LANGUAGE IDENTIFICATION BASED ON N-GRAM FEATURE EXTRACTION METHOD BY USING CLASSIFIERS

Şengül BAYRAK HAYTA[1], Hidayet TAKÇI[2], Mübariz EMİNLİ[3]

[1,3]Halic University, Department of Computer Engineering, Istanbul, Turkey
[2] Cumhuriyet University, Department of Computer Engineering, Sivas, Turkey
sengulbayrak@halic.edu.tr, htakci@cumhuriyet.edu.tr, mubarizeminli@halic.edu.tr

Abstract: *The rising opportunities of communication provided us with many documents in many different languages. Language identification has a key role for these documents to be understandable and to study natural language identification procedures. The increasing number of the documents and international communication requirements make new works on language identification obligatory. Until today, there have been a great number of studies on solving language identification problem about document based language identification. In these studies, characters, words and n-gram sequences have been used with machine learning techniques. In this study, sequence of n-gram frequencies will be used and using of the five different classification algorithms' accuracy performances will be analyzed via different sizes of documents belonging to 15 different languages. N-gram based feature method will be used to extract feature vector belonging to languages. The most appropriate method for the problem of language identification will be identified by comparing the performances of the Support Vector Machines, Multilayer Perceptron, Centroid Classifier, k-Means and Fuzzy C Means methods. During the experiments, trainining and testing data will be selected from ECI multilingual corpus.*

Keywords: *Document Based Language Identification, ECI Corpus, n-Gram Feature Extraction Method, Machine Learning Algorithms.*

1. Introduction

One of the most important advances of our time is experienced in the field of communication. The most important thing of communication is the language that is considered to remain away from these advances. The language problem is one of the most important problems to be solved on each passing day in the globalized world. Despite increasing in the amount of available documents, unfortunately, there is no opportunity to use the language of unknown documents. In order to make available sources of information more useful, language identification and language comprehension have significant duty. So, language identification is the first step of understanding the language, it has the key role in these studies.

Language identification is divided into two parts; speech identification and document identification. Spoken language identification is usually carried out by signal processing techniques and using structures that is called phonemes. In document based language identification is fulfilled with the letter sequences, words and n-gram frequencies. Language identification is the determination process of an unknown language by using features and algorithms including inside of the documents.

In our study, we focus on the identification of document based language. Language identification approaches are divided into two methods: linguistic methods and statistical methods. Linguistic method which is approach in language identification estimates the language in the documents according to the grammar rules belonging to languages. The document based approach in language identification, one of the linguistic method estimates the language according to the rules of grammar in language documentation.It makes searching according to the frequency of searches for words in the document and makes scoring them. In Turkish, vowel harmony word is given as an example for language identification. Fragmentation of the language of the document is not sufficient; its syntax structure and grammar should alse be considered. This increases the size of the complexity of language identification. This gives us a very good accuracy in the diagnosis of language linguistic approaches. The

Received on: 02.05.2013
Accepted on: 03.08.2013

best way to make language identification is to reflect the characteristics of the language expressed in statistical. In order to determine statistical language, the order of letters, the presence of certain keywords, frequencies of short words, unique and large articles or short distinctive character strings can be used.

The most commonly used method of identification in document is the statistical language identification. Statiscal language identification based on statistical analysis concept they use dimension of similarity to partition objects and they are limited with numerical data. Statistical methods make it possible to identify the language without linguist, but system training is an issue. In this type of language identification the mathematical principles to distinguish the languages are important [1]. Studies on statistical language models go back to Andrei Markov [2]. Markov modelled a case study of textual sequences in the data. Another famous work is by Claude Shannon modelling study of letter and word sequences [3]. The best way to identify document based language is fragmantation of the text to reflect the properties of the language. The terms of statistical language models are built on the weights of the terms or probabilities. According to some of the terms in this dictionary are more important than the others. Importance of the terms is sometimes given by the weight by the possibility. This kind of identification is known as machine learning language identification and machine learning algorithms are the most important terms of classification.

In this study, statistical models and machine learning methods have been used for language identification. N-gram sequences have been summarized to search for language of documents and each language has been indicated form of a vector. In our experiment 15 different languages of Latin and European origin are used. Despite the fact that identification of the languages that are close to create a statistically significant risk, the reason of prefering this is to show n-gram frequencies are successful to a great extent. In addition, documents of different sizes are used to indicate the importance of the sizes of the documents which are tested in language identification. To identify the language there have been three phase processes which are pre-processing, extract of feature vector and to the implementation of the classifiers. 1 KB and 100 KB sized documents have been selected from ECI-CDROM [4] corpus. Training data has been selected 70% and test data has been selected 30% in terms of accuracy performance. N-gram feature extraction method has been used for feature vector. Support Vector Machines, Multilayer Perceptron, Centroid Classifier, k-Means and Fuzzy C Means methods have been compared in terms of performance and the best method has been tried to be selected in terms of the problem of language identification.

