CENG-595 Neural Networks
Term Project Report

Document Classification with Neural Networks

By
Harun Reşit Zafer
50050819

Lecturer: Asst. Prof İhsan Ömür Bucak

11 June 2010

Fatih University

Department of Computer Engineering

Abstract

In this project, newspaper articles of 10 Turkish columnists have been analyzed in terms of different features such as word length, sentence length, frequency of some stop words and frequency of punctuation marks. Totally 12 features were used. Almost each feature of total 12 features has improved the accuracy of classification. The main classification method used in this experiment was **multilayer perceptron with back propagation**. **Naïve Bayes** and **Sequential Minimal Optimization** are other methods that were used for comparison with Multilayer Perceptron. A small application was implemented with Java to extract features from articles. The application uses the WEKA data mining library for classification. More than 90% classification accuracy has been obtained.

1-Introduction

Authorship analysis or **Authorship attribution** can be defined as identifying the author or determining some characteristics of the author of a given text by using natural language processing methods.

Matching a sequence of text and author according to attributes extracted from texts and known attributes of authors is called **author prediction**.

Author prediction is a sub-category of **text categorization** which is also a sub-category of **document classification**.

For English language there have been a lot of studies and accomplishments for last 35 years [1]. For Turkish similar studies are very rare. Amasyalı and Diri tried to classify documents in terms of author, gender and genre by using n-grams [1]. Taş and Görür used 35 style markers to classify newspaper articles among 30 authors [3]. Küçükyılmaz and Cambazoğlu used data mining methods on web chat texts to predict gender of users [2].

2-Problem

For this experiment, 10 different Turkish columnists which usually write under same genre (politics), has been chosen from several newspapers. 100 articles of each author are collected from the internet sites of the newspapers they write for. These articles have been used as learning data. After this data analyzed and properties of each author learnt, it was expected to be able to identify the author of a new article among 10 authors.

Preferring such a corpus has some advantages such as ease of collecting data. Since each text is written in a newspaper article and they are mostly about politics genre, this doesn't affect the results of experiment. Another advantage is avoiding time dependency. All articles are written within 12-14 months and the articles to be classified are also written recently. So time also doesn't -or almost not- affect the results of our experiment.

With these advantages a more accurate measurement for effectiveness of our method was implemented.

Selected Authors for this experiment and their newspapers can be seen from Table 1

Table 1

Author	**Newspaper**
Ahmet Turan Alkan	Zaman Gazetesi
Nedim Hazar	Zaman Gazetesi
Engin Ardıç	Sabah Gazetesi
Haşmet Babaoğlu	Sabah Gazetesi
Hıncal Uluç	Sabah Gazetesi
Perihan Mağden	Radikal Gazetesi
Yasemin Çongar	Taraf Gazetesi
Gülay Göktürk	Bugün Gazetesi
Ahmet Altan	Star Gazetesi
Dücane Cündioğlu	Yenişafak Gazetesi

