Automatic Categorization of Ottoman Poems

ETHEM F. CAN[1,*], FAZLI CAN[1,], PINAR DUYGULU[1], MEHMET KALPAKLI[2]**

[1] Computer Engineering Department, Bilkent University, Bilkent, Ankara 06800, Turkey
[2]History Department, Bilkent University, Bilkent, Ankara 06800, Turkey
[*] Present address: Computer Science Department, University of Massachusetts, Amherst, MA 01003
[**] Corresponding author (canf@cs.bilkent.edu.tr)

Abstract. Authorship attribution and identifying time period of literary works are fundamental problems in quantitative analysis of languages. We investigate two fundamentally different machine learning text categorization methods, Support Vector Machines (SVM) and Naïve Bayes (NB), and several style markers in the categorization of Ottoman poems according to their poets and time periods. We use the collected works (*divans*) of ten different Ottoman poets: two poets from each of the five different hundred-year periods ranging from the 15th to 19 th century. Our experimental evaluation and statistical assessments show that it is possible to obtain highly accurate and reliable classifications and to distinguish the methods and style markers in terms of their effectiveness.

1 Introduction

Automatic Text Categorization (ATC) methods aim to classify natural language texts into pre-defined categories and are used in different contexts ranging from document indexing to text mining [Sebastiani, 2002]. In the literature there are a variety of studies on ATC; however, quantitative studies on Ottoman literary text are not available. One reason for this is the fact that Ottoman documents are scarce in the digital environment. Initiatives such as Ottoman Text Archive Project (OTAP) and Text Bank Project (TBP) release transcribed versions of handwritten Ottoman literary texts. By considering the gap in the studies for the Ottoman language, this paper is motivated to classify a text with unknown poet or time period by employing automatic text categorization methods.

Our work contributes to research on quantitative analysis of Ottoman literature and diachronic computational studies on this language. In fact, it is the first quantitative study on Ottoman literature (a preliminary version can be seen in [Can et al., 2011]). In this study, we hypothesize that most-frequent word (MFW) style marker, quantitative text attribute, would be a good descriptor for Ottoman literary works since it is highly effective in the domain of Turkish language [Can & Patton, 2004] and the Ottoman language is primarily based on Turkish. Besides, two-word collocations (TWC) style marker is expected to provide high performance as phrases are commonly used in the Ottoman language. Accepting the Naïve Bayes (NB) as a baseline, Support Vector Machines (SVM) should provide higher performances considering the previous text categorization works in similar domains in other languages such as English [Yu, 2008].

The rest of the paper is organized as follows. In Section 2 we present a survey of related work. In Section 3 we introduce the experimental environment in terms of a brief description of the Ottoman language and

the data set we use in the experiments. In Section 4 we describe the style markers and the text categorization algorithms we use in the study. In Section 5 we present the experimental results in terms of the statistical assessment methods we use and then the results according to poets and time periods. Section 6 concludes the paper with a summary of findings and some future work pointers.

2 Related Work

In text-based data mining, statistical and machine learning methods aim to identify hidden occurrence patterns of objective text features [Witten et al., 2011]. Such patterns, which conceptually correspond to fingerprints of authors, are used for authorship attribution [Smalheiser & Torvik, 2009], author gender identification [Koppel et al., 2002], distinguishing works from each other according to intended audience [Binongo, 1994], finding the chronological order of works [Stamou, 2008], genre detection [Kanaris & Stamatatos, 2009], identifying an author's literary style development [Juola, 2007], etc. In these applications text features are referred to as style markers. Statistical methods have been used for a long time in authorship and categorization tasks; however, machine learningmethods are used in relatively more recent works. In some studies these two approaches are used together [Bagavandas et al., 2009]. [Merriam, 1989] uses a Bayes' theorem based method to classify twelve disputed Federalist Papers. [Clement & Sharp, 2003] and [Zhao & Zobel, 2005] use a similar method in their studies as well. [Houvardas & Stamatatos, 2006] employ SVM for author identification. [Joachims, 1998, Joachims, 2001] makes use of SVM in the task of text classification and observes that SVM is robust and it does not require parameter tuning for the task. [Kucukyilmaz et al., 2008] use machine learning approaches including k-nearest neighbor (k-NN), SVM, and NB to determine authors of chat participants by analyzing their online messaging texts. [Yu, 2008], as in our work, focuses on text classification methods in literary studies and uses NB, and SVM classifiers. In her study, the effect of common and function words are tested for the eroticism classification of Dickinson's poems and the sentimentalism classification of chapters in early American novels.

Style markers have been used for authorship attribution for a long time. [Holmes, 1994] gives a detailed overview of the stylometry studies in the literature within a historical perspective and presents a critical review of numerous style markers. [Grieve, 2007] has a similar study on style markers. [Juola, 2006] and [Stamatatos, 2009] present types of analysis, features, and recent developments in authorship attribution studies. [Burrows & Craig, 2001] examine two Categorization of Ottoman Poems seventeenth-century poems by using most frequent words. [Holmes et al., 2001] investigate the exact author of the Pickett Letters as a complement to traditional historical research by using top 60 frequently occurring function words. [O'Brien & Darnell, 1982] uses word collocations frequency and cover six case studies in authorship attribution. [Stamatatos et al., 1999] study categorization of ten Greek newspaper columnists using a text-processing tool (Sentence and Chunk Boundaries Detector) that segments texts into sentences. They use 22 style markers: three token-level, ten phrase-level, and nine analysis-level. A study based on writing style for identification of individuals is carried out by [Abbasi & Chen, 2008]. They develop a Karhunen-Loeve transform-based write prints method for identification and similarity detection in terms of identity.

