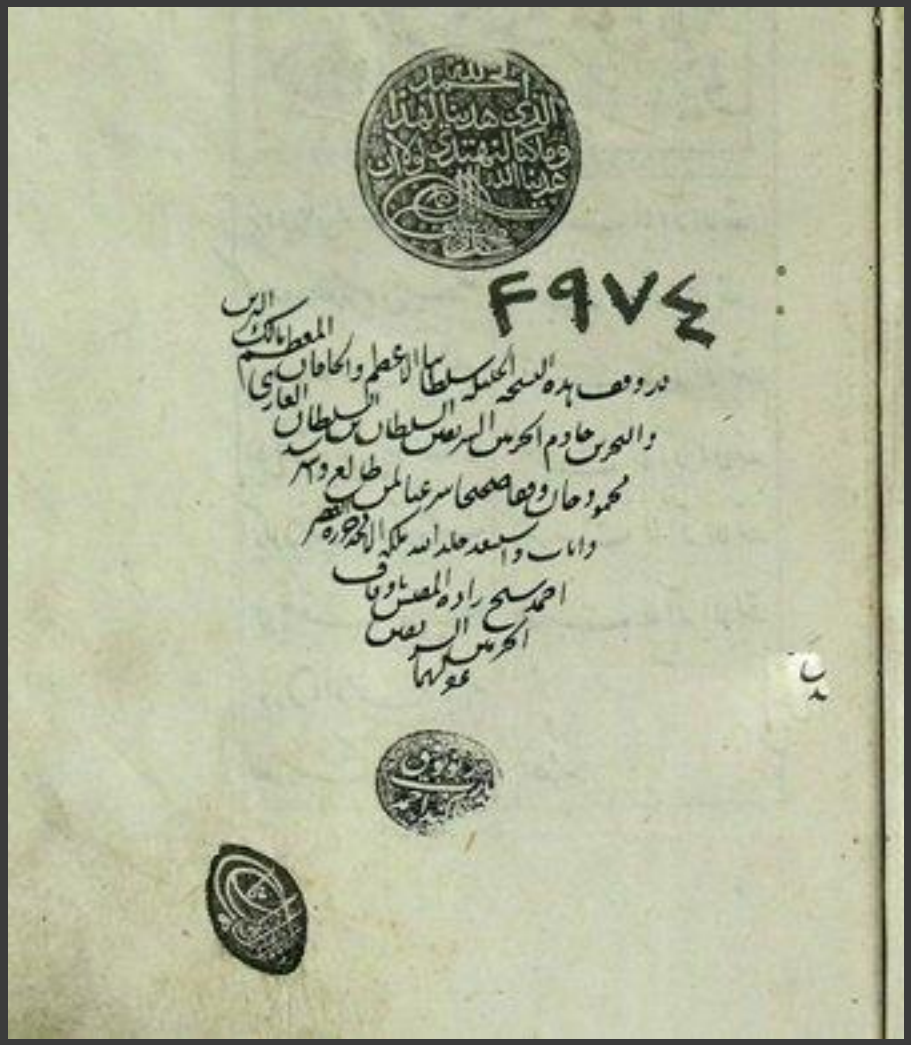
Ethem F. Can

OTTOMAN LANGUAGE STYLE ANALYSIS


F9V4

Outline

- Problem Description
- Motivation
- Related Works
- Overview of the Study
- Feature Work