1.1. Related Works

One of the simplest approaches is the use of the general words in identifying documental language[5].General words, decisive, conjunction, preposition in a language such as providing relevant information in the determination of common words in the language. Another is the use of sequences of n-gram frequencies. The best-known work on the subject Cavnar and Trenkle [6] conducted by their study that is an example of text classification methods are also valuable for language identification. Cavnar and Trenkle made many experiments and language identification based on different values of n up to 3-grams were found to be successful. Also, Suzuki [7] and his friends have been used n-gram frequencies with character encodings. Another study of n-gram based Adams and Resnik [8], dynamic web pages for all the documents have proposed a system that add language tags. To this aim, 5-grams was extracted from 220 KB training document and for 100-500 byte test document were obtained 98.68% accuracy. Another approach is using letters. 2, 3, 4, and using more letters in English, Spanish, French and Portuguese languages have on the language identification on Beesley [9] work. Dunning [10] was used machine learning algorithms for Spanish and English by using Bayesian classifier in the documents as short as 20 bytes, even as high as 92% results. Combrink and Botha [11] represented a text based language identification system for 12 languages. System is based on transition vectors. Transition vector occurs from one or more characters. The most frequent character strings are used like n-grams. System builds histograms from lots of hits for transition vectors of each language. Grefenstette[5] used 3-gram and short terms method and oversized feature set is a big problem for these two methods. The success rate of 98.96% with 3-gram method for 100 byte document. The success of short terms method carried out 98.68% accuracy 100 byte test data. Prager [12] proposed Linguini system that uses a vector space based classifier. Cosine similarity function has been used for determining language from the given feature vector. Xafopoulous and Xafopoulous friends [13] have proposed an HMM (Hidden Markov Model) based on language identification system for character strings. This system was automatic language detection on web documents. In the experiments, English, German, French, Spanish and Italian were used and for 140 byte test data 99% accuracy was reached. Takci and Soğukpınar [14] was proposed a centroid-based text categorization algorithm has been used for four languages. Their method was used 22 letters. These languages were English, French, German and Turkish. For 500 KB training data, 98% accuracy has been reached for small data. Another language identification work is Takçı and Ekinci[15] proposed 9 different languages with 60 letters that is help of minimum feature set. Using by C-SVC, MLP, LDA classifiers, The most fastest classifier was C-SVC for in this study.

2. Theoritical Background

To express language statistical, the order of letters, the presence of certain keywords, frequencies of short words (a combination of presence) is decisive and each language is represented its descriptive features. For it is extracted from the language's features and for this n-gram feature extraction method is used. Using these features in the form of a feature vector representation of documents is one of the most widely used methods [14]. The purpose of this study, to date, comparison of accuracy rate of the language identification approach based on classification methods, and so, reveal the most appropriate methods. Preprocessing and extracting features processes are the first step to identify languages.

2.1. Preprocessing Step

Machine learning methods are often designed to work with numerical data, although language identification is to study with document based data. For this reason, the document data that is brought into line with machine learning algorithms is essential to undergo a transformation. In this study, to use of n-gram feature set, texts have been converted to the frequency of n-grams at conversion stage. Despite obtaining n-gram frequencies are used in many studies' bag-of-words reporting, this study, the experimental design will be given n-gram frequency vector space model. Thus, each document as a vector of n-gram frequencies hold a document. Offering documents, statistical language models, these vectors are often also referred to as the document model or document profile. At the same time, the language or languages as a profile name that offers vectors are the language model.Two basic phases for character based n-gram preprocessing is data cleaning and transformation.

Data is filtered on data cleaning phase. Filtering can be of the document and the basis of the n-gram. Without too much information when the document filtering, often repeats in the documents are eliminated. Because of repetition of document, affect words and characters frequencies negatively. Because of repeating words that are not include information, these words are eliminated at basis on document filtering. Numerical expressions, punctuation, spaces, characters are eliminated on documents that are common on all languages.

2.2. Feature Vector Extraction with N-gram Method

N-gram is sequence of letters sequentially that formed determined by the value of n. N-gram definition can be defined that is spelled by a specific value, in accordance with a certain rule. In the text field of all the n-gram frequencies to obtain with the help of sliding windows on the text width of the n.Those obtained by counting the n-grams again involving how many pieces are each n-gram hold in a vector called profile. Trenkle and Cavnar [7] extracted n-gram profil vector on the characters with n-gram feature extraction method in their study. This study, profile vector covered all of the properties of n-gram frequency that is very complex vectors. However, it is referenced for further vector dimensionality reduction studies [5] [14] [15]. A number of adjacent letters in a formed in a couple of values as the length of the n-grams of 2-gram, 3-gram, and quad-gram are used in a word [5]. N-gram in a word can be represented as a set of overlapping. Because of previous studies give better results about language identification, in this study received the value of n is chosen 3 (3-gram, trigram)[5].Each word occurs with a different frequency and small sized data is a quick result. So, at first system makes the calculation of the n-gram frequencies to compare profiles. 2-gram, 3-gram, and quad-gram are used as the length of different numbers. For example,"student" separated to n-gram is shown on Table1.