Although there is no quantitative study on the Ottoman language, studies on contemporary Turkish do exist in literature. [Can & Patton, 2004] analyze change of writing style with time by using word lengths and most frequent words for the Turkish authors Çetin Altan and Yaşar Kemal. In another study the same

authors [Patton & Can, 2004] analyze four *İnce Memed* novels of Yaşar Kemal. In the study, they use six different style markers and determine the one that provides the best performance. Language change in Turkish and its quantification using time-separated parallel translations are studied by [Altintas et al., 2007] by using various style markers including vocabulary richness. The study shows that Turkish words have become longer by time, word stems have become significantly shorter, suffix lengths have become significantly longer for types, and the vocabulary richness based on word stems has shrunk significantly. In a recent study [Can & Patton, 2010] analyze 20th century Turkish literature in terms of change of word characteristics.

3 Experimental Environment

3.1 Ottoman Language

The Ottoman language (Osmanlıca) describes the Western Turkish dialect spoken during the period of Ottoman rule in Anatolia, Eastern Europe, much of the western portion of the Central Middle East, and North Africa. More specifically, however, it describes the literary and spoken language of the Ottoman elites, which amassed a huge vocabulary by combining Turkish words with borrowings from Arabic and Persian, and much less frequently from Western languages such as Italian, Greek, Hungarian, and Slavic.

Although the main sentence structure, morphology and syntax are Turkish, the Ottoman language is written with Arabic alphabet. This alphabet consists of thirty letters and is written from right to left. Except a few, letters are attached to each other and borrowed letters are written differently at the beginning, middle, and end of each word. The words adopted from Arabic and Persian are written with their original spelling. Besides, there is not a letter for every vowel which makes it difficult to read and understand the script.

Ottoman literature is usually understood in the restricted sense of the literature written in this elite language in forms and genres most of which were adapted from Persian and Arabic models.

3.2 Test Collection

In this study, we focus on Ottoman literary texts from ten poets and five consecutive centuries. Table 1 gives information about these texts, which is our test collection. It is created by an expert in the field: Prof. Kalpaklı, who is also from the OTAP project. The text associated with each poet is called *divan* which is an anthology of the poet's work, as it might be selected poems or all poems of the same author. The poets in this study are selected in such a way that they all together provide a good representation of the underlying literature. There are nine male and one female (Mihrî Hatun) - which is a rare case in the Ottoman literary tradition- poets from five different centuries. Some of them have smooth styles; whereas, some of them do not.

The works of the selected poets as given in Table 1 acquire almost all characteristics of the Ottoman lyric poetry [Andrews et al., 1997]. In our study, the poets whose life spanned two centuries are associated with the century they died (only exception is Mihrî Hatun since she lived in the 16th century for a short period of time).

Table 1. Ottoman literary texts-test collection.

Poet (Divan #, No. of Poems)	Century	Life Span	#Tokens	#Types
Mihrî Hatun (D_1, 245)	15th	1460-1512	34,735	9,188
Sinan Şeyhî (D_2, 221)	15th	1371?-1431	27,743	10,784
Hayalî Bey (D_3, 619)	16th	1500-1557	54,338	15,727
Revânî (D_4, 141)	16th	1475-1524	24,881	8,315
Nef'î (D_5, 224)	17th	1572-1635	51,075	14,492
Neşatî (D_6, 186)	17th	?-1674	23,799	7,984
Osmanzâde Tâ'ib (D_7, 189)	18th	1660-1724	19,610	8,772
Şeyh Gâlip (D_8, 580)	18th	1757-1799	59,301	18,506
Şânîzâde's Atâullah (D_9, 125)	19th	1771-1826	8,265	4,409
Yenişehirli Avnî (D_{10}, 425)	19th	1826-1884	54,927	18,785
Total	15th-19th	1371-1884	358,674	62,609

The texts are from the language resources OTAP and TBP. The first language resource, OTAP (http://courses.washington.edu/otap) is a cooperative international project employing computer technology and the resources of the World Wide Web to make transcribed Ottoman texts broadly accessible to international audiences. OTAP is an umbrella name for OTEP (Ottoman Text Edition Project) / OTAP (Ottoman Text Archive Project) / OHD (Ottoman Categorization of Ottoman Poems Historical Dictionary) and it is a joint effort between the University of Washington, in Seattle, WA, USA and Bilkent University in Ankara, Turkey. The second resource, TBP is a network of scholars of Ottoman literature and aims to provide an electronic transcribed texts pool for use of its members. Recently TBP has more than 120 members from all around the world. Its text pool contains Ottoman literary and historical texts from the 13th to 20th century.

3.3 Blocking

In order to prepare the data for experiments each document is split into blocks with k number of words, where k is taken as 200 to 2000 with 200-word increments. For example, if the block size (k) is 200 words, each work is divided at every 200th word; accordingly, the first 200 words constitute the first block and so on. If the number of words in the last block is smaller than the chosen block size that block is discarded. Blocking is a common approach used in stylometric studies [Forsyth & Holmes, 1996].

4 Style Markers and Classification Algorithms

In this section we first explain the style markers and classification algorithms we investigate in this study. The style marker information is needed in the presentation of the algorithms.