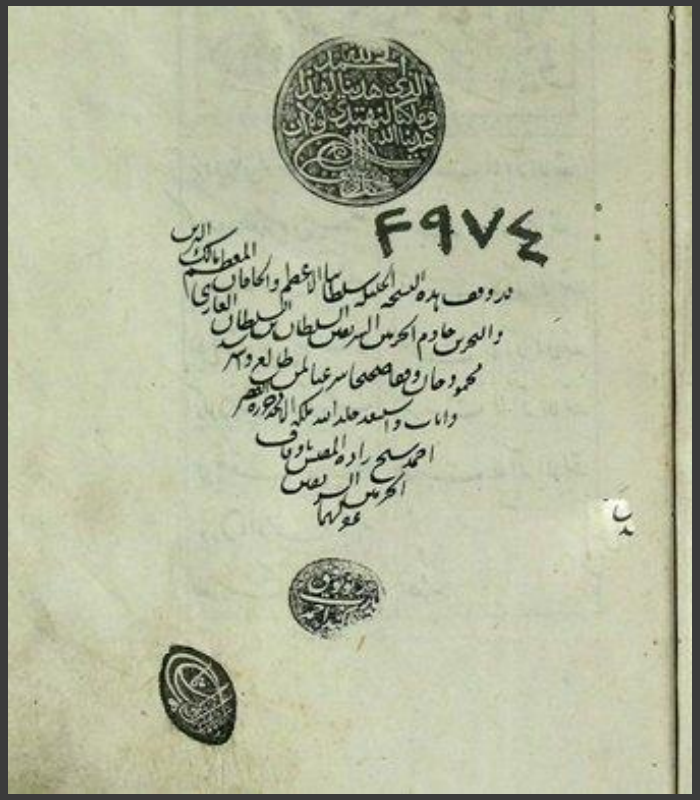
Problem Description

- Determining or Verifying the author of anonymous texts
- Automatic Text Categorization(ATC)
- Textual measurements – style markers

Motivation

- Analyzing Ottoman Language
- Determining or Verifying author's of Ottoman documents

Motivation



Motivation

- Transcripted Texts
- Obtained from Prof.Mehmet Kalpakli

Related Work

- Mendenhall, 1901
  - Used word length frequencies
  - Compute style differences

Related Word (cont'd)

- Yule, 1938
  - Introduced sentence length

Related Work (cont'd)

- Forsyth and Holmes (1996)
  - Use 5 different style markers
    - Letters,
    - Most frequent words
    - Most frequent di-grams
    - Two methods of most frequent substring

Related Word (cont'd)

- Merriam, 1989
  - Bayes' Theorem for classification

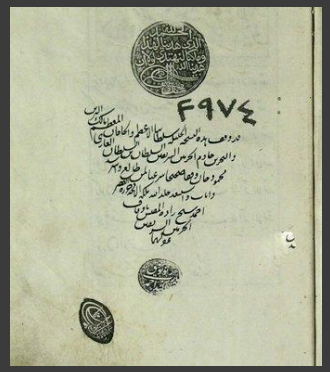
Related Word (cont'd)

- Diri and Amasyali (2003)
  - 22 different style markers, 18 authors
  - Turkish newspaper columns
  - Neural Networks
- Diri and Amasyali (2006)
  - 2-grams, 3-grams
  - Naïve Bayes, SVM, C4.5, and Random Forest
  - Increase the success rate %83 to %96

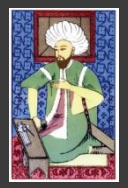
Related Word (cont'd)

- Kucukyilmaz,Cambazoglu,Aykanat,Can (2006)
  - K-NN, SVM
  - To classify gender of chat participants

Overview of the Study



beytini bedīhe naẕm itmiş kibār ü ṣıġār
gelüp aşār-ı nuḫūsetden müberrā zamānda
ile söylemişdür. Ve mezbūr Süleymān Çelebinüñ
Muḫarrik rāst geldükde temām ġazel daḫi söyler imiş.
Ṣallar düzüp aṣḥāb-ı cihādla Rūmiline geçdükde mezbūr



classification

Feature extraction

blocks

beytini
bedīhe
naẓm
itmiş
kibār ü
ṣıġār
peşend-i
bī-şümār
itmegin
Mevlīd-i
Nebī
te'līfini
kendüye
...

Dataset

- 9 authors, 9 prose
  - Serh-I Muskilat-I Baz-I Ebyat-I Urfi
  - Miratul-Muluk
  - Belagat-I Osmaniyye
  - Risale-I Garibe
  - Tabirname
  - Tezkiretulbunyan
  - Vasiyetname
  - Tazarru name-i Sinan Pasa
  - Serh-I Cezire Mesnevi

Dataset (cont'd)

- Word-Count: 148483
- Token-Count: 105315
- Type-Count: 43168
- Block-Sizes
  - 100, 200, 250, 400, 500, 750, 800, 1000, 1200, 1250, 1500, 1750, 2000

Features

- 33 Features
  - 1-14 character Token Frequency
  - 1-14 character Type Frequency
  - Average Token Length
  - Average Type Length
  - 2 word collocation
  - Token Frequency
  - Type Frequency

Features (cont'd)

- A feature vector
  - dimension of 33 + class , for each block
- %50 of the blocks used to train
- The rest for test

