# Author Attribution of Turkish Texts by Feature Mining

Filiz Türkoğlu[1], Banu Diri[1], M.Fatih Amasyalı[1]

[1] Yıldız Technical University, Computer Engineering,
34349 İstanbul, Turkey
{filizturkoglu, banu, mfatih}@ce.yildiz.edu.tr

**Abstract.** The aim of this study is to identify the author of an unauthorized document. Ten different feature vectors are obtained from authorship attributes, n-grams and various combinations of these feature vectors that are extracted from documents, which the authors are intended to be identified. Comparative performance of every feature vector is analyzed by applying Naïve Bayes, SVM, k-NN, RF and MLP classification methods. The most successful classifiers are MLP and SVM. In document classification process, it is observed that n-grams give higher accuracy rates than authorship attributes. Nevertheless, using n-gram and authorship attributes together, gives better results than when each is used alone.

**Keywords:** Author attribution, n-grams, Text classification, Feature extraction, Turkish documents.

## 1 Introduction

The goal of text categorization is the classification of documents into a fixed number of predefined categories. One of the problems in text categorization is the authorship attribution, which is used to determine the author of a text when it is not clear who wrote it. It can be used in occasions where two people claim to be the author of same manuscript or on the contrary where no one is willing to accept the authorship of a document. It is not difficult for anyone to take somebody else's work and to publish it under his or her own name. In such cases the authorship attribution methods gain importance to determine the person deserving recognition for the work [1].

Early researchers in authorship attribution used a variety of statistical methods to identify characteristics which remain approximately invariant within the works of a given author but which tend to vary from author to author [2], [3]. Mosteller and Wallace, working on Federalist Papers, used a set of function words, that is, words that are context-independent. They suggested that a small number of the most frequent words in a language (function words) could usefully serve as indicators of authorial style [4]. Yule [5] used complexity-based features such as average sentence length, average word length, type/token ratio and so forth. Recent technical advances in automated parsing and part-of-speech (POS) tagging have facilitated the use of syntactic features such as POS n-grams [3], [6]. Peng worked on a method for authorship attribution in which they modeled each author by a vector of the most frequent n-grams in the text [7]. Fung, used Support Vector Machine classifier to determine the authors of Federalist papers [8]. Kukushkina used Markov Chains for the sequence elements of a natural language text [9]. Model depends on the idea that an element of a text could be a letter or a grammatical class of a word. Stamatatos demonstrated a Multiple Regression classifier using a varied combination of syntactic style markers [10]. Stamatatos adapted a set of style markers from the analysis results of the text performed by an already existing natural language processing tool [6]. Fürnkranz described an algorithm for efficient generation and frequency-based pruning of n-gram features [11]. Cavnar described an n-gram based approach to text categorization is tolerant of textual errors [12]. In our previous work, we used only n-grams to determine the author of text, genre of the text, and the gender of the author. The success was obtained as 83%, 93%, and 96%, respectively [13].

In this paper, we focus an author attribution of Turkish texts by extracting various feature vectors and applying different classifiers. We studied the comparative performance of classifier algorithms using the Naive Bayes, Support Vector Machine, Random Forest, Multilayer Perceptron, and k-Nearest Neighbour. The effectiveness of the methods used is assessed using 10-fold cross validation. The remainder of the paper is organized as follows: In section 2, a brief description of author attribution and variant feature vectors applied to this task is mentioned. In section 3, experimental results over different datasets are shown in table and interpreted. Finally, we summarize our conclusions in section 4.

## 2 Authorship Attribution

The statistical analysis of style, stylometry, is based on the assumption that every author's style has certain features being inaccessible to conscious manipulation. Stylometry should identify features, which are expressive enough to discriminate an author from other writers [14].

In this work, we formed feature vectors from several categories of statistics that have been used previously in authorship attribution, stylistic analysis in order to compare the efficacy of each. We examined features in five main categories, which are statistical, vocabulary richness, grammatical, lexical and n-grams model. Then, we have obtained five different feature vectors from the mentioned categories. We have created five feature subsets by using the feature selector to reduce the dimension of the obtained vectors. Here, we briefly explained each of these feature vectors.