4.1 Style Markers

[Can & Patton, 2004] show that most frequent words and word lengths (in the form of token and type lengths) as style markers have remarkable performance in determining the change of writing style with time in Turkish. In another work the same authors [Can & Patton, 2010] provide consistent results with the aforementioned study (especially in terms of most frequent words). Because of their observations and since Turkish is the basis of the Ottoman language we use these text features in our study. We also use two-word collocations as another style marker, since phrases are one of the characteristic features of the Ottoman language and poets.

In the rest of this section we describe the style markers. To begin with, a few definitions are in order: A token is a word and a continuous string of letters, type is defined as a distinct word. For example, there

are thirteen tokens and eleven types in the subset which consists of the words "Niçe feryada vara nâlelerüm âh sana, Niçe bir 'arz ideyüm halümi her-gâh sana." Besides, a word that contains a dash is counted as one token, such as "Gül-izâr" (means rose-cheeked). On the other hand, "Gül izâr" (means rose cheek) is counted as two tokens, and they u a do have different meanings.

Table 2. Sixty most frequent words (MFW) for a training case with block of size 200 words (most to least frequent: bir, bu ile, ki…). In the experiments selected words or their rankings show minor variations from this. English translation of their (most frequent) meanings are also provided.

bir/one	bu/this	ile/with
ki/that	ol/be	kim/who
her/every	ne/what	ey/hey
olur/happens	gibi/like	ola/wish to be
dil/soul	yine/again	olsa/if exists
ben/me	var/there is	eyler/it does
gül/rose	oldu/happened	sen/you
ider/do	cân/dear	mi/adv.of interrog.
vü/and	içün/for	dem/moment
nice/how	olup/to become	felek/destiny
idi/was-were	olmaz/no	eyledi/made
ger/if	hem/both	cihân/earth
dil/heart	ehl/people	dahi/even
yok/absent	sana/to you	üzre/so as to
gün/day	böyle/like this	bana/to me
eyle/to make	yâr/friend	gam/sorrow
`atâ/bestowal	beni/me	olsun/so be it
durur/stops	ise/although	itdi/caused
kadar/till	tâ/until	şimdi/now
iken/while	iki/two	anâ/pain

- Most Frequent Words (MFW): We determine the sixty most frequent words appearing in the corpus excluding the test blocks (more details on test-block-exclusion during are provided in the next section). Then we obtain the normalized frequency of these words in each block. Table 2 provides the most frequent words selected for one of the 200-word blocks (English translation of their most frequent meanings are also provided). Note that as explained later we use cross validation and therefore training text changes from one cross validation step to next even for the same block size. This has a small impact on the set of the selected words or on their frequency ranking: about 90 percent of the words of this table is selected for all block sizes.
- Token & Type Length (TOL-TYL): We consider the words of length 1 to 15. We ignore the words that contain more than 15 characters, since such words are uncommon and constitute only 0.14% of the entire corpus. In this process, we count the number of occurrences of a token or a type with length of 1 to 15 individually in a block, then compute their normalized frequencies and obtain the corresponding block vector of size 15.

- Two-word Collocations (TWC): We determine the sixty most frequent two-word collocations appearing in the corpus excluding the test blocks. Then we obtain the normalized frequency of appearance of these phrases in each block and generate the corresponding numerical block vector of size 60. Table 3 provides the most frequent two-word collocations selected for one of the 200-word blocks. Like the selection of the 60 most frequent words, different training cases has a small impact on the set of the selected two-word collocations: about 90 percent of the two-word collocations of this table is selected for all cases.
- We also use all four style markers together.

Table 3. Sixty most frequent two-word collocations (TWC) for a training case with block size of 200 words (most to least frequent: her dem, bu midur, `aceb mi, var ise…). In the experiments selected two-word collocations or their rankings show minor variations from this. English translation of their (most frequent) meanings are also provided.

her dem/always	bu mıdur/is it this	`aceb mi/wonder that
var ise/if exists	her bir/each	yâ rab/o Lord
ne kadar/how many	rûz şeb/day night	bir gün/one day
her ne/whatever	şevk ile/with passion	bir dem/once
ile bir/with one	nice bir/for how long	bu gün/today
gibi bir/like one	ey dil/oy my heart	ol kadar/that much
gül gibi/like a rose	tâ ki/until that	ne var/what's the matter
ben de/me too	bir nefes/one breath	mâh-ı nev/new moon
ki bu/this one	ile ol/be with	kim ola/who's that
bu dem/this moment	lutf ile/with kindness	derd ile/with sorrow
gün gibi/apparently	haşre dek/till doomsday	nedür bu/what is this
yine bir/again one	dem ki/time that	içinde gizlidir/hidden inside
bu kadar/this much	bu kim/who is this	aşk ile/with love
ol ki/be that	böyle bir/like this	olur bu/this becomes
gel ey/come oh	dil ki/hearts that	can ile/with love
etdin beni/you've made me	virdüm hep/I always gave	bilür bilmez/guessing
bilmez sanur/thinks that doesn't know	kim istemez/everyone wants	bu gece/this night
olmasa ger/if not	dil cân/dear soul	olma zînhar/beware don't be
var iken/as it exists	ey dost/oh friend	devlet ile/with good luck
güyâ ki/imagine that	nice kim/how is that	bir mertebe/to a certain point

Since person and place names are context dependent they are manually skipped while determining the most frequent words and two-word collocations. In the experiments letters are in lowercase.