Classification Methods

- Naïve Bayes
  - Assumes that feature vectors from each class are normally distributed (so assumed Gaussian Mixture – one component per class)
  - Training data estimates mean vectors and co-variation matrices
  - Predict by using mean and co-variance matrices

Classification Methods (cont'd)

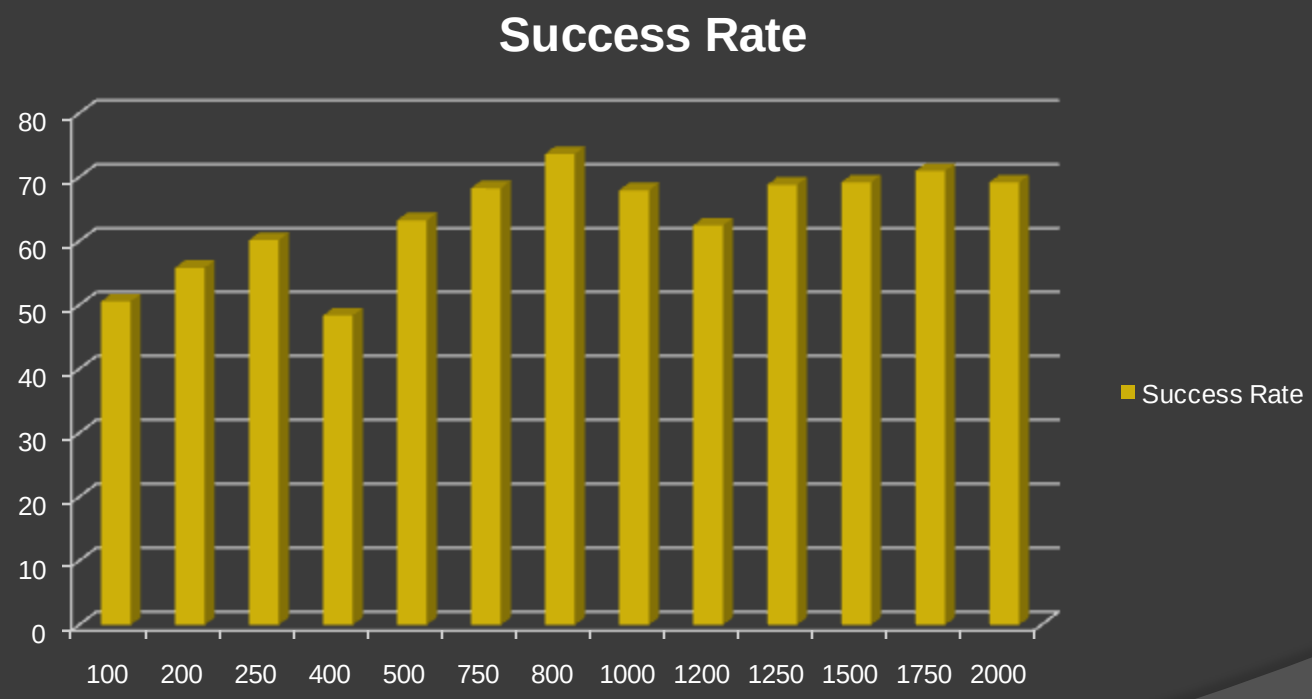
- Random Trees
  - Takes the input vector, classifies it with every tree in the forest
  - Outputs the class label that gets the majority of votes

Neural Networks

- Multi-layer perceptrons (MLP)
  - Input layer, hidden layers and output layer
  - Compute weights, in the learning phase
  - Predicts by using parameters computed

