AUTHOR GENDER IDENTIFICATION FROM TURKISH TEXT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY

BY
CEREN YAŞAR ÖNTÜRK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF
COMPUTER ENGINEERING

JULY 2019

Title of the Thesis: **Author Gender Identification From Turkish Text**
Submitted by **Ceren YAŞAR ÖNTÜRK**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

Prof.Dr. Can ÇOĞUN
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof.Dr. Erdoğan DOĞDU
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assoc.Prof.Dr. H. Hakan MARAŞ
Supervisor

Examination Date: 25/07/2019

Examining Committee Members

Committee Members

Assoc.Prof.Dr. H. Hakan MARAŞ	(Çankaya Univ.)
Prof. Dr. Hayri SEVER	(Çankaya Univ.)
Asst.Prof.Dr. Erdal ERDAL	(Kırıkkale Univ.)

STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Ceren YAŞAR ÖNTÜRK

Signature :

Date : 25/09/2019

ABSTRACT

GENDER IDENTIFICATION OF AUTHORS OF TURKISH TEXT

YAŞAR ÖNTÜRK,Ceren

M.Sc., Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Hadi Hakan MARAŞ

July 2019, 34 pages

The number of documents that are stored in a computerized environment is increasing day by day. Following the widespread use of the internet, the number of users of text-based social media applications is also expected to increase. In view of this, the content of text classification and the gender identification of authors of short texts have become an active research subject, due to the use of social media. This field has become popular since users often hide their genders in an internet environment.

A dataset is created of articles on different subjects, chosen randomly from the internet. The property of gender is used for classification in this generated dataset.

The sentence, word, character and punctuation features of these articles are utilized in a dataset created in this work. Following this, the performance of five different classification methods is compared, and the results show that the most successful method is the random forest algorithm.

Keywords: Gender Identification, Naive Bayes, Decision Tree, WEKA, Support Vector Machine, Logistic Regression, Random Forest

ÖZ

TÜRKÇE METİNLERDE YAZARIN CİNSİYET TAHMİNİ

YAŞAR ÖNTÜRK, Ceren

Yükseklisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Yöneticisi : Doç.Dr. Hadi Hakan MARAŞ

Temmuz 2019, 34 sayfa

Geçtiğimiz yıllara baktığımızda, bilgisayar ortamında depolanan belgelerin sayısı her geçen gün daha da artmaktadır. İnternetin yaygınlaşması ile birlikte metin tabanlı sosyal medya uygulamalarındaki kullanıcı sayısı da artış göstermektedir. Sosyal medyanın kullanımının aktif olması nedeniyle, kısa metinlerde yazar cinsiyetinin belirlenmesi, metin sınıflama kapsamında güncel bir araştırma konusu durumuna gelmiştir. İnternet ortamında kişiler cinsiyetlerini sakladıkların dolayı, bu çalışma alanı günümüzde popüler hale gelmiştir.

Bu çalışmada, internet üzerinden rastgele seçilmiş ve farklı konulardan oluşan makalelerden yararlanılarak veri seti oluşturulmuştur. Oluşturulan veri setinde sınıflandırma için cinsiyet özelliği kullanılmıştır.

Çalışma sırasında oluşturulan veri seti üzerinde cümle özellikleri, kelime özellikleri, karakter özellikleri ve noktalama işaretleri özelliklerinden yararlanılmıştır. Çıkan sonuçlara beş farklı sınıflandırma metodu kullanılarak, performansları birbirleriyle karşılaştırılmıştır. Çıkan sonuçlara göre en başarılı metot Rastgele Orman algoritmasıdır.

Anahtar kelimeler: Cinsiyet belirleme, Naive Bayes, Karar Ağaçları, Weka, Support Vector Machine, Rastegele Orman, Linear

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Assoc. Prof. Dr. H. Hakan MARAŞ for his supervision, special guidance, suggestions, and encouragement through the development of this thesis. Thanks are also due to Assist. Prof. Dr. Abdül Kadir Görür for his help and support.

I would like to thank Nurşah very much for her support and assistance during this thesis.

Thanks also to Mum, Dad, Yasemin and Neslihan for believing in me and for standing behind me.

I also would like to thank my dear husband Erhan ÖNTÜRK for his support and assistance during this thesis.

Finally to my miracles, Önder and Öykü: this thesis is dedicated to you.

TABLE OF CONTENTS

CHAPTERS:

APPENDICES:

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

NB	**:** Naive Bayes
SVM	**:** Support Vector Machine
RF	**:** Random Forest
KNN	**:** K Nearest Neighborhood
WEKA	: Waikato Environment of Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Knowledge is becoming more important than data, due to developing and changing environmental conditions, and a globalizing world with fewer boundaries and various marketing, research and development methods. Research and development teams have difficulties in accessing data as the internet has grown and become easier to use. The important thing in this context is the transformation of data to an information process. Data mining can be defined as a process of future prediction based on recent data that are interpreted by an expert; this information is then analyzed by numerous methods. It involves the search for connections that can provide a prediction of the future from large data silos, using a computer program. As a result, data mining extracts available information that is potentially useful but is as yet unknown and uncertain. It also includes condensing, data summary, analysis of changes, fixing deflection and certain number of technical approaches. In other words, data mining is the semi-automatic discovery of patterns, relations, changes, irregularities, rules and statistically significant structures in data.

The internet, which brings together millions of computer users and thousands of social groups, and which allows them to connect with each other directly, is a continuously growing communication network. It is defined as a social network in which users communicate with others in different cultures by defining themselves, (and the social environments in which they communicate their emotions and thoughts) virtually, with symbols symbolizing gestures and mimicking movements used in normal social life. Since the internet is considered to be a global environment, the very large number of social networks that have developed cannot be ignored.

Authorship studies have begun to gain importance following the widespread use of social networks. One of the reasons for this is that users in this environment often hide their own information and use fake accounts. Work on author identification has been done in order to provide solutions to these problems that have arisen through the evolution of internet and shared information. The first work on author identification in the Turkish language was done in 1999, and this field has been expanding recently.

1.2 OBJECTIVES

One of the problems in the classification of a document that is encountered during the author identification process is that of making predictions about an unknown or suspected author of a document. Another problem is that of determining the author's gender or the type of the document. Due to the development of social media in which users can hide their genders, research on this topic is a necessity.

It is possible that certain authors always ask questions, use punctuation to emphasize certain sentences, or repeat the same words; every author has a different personality, and the specifications of their writings may therefore be also different. According to this idea, the type of text written may differ based on the author's gender.

1.3 CORPUS

The dataset used in this study was drawn from daily newspaper articles about politics, sports, human interest and news. The dataset is based on 840 articles from different authors, which were randomly selected from newspapers while creating the dataset. An example of one these articles can be seen in Appendix B.

1.4 ORGANIZATION OF THESIS

The rest of this document is organized as follows.

Chapter 2 discusses prior research that can help to identify an author’s gender.

Chapter 3 describes how the dataset was created, through the author’s selections of sentences, words, punctuation and character-based features.

Chapter 4 explains the classifiers and how these are used together.

Chapter 5 presents the results and a discussion of the datasets that were used.

CHAPTER 2

LITERATURE SURVEY

There are many successful studies in the literature concerning text classification. Examples include the identification of an author’s gender, e-mail classification, definitions of text topics and recognition of authors. In these studies, various text representation methods have been put forward for numerous types of problem. The most widely used methods involve word/word group frequency, n-gram frequency, word cluster, hidden zone index and functional words. However, no detailed comparison of text representation methods for various types of problems has been carried out.