4.2 Classification Algorithms

We employ two machine learning-based classifiers: Naïve Bayes (NB): a generative classifier and Support Vector Machines (SVM): a discriminative classifier [Duda et al., 2000, Vapnik, 1995]. The use of fundamentally different classifiers provides us a wide test spectrum to investigate the performance of

machine learning methods in ATC of Ottoman literary texts. Furthermore, NB and SVM are commonly used in similar studies. For example, [Yu, 2008] indicates that SVM is among the best text classifiers. In the same work it is also indicated that NB is a simple but effective Bayesian learning method and often used as a baseline. Based on these observations we focus on these two methods and investigate and compare their performances in ATC of Ottoman literary texts.

In Naïve Bayes classifier each feature is assumed to be independent of every other feature. Even though NB is based on a simple probabilistic schema with good results, in real life cases this assumption might not be valid or this may lower its success rate. In this study we employ the model used in [Zhao & Zobel, 2005].

Support Vector Machines (SVM) classifier does not use the assumption that features are independent; it constructs a hyper-plane using a set of support vectors in a high dimensional space and tries to find decision boundary among the classes by making the separation or margin among them larger [Joachims, 1998, Joachims, 2001, Vapnik, 1995]. In SVM we employ two different kernel functions; polynomial (poly), and radial-basis-function (rbf) kernels. In the experiments of SVM with the polynomial kernel (SVM-poly) we run tests when the degree is set to 1, 2, 3, 4, and 5. With the radial-basis-function kernel (SVM-rbf), we set γ (width of the kernel) to 0.6, 0.8, 1.0, 1.2, and 1.4. Considering the regularization parameter we use a default value (1.0) for all experiments. Similar settings for SVM are used in [Joachims, 1998] for text classification and successful results are obtained.

In our study, for the construction of training and test corpora, we prefer cross validation (leave-one-out) in which division of data is not important compared to splitting the corpus as training and test set. The experimental results are then averaged across all iterations of cross validation. Since each element in the corpus is used in training and test set at least once. We use OpenCV library (http://sourceforge.net/projects/opencvlibrary/) that is based on LibSVM [Chang & Lin, 2011] to train the classifiers. For the construction of training and test corpora, we prefer K-fold cross validation in which division of data is not important compared to splitting the corpus as training and test set. In our study, we use ten for K.

Instead of extracting two single lists of words for most frequent words and two-word collocations using the whole corpus, we extract individual lists of words for each iteration of the cross validation from the training corpus so that test corpus do not have any effect in the selection of the most frequent words and two-word collocations. In other words, for each test block we determine the most frequent words and two-collocations by only considering the contents of the training blocks. In this way unbiased feature selection is guaranteed.

5 Experimental Results

5.1 Statistical Evaluation Approach

We conduct a two way analysis of variance (ANOVA) in order to see if the classification performances of the tested cases are significantly different from each other. When the main effects of the factors, style markers and machine learning algorithms, are statistically significantly different in explaining the variance of classification accuracy, we conduct post-hoc multiple comparisons using Scheffe's correction [Scheffe, 1953] for the levels of each factor.

Matlab's Statistical Toolbox is used to conduct the ANOVA and multiple comparison tests. (Later, in Fig. 3 and 5 we provide the multiple comparison results for the machine learning algorithms in poet and time period categorization for $\rho < 0.05$.)

5.2 Classification by Poet

In the experiments we analyze the performances of the classifiers with MFW, TOL, TWC, and TYL. For each case, we obtain a highly effective performance.

In Table 4, we provide poet classification accuracies of the style markers MFW, TOL, TWC, and TYL with the machine learning methods NB, and two versions of SVM for different block sizes. The table shows that for MFW with SVM-poly, we obtain the best accuracy score when the polynomial degree is 1; similarly, we obtain the best accuracy score for SVM-rbf when γ is 1.2. In the table the values of these parameters that provide the best performances of TOL, TWC, and TYL are also given.

Table 4. Poet classification accuracies of MFW, TOL, TWC, and TYL with NB, SVM-poly, and SVM-rbf for different block sizes. The paramters, polynomial degree for SVM-poly and γ for SVM-rbf that yield the listed results, are also provided.

Block Size	MFW			TOL			TWC			TYL		
	NB	SVM-poly deg=1	SVM-rbf γ=1.2	NB	SVM-poly deg=5	SVM-rbf γ=1.2	NB	SVM-poly deg=1	SVM-rbf γ=1.0	NB	SVM-poly deg=5	SVM-rbf γ=1.2
200	63.13	73.36	**74.75**	30.98	28.25	26.99	34.85	34.23	35.76	37.12	34.65	35.78
400	73.67	84.50	**84.97**	35.35	34.35	36.11	45.00	42.34	43.63	37.89	35.78	35.32
600	81.71	87.88	**88.37**	43.66	40.33	41.48	49.91	48.99	49.57	46.28	40.78	37.93
800	85.21	**91.00**	90.49	45.69	43.52	44.90	57.07	56.44	57.58	49.65	44.34	45.42
1000	85.47	91.08	**91.42**	44.77	42.49	43.55	61.96	61.22	61.50	50.00	45.35	48.23
1200	86.55	**91.32**	**91.32**	51.53	49.16	49.80	63.72	65.77	64.73	56.93	48.29	47.82
1400	**91.66**	91.28	91.12	51.36	47.02	48.46	64.45	69.21	69.69	59.55	50.24	48.56
1600	89.64	**91.22**	91.12	50.50	48.49	48.64	68.30	68.53	69.90	55.08	49.40	46.65
1800	88.57	**92.51**	**92.51**	54.52	55.38	56.88	69.96	69.42	68.10	64.22	56.76	56.32
2000	87.05	**92.80**	**92.8**	57.78	52.4	55.44	71.93	71.42	71.42	59.17	53.63	54.53
Avg.	83.27	88.70	**88.89**	46.60	44.14	45.23	58.72	58.76	59.19	51.59	45.92	45.66

