# Taking Advantage of Turkish Characteristic Features to Tackle with Authorship Attribution Problems for Turkish

Neslihan Şirin Saygılı, Tassadit Amghar, Bernard Levrat
Computer Science Laboratory
University of Angers
Angers, France
e-mail:
{neslihansirin.saygili,amghar,levrat}@{etud,info}.univ-angers.fr

Tankut Acarman
Computer Engineering Department
Galatasaray University
Istanbul, Turkey
e-mail: tacarman@gsu.edu.tr

***Abstract*—The rapid increase in the number of the electronic and online texts, such as electronic mails, online newspapers and magazines, blog posts and online forum messages has also accelerated the studies carried out on authorship attribution. Although the studies are not as abundant as in English language, there have been considerable studies on author identification in Turkish in the last fifteen years. This paper includes two parts; first part is a quick review of Turkish authorship attribution studies. The review is focused on the stylometric features that enable authors to be distinguished one from another. In the second part, we analyze the main characteristics of the Turkish language and depict our first experiments on Turkish corpora. In these lasts, we experiment different kind of n-gram and word structure, taking advantages of Turkish characteristic features by the frequent usage of gerunds in Turkish language, and use Support Vector Machines as learning algorithm.**

***Keywords-authorship attribution; Turkish language; stylometry; n-gram; gerunds; Support Vector Machines.***

## I. INTRODUCTION

Authorship attribution studies based on statistical methods have begun in the late 19th century where Mosteller and Wallace's impressive 'Federalist Paper' study [1] renewed interest on this issue. The aim behind automatic authorship attribution task is the identification of the author of a text among several ones using for that different characteristics in which stylistic features predominate, depending on the methodology used for achieving the task.

Over the past two decades three research domains have played an important role in development of authorship attribution methods: information retrieval, machine learning and natural language processing. To consider roughly the contributions of each of these domains we can say that information retrieval provides efficient methods for modeling and processing huge number of documents, machine learning furnishes ways to extracts the most suitable set of features characterizing a great volume of data to be used for a specific task, and natural language processing gives models suited to cope with natural language data. Furthermore, the remarkable increase of available electronic text amount (e.g., emails, blogs, online forum messages, source code, etc.) greatly expanded the range of applications of authorship attribution among which cites criminal law, intelligence and computer forensic [2].

Quantitative authorship detection earlier studies began in the 18th century with works on, plays supposed to be authored by William Shakespeare [3] [4]. Two periods could be distinguished in Authorship attribution methodologies:

The first one is dominated by linguistic, stylometry and computer-assisted studies. Computer-assisted means computer programs only calculate some metrics and human decides the final authorship attribution result. T. Yule proposed a metric called vocabulary richness which points out the probability of any randomly selected pair of words will be identical [5]. Ellegard proposed distinctiveness ratio that indicates how far the author is from the average usage of a word [6]. Later, in 1964, Mosteller and Wallace's work was based on Bayesian statistical analysis [1].

Until 1990, the authorship attribution methodologies were computer-assisted instead of computer-based. Computer-based means computer programs both calculate metrics and decide the final authorship attribution result. After developments of information retrieval, machine learning and natural language processing authorship attribution proceeded to second phase. The second phase consists of computer-based studies rather than computer-assisted studies. Increment of available electronic texts reveals the potential of authorship attribution usages in various applications such as criminal law, intelligence, civil law, computer forensic, and literary research [2]. In addition to this, from machine learning perspective, authorship attribution is regarded as a multiclass single label text categorization task [7].

Before this study, we made a short survey of Turkish authorship attribution studies from the point of stylometry. The main goal in this paper is to enrich stylometric features set used in the works described in the review, and to use them in our first experimental approaches. Regarding to these goals the paper follows the following plan: Section II

 ISBN: 978-1-61208-513-5 

tries to characterize Turkish language in its major characteristics which, distinguish it from other languages like English or French, Section III is a quick review of stylometric features used in authorship attribution, Section IV depicts the experimental processing. Section V is a conclusion where we give some lights on the continuation of this ongoing research.

## II. TURKISH LANGUAGE

Turkish belongs to the Turkic family of Altaic languages and as such deeply differs from most natural languages on which natural language processing researcher mostly bears on. This is the reason why it is interesting to analyze its main characteristics in the aim of adapting generally used methodology to its idiosyncrasies. First of all, Turkish is an agglutinative language, where functions and derivative of words are mainly indicated by suffixes added to the end of the words where languages like English generally mark the function by the position of the words in the sentences and have comparatively less derivative. To give an idea of this, in corpora words occurrences are formed by productive affixations of multiple suffixes from about 30 K root words.

Oflazer gave wide coverage to the challenges of Turkish regarding with natural language processing in his study [8]. There are a variety of difficult features of Turkish in terms of the natural language processing such as agglutinating morphology, vowel harmony and free constituent order in syntax. Derivational morphemes are frequently used in the Turkish language. Frequent uses of derivational morphemes provide the language productivity. Practically, infinite vocabulary raises interesting issues to be considered in any natural language processing applications. There are some difficulties of Turkish language on the natural language processing applications below.