### 2.1 Corpus

In this work, using the same corpus in the study of Diri, 630 documents written by a single author are obtained from 35 texts per 18 different authors that are writing on different subjects like sport, popular interest and economics. All documents were originally downloaded from a Turkish daily newspaper www.hurriyet.com.tr and www.vatanim.com.tr [15]. In order to determine the authorship attribution performance when employing the homogeneous and heterogeneous documents, and different dataset sizes, this corpus is divided into 3 parts: Dataset I, Dataset II, Dataset III.

## 2.2 General Feature Vector (gfv)

**Statistical features:** Early stylometric studies introduced the idea of counting features in a text and applied this to word lengths and sentence lengths [14]. Other token-level features are word count, sentence count, character per word count, punctuation counts, etc. We used a set of 10 style markers.

**Vocabulary richness features*:*** Many studies found different statistics to determine the richness of an author's vocabulary. These features points out an author's creativity. We applied three different features, which are type/token ratio, hapax legomena and hapax dislegomena. Type/token ratio is presented as V/N where V is the size of the vocabulary of the text, and N is the number of tokens of the text. Hapax legomena refers to words that only once occur in a given body of text. The most frequent words are expected all texts and rarely used ones provide greatest information. Hapax legomena estimates the probability that an author will produce a new rewrite rule that she/he has not yet used before. Hapax dislegomena count is defined as the number of twice-occuring words.

**Grammatical features:** When extracting these features, the developed Turkish Word Database (TWD) that has been based on the dictionary of Turkish Language Society with 35,000 words, is used. In the Turkish language possible grammatical word types are adjective, noun, verb, particle, pronoun, adverb, conjunction, and exclamation type. The system automatically detects the type of the word by implementing the Turkish Grammatical Rules module on the sentence [15].

**Function word features:** Function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with order words within a sentence. Function words may be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles. The first research on attribution using function words was that of Burrows [16]. There is no study on function words over attributing authors of Turkish texts. Thus, we formed a list of function words from TWD. When we calculate the frequencies of these words, number of function words at least once-occurring in whole corpus is 620 function words. We constructed a feature vector from the frequencies of these 620 function words in Turkish. Some of these function words are ‘neden-*why*’, ‘ayrıca-*furthermore*’, ‘belki-*maybe*, ‘daima-*always*'' etc.

We joined statistical, richness, grammatical features and function words, and obtained a 641-dimensioned feature vector. It is called *gfv*.

## 2.3 N-gram Model

An n-gram is an n-character fragment of a longer string. In literature, the n-gram term is included the notion of any co-occurring set of characters in a string [12]. We have handled the text as a whole and we have extracted the bi-grams and the tri-grams.

While forming the bi-grams and the tri-grams of the corpus, the numbers of occurrences of each feature are counted. At the end of this process we have observed that the number of different bi-grams and tri-grams are too much. In order to avoid the combinatorial explosion in the feature vectors, which consist of bi-grams and tri-grams, we used a threshold value (greater than 75) to reduce the number of features.

Infrequent features are removed from the feature vectors. The dimensions of the bi-gram *bgfv*, and tri-gram *tgfv* feature vectors are 470 and 1037 respectively.

After that, we combined *bgfv* and *tgfv* producing a new feature vector. The dimension of this vector, *btgfv*, is 1507.

Finally, we put together *gfv* and *btgfv* and obtained a 2148-dimensioned new vector, which is called *gbtgfv*.

### 2.4 Feature Selection

Features used to describe samples may not be necessarily all relevant and beneficial for the inductive learning and reduce the quality of induced model. A high number of features may slow down the process while giving similar results as obtained with much smaller feature subset. To learn the effect of high-dimensioned feature set over success ratio, we used CfsSubsetEval function, which is implemented in WEKA package (available at www.cs.waikato.ac.nz/ml/weka).

We reduced features of general feature vector, *gfv* and obtained a new vector, *rgfv*. Its dimension is 24 for Dataset I: 9 function words, 9 statistical, 4 grammatical and 2 richness features; 17 for Dataset II: 5 function words, 7 statistical, 5 grammatical features; 40 for Dataset III: 25 function words, 9 statistical, 4 grammatical, and 2 richness features. Same process was applied for Bi-gram feature vector, *bgfv*, and was formed *rbgfv*, which has 25 features for Dataset I, 20 for Dataset II and 63 for Dataset III. We decreased dimension of Tri-gram feature vector, *tgfv*, and obtained *rtgfv*. It has left 60 most distinguishing features for Dataset I, 33 for Dataset II and 101 for Dataset III. When features decreased from *btgfv* (combination of bi-gram, tri-gram features), *rbtgfv* is obtained. For Dataset I, vector has 61 features, for Dataset II it has 26 features, and for Dataset III it has 101 features. We decreased dimension of *gbtgfv*, and obtained *rgbtgfv*. It has left 69 most distinguishing features for Dataset I, 30 for Dataset II and 103 for Dataset III. All used feature vectors are shown at Table 1.