Table 1. An Example of N-gram Sequence

bi-grams	:	st	tu	ud	de	en	nt
tri-grams	:	stu	tud	ude	den	ent	
quad-gram	:	stud	tude	uden	dent		

In our study, a total of 60 letters including in a mixed of 15 different languages (Turkish-tur, English-eng, German-ger, Dutch-dut, French-fre, Italian-ita, Algerian-cze, Spanish-spa, Portuguese-por, Norwegian-nor, Maltese-mal, Latin-lat, Lithuanian-lit, Swedish-swe, Andoain-gae) used letters of the alphabet. Letter feature set consist of 60 letters that are recognized by 26 letters of the English alphabet and other languages' special letters on language identification system elements [15]. In other words, our feature set is a combination of 15 languages alphabets that low discriminant features are not included in the set of letters issued during feature extraction.

Table 2. Letter Feature Set on Creation of Feature Vector

'a', 'à', 'á', 'â', 'ã', 'ä', 'å', 'æ', 'b' , 'c', 'ç', 'd', 'e', 'è', 'é', 'ê', 'ë', 'f', 'g', 'ğ', 'h', 'i', 'ı', 'ì', 'í', 'î','ï', 'j', 'k', 'l', 'm', 'n', 'ñ', 'o', 'ò', 'ó', 'ô', 'õ', 'ö', 'p', 'q', 'r', 's', 'ş', 't', 'u', 'ù', 'ú','û', 'ü', 'v','w', 'x', 'y', 'ÿ', 'ß', 'ø', 'z'

Maximum number of letter feature set that may occur is 60x60x60 = 216000 3-gram frequencies. However, due to the problem of performance the representative value at the beginning of the study, which is a better representation of the classes of n-gram frequencies have been selected in the first 300 3-gram frequencies, so that performance problem has been solved appropriately. Values of the first 300 3-gram frequencies have been chosen with Fisher filtering [16] method through Tanagra is machine learning tool. The

first 300 attributes have been obtained from 100 KB training document.

For each languages have been used a total of 100 KB sizing documents for training and test data from ECI corpus [4] and 100 KB documents have been divided into 1 KB segments so, the values of 3-gram frequencies are created sequentially. For each languages have been obtained language profiles. After that, for 15 languages have been found 15 language profiles. As shown in Table 3 which is based on the most important 3-gram frequencies by sequences and languages as 100 KB training documents data. For example, 'the' 3-gram sequence is for English, 'lar' 3-gram sequence is for Turkish the most significant frequencies.

Table 3. Sequences of the Decisive 3-gram Frequencies by Languages

the	1483	ENG
kan	1168	MAL
zio	212	ITA
ist	417	GER
die	212	DUT
lar	760	TUR
ada	196	SPA
fro	343	GAE
som	71	SWE
jeg	535	NOR
ndo	302	POR
ost	320	CZE
que	314	LAT
ent	663	FRE
usi	329	LIT

2.2.1. Profile Based Method with 3-gram Frequencies

Machine learning methods are two basic approaches that are representing a class's is a centroid value or a class that represents of many examples cases. The fact that a single instance of a class that the method known as profile based method. Profile of the class is a value near the center is capable of representing the values. This method is known as centroid based methods that centroid values represent all classes and so many of this kind classifiers training data set is presented with a single vector.

2.2.2. Instance Based Method with 3-gram Frequencie

Contrary to profile based method, instance based method refers to is a class that represents a large number of instances. To tell the training phase, a class is reported vectors that class as a model presents many instances and collectively represent the class. Many machine learning algorithms use this method. Cost of the transaction is more than other methods but data presentation ability is greater than other methods and more suitable for complex data.

2.3. Using Classificion Methods

N-gram method is a simple and reliable method used to classify documents. The basic idea is to define a document in the formation of n-gram frequencies. N-gram frequency approach works regardless of the language. So, a particular language does not require a detailed structure of grammar or a dictionary. Possible to achieve similar results by using all the letters or syllables statistics.

After extracted 3-gram profiles through train and test documents, classification algorithms are applied to these data. Classification algorithms are described comparatively. Our study, Centroid Based, k-Means, Fuzzy C Means classifiers are applied with profile based method and Support Vector Machine, Feed Forward Back Propagation classifiers are applied by instance based method.