The first study of author recognition was done by Stamatatos et al. [1]. These authors carried out a study to determine the authors of documents by using various combinations of syntactic style features. Stamatatos and Kokkinakis also carried out studies of two datasets in Greek drawn from 10 authors, and obtained correct classification rates of 85% and 97%, respectively. Stamatatos et al. used different specifications, and their highest rate of identification was 81%, using 10 different classes.

Diri and Amasyalı carried out studies to determine the types and genders of authors of Turkish documents using the n-gram method [2]. A dataset was created from a collection of articles from different authors about sports, human interest stories, politics and news, published on the web and in specific newspapers. The dataset contained 630 documents, including 35 articles from 18 different authors, of which 28 were used for training, and seven for testing. Specification vectors containing 2-grams and 3-grams were created for this dataset. Four vectors used for classification were also derived from these specification vectors. Following this, the Naïve Bayes (NB), Support Vector Machine (SVM), C 4.5 and Random Forest (RF) classification methods were used to compare each of these specification vectors to determine which was most successful. As can be seen from Table 1, the most successful result was NB, with a score of 83.3%. A classification process that involved reducing specifications was more successful than the application of personal classification. Diri and Amasyalı obtained a success rate of 93.6% in author identification studies using six classes.

Table.1. Results of work by Diri [2]

Ng-ind		Feature count							
Dataset 1		**100**	**142**	**200**	**257**	**300**	**324**	**400**	**500**
Success rate %	**2-gram**	71.3	-	78.6	**78.8**	80	-	82.5	78.8
	3-gram	75	-	82.5	-	78.8	**80**	82.5	78.8
	4-gram	91.3	**85**	88.8	-	88.8	-	88.8	85
Dataset 2		**100**	**142**	**200**	**257**	**300**	**324**	**400**	**500**
Success rate %	**2-gram**	70	-	71.3	**70**	72.5	-	75	71.3
	3-gram	75	-	73.8	-	73.8	**70**	75	73.8
	4-gram	75	**70**	73.8	-	75	-	77.5	73.8
Dataset 3		**75**	**100**	**200**	**208**	**217**	**300**	**400**	**500**
Success rate %	**2-gram**	-	75	75	-	**81.3**	79.2	81.3	75
	3-gram	-	70.8	85.4	**83.3**	-	87.5	93.8	93.8
	4-gram	**83.3**	85.4	85.4	-	-	87.5	93.8	87.5

Murugaboopathy et al. carried out several studies on gender prediction in text [3], using data gathered from e-mail attachments. The word counts of these texts were separated into groups; for example, more than 50 words and more than 1000 words ormore than 50 words, more than 100 words and more than 200 words. They obtained a new dataset by using character-based, word-based, syntactic and structural features, and function words. They used the SVM and decision tree algorithms on this dataset, and their results showed that SVM was more useful than the decision tree approach in this study. For gender identification studies, researchers generally use an n-gram model, and use the NB, support vector machine (SVM), RF and K nearest neighbor (KNN) algorithms on datasets.

Peng et al. conducted trials on data in Greek, Chinese and English, carrying out author identification using an n-gram method by using one n-grams to 10 n-grams for the dataset [4]. They used the same dataset after applying an absolute smoothing technique, and obtained maximum results of 74% and 90% using 3-grams. Peng et al. also carried out studies in Greek, Japanese, English and Chinese in order to prove the independency of the n-gram from the language by using a character-level n-gram model. They obtained an 81% success rate for six classes using the NB classification method based on a character-level n-gram model.

Cavnar achieved 80% success using an n-gram-based study of author identification. [5]

Atoosa Mohammad Rezaei created the clustering method by using five different specifications for the gender identification problem [6]. Character-based, word-based, syntactic and structural features of clusters were used together with a function word in the text. After the clustering method was applied, two machine learning algorithms were used on the new data clusters; SVM and NB were chosen, since these two algorithms are the strongest classifiers. As shown in Table 2, the accuracy achieved was relatively high.

Table 2. Results of Rezaei's work [6]

Model	**Accuracy**

SVM-Linear	76.7%
SVM-Poly2	77.68%
SVM-Poly3	77.08%
SVM-RBF C=1, Gamma=0.01	76.21%
SVM-RBF C=3, Gamma=0.5	78%

Jonathan and Keselj classified authors using n-gram analysis to determine their genders [7]. A dataset consisting of compositions by English students was used. In these experiments, the character and word levels, and the grams of the objects in the cluster were used. The same results were obtained for all of the methods used, and a maximum success rate of 81% was achieved.

Nowson and Oberlander extracted the n-grams of a dataset to perform gender determination using the n-gram method, and conducted experiments on English data [8]. Their datasets were created using the writings of 71 different authors, of whom 47 were female and 24 were male. Each dataset consisted of eight male and 15 female writers. In the study, the SVM method, which is included in the WEKA sorting tool, was used to classify the documents. A 93% success rate was achieved for classification with SVM.

Fung used the SVM classifier to authorize Federalist publications. In the study, these publications were separated by a plane in a three-dimensional space based on the words ‘as’, ‘of’ and ‘on’. SVM was implemented using a set of functional words to separate the publications from each other.[9]

Na Cheng et al. carried out studies of gender identification based on e-mails [10]. They extracted 68 psycho-linguistic features for analysis, and used decision tree and SVM algorithms to identify the gender of the author. They achieved a success rate of 82.2%, as can be seen in Table 3.

Table 3. Comparison of results from Na Cheng et al.
based on decision tree and SVM

Classifier	Minimum words per e-mail		
	50	100	200
Decision tree	73.38	80.43	78.93
SVM	80.08	82.20	81.03

Farrell carried out studies on blogs that he found on the internet [11], and tried different algorithms on this new dataset using the WEKA program. Three different configurations were tried for these data, for each of these algorithms. The first of these configurations involved using unigrams, words/POS, including punctuation, ignoring all stop words in the n-grams, and tracking feature hits methods on blogs. The second involved the use of bigrams, words/POS, including punctuation, ignoring all stop words in the n-grams, and tracking feature hits methods. The last configuration involved using trigrams, words/POS, including punctuation, ignoring all stop words in the n-grams, and tracking feature hits. An examination of the results of this study shows that they are close to each other, as can be seen in Table 4.

Table 4. Results of work by Farrel [11]

90% Training **10% Testing**	**Recall**	**Recall** **Precision**	**F-Score**
Naive Bayes female	.82	.74	.78
Naive Bayes male	.77	.82	.79
SVM female	.76	.69	.72
SVM male	.70	.79	.74
WEKA naive Bayes female	.82	.74	.78
WEKA naive Bayes male	.77	.82	.79
WEKA SVM female	.87	.40	.55
WEKA SVM male	.61	.94	.74
80% Training **20% Testing**			
Naive Bayes female	.71	.37	.49
Naive Bayes male	.57	.84	.68
SVM female	.63	.49	.55
SVM male	.58	.72	.64
WEKA naive Bayes female	.71	.37	.49
WEKA naive Bayes male	.57	.84	.68

WEKA SVM female	.61	.48	.54
WEKA SVM male	.57	.69	.62

CHAPTER 3

METHODOLOGY

3.1 FEATURES OF THE ARTICLE

For this study, Turkish texts on various topics were drawn from daily newspapers. The resulting dataset consisted of 840 articles by different authors.

When analyzing this data, each article was considered separately. In this step, the articles were grouped according to their different characteristics. There were four distinctive features: sentence, word and character properties and punctuation marks. A diagram of this process can be seen in Figure 1.