For MWF for all block sizes SVM-poly and -rbf provide better results than NB. Both versions of SVM have similar results. For TOL for almost all block sizes NB provides slightly better results than SVM-poly and -rbf. Scores of SVM-rbf are slightly better than the scores of SVM-poly. For TWC all methods yield similar accuracy scores. For TYL for all block sizes NB provides a slightly better performance that those of the SVM classifiers and both versions of SVM have similar performances. From the table we can see that for MFW the difference between NB and SVM classifiers are noticeable for the other cases NB and SVM classifiers performances are mostly compatible with each other.

In Fig. 1 the average style marker classification rates are provided. The averages are for individual block sizes and are obtained by using the results listed in Table 4. The purpose of this figure is to show the performances of the style markers with a variety of learning methods (in our case they are NB and two

versions of SVM). As can be seen MFW provides a performance which is consistently better than those of the other style markers. TWC provides the second best performance, and TYL and TOL (in that order) follow TWC. Considering individual style markers or all-style markers, the performance gets better when the block size gets larger since larger blocks contain more evidence about what they are.

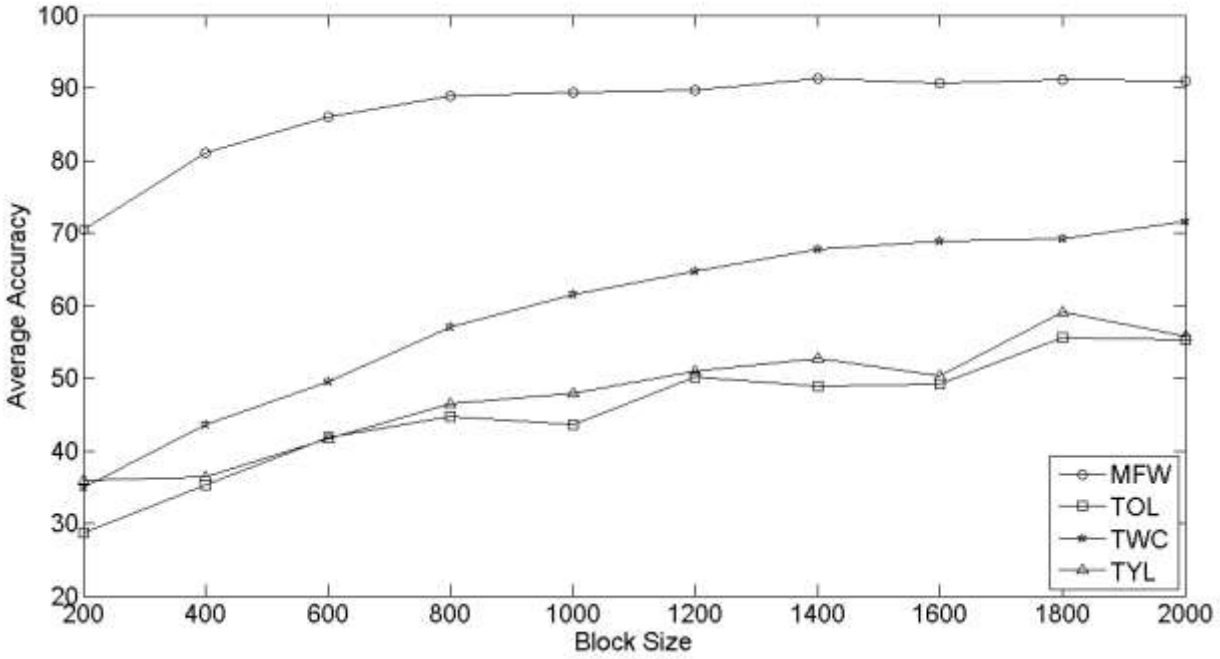

Figure 1. Poet average correct classification accuracies for four (MFW, TOL, TWC, and TYL) style markers using results of NB, SVM-poly, and SVM-rbf classifiers.

We also combined all four style markers and the results are given in Table 5. Both versions of SVM provide a consistently better performance than NB. Furthermore, SVM provides a slightly better performance than that of MWF, which yields the highest correct classification rate. However, NB provides a performance which is substantially lower than its best case which is again obtained with MFW.

Table 5. Poet classification accuracies using all style markers.

Classifier	200	400	600	800	1000	1200	1400	1600	1800	2000	Avg.
NB	40.53	49.18	41.59	49.06	57.60	55.56	60.36	59.65	57.41	62.18	53.31
SVM-poly (d=1)	**76.18**	85.72	**87.83**	**91.64**	92.84	93.77	**92.57**	92.57	93.57	92.34	89.94
SVM-rbf (γ=1.2)	76.06	**86.07**	87.66	**91.64**	**92.85**	**93.78**	**92.57**	**92.98**	**93.58**	**92.40**	**89.96**

In practical applications one may want to use a majority vote approach for the classification of a document based on the success rate with its blocks. In such an approach, it would be important to consider document level accuracy by considering the categorization accuracy obtained by not one but with all blocks. With this in mind, we also considered block level classification accuracy with smaller block sizes. In Fig. 2, we provide an example confusion matrix of poet categorization experiment (with SVM-poly, using all style markers together when the block size is 200). The figure illustrates the

predictions of the unlabeled blocks in the test phase. The prediction rates are mapped to a gray-scale color domain ("absolute" black represents 100%, and white represents 0%). In the figure, the diagonal cells give the ratio of the correctly classified blocks. As can be seen the majority of the blocks are classified to the right document. The first line of the matrix (D_1) shows that the ratio of correct classifications for D_1 which is 0.78 (78%). However, 11% of the blocks are incorrectly classified as belonging to D_2. As intuitively expected, texts of somewhat contemporaneous poets tend to mix with each other rather than texts of non-contemporaneous poets. In all cases the majority voting approach gives the correct result, i.e., the majority of blocks of a work goes to its own category.