2.3.1. Centroid Based Classifier

Centroid based classifier that is the vector space model based and high performing method for document classification. Vector space model is the provision of opportunities provided by the document that takes each document d term space. The size of each document vector of the frequency of 3-gram weight. Classes are hold in centroid vectors that is offered class elements are an average value and the value of this medium is considered to represent the all classes. The training set contains different categories of k that is centroid vectors to obtain the training data [14]. In our study k is 15.

$$\vec{C} = \frac{1}{|S|}\sum_{d \in S} \vec{d} \tag{1}$$

The similarity between the centroid vectors with unknown languages of the document is to be identificated as follows.

$$sim(\vec{x}, \overrightarrow{C_{j,j\in k}}) = \frac{\vec{x}.\overrightarrow{C_j}}{\left\|\vec{x}\right\|_2 * \left\|\overrightarrow{C_j}\right\|_2} \tag{2}$$

The highest similarity to the document language, based on the value that is assigned to the centroid value after similarity calculation that is used cosine method. Selection of maximum similarity and there are presented the following equation.

$$arg\ max\ (cos_j(d1,d2)) = d1*d2 / ||d1||\ ||d2|| \tag{3}$$

In our study, cosine method has been used for centroid document classifier to classify an unknown category. In order to find out which documents belong

to the category of a test with the test document, the similarities between the centroid vectors are calculated sequentially. So, test document is assigned to the nearest category. The closest category is utilized to obtain the maximum similarity as formula 3.

2.3.2. Support Vector Machines (SVM)

Vapnik has been developed support vector machines algorithm[17] that is given optimal results in signal processing, artificial learning and data mining fields[17][1]. Basically, SVM interested in 2 classes problems. Knerr and his friends have been suggested one-against-one method for classify multi-class data [18]. In this method, n(n−1)/2 classifiers are generated with n class and each of data is trained with two classes. By this method, the problem of multi-class problem gets converted into two classes. Another method is for multi-class data classification [19] that is specified for n and so, SVM become for n classes. i.th SVM, as the data i own class using the class data, all of the data from the other classes as if they belonged to the second class agrees. So +1 label to, data, while all the data belonging to other classes and training. This way n gives -1 the label makes for a SVM.

In this study, SVM algorithm, more than two classes C-SVC algorithm is used by applications on the performance analysis. SVM is a comprehensive supervised classification algorithm [20] and is used in document based language identification study [21].

2.3.3. k-Means Clustering Algorithm

k-Means clustering solves the problem by unsupervised learning that is the simplest method. Algorithm's general logic is as the input parameter to divide the k-cluster through data object of a data set consists of n. The aim of this algorithm is obtained at the end of the partitioning process, intra-cluster similarity is to ensure that the maximum and minimum inter-cluster similarity. The performance of the method is effected the number of clusters k, the initial cluster centers are selected as the criteria for the measurement of values and similarities [22].

Method's steps are showed following:

Step 1: Before clustering, to determine initial cluster centroid value for number of k cluster. c = $\{c_1, c_2, \ldots c_k\}$. For this, objects are selected between random point (k = 15).

Step 2: Training set in the each data, $x_i = \{x_1, x_2, \ldots x_p\}$ are included in the nearest or similar cluster with selected initial cluster centers. To calculate cluster that is used similarity formula (p = 300).

$$cos\ \ x_i, center\ \ c_j\ \ = \frac{x_i}{||x_i||} * \frac{center(c_j)}{|\ center\ \ c_j\ \ |}$$

(i={1,2,3...n},j={1,2,3...k}) (4)

Step 3: It is composed of a cluster center points of clusters are changed with the average values of all the objects.

$$center(c_j) = \frac{\sum_{i=1}^{n_j}(x_i)}{n_j} \quad (5)$$

($x_i \in c_j$) and n_j, j.cluster data number

Step 4: Repeat Step 2 and Step 3 while unchanging cluster centers for language identification.

2.3.4. Artificial Neural Networks (ANN)

Generally, Artificial Neural Networks (ANN) classifier can be defined as a system that is designed to simulate the working principle of the human brain. At learning process, in order to achieve the desired purpose learning algorithm that includes the renewal of the ANN weights. Power of ANN algorithm about calculation and data mining takes its ability to learn and generalization that encountered in the process of training or learning is defined as the inputs of ANN to produce appropriate responses. This cutting-edge feature, ANN shows the ability to solve complex problems [23].

Our study while text-based language identification, feed forward backpropagation artificial neural network learning algorithm used in the classification. Feed Forward Back Propagation Algorithm has been developed for nonlinear cases. Back Propagation is a specific technique for implementing weight for a multilayer perceptron. The basic idea is to efficiently compute partial derivatives of an approximating function realized by the network with respect to all the element of the adjustable weight vector for a given value of input vector [24]. Owing to nonlinear activation functions, it can be easily classify nonlinear data. Since Feed Forward Back Propagation Algorithm is an appropriate method for classification, it can be used for language identification. Input layer's neuron number is 300 that are equal to the feature vector size. Hidden layer neuron number is 16 and output layer neuron number is 15 that are number of language classes. Feed Forward Back Propagation learning algorithm's layer architecture is given as Figure1.