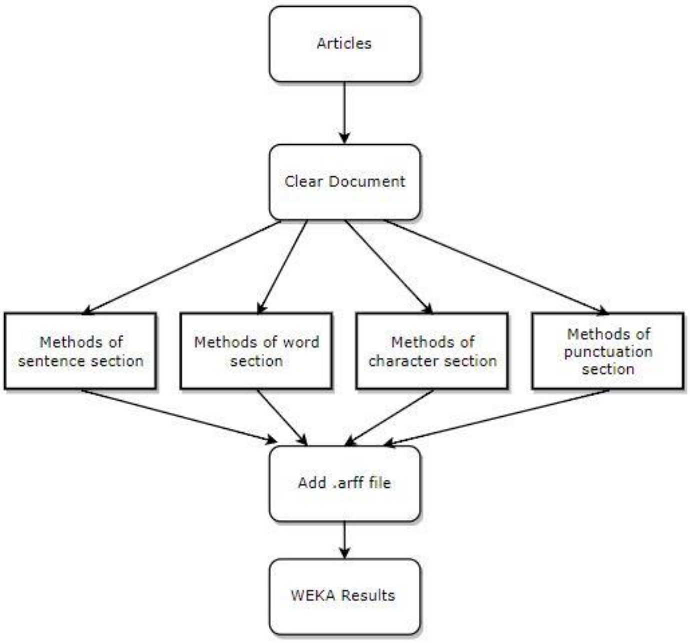

Articles
Clear Document
Methods of sentence section
Methods of word section
Methods of character section
Methods of punctuation section
Add .arff file
WEKA Results

Fig.1. Schema of the first stage

At the sentence properties stage, the properties of the sentences were examined. For example, different results were obtained by looking at the various features, as can be seen in Table 5, such as the number of words a document contained, how many words the article contained, how many paragraphs were included, and the total number of syllables. Average values of all features were calculated.

Table 5. Sentence properties

Feature	**Description**
Number of words	Total words
Average word length	Total characters / number words
Number of different words	Total different words
Average number of different words	Number of different words / Total words

Number of sentences	Total sentences
Number of syllables	Total syllable
Average number of syllables	Total syllable/ Total words
Number of short words	Total short words
Average number of short words	Total short words/ Total words
Number of long words	Total long words
Average number of long words	Total long words/ Total words
Average number of words in a sentence	Total words/ Total sentences
Number of lines	Total number of lines
Number of paragraphs	Total number of paragraph
Average number of lines in a paragraph	Total lines/ Total paragraphs
Average number of sentences in a paragraph	Total sentences/ Total paragraphs

In the word properties section, there are various methods like finding the type of word, such as a noun, adjective or adverb, or whether the word contains first, second or third person information. A list of these features is given in Table 6.

Derman and Tekeli's library is used here; these authors worked on a C# version of Zemberek, called the Zemberek Project [12]. Zemberek is an easy-to-use and open library for anyone working in the field of natural language processing in Turkish. This library is used to break up the words into their types. The operational logic is shown in Figure 2.

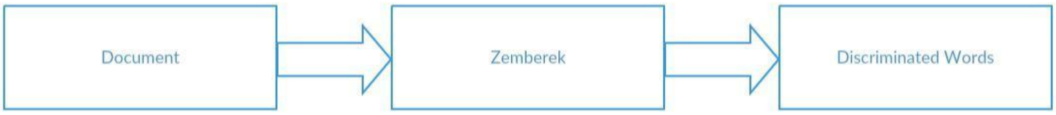


Fig. 2. Usage of Zemberek

Table 6. Word properties

Feature	**Description**
Number of Noun	Total number of noun
Number of Verb	Total number of verb
Number of Adjective	Total number of adjactive
Number of Preposition	Total number of preposition
Number of Adverb	Total number of adverb
Number of Conjuction	Total number of conjuction
Average Number of Noun	Total noun / total words
Average Number of Verb	Total verb / total words
Average Number of Adjective	Total adjective / total words
Average Number of Preposition	Total preposition / total words
Average Number of Adverb	Total adverb / total words
Average Number of Conjuction	Total conjuction / total words
butWords	AMA, FAKAT, LAKİN ,ANCAK
stopWords	A, ACABA , ALTI …
firstPerson	BEN, BENİ, BANA, BENDE, BENDEN, BENİM
secondPerson	SEN, SENİ, SANA, SENDE, SENDEN, SENİN
thirdPerson	O, ONU, ONA, ONDA, ONDAN, ONUN
women	HANIM, KIZ, KADIN, BAYAN, DİŞİ

men	ERKEK, BEY, OĞLAN, BAY, ERİL

In the character properties section, the characters in the articles are considered to be vowel, consonant, characters, and upper and lower case characters. The list of the features is seen on Table 7.

If we make generalizations within the articles, capital letters are often used by both ladies and gentlemen, not only at the beginning of the sentence but also when emphasizing in the sentence. In addition to this, personal pronouns are another commonly used item in narration.

Because of the structure of the Turkish language, consonants are used more frequently. The variability of the number of consonants affects the variability of the number of words in the same order. As a result, this situation is being used as an important discriminator.

Feature	**Description**
Number of Vowel	Total number of vowel
Number of Consonant	Total number of consonant
Number of Character	Total number of character
Number Of Letter	Total number of letter
Number Of Space	Total number of space
Number of special Character	Total number of special character
Number Of Upper Case	Total number of upper case
Number Of Lower Case	Total number of lower case
Number of Numeral	Total number of numeral
Average of Vowel	Total vowel / Total character
Average of Consonant	Total consonant / Total character
Average of Letter	Total letter / Total character

Table 7. Character properties

Average of Space	Total space / total character
Average of Special Character	Total special character / Total character
Average of Upper Case	Total upper case / Total character
Average of Lower Case	Total lower case / Total character
Average of Numeral	Total numeral / Total character
Average of Char in a Word	Total character / Total word
Average of Char in a Sentence	Total character / Total sentence

In terms of the properties of punctuation marks, one marker involves how frequently punctuation marks are used in articles. A feature list of punctuation properties is given in Table 8.

The article writers often did not comply with the formal rules of the Turkish language, and wrote colloquially, without the use of many punctuation marks. From the results, it can be seen that the period (.), comma (,), question mark (?) and exclamation point (!) are used frequently, while other punctuation marks are hardly ever used.

Table 8. Punctuation properties

Feature	**Description**
Number Of Punctuation	Total number of punctuation
Number Of Point	Total number of .
Number Of Comma	Total number of ,
Number Of Semicolon	Total number of ;
Number Of Colon	Total number of :
Number Of Question Mark	Total number of ?
Number Of Exclamation Mark	Total number of !
Number Of ThreePoints	Total number of …
Number Of DoubleQuotes	Total number of “
Number Of Sharp	Total number of #
Number Of Dolar	Total number of $
Number Of Percent	Total number of %

