlike the Bernoulli model, in the multi-nominal model, words that don’t exist (\(s_j = 0\)) in a document don’t affect the probability (\(p^n = 1\)).

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

\[ P(K_i \mid D) = \prod_{u=1}^{len(D)} P(u \mid K_i) \]  \(10\)

Here, \(n_i\) is the \(i\)-th word in document \(D_i\).

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>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>BernoulliNB</td>
<td>0.65</td>
<td>0.017</td>
</tr>
<tr>
<td>MultinomialNB</td>
<td><strong>0.86</strong></td>
<td><strong>0.009</strong></td>
</tr>
<tr>
<td>LinearSVC</td>
<td>0.82</td>
<td>0.89</td>
</tr>
<tr>
<td>Perceptron</td>
<td>0.86</td>
<td>0.019</td>
</tr>
</tbody>
</table>

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.
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

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