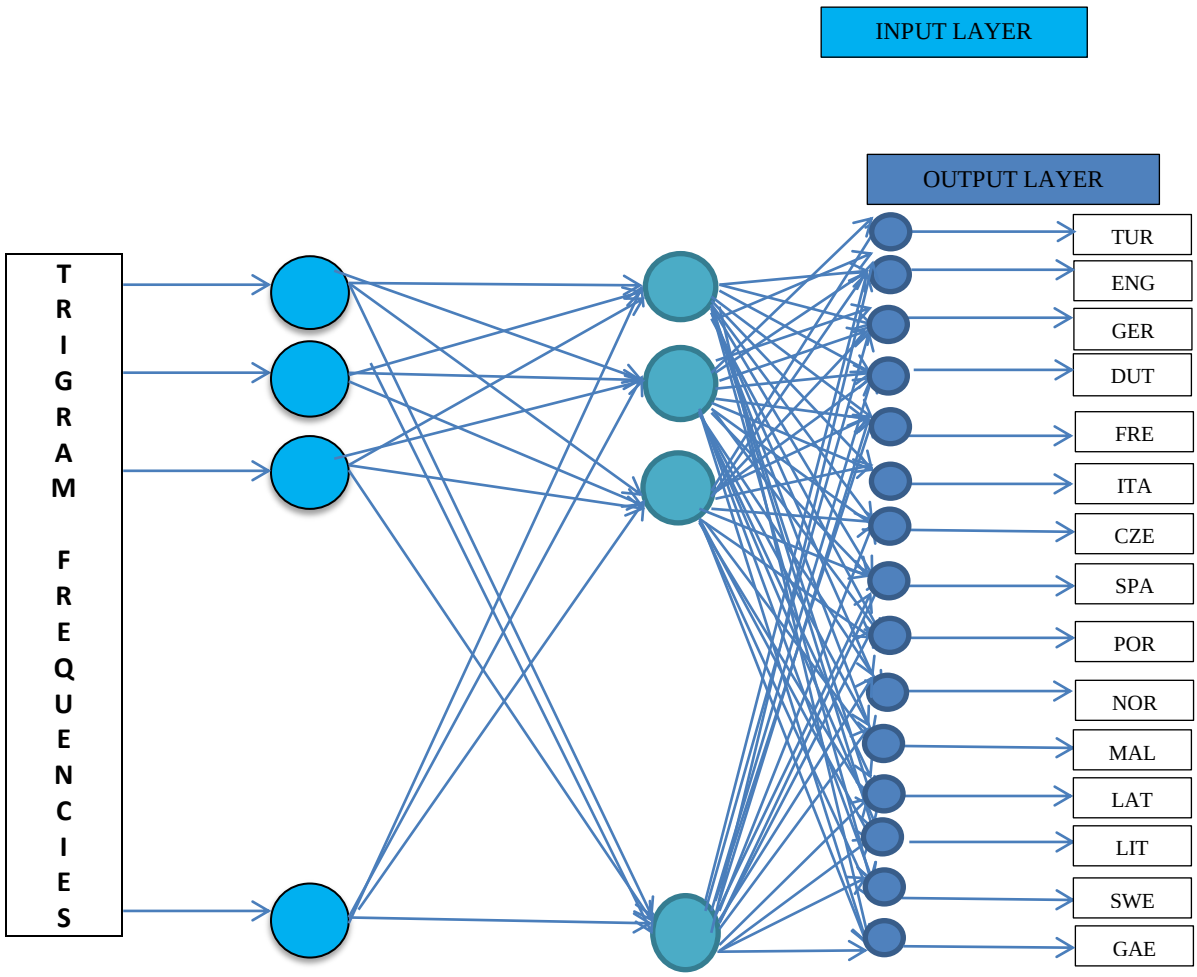

Figure 1. Feed Forward Back Propagation Algorithm Layer Architecture

2.3.5. Fuzzy C Means Algorithm (FCM)

Fuzzy logic can be defined as a solid mathematical order to establish for studying the uncertainities expression. Statistics and probability theory works certainity, but people live in certain events; the media is filled with more uncertainty. Therefore, to understand the ability of human beings that should be worked with uncertainty [22]. Fuzzy C Means (FCM) is a suitable method in some cases for objects can not assign to a precisely cluster.

Essential steps for FCM algorithm:

Step 1: To be selected number of cluster and fuzzication index value (m).The initial values is determined for U matrix that is showing the degree of membership for by the matrix.

Step 2: Assuming cluster centroid(c) values are generated randomly , using by these values degrees of membership of account is as follows (m=2)

$$u_{ji} = \sum_{l=1}^{c} \left(\frac{\|c_j, x_i\|}{\|c_l, x_i\|} \right)^{-(2)/(m-1)}$$

$i = \{1, 2, 3...n\}, n = 300,$

$j = \{1, 2, 3...c\}, c = 15$ (6)

Step 3: c values of the cluster prototype is updated according to the Eqn(6).

$$c_j = \frac{\sum_{i=1}^{n} u_{ji}^{m} x_i}{\sum_{i=1}^{n} u_{ji}^{m}} \quad (7)$$

Step 4: If $\left\| C^{(t)} - C^{(t-1)} \right\| < \varepsilon$ iteration is stopped,

otherwise return to step 2. (8)

After Fuzzy C Means (FCM) algorithm has been applied that language is determined by the degree of membership of the clusters. For each language is calculated the other languages' degree of membership and result for the highest membership degree of the cluster that are included in the language. A language may be associated with more

than one cluster. Languages associations are given membership degrees.