**Table 1.** General feature vector.

| Vector name | Explanation (Num. of features at Dataset I-II-III) |
|---|---|
| gfv | General Feature Vector (641) |
| rgfv | Reduced General Feature Vector (24-17-40) |
| bgfv | Bi-gram Feature Vector (470) |
| rbgfv | Reduced Bi-gram Feature Vector (25-20-63) |
| tgfv | Tri-gram Feature Vector (1037) |
| rtgfv | Reduced Tri-gram Feature Vector (60-33-101) |
| btgfv | Combined Bi-gram and Tri-gram Feature Vector (1507) |
| rbtgfv | Reduced Combined Bi-gram and Tri-gram Feature Vector (61-26-101) |
| gbtgfv | Combined gfv and btgfv (2148) |
| rgbtgfv | Reduced Combined gfv and btgfv (69-30-103) |

## 3 Experimental Results

In this work, we used WEKA's 5 different classification algorithms, which are Naive Bayes, Support Vector Machine, Random Forest, k-Nearest Neighbor, and Multilayer Perceptron in text classification. All experiments are done with WEKA's default parameters.

In our experiments, we showed whether the modeling of Turkish texts with statistical, richness, grammatical features, function words and n-gram is a successful approach or not for determining the author of documents. We ran 10-fold cross-validation experiments on our all datasets using various combinations of feature types and five classification algorithms.

**Dataset I:** This set consists of 630 singly-authored documents written by 18 different authors, with 35 different texts written on different topics. We applied our 5 different classifiers and get the accuracy rates (%) shown in Table 2. The best performance in Dataset I, 92.5%, is obtained from *gbtgfv* with SVM. On our corpus, NB, RF and k-NN give better results when the feature selection process is applied, while SVM and MLP give weaker.

In Dataset I, as we calculate average values, *rgbtgfv* gives highest accuracy rate overall vectors and the most successful classifier is SVM. NB, RF and k-NN are achieved their best performances with *rgbtgfv* feature vector. Their success ratios are 85.6%, 82.0% and 79.0% respectively.

**Table 2.** Classification Results of Dataset I.

| | gfv | bgfv | tgfv | btgfv | gbtgfv | rgfv | rbgfv | rtgfv | rbtgfv | rgbtgfv | **avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NB | 66.5 | 69.4 | 70.2 | 78.1 | 78.1 | 75.4 | 78.4 | 80.2 | 85.1 | 85.6 | 76.7 |
| SVM | 80.0 | 88.1 | 91.6 | 92.2 | **92.5** | 70.3 | 73.3 | 83.8 | 88.1 | 88.4 | **84.8** |
| RF | 48.0 | 51.6 | 42.5 | 46.0 | 45.7 | 69.5 | 78.3 | 69.0 | 77.6 | 82.0 | 61.0 |
| k-NN | 23.6 | 64.1 | 51.7 | 60.5 | 53.7 | 66.6 | 71.4 | 68.9 | 78.4 | 79.0 | 61.8 |
| MLP | 8.5 | 89.0 | 89.2 | 92.4 | 90.3 | 72.2 | 77.8 | 81.3 | 88.4 | 86.3 | 77.5 |
| **avg** | 45.3 | 72.4 | 69.0 | 73.8 | 72.1 | 70.8 | 75.8 | 76.6 | 83.5 | **84.3** | 72.4 |

**Dataset II:** This set consists of 315 singly-authored documents written by 9 different authors, with 35 different texts written on the same topic. We applied our 5 different classifiers and get the accuracy rates % shown in Table 3.

The best performance in Dataset II, 95.4%, is achieved from *gbtgfv* with MLP. NB, RF and k-NN give better results when the feature selection process is applied, while SVM and MLP give weaker. When we calculate average values in Dataset II, *rbtgfv* gives highest accuracy rate overall vectors and the most successful classifier is MLP. NB, SVM, RF and k-NN are achieved their best performances with *rbgfv*, *gbtgfv*, *rgbtgfv* and *rbtgfv* feature vectors respectively. Their success ratios are 90.8%, 94.6%, 91.7% and 89.5% respectively.