Divan	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
D_1	0.78	0.11	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.06
D_2	0.00	0.77	0.08	0.00	0.08	0.00	0.00	0.08	0.00	0.00
D_3	0.04	0.08	0.79	0.04	0.00	0.00	0.00	0.04	0.00	0.00
D_4	0.08	0.08	0.00	0.69	0.08	0.08	0.00	0.00	0.00	0.00
D_5	0.00	0.00	0.00	0.04	0.79	0.04	0.08	0.04	0.00	0.00
D_6	0.07	0.00	0.00	0.07	0.00	0.71	0.07	0.00	0.07	0.00
D_7	0.00	0.00	0.00	0.00	0.00	0.10	0.80	0.10	0.00	0.00
D_8	0.00	0.03	0.00	0.00	0.03	0.07	0.03	0.79	0.03	0.00
D_9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.70	0.10
D_{10}	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.04	0.79

Figure 2. Poet classification confusion matrix for the SVM-poly classifier using all style markers together when block size is 200 (some row sums are not equal to 1.00 due to rounding).

Summary and Statistical Analysis

If we look at the experimental results we see that in general SVM is a more accurate classifier. For example, with the stand-alone use of the four style markers the SVM classifiers provide a performance compatible to (with TOL and TYL), or better than (with MFW) that of NB. When all four style markers are used together, SVM provides a substantially better performance than NB. This can be attributed to the fact that SVMs are robust with respect to large dimensionalities.

Fig. 3 provides the multiple comparisons of the machine learning algorithms in poet categorization for $\rho < 0.05$ using Scheffe's method (the values in the comparisons are the scores in Table 4 and 5). In the figure, if the vertical dashed lines appearing at the edges of the horizontal lines (machine learning algorithms) cut another horizontal line then the groups of the lines are not significantly different, otherwise they are significantly different. According to comparisons, the SVM classifiers with different kernels are not significantly different from each other, but they are significantly different from the NB classifier.

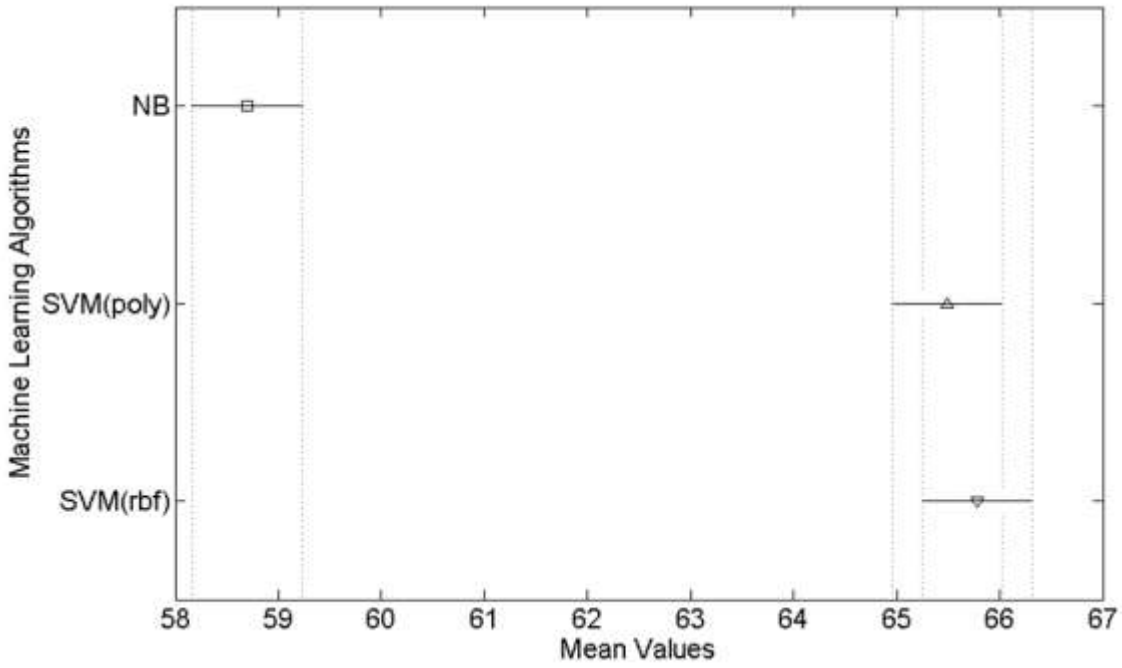

Figure 3. Multiple comarison results using Scheffe's method for machine learning algorithms in poet categorization.

5.3 Classification by Time Period

In the classification of texts by time period (century) with individual style markers, MFW (Most Frequent Words) provides the best classification scores (up to 94%) with the SVM classifier. TWC provides the second best performance, and TOL and TYL follow the style marker TWC. SVM mostly performs better than NB with MFW. For TOL and TYL, NB provides slightly more accurate results than SVM. The NB and SVM classifiers have almost the same performance with TWC.