3-Method

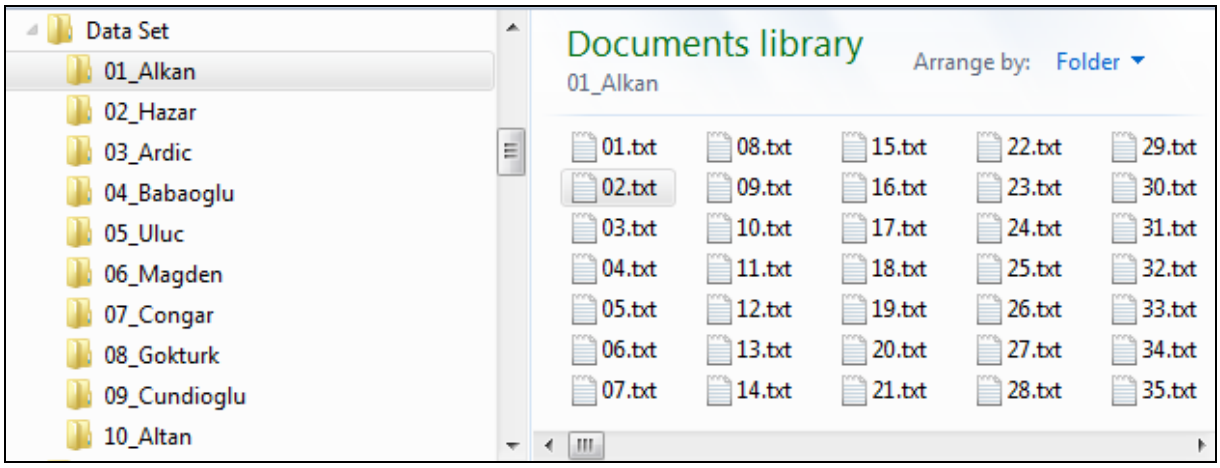
3.1. Preprocessing

Specific features from articles were extracted by the application implemented for this project. These features can be seen from Table 2.

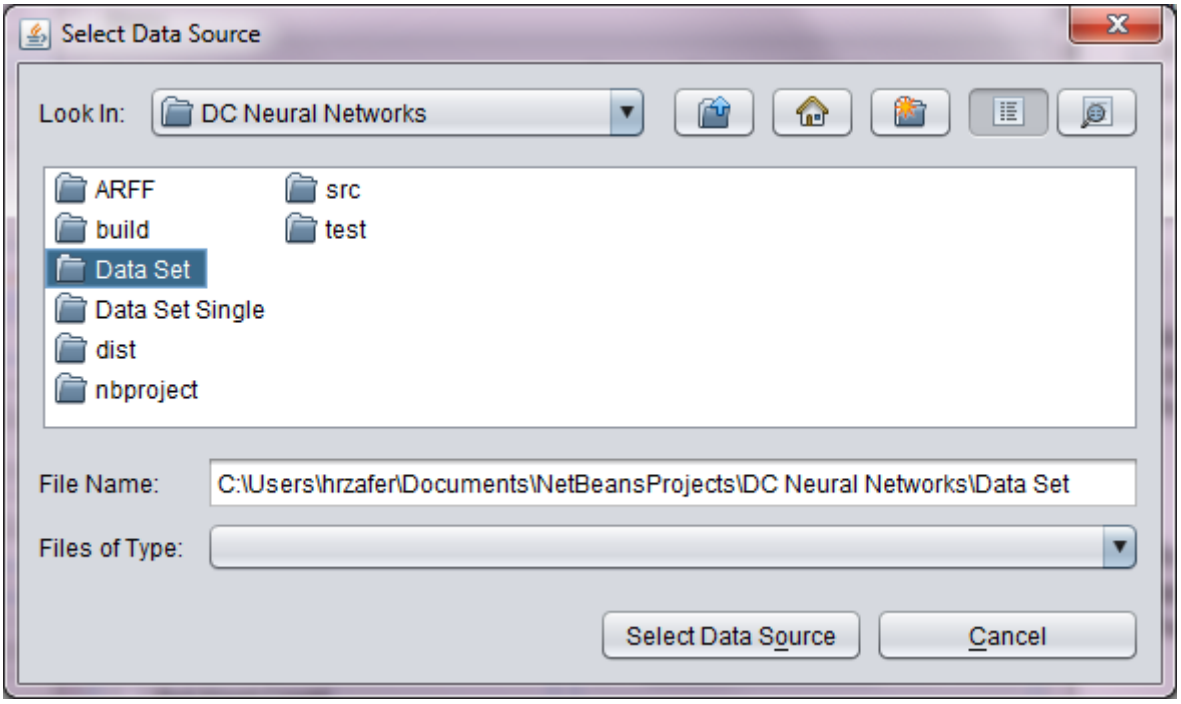
Table 2

Feature	**Description**
Word Length Average	Average word length of the text
Word Length Variance	Variance of word length in the text
Sentence Length	Average word count of sentences
Paragraph Count	Number of paragraphs in the text
Number Count	Number of numeric tokens in the text
Why word count	Number of the words “neden”, “niçin”, “niye”
First singular person count	Number of words derived from the pronoun “ben” such as “benim”, “bence”, “bana” etc.
Comma count	Number of commas in the text
Semicolon Count	Number of semicolon
Question mark count	Number of question marks
Exclamation mark count	Number of exclamation marks
Ellipsis count	Number of ellipsis
Colon count	Number of colons
But word count	Number of words “ama”, “ancak”, “fakat”