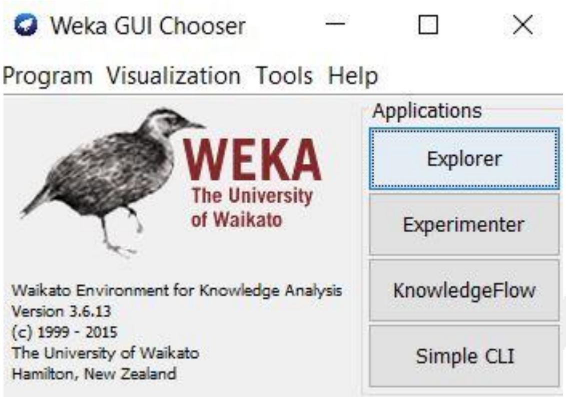
Number Of And	Total number of &
Number Of Left Bracket	Total number of (
Number Of Right Bracket	Total number of)
Number Of Star	Total number of *
Number Of Plus	Total number of +
Number Of Minus	Total number of -
Number Of Slash	Total number of /
Number Of Smaller	Total number of <
Number Of Bigger	Total number of >
Number Of Equal	Total number of =
Number Of At	Total number of @
Number Of Left Brace	Total number of {
Number Of Right Brace	Total number of }
Number Of Line	Total number of |
Number Of Tilde	Total number of ~
Averages of all punctuations	Total number of punctuation marks / Total characters

CHAPTER 4

MACHINE LEARNING

4.1 WEKA

After the data has been pre-processed, it can then be classified using WEKA; this is a package used for machine learning, which is an important topic in computer science. It was developed as an open source Java code at the University of Waikato, and is distributed under the GPL license. It is named after the university at which it was developed, and stands for Waikato Environment for Knowledge Analysis [13]. A screenshot from this application is shown in Figure 3.



Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand

Fig.3. Screenshot from WEKA

WEKA reads data from a simple file and assumes that the stochastic variables in the data are numerical or nominal values. Data can also be retrieved from a database, but in our case the data are in the form of a file.

Many libraries for machine learning and statistics are available in WEKA, for example for data preprocessing, priming, sorting, grouping, feature selection and property extraction. There are also tools that allow visualization of the results of these operations.

In this study, WEKA version 3.6.13 was used. This program supports many algorithms for classification, clustering and association rules. WEKA supports text-based file types including arff, arff.gz, names, data, csv, c45, libsvm, dat, bsi, xrff, xrff.gz; it also supports the use of databases and URLs at which data are stored.

4.2 Decision Tree

A decision tree is a type of supervised machine learning in which data are constantly divided based on a given parameter. The tree is composed of two entities: decision nodes, which indicate where the data are divided, and leaves, which are the final results [14]

There are several advantages to using a decision tree; for example, it is easy to understand and interpret, meaning that the user can understand the consequences of decision tree learning. In most alternative techniques, the data becomes available after a little processing. The preprocessing stage of the decision tree is shorter and simpler than in other alternative methods, and this method can be used to process both numerical and class data. Most machine learning algorithms are useful for either numerical applications or classification problems, while decision tree learning can be used in both. It also has a low calculation complexity. Due to its simplicity and speed, a large amount of data can be processed in a short time, and this approach becomes preferable over alternative methods as the amount of data increases.

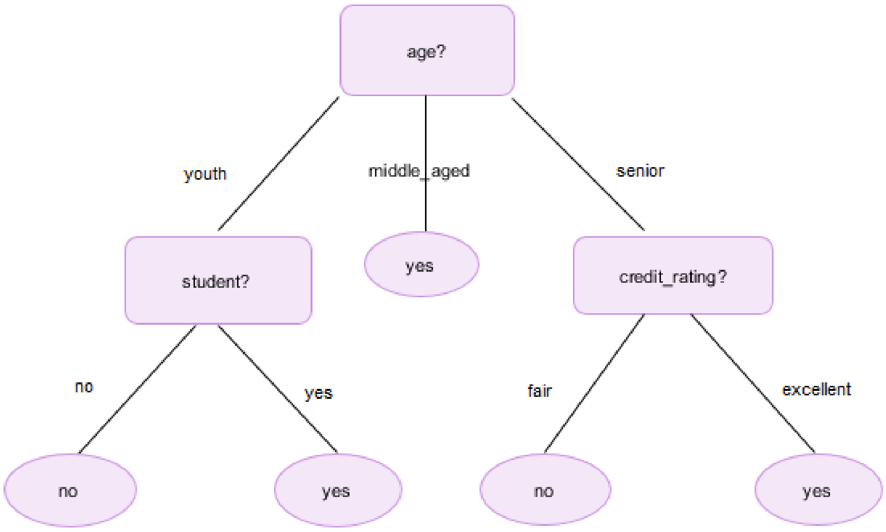


Fig.4. Example of a decision tree

There are several types of algorithms that can be used in decision tree learning. Two of these are analyzed here: the ID3 and RF methods.

4.2.1 ID3 Algorithm

The ID3 algorithm was developed by J. R. Quinlan to produce a decision tree from a dataset. This algorithm uses the top-down and greedy search techniques, and is based on the concepts of entropy and information gain.

Entropy reflects the uncertainty in an outcome and the possibility of an unexpected occurrence. If all the samples are regular, entropy becomes zero; or if the samples are equal, entropy becomes one. Entropy is defined as in Equation (1).

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1}$$

Entropy is calculated not only for the target, but also for the properties. However, the entropy is calculated for the properties. The target is also taken into consideration. Another expression for entropy is given in Equation (2).

$$E(T, X) = \sum_{c \in X} P(c) E(c) \tag{2}$$

Information gain is based on the removal of all entropy after a dataset is traced on a feature. In the case of low entropy, the importance of a property is increased in the ID3 algorithm, while the reverse is true for entropy approaching one. In information gain, however, the opposite holds, and entropy can be thought of as the opposite of information gain. When building the decision tree, the property with the highest information gain is selected. The information gain can be expressed as in Equation (3).

$$Gain(T,X) = Entropy(T) - Entropy(T,X) \tag{3}$$

4.2.2 Random Forest

The RF algorithm is a community learning method, and is an algorithm that aims to increase the classification value by generating more than one decision tree during the classification process. Individually constructed decision trees are merged together to form larger decision trees. These decision trees are subdivisions randomly selected from the dataset to which they are connected.

RF offers excellent validity, and gives more accurate results than many other methods. It also works very quickly, and uses datasets that have a distributed or unbalanced distribution, with thousands of categorical variables and a large number of class labels.

Of the existing machine learning methods, RF provides unique predictive validity and model interpretability. It also provides better generalizations and makes available estimates because random sampling and community methods give improved properties for these techniques.

The RF algorithm takes two parameters that are set by the user:

m: The number of variables used in each node to determine the best partition.

n: the number of trees to improve

Initially, boot samples are created from two thirds of the training dataset. One third of the training dataset, referred to as out-of-bag (OOB) data, is used to test for errors. At each node, m variables are selected randomly from the total set, and the best branch is determined among these variables. The number of m variables taken equal to the square of the number of m variables generally gives the result that is closest to the optimum. The GINI index is calculated in order to measure the homogeneity of the classes; a lower GINI index means that the class is more homogeneous. A branch is successful when the GINI index of a lower node is smaller than the index of the upper node.

The GINI index uses the following notation:

T: Dataset

n: Selected data

p_j: The ratio of the selected data.

The GINI index is expressed by Equation (4).

$$Gini(T) = 1 - \sum_{j=1}^{n} (p_j)^2 \tag{4}$$

After the Gini index has been determined, the classes of the test datasets are determined based on its value. The most appropriate classification is made on the basis of the overall results.

4.3 Naive Bayes

This is a classification algorithm that acts as an estimator and a descriptor in analyzing the relationship between variables.

NB is based on the principle of learning data; in other words, the data used in the training stage calculates how many times each output has arrived in order to learn the model. This value is called the priority probability. In these calculations, the combination of each independent variable with its dependent variable determines the frequency at which the event occurs. This frequency is used to estimate from the dataset [15].

NB is a well-known and frequently used algorithm for text categorization. A training dataset is determined for the target function, and new samples are presented that are defined by their values, where the learner estimates the target value or class.