Results

- Naïve Bayes



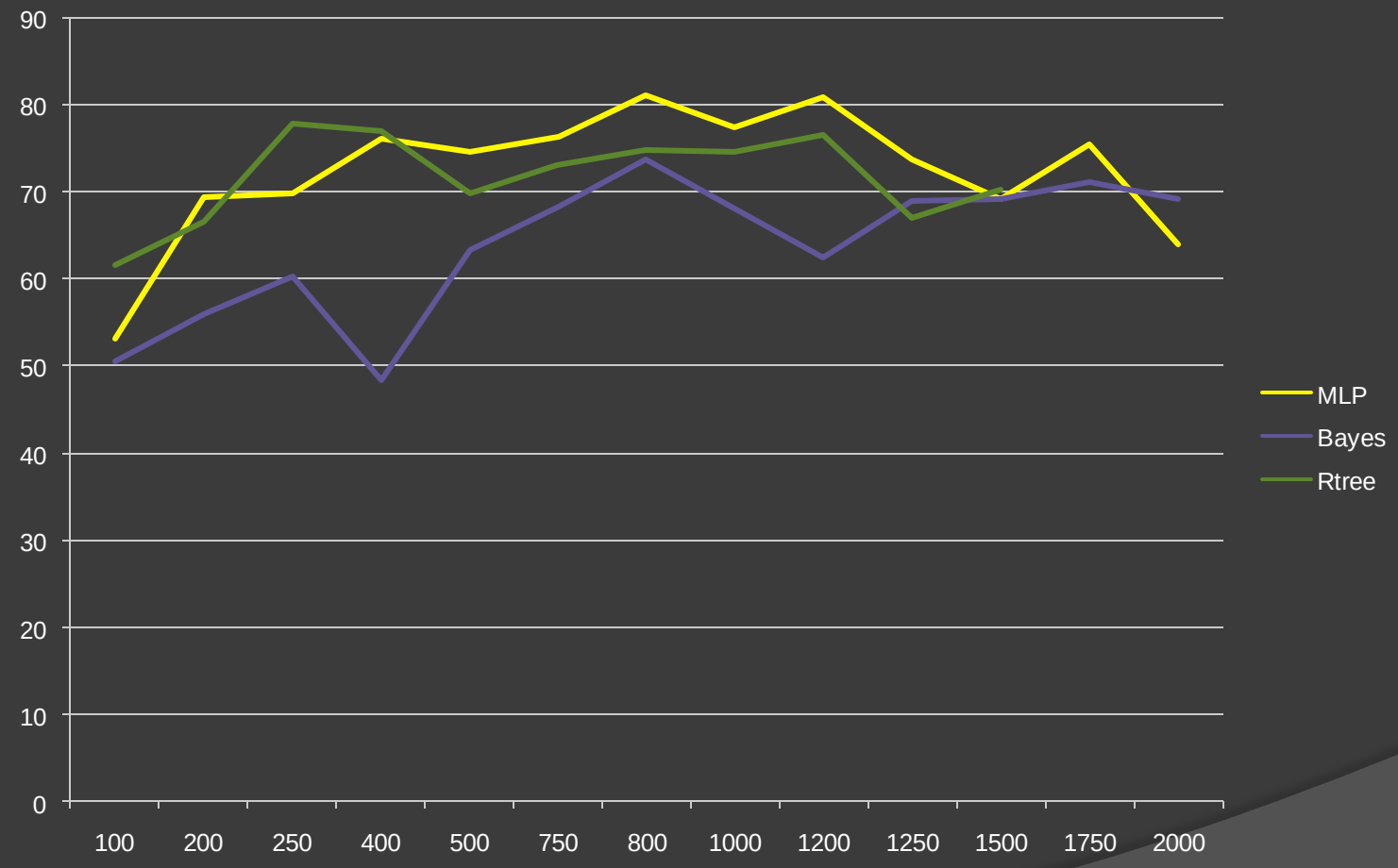
Results (cont'd)

- Random Trees

Results (cont'd)

- Neural Networks (MLP)

Results (cont'd)



Feature Work

- Re-evaluate used classifiers with different parameters
- New Classifiers ; k-NN, SVM
- Attribute Selection & Instance Selection
- New Features
  - Vocabulary Richness, Character Collocation…

References

- Amasyali, M. F., Diri, B. Automatic text categorization in terms of author, genre and gender. Natural Language Processing and Information Systems, Proceedings Lecture Notes in Computer Science 3999:221-226 2006
- K. Burns, “Naïve Bayesian inference in disputed authorship: A case study of cognitive errors and new system for decision support”, Information Sciences, vol 176, no.11, 2006, pp 1570-1589
- Can, F., Patton, J. M. “Change of writing style with time.”, Computer and the Humanities. Vol.38, No.1 (2004), pp.61-82
- Diri B., Amasyali MF, (2003), “Automatic Authorship Attribution in Turkish Language”, ICANN/ICONIP 2003
- Grieve. J., “Quantativ Authorship Attribution: An Evaluation of Techniques”, Literary and Linguistic Computing, Vol.22, No.3, 2007

- Thanks
- Questions.