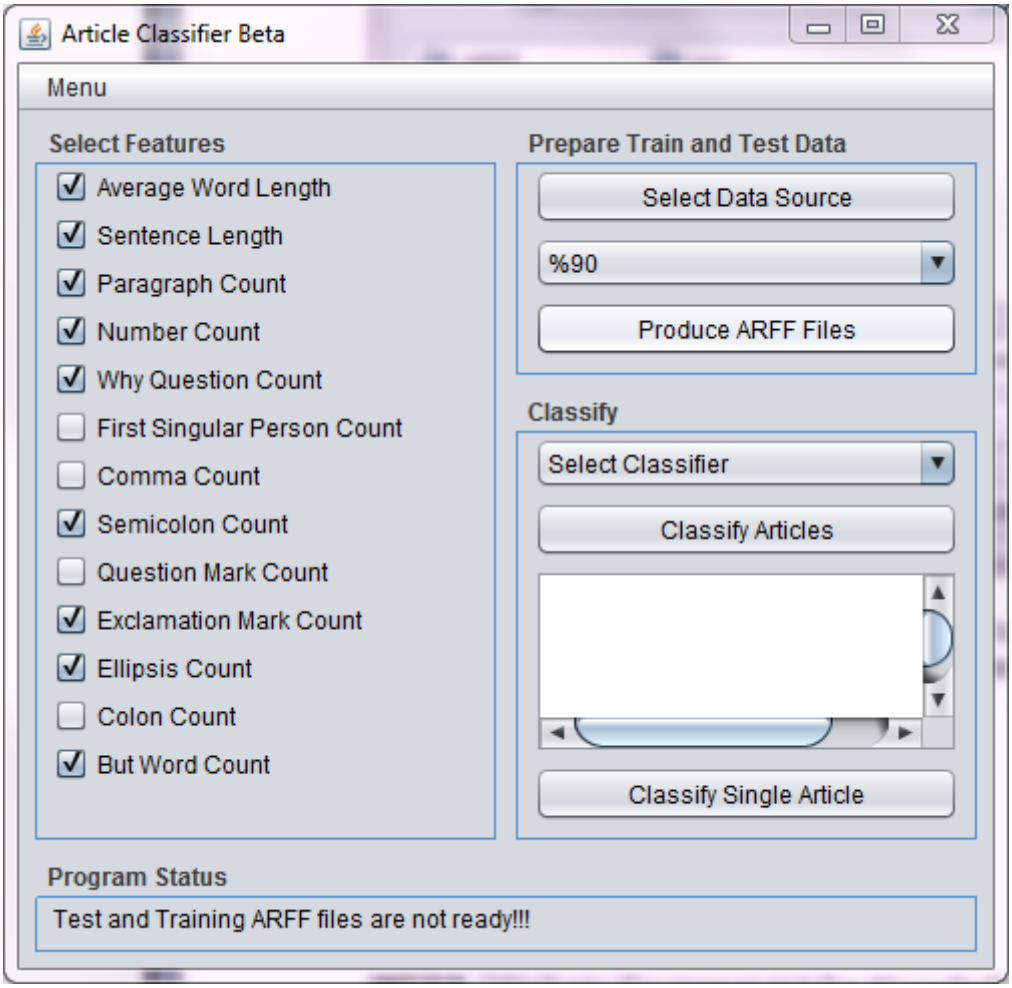
All articles are saved as txt files in folders that belong to authors separately. Here is the folder structure:



User chooses this directory through the graphical file selector of the application. This folder normally is kept under the home folder of application but any other directory with valid structure and content can be chosen. This gives the ability of using the application with different authors with collected articles.



After this, user selects preferred features to be extracted from each article. Then the ratio of training and test data is selected. Finally by clicking the "Produce ARFF files" ARFF files are produced.



For example if the ratio is %90, 90 of articles for each author is used as training data and 10 articles is used as test data.