Classification is an important data mining problem. The input is the training dataset. Each datum in this dataset has many attributes. Attributes with numerical fields are called numerical attributes, attributes with non-numerical fields are called categorical attributes. A further important attribute is known as a class tag. Classification aims to create a short model within the untagged records that can be used to predict the class label. Many classification models such as NB, KNN , Decision Trees and Artificial Neural Networks have been developed with this aim.

The NB classifier is a simple and rapid technique for classifying categorical data. Bayesian classification is a classical variable-dependent method that uses a certain probability distribution for the training data. NB analyzes the relationship between each independent feature in the training set and the conditional probability within each of these relationships. Forecasting is carried out to classify a new situation by combining the effects of the independent variables on the dependent variables.

The classification steps used in the NB technique are as follows: during the training process, the previous probability of each outcome is determined as the number of times the output has arrived in the training set within the relevant category. For example, if there are five conditions and the first result passes twice, the probability of the result is 0.4. In addition to the previous possibilities, this approach also calculates how many times each independent attribute passes within each dependent attribute. The value of this frequency is used to calculate the conditional probability values generated by multiplying each of the calculated probability values by the product.

The NB algorithm is one of the most well-known classification algorithms, and numerous researchers have studied the theoretical and experimental results of this approach. It is used extensively in data mining applications, and has yielded surprisingly good results in many applications. Nevertheless, the assumptions made in this technique are insufficient, as all features are treated as equal in the learning stage, which may not be realistic; for example, when estimating whether or not a given person has diabetes, blood pressure is more important than the person’s height. For this reason, the performance of the NB algorithm can be improved with mitigating assumptions.

The Bayesian classifier is expressed as follows.

X is an example of a dataset for which the class is unknown, consisting of X = {X1, X2, X3, ….,XN }.

The class values are assumed to be C1, C2, C3, ..., Cn. The probability of the test data to be assigned to the class is calculated as shown in Equation (5):

$$P\left(\frac{C_j}{X}\right) = \frac{P(\frac{X}{C_j}).P(C_j)}{P(X)} \tag{5}$$

The class with the greatest value is the test result. The greatest value can be expressed as in Equation (6).

$$\arg max_{ci}\{P(X \mid C_i)P(C_i)\} \tag{6}$$

4.4 Logistic Regression

In linear regression analysis, the independent and dependent variables are specified numerically (in the form of continuous or intermittent numerical values, as measured). For example, if a relationship between age and blood pressure is sought, both age and blood pressure need to be specified numerically, and cannot be specified as qualifications. If the dependent variable is specified as the attribute, the relationship between the argument and the variable can be found by logistic regression [16].

Table 9 compares the linear and logistic regression methods.

Table 9. Linear and logistic regression methods

Variables	**Linear regression analysis**	**Logistic regression analysis**
Dependent	Continuous digital Circular digital	Quality
Independent	Continuous digital Circular digital	Continuous digital Circular digital Quality (each argument may have another form of measurement)

Attribute dependent variable:

Binomial: For example: alive/dead, effective/ineffective
Multinomial: For example: working—not working—retired
Ordinal: For example: very effective—moderately effective—ineffective
Logistic regression analysis can be applied in all of these cases.

The method to be applied depends on the category number of the dependent variable. The most common case is that the dependent variable has two categories, as shown in Table 10.

Table 10. Methods for dependent and independent variables

Number Dependent Variable Categories	Number Independent Variables	Number Independent Variable Categories	Method to be Applied

2	1	2	Binominal Logistic Regression
2	1	2+	Binominal Logistic Regression
2	2+	Various	Many Variable Logistic Regression
2+ unordered	Single / Multi	Various	Multinominal Logistic Regression
2+ ordered	Single / Multi	Various	Ordinal Logistic Regression

The goal of the logistic regression method is to find the simplest model that can predict the result of the dependent variable. The suitability of the model obtained as a result of the logistic regression analysis is confirmed using the model chi-square test, and whether or not each independent variable is meaningful in the model is tested using the Wald statistic.

The logistic regression function is given in Equation (7):

$$\mathit{\Pi}(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \tag{7}$$

Another form of this expression is given in Equation (8):

$$\mathit{\Pi}(x) = [1 + \exp(-\beta_0 - \beta_1 X)]^{-1} \tag{8}$$

In the logistic regression method, none of the assumptions of linear regression analysis are used. This provides significant flexibility, and logistic regression has therefore become the preferred method.

However, several points need to be taken into account when using logistic regression analysis in research. Firstly, all of the relevant arguments must be included in the model. If some variables are not included in the model, this may cause the error term to increase, and the model may be insufficient.

Secondly, incompatible arguments should be excluded. If causally unfit variables are included in the model, this makes the model more complicated and possibly too difficult to analyze. Observations for each individual should be made only once, and repetition of measurements should not occur.

The measurement error in the independent variables must also be small, and there should be no lost data. Errors cause bias and the inadequacy of the model in estimating the coefficients. There should not be multiple connections (multicollinearity) between arguments, and the independent variables should not be related to each other. In addition, extreme values should not appear in the data, since in the same way as in linear regression, these values can significantly affect the result.

If there is a large difference between the expected and observed variances of the dependent, then the model is inadequate and needs to be redefined. Possible reasons for this may be that the sample was not randomly selected or that there is a serious problem with the research scheme.

4.5 Support Vector Machine (SVM)

This algorithm was developed by Vapnik and subsequently standardized by Cortes and Vapnik [17]. It is a learning method used in data analysis, pattern recognition, classification and regression analysis. It generates a mapping function between the input and output using training data [17].

SVM creates one or more hyperplanes in a multidimensional space. One property of hyperplanes is that they have a maximum margin.

This approach can be used to solve classification problems, and involves a relatively flexible representation of class boundaries and automatic complexity control. SVM is a machine learning algorithm with a single global minimum that can be found in polynomial time. It is easy to use, and its generalization performance is good. Many problems can be solved by applying this algorithm with very small changes.

SVM separates two groups by drawing a border between them in a plane to create the classification. The location of this border should be the farthest possible point from the members of the two groups, and SVM is used to determine where it is drawn.

The data can be divided into two types: linear and nonlinear. The linear data can be linearly separated, as can seen in Figure 5. This involves a process of choosing the line with the highest margins from the infinite number of lines that can separate the data.

Nonlinear data cannot be linearly separated, as shown in Figure 6. However, the original study data can be transformed into a higher dimension using nonlinear maps.

