Automatic Categorization of Ottoman Literary Texts by Poet and Time Period

Ethem F. Can, Fazli Can, Pinar Duygulu and Mehmet Kalpakli

Abstract Millions of manuscripts and printed texts are available in the Ottoman language. The automatic categorization of Ottoman texts would make these documents much more accessible in various applications ranging from historical investigations to literary analyses. In this work, we use transcribed version of Ottoman literary texts in the Latin alphabet and show that it is possible to develop effective Automatic Text Categorization techniques that can be applied to the Ottoman language. For this purpose, we use two fundamentally different machine learning methods: Naïve Bayes and Support Vector Machines, and employ four style markers: most frequent words, token lengths, two-word collocations, and type lengths. In the experiments, we use the collected works (*divans*) of ten different poets: two poets from five different hundred-year periods ranging from the 15th to 19th century. The experimental results show that it is possible to obtain highly accurate classifications in terms of poet and time period. By using statistical analysis we are able to recommend which style marker and machine learning method are to be used in future studies.

E. F. Can · F. Can (✉) · P. Duygulu
Department of Computer Engineering, Bilkent University,
06800 Ankara, Turkey
e-mail: canf@cs.bilkent.edu.tr

E. F. Can
e-mail: efcan@cs.bilkent.edu.tr

P. Duygulu
e-mail: duygulu@cs.bilkent.edu.tr

M. Kalpakli
Department of History, Bilkent University, 06800 Ankara, Turkey
e-mail: kalpakli@bilkent.edu.tr

E. Gelenbe et al. (eds.), *Computer and Information Sciences II*,
DOI: 10.1007/978-1-4471-2155-8_6, © Springer-Verlag London Limited 2012

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 [1]. In the literature there are a variety of studies in ATC; however, studies on historical manuscripts are rare. One reason for this is the fact that old documents are scarce in the digital environment. Resources such as Ottoman Text Archive Project (OTAP) and Text Bank Project (TBP) release transcribed versions of handwritten Ottoman literary texts [2]. There are millions of pages of texts in Ottoman that are to be analyzed and classified after transcription [3]. 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 and ultimately show the achievability of effective automatic categorization of historical Ottoman texts so that it can be employed when these documents are transcribed.

The contributions of this study are the following. We provide the first style-centered ATC study on the Ottoman language. Within the context of this language, we evaluate the performance of two different machine learning methods in ATC by using four style markers. By using statistical analysis we are able to recommend which machine learning method and style marker are to be used in future studies. The availability of huge amount of text in the Ottoman language, especially the Ottoman archives [3], confirms the practical importance and implications of our study.

2 Related Work

In ATC, style markers are used in analyzing the writing styles of authors. Holmes [4] gives a detailed overview of the stylometry studies in the literature within a historical perspective and presents a critical review of numerous style markers. Statistical methods have been used for a long time in authorship and categorization tasks and machine learning methods are used in relatively more recent works. A Bayes' theorem-based algorithm is firstly used to classify twelve disputed Federalist Papers in [5]. McCallum and Nigam [6] compare a multivariate Bernoulli model, and multinomial model. SVM (Support Vector Machines) is another machine learning method used in authorship attribution studies. Joachims 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 [7]. Kucukyilmaz et al. [8] use machine learning approaches including k-nearest neighbor (k-NN), SVM, and Naïve Bayes (NB) to determine authors of chat participants by analyzing their online messaging texts. Yu [9] focuses on text classification methods in literary studies and uses NB, and SVM classifiers. In her work, the effects of common and function words are investigated.