In Table 6, the classification accuracies using all style markers for different block sizes and machine learning methods are provided. When we use all style markers, the performance of NB decreases with respect to its performance with the individual style markers. However, SVM provides a better performance when we compare the experiments focusing on individual style markers. As in the poet categorization, block size effect has a similar pattern on time categorization; i.e. the larger block size provides better performance.

Table 6. Time period classification accuracies using all style markers.

Classifier	200	400	600	800	1000	1200	1400	1600	1800	2000	Avg.
NB	42.55	53.50	60.06	62.35	63.66	66.21	64.71	67.87	65.33	66.19	61.24
SVM-poly (d=1)	78.42	**83.14**	**85.56**	**90.63**	**90.46**	**90.42**	91.21	**92.16**	**95.53**	**94.37**	**89.19**
SVM-rbf (γ=1.2)	**75.98**	83.00	84.73	85.95	88.64	89.45	**91.93**	90.24	94.84	90.31	87.50

Confusion matrix of time period categorization for SVM-poly using all style markers together when the block size is 200 is provided in Fig. 4. The majority of the blocks for the given case are classified to the

correct time period (indicated by the diagonal cell contents). As intuitively expected, texts coming from adjacent time periods tend to mix with each other rather than texts with distant time periods.

Century	15th	16th	17th	18 th	19 th
15th	0.83	0.14	0.00	0.03	0.00
16th	0.09	0.77	0.11	0.03	0.00
17th	0.00	0.08	0.78	0.14	0.00
18th	0.00	0.03	0.05	0.82	0.11
19th	0.00	0.00	0.08	0.19	0.73

Figure 4. Time period confusion matrix for the SVM-poly classifier using all style markers together when block size is 200 (for one case row sum is not equal to 1.00 due to rounding).

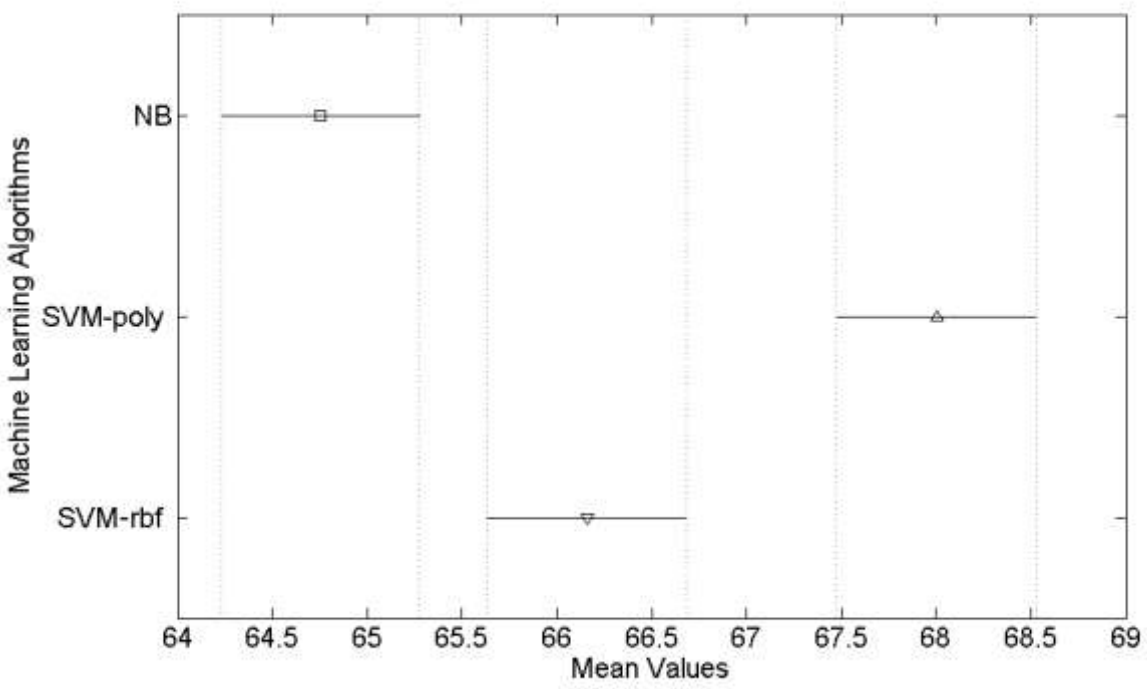

Figure 5. Multiple comparison results using Scheffe's method for machine learning algorithms in time period categorization.

Summary and Statistical Analysis

Summary and Statistical Analysis When we consider the time period experiments, MFW and SVM, respectively, appear as the most appropriate style marker and the machine learning method.

Fig. 5 provides the multiple comparisons of the machine learning algorithms in period categorization for $\rho < 0.05$ using Scheffe's method (the values in the comparisons are the scores in Table 4 and 5). According to comparisons, they are significantly different from each other for combinations of all pairs.

6 Conclusion and Future Work

We contribute to research on quantitative and diachronic analysis of the Ottoman language. The experimental results confirm our hypotheses: SVM is a more accurate classifier compared to NB in categorization tasks; similarly, MFW outperforms the other style markers; furthermore, with no doubt TWC is the second best style marker for categorization. We show that it is possible to distinguish poets from each other and the same is also true for time periods. In this process we obtained remarkable results, almost 90% accuracy, even with small block sizes.