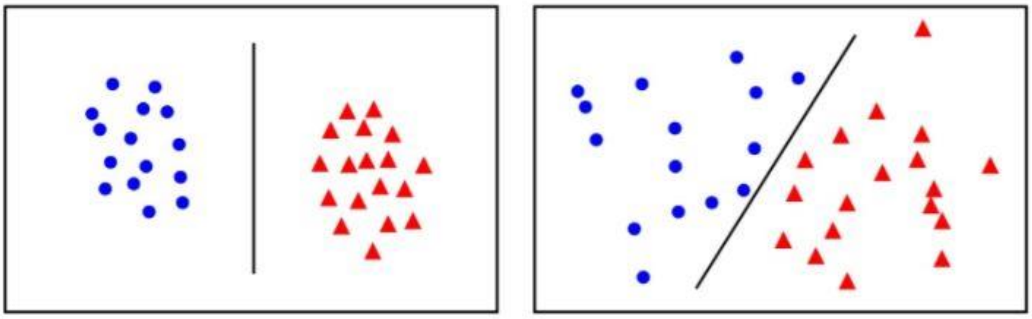


Fig.5. Linearly separated data

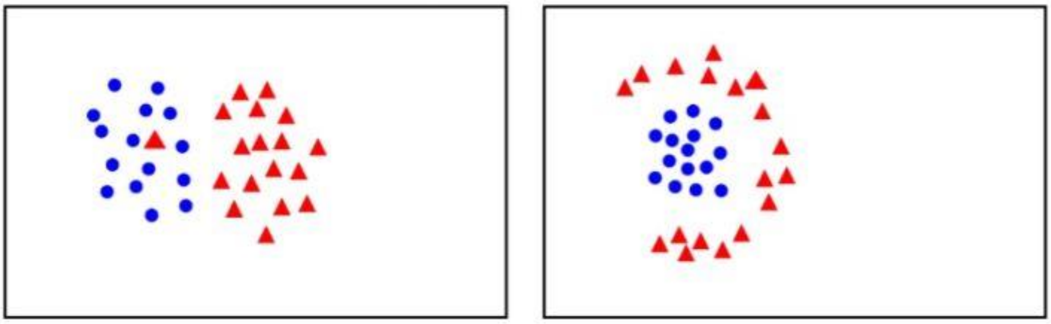


Fig.6. Nonlinearly separated data

It is possible to offset the distance between the two classes above. Each point in this plane can be defined by Equation (9):

$$D = \{(x_i, c_i) \mid x_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^{n} \qquad (9)$$

The above notation can be read as follows. For every x and c, the vector x is a point in the space and c is the value indicating that this point is −1 or +1. These points range from i = 1 to n. In other words, this representation refers to the points of the previous equation.

If this representation is on a hyperplane, every point in the graph is represented as follows:

wx – b = 0

where w is the normal vector perpendicular to the hyperplane, x is independent parameter of the point, and b is the shear rate. This equation makes it possible to simulate ax + b to the right equation.

Again, according to the above equation, the value b / || w || gives the difference between the two groups in terms of distance, and this difference is called the offset. In order to obtain the highest value of this distance based on the difference equation, the quantity 2 / || w || in the equation above gives three values: 0, −1 and +1, as shown in the above equation. That is, the distance between the lines is two units.

The two equations are obtained as follow:

wx – b = -1

wx + b = 1

These equations’ results are from finding the highest values obtained the truths. It is also assumed that the problem is linearly separable based on these equations.

The hyperplane between the two groups cannot be unidirectional, as shown by the following example. Although two possible hyperplane possibilities are shown in Figure 7, the SVM method selects the one with the greatest tolerance (offset).

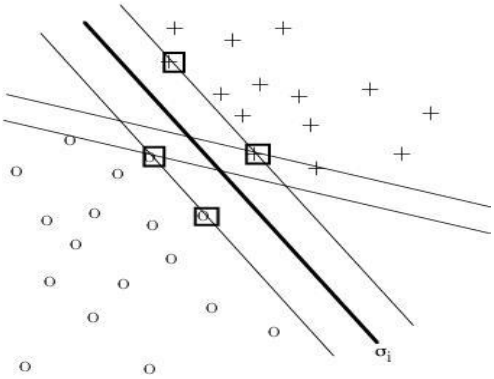


Fig.7. Hyperplane possibilities

CHAPTER 5

EXPERIMENTS AND RESULTS

A total of 840 articles by different authors were used as a basis for the experimental results. A example of one of these articles is given in Appendix B. Using these articles, 125 different attribute end-points were taken for each material. The results are in arff format, as used by WEKA. The 125 properties include sentence, word, character and punctuation features.

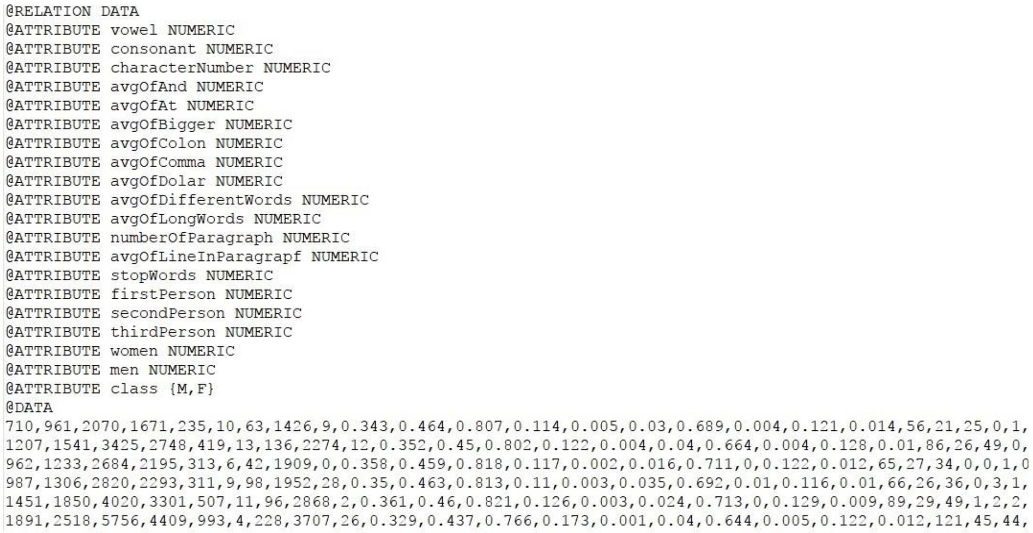
The Zemberek library was used to separate words into their types. This library is very useful for finding roots, annexes or types of words.

Zemberek examines a given word morphologically using a very simple system. First, it identifies possible candidates for the root of the given word. It then tries to attach this root in the appropriate order. If, during this procedure, the root is able to obtain the same result, then the appropriate roots and suffixes are also found.

Another specific feature is a class of special words in Turkish called “stop words”, which are listed in Appendix A.

The result of these methods is printed to a text file, and each result type is then converted to .arff format, as shown in Figure 8. All results are numerical values.

Following the format conversion, the results are used determine the gender of the author using the classification packs contained in WEKA (NB, decision tree, RF, logistic, SVM). Based on this analysis, it is then determined whether the result is positive or which algorithm is most suitable for this project.



```
@RELATION DATA
@ATTRIBUTE vowel NUMERIC
@ATTRIBUTE consonant NUMERIC
@ATTRIBUTE characterNumber NUMERIC
@ATTRIBUTE avgOfAnd NUMERIC
@ATTRIBUTE avgOfAt NUMERIC
@ATTRIBUTE avgOfBigger NUMERIC
@ATTRIBUTE avgOfColon NUMERIC
@ATTRIBUTE avgOfComma NUMERIC
@ATTRIBUTE avgOfDolar NUMERIC
@ATTRIBUTE avgOfDifferentWords NUMERIC
@ATTRIBUTE avgOfLongWords NUMERIC
@ATTRIBUTE numberOfParagraph NUMERIC
@ATTRIBUTE avgOfLineInParagrapf NUMERIC
@ATTRIBUTE stopWords NUMERIC
@ATTRIBUTE firstPerson NUMERIC
@ATTRIBUTE secondPerson NUMERIC
@ATTRIBUTE thirdPerson NUMERIC
@ATTRIBUTE women NUMERIC
@ATTRIBUTE men NUMERIC
@ATTRIBUTE class {M,F}
@DATA
710,961,2070,1671,235,10,63,1426,9,0.343,0.464,0.807,0.114,0.005,0.03,0.689,0.004,0.121,0.014,56,21,25,0,1,
1207,1541,3425,2748,419,13,136,2274,12,0.352,0.45,0.802,0.122,0.004,0.04,0.664,0.004,0.128,0.01,86,26,49,0,
962,1233,2684,2195,313,6,42,1909,0,0.358,0.459,0.818,0.117,0.002,0.016,0.711,0,0.122,0.012,65,27,34,0,0,1,0
987,1306,2820,2293,311,9,98,1952,28,0.35,0.463,0.813,0.11,0.003,0.035,0.692,0.01,0.116,0.01,66,26,36,0,3,1,
1451,1850,4020,3301,507,11,96,2868,2,0.361,0.46,0.821,0.126,0.003,0.024,0.713,0,0.129,0.009,89,29,49,1,2,2,
1891,2518,5756,4409,993,4,228,3707,26,0.329,0.437,0.766,0.173,0.001,0.04,0.644,0.005,0.122,0.012,121,45,44,
```