3. Experimental Studies

In our study, documents of 15 languages (Turkish-tur, English-eng, German-ger, Dutch-dut, French-fre, Italian-ita, Algerian-cze, Spanish-spa, Portuguese-por, Norwegian-nor, Maltese-mal, Latin-lat, Lithuanian-lit, Swedish-swe, Andoain-gae) have been selected from ECI-CDROM corpus. Documents have been selected two different sizes that are 1 KB and 100 KB. Training data has been selected 70% and test data has been selected 30% in terms of accuracy performance. Training and test data vectors are shown in Table 4 as two different sizes documents.

Table 4. Training and Testing Data Vectors Size

	Training File(%70)	Vector Size(For Training)	Test File(%30)	Vector Size(For Testing)
1KB	105 documents	105x300	45 documents	45x300
100 KB	1050 documents	1050x300	450 documents	450x300

In the experimental study 1 KB sized document files are represented with 45x300 vectors for training, 105x300 vector for testing and 100 KB sized documents have been represented as 1050x300 vector for training and 450x300 vector for testing vector space. These vector space datahas been applied as input for Centroid Based, Feed Forward Back Propagation, SVM, k-Means, FCM machine learning algorithms [25]. Accuracy rates of machine learning algorithms have been tested after training process. In this study, SVM, k-Means classifiers accuracy rate have been taken by Tanagra machine learning tool [16] and Feed Forward Back Propagation accuracy rate have been taken by MATLAB tool [26]. Accuracy rate of training and testing have been given in Table 5 according to documents files.

Table 5. Accuracy Rates of Training and Testing

Classifiers	**ACCURACY RATES**			
	1K Train	**1K Test**	**100K Train**	**100K Test**
Centroid Based	91.3	93.5	93.3	93
SVM	86.6	61	100	85
Feed Forward Back Propagation	73.5	91	92	77
k-Means	56	76.4	61.3	64
FCM	81	86.6	89.3	77

In our work, Centroid Based, SVM, Feed Forward Back Propagation classifier methods have been received from the best success as it is indicated in Table 5. The test results which are obtained in terms of the methods that are applied according to file sizes of the languages, are given in Table 6, Table7.

Table 6. Accuracy Rate of Classification Methods for Size of 1 KB Documents (45 documents)

	cze	dut	eng	fre	gae	ger	ita	lat	lit	mal	nor	por	spa	swe	tur
Centroid Classifier	100%	100%	67%	100%	67%	100%	67%	100%	100%	67%	67%	100%	67%	100%	100%
SVM	100%	60%	50%	57%	54%	60%	50%	55%	30%	40%	80%	70%	76%	71%	57%
Feed Forward	100%	67%	60%	66%	100%	100%	100%	100%	100%	100%	100%	60%	75%	66%	75%
k-Means	100%	33%	67%	33%	33%	33%	67%	100%	33%	100%	33%	33%	33%	33%	33%
FCM	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	67%	67%	33%	33%

Table 7. Accuracy Rate of Classification Methods for Size of 100 KB Documents (450 documents)

	cze	dut	eng	fre	gae	ger	ita	lat	lit	mal	nor	por	spa	swe	tur
Centroid Classifier	100%	100%	100%	100%	83%	93%	83%	97%	97%	83%	83%	97%	83%	97%	100%
SVM	93%	88%	92%	88%	69%	94%	89%	72%	24%	16%	88%	88%	91%	79%	90%
Feed Forward	91%	94%	93%	88%	68%	94%	92%	90%	70%	60%	89%	94%	75%	90%	90%
k-Means	84%	1%	32%	74%	100%	90%	19%	98%	91%	13%	82%	69%	12%	100%	88%
FCM	93%	97%	97%	93%	83%	93%	93%	83%	83%	97%	93%	30%	30%	30%	66%

Table 6 and Table 7 have been obtained from the results of experimental accuracy rate respectively given graphically in Figure 2 and Figure 3.