**Dataset III:** This set consists of 315 singly-authored documents written by 9 different authors, with 35 different texts written on different topics. We applied our 5 different classifiers and get the accuracy rates % shown in Table 4. Maximum success ratio in Dataset III, 96.9%, is obtained from *btgfv* with MLP. Decreasing the number of classes in dataset and selecting documents of authors that are writing on different topics, gave us the strongest performance overall datasets, obtained on an easier data set. *rgbtgfv* is the most distinguishing feature vector and Multilayer Perceptron is more successful than others, when we calculate average values for Dataset III. NB, RF and k-NN are achieved their best performances with *rgbtgfv* feature vector. Their success ratios are 91.1%, 87.9% and 90.5% respectively.

**Table 3.** Classification Results of Dataset II.

| | gfv | bgfv | tgfv | btgfv | gbtgfv | rgfv | rbgfv | rtgfv | rbtgfv | rgbtgfv | **avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NB | 65.7 | 77.1 | 71.1 | 75.9 | 76.5 | 84.1 | 90.8 | 85.4 | 88.9 | 89.8 | 80.5 |
| SVM | 83.8 | 92.1 | 91.7 | 93.3 | 94.6 | 79.7 | 89.5 | 87.0 | 90.2 | 91.1 | 89.3 |
| RF | 56.2 | 67.3 | 50.8 | 61.3 | 61.6 | 78.4 | 89.5 | 80.6 | 89.8 | 91.7 | 72.7 |
| k-NN | 34.2 | 66.7 | 50.8 | 58.7 | 55.9 | 73.3 | 86.0 | 79.0 | 89.5 | 79.0 | 67.3 |
| MLP | 85.0 | 91.4 | 91.0 | 95.2 | **95.4** | 81.0 | 89.2 | 86.3 | 92.4 | 92.4 | **89.9** |
| **avg** | 65.0 | 78.9 | 71.1 | 76.9 | 76.8 | 79.3 | 89.0 | 83.7 | **90.2** | 88.8 | 80.0 |

**Table 4.** Classification Results of Dataset III.

| | gfv | bgfv | tgfv | btgfv | gbtgfv | rgfv | rbgfv | rtgfv | rbtgfv | rgbtgfv | **avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NB | 78.4 | 79.7 | 81.0 | 86.0 | 87.3 | 84.1 | 87.9 | 90.2 | 89.5 | 91.1 | 85.5 |
| SVM | 87.0 | 94.6 | 95.2 | 96.8 | 96.8 | 79.7 | 90.5 | 95.2 | 94.3 | 96.5 | 92.7 |
| RF | 65.0 | 68.3 | 63.5 | 64.4 | 67.0 | 78.4 | 81.2 | 82.5 | 85.4 | 87.9 | 74.4 |
| k-NN | 35.2 | 70.8 | 53.2 | 58.7 | 54.3 | 73.3 | 82.5 | 86.7 | 84.4 | 90.5 | 69.0 |
| MLP | 89.2 | 95.6 | 95.2 | **96.9** | 94.5 | 81.0 | 90.8 | 95.2 | 94.9 | 94.3 | **92.8** |
| **avg** | 71.0 | 81.8 | 77.6 | 80.6 | 80.0 | 79.3 | 86.6 | 90.0 | 89.7 | **92.1** | 82.9 |

## 4 Conclusion

In this study, we have investigated the authorship attribution of Turkish texts using various feature vectors on different datasets. Dataset I consists of 630 singly-authored documents written by 18 different authors, with 35 different texts written on different topics. Dataset II is formed from 315 singly-authored documents written by 9 different authors, with 35 different texts written on the same topic, while Dataset III consists of 315 singly-authored documents written by 9 different authors, with 35 different texts written on different topics. Function words, lexical, statistical, grammatical, and n-gram features are automatically extracted from documents and formed different feature vectors. We applied 5 classification algorithms that are Naive Bayes, Random Forest, Multilayer Perceptron, Support Vector Machine and k-Nearest Neighbour.

First, we experiment these feature vectors with each classification algorithm on each dataset. Applying feature selection method, we obtained decreased dimension of these vectors and remained with more distinguishing features. We also classified our documents using these features. The average performance of Dataset I, II and III are 72.4%, 80.0% and 82.9%. We observed a reduction in classification performance when increasing class count, as Dataset II and III gave better results than Dataset I.