Fig.8. . Example of.arff format

In Weka, the 10-fold feature is used for cross-validation as a default. In addition, 30-, 50-, 70- and 90-fold samples are prepared for each classification. Table 11 shows the results for each classifier and cross-validation.

Table 11. Success of classifications applied without feature reduction

Classifier	10 folds	30 folds	50 folds	70 folds	90 folds
ID3	65.59	66.78	66.78	68.21	67.61
NB	60.71	61.66	61.19	61.42	61.30
Logistic	76.33	75.71	75.95	75.59	75
Random forest	78.45	80	79.16	**80.47**	80.47
SVM	73.33	72.85	73.69	73.33	72.97

The n-fold content that appears in this table refers to dividing the dataset into several portions for testing. For example, when the 10-fold feature is selected, of the 840 articles used, 830 are used for training and 10 for testing.

According to the selected n-fold features, there are fluctuations in the results, although examining the rates shows no significant difference.

The classification results show that an RF classification algorithm with a ratio of 80.47% achieves the highest classification for all n-fold properties. The second highest regime is the logistic classification algorithm with 76.33%, and another strong classification algorithm is SMO with 73.69%.

The decision tree algorithm has a classification ratio of 68.21%, and the NB algorithm 61.66%. Although these rates are higher than the average, they do not give correct results.

Feature extraction involves a size reduction operation. Entrances to a system are not an entire piece of information, but rather the removal of some of the material that constitutes this information, and the system is built on these attributes. In this way, the size of a complicated dataset is reduced to a simpler problem. A properly constructed feature extraction and the design of a system that is suitable for these features can affect the success and performance of the result.

Attribute selection can be carried out using the "Select Attributes" tab. The "Attribute Evaluator" and "Search Method" methods are first selected; the former specifies the evaluator method to be used in order to determine the degree of conformity of each subset of attributes, while the latter specifies how to search within the attributes. The value of chi squared is calculated for the attribute evaluator in order to find selected attributes. Only the “number of lower case letters” property is not selected, and this property is removed from the data. The results are shown in Table 12.

Table 12. Success of classifications applied when the feature is reduced

Classifier	10 folds	30 folds	50 folds	70 folds	90 folds
ID3	65.59	66.78	66.78	68.21	67.61
NB	60.59	61.9	60.83	61.3	61.07
Logistic	76.42	75.59	75.59	75.35	75.47
Random forest	79.28	78.45	**80.35**	79.16	79.4

SVM	73.09	72.85	73.57	73.33	73.09

It can be seen from these results that RF classification algorithm gives better results than other algorithms with a success rate of 80.35%. The next best classification is the logistic algorithm with 76.42%.

A comparison between classification with and without feature reduction is given in Table 13.

Table 13. Classification with and without feature reduction

Classifier	10 folds	30 folds	50 folds	70 folds	90 folds
ID3	0	0	0	0	0
NB	-0.12	0.24	-0.36	-0.12	-0.23
Logistic	0.09	-0.12	-0.36	-0.24	0.47
Random forest	0.83	-1.55	1.19	-1.31	-1.07
SVM	-0.25	0	-0.12	0	0.12

There are two further important measurement criteria in information retrieval studies [18]: precision, i.e. how much of the information is related to the desired information, and recall, i.e. how much information has to be brought in.

Precision (p) is calculated as the ratio of the correct results to the total information provided. The formula for precision is given in Equation (10).

$$Precision = \frac{\{related \quad return\} \cap \{all \quad data \quad extraction\}}{\{all \quad data \quad extraction\}} \qquad (10)$$

Recall (r) is calculated as the ratio of the correct results that should return to the total correct results. The formula for recall is given in Equation (11).

$$Recall = \frac{\{related \quad return\} \cap \{all \quad data \quad extraction\}}{\{related \quad data \quad extraction\}} \tag{11}$$

In the above definitions, the f_1-measure is the harmonic mean of these values, as shown in Equation (12).

$$F_1 - Measure = 2\frac{Precision.Recall}{Precision + Recall} = 2\frac{p.r}{p + r} \tag{12}$$

It is also possible to link the F-measure to a β value. In this case, the F_1-measure is replaced by the F_β-measure, as shown in Equation (13).

$$F_\beta - Measure = (1 + \beta_2)\frac{Precision.Recall}{\beta^2.Precision + Recall} = (1 + \beta_2)\frac{p.r}{\beta^2.p + r} \tag{13}$$

Table 14 shows that the results are very close to each other, except for NB and decision tree for 10-fold cross-validation.

Table 14. Results for several algorithms without feature reduction

Algorithm	**Precision**	**Recall**	**F-measure**
ID3	0.655	0.656	0.656
Random Forest	0.787	0.785	0.781
Logistic Regression	0.763	0.763	0.763
NB	0.655	0.607	0.6
SVM	0.733	0.733	0.729

The results after feature extraction are given in Table 15.

Table 15. Results for several algorithms after feature reduction

Algorithm	**Precision**	**Recall**	**F-measure**
ID3	0.655	0.656	0.659
Random Forest	0.796	0.793	0.789
Logistic Regression	0.765	0.764	0.764
NB	0.655	0.606	0.599
SVM	0.73	0.731	0.727

As can be seen from Tables 14 and 15, the results are very close to each other. This means that feature reduction relatively unimportant in improving the success of the results.

The most successful of the classifiers used is RF, when both without feature reduction and after feature reduction. However, the success rates of the logistic regression and SVM classifiers are also observed to be high. The results show that the RF algorithm is most suitable for gender identification in Turkish texts.

REFERENCES

1. Argamon, S., & Koppel, M. (2012). A systemic functional approach to automated authorship analysis. JL & Pol'y, 21, 299.
2. Diri, B., & Amasyalı, M. F. (2003, June). Automatic author detection for turkish texts. In Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP) (pp. 138-141).
3. Murugaboopathy, N. S. G., Hariharasitaraman, S., & Babu, T. R. (2013). Appropriate gender identification from the text. International Journal of Emerging Research in Management and Technolog, 58-61.
4. Peng, F., Schuurmans, D., Keselj, V., Wang, S., “Language Independent Authorship Attribution using Character Level Language Models”, EACL, 267-274 (2003).
5. Cavnar, W. B. ve Trenkle, J. M., (1994), “N-gram-based text categorization”, Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval.Information Systems Project Management, Jolyon E. Hallows, AMACOM Pres.
6. Rezaei, A. M. (2014). Author Gender Identification from Text (Doctoral dissertation, Eastern Mediterranean University (EMU)-Doğu Akdeniz Üniversitesi (DAÜ)).
7. Argamon S., Fine J., Koppel M., Shimoni A. R., (2003), Automatically Categorizing Written Texts by Author Gender”, 1Dept. of Computer Science, Bar-Ilan University Ramat Gan 52900, Israel.
8. Nowson S., Oberlander J., (2006), “Openness and gender in personal weblogs ”, School of Informatics, University of Edinburgh,2 Buccleuch Place, Edinburgh, EH8 9LW
9. Fung, G. ve Mangasarian, O. L., (2002), Incremental Support Vector Machine Classification Second SIAM International Conference on Data Mining.