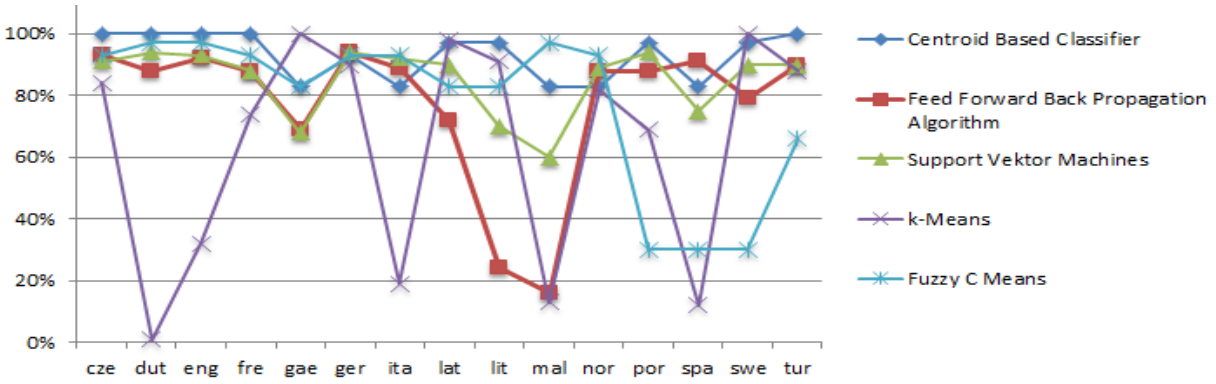

Figure 2. Accuracy Analysis of the 1 KB (45 documents) Document on the Test Data

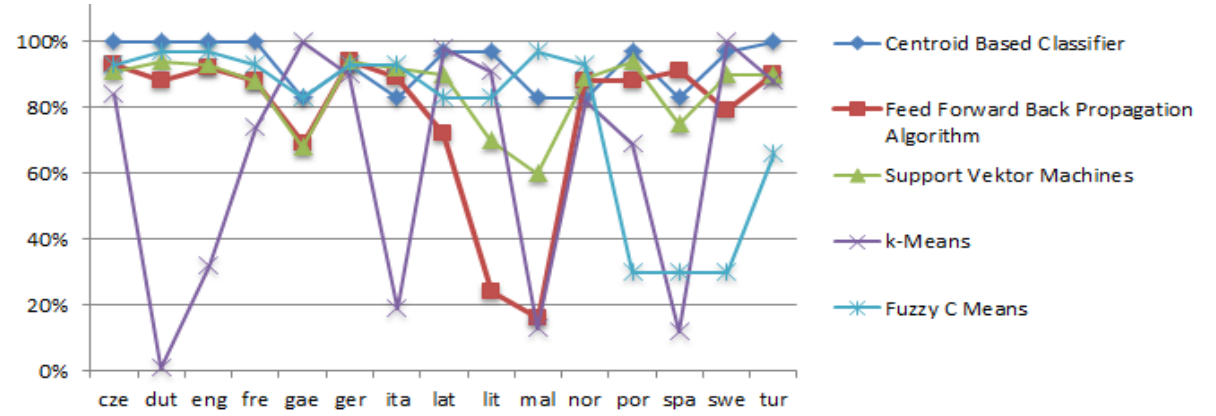

Figure 3. Accuracy Analysis of the 100 KB (450 documents) Document on the Test Data

In testing process, as shown in Figure 2 and Figure 3, the above methods have been given according to different languages accuracy rates because of language identification accuracy rate is based on the language. SVM and Centroid Based classifier have provided with our experiment the best identification performance and k-means algorithm has been observed to give the lowest performing in terms of experiment classification and FCM has provided a better result than k-Means. The reason for this is that fuzzy clustering has been done instead of classical clustering.

4. Conclusion

In this study, a private set of letters have been used in a mixed alphabet. Because of the mixed alphabet, 3-gram frequencies have been decreased with feature extraction. So, reduced feature selection, due to the increasing number of 3-gram frequencies of the mixed alphabet. Therefore, providing the capture of critical information.

It has been aimed to design a high performance system to be able to classify according to the document of unknown language. For this, two important things have been done to improve performance. The first one is done low dimensional feature space. 300-dimensional feature vector space has been focused on property of low dimensional feature space by passing preprocessing stage. Documents of the each language have consisted of 60 letters that is common to all languages combinations of 300 dimensional attribute values have been represented in the frequency. The reason for this is to provide increased importance of language representing qualifiers.

Our experiment could be regarded as the first among the similar ones since, 5 different classifiers have been studied in 15 different languages with different documents sizes. Empirically, after all with the order of Centroid Based, SVM, Feed Forward Back Propagation, FCM, k-Means have been the best classifiers.

This study based on documents has been represented with 3-gram feature extraction method by statistical language identification models. Language identification system which has been developed on the basis of this method, the author recognition, machine translation systems, and automatic response systems can be used in applications such as spam blocking, and intrusion

detection. Document based language identification system is aimed to transform the mobile applications in future studies.

5. REFERENCES

[1] T. Dunning,"Statistical identification of language", *Technical Report CRL Technical Memo MCCS-94-273, University of New Mexico*, ESANN.99, Belgium, 1994.

[2] Y. Ueda,, S. Nakagawa,, "Prediction for phoneme/syllable/word-category and identification of language using HMM.", International Conference on Spoken Language Processing, volume 2, pages 1209-1212,1990.

[3] C. Manning,, H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, Boston, 1994.

[4]European Corpus Initiative Multilingual Corpus (ECI/MCI)(2005),http://www.elsnet.org/resources/eciCorps. html, Page last modified 29-03-2005.

[5] G. Grefenstette,, "Comparing two language identification schemes", *Proceedings of JADT 3rd International Conference on Statistical Analysis of Textual Data*, 1995.