We can say that there is a close relationship between performance and number of class.

To determine the capability of identifying authorship for heterogenous documents, we compare the results of Dataset II and Dataset III. When employing the different type of documents, success increases as in Dataset III as we expected. Evaluated average accuracy results in each dataset to determine most successful classifier, feature vector and classification is shown Table 5.

**Table 5.** Comparing classification problems.

| | Most Successful | | | Avg.Succ. Ratio |
|---|---|---|---|---|
| | Classifier | Feature Vector | Classification | |
| Dataset I | SVM 84.8% | *rgbtgfv* 84.3% | SVM - *gbtgfv* 92.5% | 72.4% |
| Dataset II | MLP 89.9% | *rbtgfv* 90.2% | MLP - *gbtgfv* 95.4% | 80.0% |
| Dataset III | MLP 92.8% | *rgbtgfv* 92.1% | MLP - *btgfv* 96.9% | 82.9% |
| Avg.of 3 Datasets | SVM 88.9% | *rgbtgfv* 88.4% | - | - |

In general, MLP and SVM give good performance; while *rgbtgfv* and *rbtgfv* are tend to be more successful than other vectors. On our corpus, NB, RF and k-NN give better results when the feature selection process is applied, while SVM and MLP give weaker. SVM is the best classifier while *rgbtgfv* is the most distinguishing feature vector according to the average result of 3 datasets. The best performance in Dataset I, 92.5%, is obtained from *gbtgfv* with SVM algorithm. The best result in Dataset II, 95.4%, is achieved from *gbtgfv* with MLP while maximum success ratio in Dataset III, 96.9%, is obtained from *btgfv* with MLP.

As a result, in authorship attribution of Turkish documents, it is observed that n-grams are more successful than authorship attributes. However, combination of n-grams and authorship attributes performs better results than using them separately. We can say that, this work is the most successful and extensive study made for authorship attribution of Turkish documents.

## References

1. Geritsen, C.M.: Authorship Attribution Using Lexical Attraction, Master Thesis Department of Electrical Engineering and Computer Science, MIT (2003)
2. Holmes, D.: The Evolution of Stylometry in Humanities Scholarship Literary and Linguistic Computing, (1998) 13, 3, 111-117
3. Koppel, M., Schler, J.: Exploiting Stylistic Idiosyncraises for Authorship Attribution, IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico (2003)
4. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist Reading, MA:Addison-Wesley (1964)
5. Yule, G.U.: On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship, Biometrica (1938) 30, 363-390
6. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-Based Authorship Attribution without Lexical Measures, Computers and the Humanities (2001) 193-214

7. Peng, F., Schuurmans, D., Keselj, V., Wang, S.: Language Independent Authorship Attribution using Character Level Language Models, 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest (2003) 267-274
8. Fung, G., Mangasarian, O.:The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization, Proceedings of the 2003 Conference of Diversity in Computing, Atlanta, Georgia, USA (2003) 42-46
9. Kukushkina, O.V., Polikarpov, A.A., Khemelev, D.V.: Using literal and grammatical statistics for authorship attribution, In Problemy Peredachi Informatsii, Volume 37(2) (2000)
10. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Authorship Attribution, Nineth Conf. European Chap. Assoc. Computational Linguistics, Bergen, Norway (1999)
11. Fürnkranz, J.: A Study using n-gram Features for Text Categorization, Austrian Research Institute for Artifical Intelligence (1998)
12. Cavnar, W.B.: Using an n-gram-based Document Representation with a Vector Processing Retrieval Model. In Proceedings of the Third Text Retrieval Conference(TREC-3) (1994)
13. Amasyalı, M.F., Diri, B.: Automatic Turkish Text Categorization in Terms of Author, Genre and Gender, NLDB, Klagenfurt, Austria (2006) 221-226
14. Diederich, J., Kindermann, J., Leopold, E., Paass G.: Authorship attribution with Support Vector Machines, Poster presented at The Learning Workshop (2000)
15. Diri, B., Amasyalı, M.F.: Automatic Author Detection for Turkish Texts, Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP) (2003) 138-141
16. Burrows, J.: Word patterns and story shapes: The statistical analysis of narrative style, Literary and Linguist Comput (1987) 2:61-70