Table 1 Ottoman literary texts used in this study

Text (no. of poems)	Century	Life span	No. of tokens	No. of types
Mihrî Hatun’s divan (245)	15th	1,460–1,512	34,735	9,188
Sinan Şeyhî’s divan (221)	15th	1371?–1431	27,743	10,784
Hayalî Bey’s divan (619)	16th	1,500–1,557	54,338	15,727
Revânî’s divan (141)	16th	1,475–1,524	24,881	8,315
Nef’î’s divan (224)	17th	1,572–1,635	51,075	14,492
Neşatî’s divan (186)	17th	?–1,674	23,799	7,984
Osmanzâde Tâ’ib’s divan (189)	18th	1,660–1,724	19,610	8,772
Şeyh Gâlip’s divan (580)	18th	1,757–1,799	59,301	18,506
Şânîzâde’s Atâullah’s divan (125)	19th	1771–1826	8,265	4,409
Yenişehirli Avnî’s divan (425)	19th	1826–1884	54,927	18,785
Total	–	–	358,674	62,609

To the best of our knowledge there is no previous categorization study on the Ottoman language; however, there are studies on contemporary Turkish. Can and Patton [10] 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 they analyze the *Ince Memed* tetralogy of Yaşar Kemal [11].

3 Corpus and Experimental Design

In this study, we focus on Ottoman literary texts of ten poets and five consecutive centuries. Table 1 gives information about these texts. 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 poet. 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 poets from five different centuries. The works of the picked poets given in Table 1 acquire almost all characteristics of the Ottoman lyric poetry [12]. 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 sixteenth century for a relatively short period of time).

Each document is split into blocks with *k* number of words, where *k* is taken as 200–2,000 with 200-word increments. 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 [13].

Can and Patton [10] 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. 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. Accordingly, the style markers used in the study are: Most

Frequent Words (MFW), Token Lengths (TOL), Two-word Collocations -two consecutive words- (TWC), and Type Lengths (TYL). In our study, a token is a word and a continuous string of letters. Besides, a word that involves a dash is counted as one token. Only word of length one is '*o*' (the third person and singular pronoun). Type is defined as a distinct word. For example, the following line from Yenişehirli Avnî: '*Yâhû ne kâtib ol ne mühendis ne veznedâr*' contains eight tokens and six types.

We employ two machine learning-based classifiers: Naïve Bayes (NB)- a generative classifier and Support Vector Machines (SVM)- a discriminative classifier [14, 15]. 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 [9] indicates that SVM is among the best text classifiers. In the same work it is also indicated that NB is a simple but effective machine learning method and often used as a baseline.

In this study we employ the model used in [16] for NB. In SVM we employ two different kernel functions; polynomial (poly or p), and radial-basis-function (rbf) kernels. We refer to these methods as SVM-poly (or SVM-p) and SVM-rbf, respectively. These choices are motivated by the successful results obtained by them in [17]. 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. In the experiments with SVM for the polynomial kernel we run tests when the degree is set to 1, 2, 3, 4 and 5. For the radial-basis-function kernel, we set γ (width of the kernel) to 0.6, 0.8, 1.0, 1.2, and 1.4. Similar settings for SVM are used in [17] for text classification and successful results are obtained.

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 [18] for the levels of each factor (an abridged presentation is provided in the next section).

4 Experimental Results

4.1 Classification by Poet

In Table 2, 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

Table 2 Poet classification accuracies of MFW, TOL, TWC, TYL with NB, SVM-poly, and SVM-rbf for different block sizes

	MFW			TOL			TWC			TYL		
Block Size	NB	SVM-p deg = 1	SVM-rbf $\gamma = 1.2$	NB	SVM-p deg = 5	SVM-rbf $\gamma = 1.2$	NB	SVM-p deg = 1	SVM-rbf $\gamma = 1.0$	NB	SVM-p deg = 5	SVM-rbf $\gamma = 1.2$
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	36.1	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
1,000	85.47	91.08	91.42	44.77	42.49	43.55	61.96	61.22	61.50	50.00	45.35	48.23
1,200	86.55	91.32	91.32	51.32	49.16	49.80	63.72	65.77	64.73	56.93	48.29	47.82
1,400	**91.66**	91.28	91.12	51.36	47.02	48.46	64.45	69.21	69.69	59.55	50.24	48.56
1,600	89.64	91.22	91.12	50.50	48.49	48.64	68.30	68.53	69.90	55.08	49.40	46.65
1,800	88.57	92.51	92.51	54.52	**55.38**	**56.88**	69.96	69.42	68.10	**64.22**	**56.76**	**56.32**
2,000	87.05	**92.80**	**92.80**	**57.78**	52.40	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

The parameters, polynomial degree (deg) for SVM-poly (SVM-p) and γ for SVM-rbf that yield the listed results, are also provided these parameters that provide the best performances of TOL, TWC, and TYL are also given.