10. Cheng, N., Chen, X., Chandramouli, R., & Subbalakshmi, K. P. (2009). Gender identification from E-mails. CIDM, 9, 154-158.
11. Farrell, P.D. (2015). Automating Author Gender Identification From Blogs(University of North Carolina at Chapel Hill).
12. Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. Structure, 10, 1-5.
13. https://en.wikipedia.org/wiki/Weka_(machine_learning) (28 September 2018)
14. http://bilgisayarkavramlari.sadievrenseker.com/2012/04/11/karar-agaci-ogrenmesi-decision-tree-learning/ (01 October 2018)
15. Doğan, S. (2006). Türkçe dokümanlar için n-gram tabanlı sınıflandırma: Yazar, tür ve cinsiyet (Doctoral dissertation, YTÜ Fen Bilimleri Enstitüsü).
16. https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/ (14 December 2018)
17. https://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners (17 December 2018)
18. https://www.e-adys.com/makine_ogrenmesi/model-degerlendirme-siniflandirma/ (14 December 2018)
19. Levent, V. E., & Diri, B. (2014). Türkçe Dokümanlarda Yapay SinirAğları İle Yazar Tanıma. Akademik Bilişim.
20. Alowibdi, J. S., Buy, U. A., & Yu, P. (2013, August). Language independent gender classification on Twitter. In Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining (pp. 739-743). ACM.
21. Ciot, M., Sonderegger, M., & Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing(pp. 1136-1145).
22. Kolyiğit, Ö. (2013). Türkçe dokümanlar için yazar tanıma (Master's thesis, Adnan Menderes Üniversitesi, Fen Bilimleri Enstitüsü).
23. Amasyalı, M. F., & Diri, B. (2006, May). Automatic Turkish text categorization in terms of author, genre and gender. In International Conference on Application of Natural Language to Information Systems (pp. 221-226). Springer, Berlin, Heidelberg.

24. Amasyalı, M. F., Balcı, S., Mete, E., & Varlı, E. N. (2012). Türkçe metinlerin sınıflandırılmasında metin temsil yöntemlerinin performans karşılaştırılması/a comparison of text representation methods for Turkish text classification. EMO Bilimsel Dergi, 2(4).

APPENDIX A

TURKISH STOP WORDS

acaba, ama, ancak, artık, asla, aslında, az, bana, bazen, bazı, bazıları, bazısı, belki, ben, beni, benim, beş, bile, bir, birçoğu, birçok, birçokları, biri, birisi, birkaç, birkaçı, birşey, birşeyi, biz, bize, bizi, bizim, böyle, böylece, bu, buna, bunda, bundan, bunu, bunun, burada, bütün, çoğu, çoğuna, çoğunu, çok, çünkü, da, daha, de, değil, demek, diğer, diğeri, diğerleri, diye, dolayı, elbette, en, fakat, falan, felan, filan, gene, gibi, hangi, hangisi, hani, hatta, hem, henüz, hep, hepsi, hepsine, hepsini, her, her biri, herkes, herkese, herkesi, hiç, hiç kimse, hiçbiri, hiçbirine, hiçbirini, için, içinde, ile, ise, işte, kaç, kadar, kendi, kendine, kendini, ki, kim, kime, kimi, kimin, kimisi, madem, mı, mi, mu, mü, nasıl, ne, ne kadar, ne zaman, neden, nedir, nerde, nerede, nereden, nereye, nesi, neyse, niçin, niye, ona, ondan, onlar, onlara, onlardan, onların, onu, onun, orada, oysa, oysaki, öbürü, ön, önce, ötürü, öyle, sana, sen, senden, seni, senin, siz, sizden, size, sizi, sizin, son, sonra, şayet, şey, şimdi, şöyle, şu, şuna, şunda, şundan, şunlar, şunu, şunun, tabi, tamam, tüm, tümü, üzere, var, ve, veya, veyahut, ya, ya da, yani, yerine, yine, yoksa, zaten, zira

APPENDIX B

SAMPLE TEXT

Öyle özel zamanlarda çıkıyor ki aramızdan halk kahramanları...

Tam da ihtiyacımız varken... Tam da insanlıktan umudumuzu kesmişken...

Hatırlayın, bir ninemiz vardı hani...

Çanakkaleli...

Kendisine bağlanacak emekli maaşını geri çevirmişti;

"Ben devletime borçluyum aslında, ihtiyacım yok" diyerek.

Tek göz odada bir tas çorbaya kaşık sallayıp kıt kanaat geçinmeye çalışırken hem de... O sıralar gazeteler, 'Vergi yüzsüzleri açıklansın mı, açıklanmasın mı?' diye tartışıyordu. Zamanlamaya bakar mısınız!

301 maden işçimizin yüreklerimizi kömür karası bir kasvete buladığı günlerdi. Bir genç madenci çıktı. Ambulansa bindirilirken, sedyedeki beyaz çarşafı görüp "Çizmelerimi çıkarayım mı?" diye sordu.

Sahiplerinin, 300-500 bin liraya bir yaşam odasını bile esirgediği o cehennem madeninden çıkarken hem de...

Ve Şerife Nine... Terörün, şiddetin, vahşetin kol gezdiği haber bültenlerine nasıl da doğdu güneş gibi... Hastaydı, ayakta duracak hali yoktu ama geldiği hastanede çamurlu ayakkabılarını kapıda çıkarıp öyle girdi polikliniğe... Herkesin ayak izini, parmak izini bu dünyaya bırakmak için ölesiye yarıştığı, PKK'nın hastanelere roket atıp ambulansları kaçırdığı günlerde hem de...

Sağlık Bakanlığı harika bir jest yaptı. Şerife Cesur'un ismini Kırıkkale'nin Sulaklık ilçesindeki o devlet hastanesine verdi.

Koca koca işadamları, sanayiciler; yüzlerce milyon dolar bağışlayıp isimlerini ancak bir hastanenin küçük bir servisinin kapısındaki pirinç tabelada görebiliyorlar. Şerife Cesur'un ismi ise şimdilerde bir ilçe hastanesinin giriş kapısının üzerinde duruyor. Neden?

Çünkü hepimizin hastalanmaya yüz tutmuş ruhlarını, 'sadece ayakkabılarını çıkartarak' tedavi etti de ondan...

Ayaklarına sağlık ninem...

NOT: Belki geç oldu ama ambulanstaki o madenci gencin isminin, devlet tarafından işletilen bir maden tesisine verilmesini dört gözle bekliyorum

APPENDIX C

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: YAŞAR ÖNTÜRK, Ceren

Nationality: Turkish

Date and Place of Birth: 14 December 1986, Samsun

Phone: +90 545 787 24 09

email: yasar.ceren@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
BS	Çankaya Univ. Computer Engineering	2010
High School	Samsun Atatürk Anadolu Lisesi	2004

WORK EXPERIENCE

Year	Place	Enrollment
2012 -	T.C. Millî Savunma Bakanlığı	Computer Engineer

FOREIGN LANGUAGES

Advanced English, Beginner German