Another contribution of the study is the Ottoman text categorization collection we constructed by using the OTAP and TBP Ottoman language resources. It is prepared by a literary scholar: the poets and their works provide a good representation of the underlying Ottoman literature. The same set of literary works, our test collection, can be used by other researchers in similar studies. For such cases our results provide a baseline for comparison. Our results can be used for the construction of tools in finding hidden patterns in text and understanding author and time period of Ottoman literary works. For such tools in other languages one may refer to the JGAAP [Juola, 2006], MONK [Guzmán-Cabrera et al., 2009], and Nora [Plaisant et al., 2006] projects.

SVM and MFW yield performances that are mostly statistically significantly different from their counterparts. Based on these observations we recommend their use in future related studies. Furthermore, the majority vote-based classification of large documents gives excellent results. Additional experiments with various combinations of style markers can be fruitful to further improve the classification accuracy.

In future work, diachronic investigation of the Ottoman language and supporting visual investigation of Ottoman script by lexical statistical information look interesting. For language change studies, an approach similar to the one defined in [Altintas et al., 2007] can be employed by using frequently (re)written famous stories, such as (Leyla and Mecnun), in different centuries; or *nazires*, different interpretations, of the poems of famous Ottoman poets.

Acknowledgements

This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under the grant number 109E006. Any opinions, findings and conclusions or recommendations expressed in this article belong to the authors and do not necessarily reflect those of the sponsor. We thank Walter G. Andrews, Jon M. Patton, and Selim Aksoy for their helpful pointers.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1-29.

Altintas, K., Can, F., & Patton, J. M. (2007). Language change quantification using time-separated parallel translations. *Literary and Linguistic Computing*, 22(4):375-393.

Andrews, W. G., Black, N., & Kalpaklı, M. (1997). *Ottoman Lyric Poetry*. University of Texas Press, Austin, Texas, USA.

Bagavandas, M., Hameed, Abdul, & Manimannan, G. (2009). Neural computation in authorship attribution: The Case of selected Tamil articles. *Journal of Quantitative Linguistics*, 16(2):115-131.

Binongo, J. N. G. (1994). Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: A statistical expression. *Literary and Linguistic Computing*, 9(4):267-279.

Burrows, J. F., & Craig, H. (2001). Lucy Hutchinson and the authorship of two seventeenth-century poems: A computational approach. *The Seventeenth Century*, 16:259-282.

Can, E. F., Can, F., Duygulu, P., & Kalpakli, M. (2011). Automatic categorization of Ottoman literary texts by poet and time period. ISCIS 2011: 51-57

Can, F., & Patton, J. M. (2004). Change of writing style with time. *Computers and the Humanities*, 38(1):61-82.

Can, F., & Patton, J. M. (2010). Change of word characteristics in 20th century Turkish literature: A statistical analysis. *Journal of Quantative Linguistics*, 17(3):167-190.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27.

Clement, R., & Sharp, D. (2003). Ngram and Bayesian classification of documents. *Literary and Linguistic Computing*, 18(4):423-447.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification, 2nd ed*. Wiley-Interscience, USA.

Forsyth, R. S., & Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, 11(4):162-174.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(2):251-266.

Guzmán-Cabrera, R., Montes-y-Gómez, M., Rosso, P., & Pineda, L. V. (2009). Using the Web as corpus for self-training text categorization. *Information Retrieval*, 12(3):400-415.

Holmes, D. I. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87-106.

Holmes, D. I., Gordon, I., & Wilson, C. 2001. A widow and her soldier: Stylometry and the American Civil War. *Literary and Linguistic Computing*, 16(4):403-420.

Houvardas, J., & Stamatatos, E. 2006. N-Gram feature selection for authorship identification. *Lectures Notes in Computer Science*,4183:77-86.

Joachims, T. 1998. Text Categorization with support vector machines: Learning with many relevant features. In ECML-98 : European conference on machine learning, pages 137-142.

Joachims, T. 2001. A statistical learning model of text classification for support vector machines. In Proceedings of the 24th ACM SIGIR Conference, pages 128-136.

Juola, P. (2006). Authorship attribution. *Foundation and Trends in Information Retrieval*, 1(3):233-334.

Juola, P. (2007). Becoming Jack London. *Journal of Quantitative Linguistics*, 14(2-3):145-147.

Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management*, 45(5):499-512.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*,17(4): 401-412.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management*, 44(4):1448-1466.

Merriam, T. (1989). An experiment with the Federalist papers. *Computers and the Humanities*, 23(3):251-254.

O'Brien, D. P., & Darnell, A. C. (1982). *Authorship Puzzles in the History of Economics: A Statistical Approach*. Macmillan, London.

Patton, J. M., & Can, F. (2004). A stylometric analysis of Yaşar Kemal's *İnce Memed* tetralogy. *Computers and the Humanities*, 38(4):457-467.

Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M., G., Smith, M., N., Clement, T., & Lord, G. (2006). Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In Proceedings of ACM/IEEE Joint Conference on Digital Libraries JCDL, pages 141-150.

Scheffe, H. (1953). A Method for judging all contrasts in the analysis of variance. *Biometrica*, 40(1-2):87-110.

Sebastiani, F. (2002). Machine learning in automatic text categorization. *ACM Computing Surveys*, 34(1):1-47.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, 43:287-313.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538-556.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (1999). Automatic authorship attribution. In Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics, pages 158-164.

Stamou, C. (2008). Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, 23(2):181-199.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*. Morgan Kaufmann, Burlington MA, USA.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327-343.

Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*, 3689:174-189.