For MWF for all block sizes SVM-poly and -rbf provide better results than those of NB. Both versions of SVM have similar results. For TOL for almost all block sizes NB provides slightly better results than those of 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.

Statistical Analysis We do the multiple comparisons of the style markers and machine learning algorithms in poet categorization for $\rho < 0.05$ using Scheffe's method. According to comparisons, TOL and TYL are not significantly different from each other; whereas, other pairs of style markers are significantly different from each other. Considering the machine learning algorithms, the SVM classifiers with different kernels are not significantly different from each other, but they are significantly different from the NB classifier.

4.2 Classification by Time Period

In addition to classifications of texts by poet, in this study we also study classifications of texts by time period. In the corpus, there are ten *divans* from 15th to 19th centuries (two *divans* per century). In the classification of texts by time period, 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.

Statistical Analysis As in the poet classification section, we do the multiple comparisons of the style markers and machine learning algorithms in period categorization for $\rho < 0.05$ using Scheffe's method (they are obtained by using the results as in Table 2). According to comparisons, TOL and TYL are not significantly different; whereas, other pairs of style markers are significantly different from each other. Moreover, considering the machine learning algorithms, they are significantly different from each other for combinations of all pairs.

5 Conclusion

We present the first style-centered ATC study on Ottoman literary texts particularly on collected poems (divans) of ten different Ottoman poets from five different centuries. The statistical tests show that SVM and MFW yield performances that are mostly statistically significantly different from their counterparts. Based on these observations we recommend the use the SVM classifier and MFW style marker in future related studies on this language.

The availability of huge amount of text to be digitized in the Ottoman language confirms the practical importance and implications of our results. We hope that our work and results would serve as an incentive for more research using these documents.

Acknowledgments 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.

References

1. Sebastiani, F.: Machine learning in automatic text categorization. ACM Comput. Surv. **34**(1), 1–47 (October 2002)
2. Ottoman Text Archive Project. http://courses.washington.edu/otap/ (2011)
3. Başbakanlık Devlet Arşivleri, T.C.: http://www.devletarsivleri.gov.tr (2011)
4. Holmes, D.I.: Authorship attribution. Comput. Human. **28**(2), 87–106 (October 1994)
5. Merriam, T.: An experiment with the federalist papers. Comput. Human. **23**(3), 251–254 (1989)
6. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization (1998)
7. Joachims, T.: A statistical learning model of text classification for support vector machines. In: Proceedings of the 24th ACM SIGIR conference, 128–136 (2001)
8. Kucukyilmaz, T., Cambazoglu, B.B., Aykanat, C., Can, F.: Chat mining: Predicting user and message attributes in computer-mediated communication. Inf. Process. Manag. **44**(4), 1448–1466 (2008)
9. Yu, B.: An evaluation of text classification methods for literary study. Lit. Ling. Comp. **23**(3), 327–343 (2008)
10. Can, F., Patton, J.M.: Change of writing style with time. Comput. Human. **38**(1), 61–82 (2004)
11. Patton, J.M., Can, F.: A stylometric analysis of Yaşar Kemal's *İnce Memed tetralogy*. Comput. Human. **38**(4), 457–467 (2004)
12. Andrews, W.G., Black, N., Kalpakli, M.: Ottoman lyric poetry. University of Texas Press, Austin, Texas, USA (1997)
13. Forsyth, R.S., Holmes, D.I.: Feature-finding for text classification. Lit. Ling. Comput. **11**(4), 162–174 (June 1996)
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification (2nd edn.). Wiley-Interscience, New York (2000)
15. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
16. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. Lect. Notes Comput. Sci. **3689**, 174–189 (November 2005)
17. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: ECML-98, 137–142 (1998)
18. Scheffe, H.: A method for judging all contrasts in the analysis of variance. Biometrica **40**, 87–104 (1953)