3.2. Classification Algorithms

Multilayer Perceptron with Back Propagation: Multilayer perceptron is an artificial neural network. It is more powerful than perceptron since it can classify data that is not linearly separable. Multilayer perceptron is a feedforward neural network. This means information always moves one direction and never goes backwards. Back propagation is a supervised learning method that teaches artificial neural networks how to perform a given task. It was first presented by Arthur E. Bryson and Yu-Chi Ho in 1969 [10][11]. But its real importance wasn't understood until the work of f David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams in 1986. It is one of the most important methods of Machine Learning.

Naïve Bayes: *A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".* [7]

Sequential Minimal Optimization: *In mathematical optimization and machine learning, Sequential Minimal Optimization (SMO) is an algorithm for solving large quadratic programming (QP) optimization problems, widely used for the training of support vector machines. First developed by John C. Platt in 1999, SMO breaks up large QP problems into a series of smallest possible QP problems, which are then solved analytically*. [9]

3.3. WEKA

WEKA *(Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the* ***GNU General Public License****. [6]*

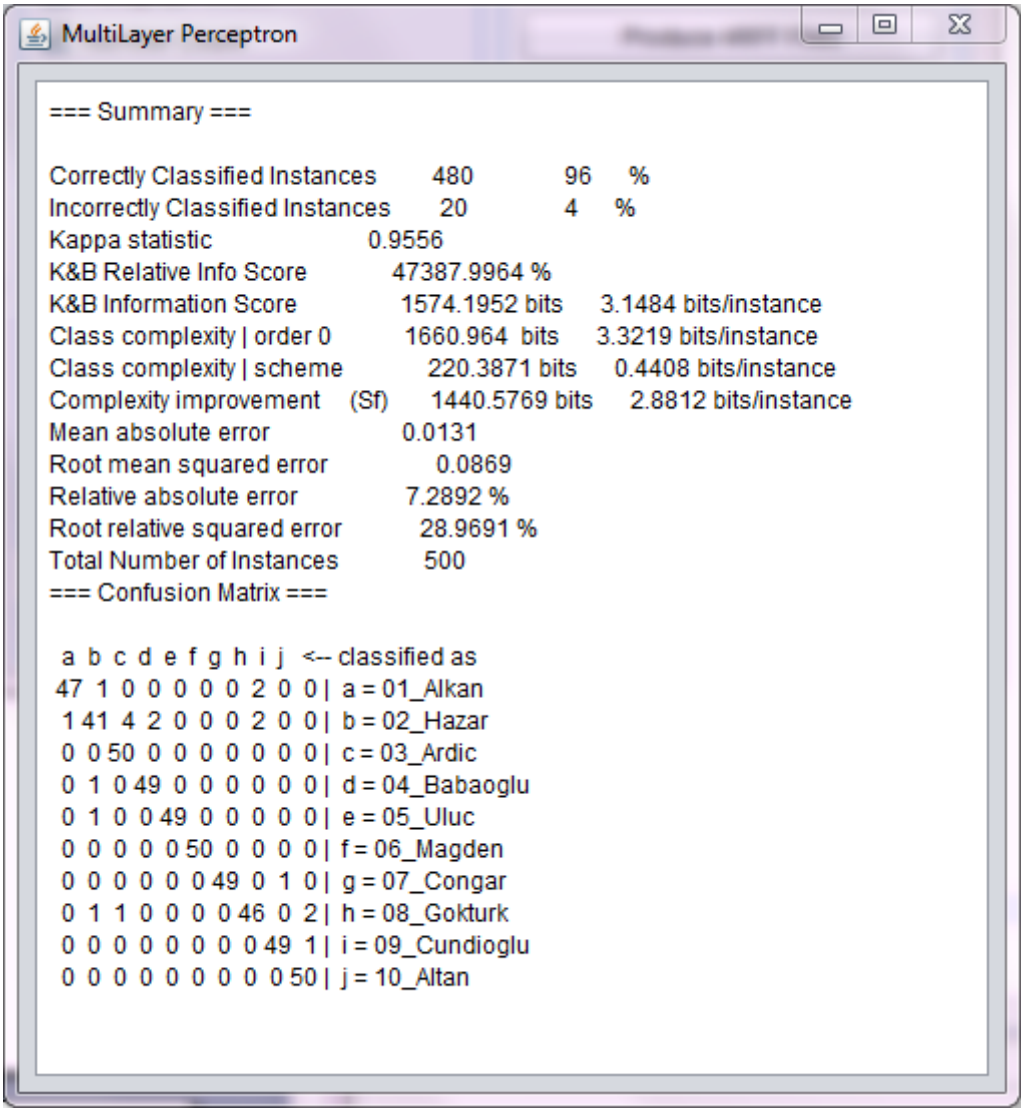
WEKA data mining tool is actually a software library which is a collection of Java classes. This library can both used through a Graphical user interface and form the command prompt. Since it is written in Java, it is also possible to use WEKA classes in particular Java projects.[8]