[6] W.B. Cavnar, J.M. Trenkle, "N-gram-based text categorization", *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas*, US 161–175,1994.

[7] I. Suzuki, Y. Mikami, A. Ohsato, Y. Chubachi, **"**A language and character set determination method based on N- gram statistics", *ACM Transactions on Asian Language Information Processing(TALIP)*, Volume 1 Issue 3, Pages 269 – 278, 2002.

[8] G. Adams, and P. Resnik,, "A language identification application built on the Java client-serandr platform", *In J. Burstein & C. Leacock (Eds), From Research to Commerc ial Applications: Making {NLP} Work in Practice Somerset, New Jersey: Association for Computational Linguistics*, pp. 43 – 47, 1997.

[9] K. Beesley, "Language Identifier: A computer program for automatic natural language identification of on-line text", *Proceedings of the 29th Annual Conference of the American Translators Association*, 47 – 54, 1988.

[10] T. Dunning, "Statistical identification of language", *Technical Report, New Mexico State Uniandrsity, CRL MCCS*-94-273, 1994.

[11] H. Combrinck, and E. Botha, "Automatic language identification: Resisting complexity", *South African Computer Journal*, 27, 18 – 26, 1995.

[12]J.M.Prager, Linguini, "Language identification for multilingual documents"., *Journal of Management Information Systems*, 16(3), 71 – 101, 1999.

[13] A. Xafopoulos, C. Kotropoulos, G. Almpanidis,, and Pitas, I., "Language identification in web documents using discrete hidden Markov models", Pattern Recognition, 37(3), 583 – 594, 2004.

[14] H. Takçı,, İ. Soğukpınar,, "Centroid-Based Language Identification Using Letter Feature Set", *Lecture Notes in Computer Science, (CICLING 2004) Springer-Verlag*, Vol.2945/2004, pages 635-645, February 2004.

[15] H. Takçı, E. Ekinci, "Minimal feature set in language identification and finding suitable classification method with it", SCIVerse Science Direct, Procedia Technology 1, pages 444-448, 2011.

[16]TanagraTool,http://eric.univlyon2.fr/ricco/tanagra/en/tanagra.html.

[17] C. Cortes, V. Vapnik, " Support-vector networks", *Machine Learning*, Vol: 20(3), pp.273–297, 1995.

[18] S. Knerr, L. Personnaz, G. Dreyfus, Single-layer learning revisited: A stepwise procedure for building and training a neural network, *in Neurocomputing: Algorithms, Architectures and Applications*, J. Fogelman, Ed. New York: Springer-Verlag, 1990.

[19] E. Mayoraz, and E. Alpaydin, "Support vector machines for multi-class classification," *in IWANN*, Vol: 2, pp. 833–842, 1999.

[20] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support andctor machines, 2001.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[21] T. Joachims, *Learning to Classify Text using Support Andctor Machines*. Kluwer, Boston, 2002.

[22] Z. Şen, *Mühendislikte Bulanık (Fuzzy Mantık) ile Modelleme Prensipleri*, Su Vakfı, İstanbul, 2004.

[23] Ç. Elmas, *Yapay Sinir Ağları Kuram, Mimari, Uygulama*, Seckin Yayıncılık, Ankara, 2003.

[24] S.Saarinen, R.B.Bramley, and G. Cybenko, 1992. "Neural Networks, backpropagation and automatic differentiation", in Automatic Differantiation of Algorithms: Theory, Implementation, and Application, A.Griewank and G.F.Corliss, eds.,pp.31-42,Philedelphia:SIAM.

[25] Ş. Bayrak, H. Takçı, M. Eminli, 2012."Makine Öğrenme Yöntemleriyle N-Gram Tabanlı Dil Tanıma", Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu, ELECO.

[26] http://www.mathworks.com/products/matlab/nntol

Şengül Bayrak was born in 1986. She was graduated from Halic University, Computer Engineering Department, 2009. She was graduated master degree from Halic University, 2011. She has worked as a Research Assistant Halic University since 2009. Her studies are about Data Mining, Language Identification, Database Management, Data Security.

Hidayet Takçı was born in 1974. He was licence degree from Trakya University, Computer Engineering, 2002. He was graduated master degree and PhD degree from Gebze Institute of Technology Computer Enginerering degreeHe worked as a lecturer in Gebze Institute of Technology, in Department of Computer Engineering, 2002-2011. He has worked Department of Computer Engineering, Cumhuriyet University as Vice President & Head of the Department Software Engineering since 2011. His experiments are about Data Mining and Text Mining, Neural Networks and Applications, Language Identification, Author Identification, Criminal Data Mining, Information Crime.

Mübariz Eminli was born in 1948. He was licence degree from Azerbaycan Technical University, Otomation and Computer Technologies Faculty, 1971. He was graduated master degree and PhD from Azerbaycan Technical

University Informatic and Computer Technologies Department. He has worked as a lecturer at Halic University Department of Computer Engineering since 2009. His experiments are Fuzzy Clustering, Fuzzy Modelling, Artificial Neural Network, Data Mining.