Here is a code section from this project that uses WEKA classes.

```
public static String evaluate(String classifier, Instances train, Instances
test){
        Evaluation eval=null;
        if (classifier.equals("NaiveBayes")){
           eval = WekaUtil.getEvaluation(train, test, new NaiveBayes());
        }
        else if(classifier.equals("SMO")){
            eval = WekaUtil.getEvaluation(train, test, new SMO());
        }
        else if(classifier.equals("MultiLayer Perceptron")){
            eval = WekaUtil.getEvaluation(train, test, new
MultilayerPerceptron());
        }
        else{
            throw new IllegalArgumentException("Invalid classifier name");
        }
        return WekaUtil.printConfusionMatrix(eval);
    }
```

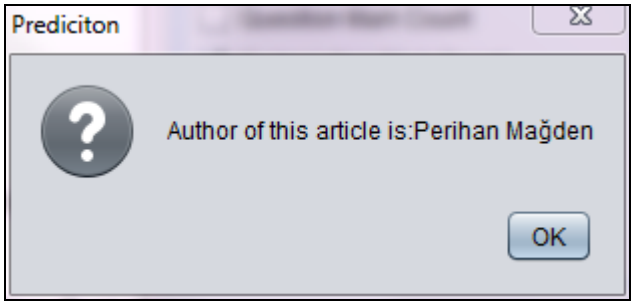
3.3. Using the Application

After the ARFF files production, two files, train.arff and test.arff are produced under a specific folder. After that step user can select a classifier and by clicking the “Classify Articles” button, multiple articles are classified. Here is a screen view of classification that uses % 50-% 50 train-test ratios.



As seen from the window multilayer perceptron with back propagation has %96 percent of success.

Another utility of application is that it allows user to classify single articles that is out of the Dataset. By copying a new article that belongs any of 10 author and pasting the text into the text are on the application window it is possible to classify this article. When user clicks “Classify Single Article” button application predicts the author of given text and prompts user with a dialog box as below.



4-Experiment Results

Classifications results for each classification method with different train-test ratios can be seen from the Table 3.

Table 3

Training/Test	Multilayer Perceptron	Naive Bayes	SMO
90	93%	87%	91%
80	95%	89%	93%
70	95%	90%	92%
60	95,50%	90%	91,75%
50	96%	90,20%	92%

The anomaly of the table above is that when we reduce training data and increase the amount of test data, slightly better results were obtained. From these results we can say that at most 50 articles for an author are enough to train a learning algorithm.

A sample confusion matrix for Multilayer perceptron with %90-%10 train-test ratio is below:

```
  a  b  c  d  e  f  g  h  i  j   <-- classified as
 10  0  0  0  0  0  0  0  0  0 |  a = 01_Alkan
  0  7  3  0  0  0  0  0  0  0 |  b = 02_Hazar
  0  0 10  0  0  0  0  0  0  0 |  c = 03_Ardic
  0  0  0 10  0  0  0  0  0  0 |  d = 04_Babaoglu
  0  1  0  0  9  0  0  0  0  0 |  e = 05_Uluc
  0  0  0  0  0 10  0  0  0  0 |  f = 06_Magden
  0  0  0  0  0  0  9  0  1  0 |  g = 07_Congar
  0  0  1  0  0  0  0  8  0  1 |  h = 08_Gokturk
  0  0  0  0  0  0  0  0 10  0 |  i = 09_Cundioglu
  0  0  0  0  0  0  0  0  0 10 |  j = 10_Altan
```

All classification algorithms are less successful with especially two authors: Nedim Hazar and Gülay Göktürk. It can be said that these features are not appropriate ones for these authors.

Table 4 shows a comparison for some features and feature groups. These results were obtained by using 80%-20% train-test ratios.

Table 4

Data Set	Multilayer Perceptron	Naïve Bayes	SMO
Word length average + variance (group-1)	30%	33%	30%
Group-1 + Sentence Length (group-2)	51%	45%	45%
Group-1 + Paragraph count (group-3)	50%	47%	41%
Punctuation Only	82%	79%	71%

Stop Words Only	23%	21%	21%
Stop Words + Numbers (group-4)	28%	25%	24%
All features except group-4	94.5%	91.5%	90.5%
All features	95%	89%	93%

From the table we can say that counting stop words such as "neden", "niçin", "ama" or words related with the first singular pronoun "ben" doesn't give good results. Also counting numeric tokens in the text such as "2" or "1990" is not a good method. By removing these features obtained results are not worse, even better with some methods.

Multilayer perceptron with back propagation is slightly better than other two classifiers. However training period is remarkably longer for this method. But once it is trained the model can be saved and classification can be done as quickly as other methods.

SMO and **Naïve Bayes** usually gave similar and satisfying results with all features.

5-Conclusion and Future Work

According to results of this experiment it is seen that even just punctuation marks give 80% accuracy and are good features. Other stylistic features such as word and sentence length and paragraph count also affect results remarkably. On the other hand chosen stop words didn't give expected results. They sometimes even reduced the accuracy. This doesn't mean stop words are bad features; maybe better stop words combinations could be selected.

In addition it was observed that there is a threshold for training data amount that more data than this amount is not necessary to obtain satisfying results. For this experiment it can be said that 50 articles are enough to resolve an author's writing style.

Lastly, it is seen that every feature doesn't work well for every author. The feature set used in this experiment worked well for most of the authors. However for Nedim Hazar the accuracy was remarkably low. So for this author there are some other features to be found that differentiates him. From this point, it can be said that for every author there are some special features that defines this author best.

6-References and Resources

[1] M. Amasyalı and B. Diri, "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender," *Natural Language Processing and Information Systems*.

[2] T. Kucukyilmaz, B. Cambazoglu, C. Aykanat, and F. Can, "Chat Mining for Gender Prediction," *Advances in Information Systems*, p. 274–283.

[3] T. Tufan and A. GÖRÜR, "Author Identification for Turkish Texts," *jas.cankaya.edu.tr*.

[4] "Document classification.", http://en.wikipedia.org/wiki/Document_classification, last visited 10.06.2010

[5] OKTAY M., KURT A., KARA M., An Extendible Frequency Analysis Tool for Turkish, "Türkçe İçin Bir Sıklık Analizi Programı", 38th International Congress of Asian and North African Studies (ICANAS 38), ANKARA/TÜRKİYE, Sep. 2007.

[6] "WEKA", http://en.wikipedia.org/wiki/Weka_%28machine_learning%29, last visited 10.06.2010

[7] "Naive Bayes classifier", http://en.wikipedia.org/wiki/Naive_Bayes_classifier, last visited 10.06.2010

[8] "Class NaiveBayesMultinomial", http://weka.sourceforge.net/doc/weka/classifiers/bayes/NaiveBayesMultinomial.html, last visited 10.06.2010

[9] "Sequential Minimal Optimization", http://en.wikipedia.org/wiki/Sequential_Minimal_Optimization, last visited 10.06.2010

[10] Stuart Russell and Peter Norvig. *Artificial Intelligence A Modern Approach*. p. 578. "The most popular method for learning in multilayer networks is called Back-propagation. It was first invented in 1969 by Bryson and Ho, but was more or less ignored until the mid-1980s."

[11] Arthur Earl Bryson, Yu-Chi Ho (1969). *Applied optimal control: optimization, estimation, and control*. Blaisdell Publishing Company or Xerox College Publishing. pp. 481.