- Spelling correction: the methods using finite vocabulary are not appropriate for Turkish.
- Tag set design: finite tags set numbers of techniques are not suitable for Turkish.
- Statistical language modeling: there is high rate of unknown words for Turkish.
- Syntactic modeling: Turkish derivational morphemes complicate the modeling.
- Statistical translation: based on morphological structure translation gives better results [8].

A remarkable point is stemming. Texts are expressed as a dimensional space with a number of at least one time occurring words in the different documents. Using derived words increases the size of the dimensional space. Thus, stemming is one of the frequently used methods. Stemming has been applied successfully to many different languages. Nevertheless, this approach is less feasible to an agglutinative language, because agglutinative languages require a more detailed level of morphological analysis. Complex morphological techniques are required that remove suffixes from words according to their internal structure [9]. Another supporting idea is that, stemming in Turkish could not provide the desired result. Turkish has a complex morphological structure, for instance derived words may be incorporated into different classes as morphological and semantic.

## III. MOTIVATION AND REVIEW OF STYLOMETRIC FEATURES

The prevalence of electronic documents initiates a large number of natural language processing studies all around the world. Precisely, there is a variety of English language processing concerning authorship attribution. Unfortunately, the numbers of Turkish authorship attribution studies are less than English studies; Turkish studies have been made for the last fifteen years. Starting this point, our first step is to review Turkish authorship attribution studies. Because these kinds of studies are crucial for Turkish language, which lacks natural language processing compare with English and other commonly treated languages. One of the motivations is to obtain the more important characteristics of Turkish by analyzing these studies.

Stylometry is the application of the study of linguistic style by which a person can make a decision about another person by its writing style. It focuses on readily computable and countable language features, such as sentence length, phrase length, word length, vocabulary frequency, distribution of words of different lengths. Stylometric features can be separated into three main groups in this review; lexical, character based and syntactic features.

Firstly, lexical features can be categorized to token-based, vocabulary richness, vectors of word frequencies and word n-gram model. Token-based features are based on the number of tokens or the length of tokens. Some token-based features are average word length, average sentence length, average number of sentences, and average number of words. Vocabulary richness can be defined as attempts to quantify the diversity of the vocabulary of a text. Vectors of word frequency are described bag-of-words text representations where a text is represented as the bag of its words, each one having a frequency of occurrence disregarding grammar and even word order. N-gram is defined as an adjacent sequence of n items from a given sequence of text or speech, in which the n should be an integer greater than zero. Due to the fact that there is a huge number of lexical features and no restrictions about the field of applications explain why a large number of Turkish authorship attribution methods prefer lexical features. Among 11 studies focused on Turkish, token-based features and frequencies of words have been used six times, the word richness five times and a model of word n-grams three times.

Secondly, a variety of character level measures can be used, such as alphabetic character counts, digit character counts, upper case, lower case character counts, punctuation mark counts, etc. Reference [10] suggested that an author has similar character frequency in her/his all texts. So, this study shows that character frequency based features give successful results in Turkish authorship attribution. Beside, [11] indicates that character level n-grams are suitable models to solve different Turkish text classification problems.

Lastly, the basic idea of the syntactic approach is that the author unwittingly tends to use similar syntax in her/his all

 ISBN: 978-1-61208-513-5 

articles. A widespread syntactic approach is using Part-of-Speech (POS) tagging. POS tagging is the process of labeling each word in a sentence as corresponding to its adequate part of speech. As is known that Turkish is an agglutinative language, which has a complex morphological structure. This morphological complexity of the Turkish language causes numerous different words appear in surface structures in the text. POS tags of the words can change from each other by using several suffixes. Herewith, it is more difficult to determine the final POS tag of a word using the root than in English language. Nevertheless, POS tagging have been used in four Turkish studies.

This is the first review of Turkish authorship attribution studies and the main point of the study that obtains more successful stylometric features for Turkish language. According to the results of reviewed Turkish author detection studies; word length; character n-gram and word n-gram models are the most successful features.

## IV. NEW CHARACTERISTIC APPROACHES FOR TURKISH

According to the previous section, word length and n-gram models are more important features than other stylometric features for Turkish studies. In addition to this, we assume that highlighting the characteristic features of Turkish will produce favorable results. For that purpose, we conducted two experiments. We have three datasets, which are consisting of Milliyet, Kıbrıs [12], and Radikal newspapers articles. Kıbrıs dataset has 7 authors, 50 articles for each author and average word count per article is 535. Milliyet has 9 authors, 50 articles for each author and average word count per article is 461. Radikal has 7 authors, 250 articles for each author and average word count per article is 836. 80% of each dataset is used as training data and 20% of each dataset is used as test data.

On the implementation side, we used scikit-learn [13], which is a powerful python machine-learning library. Scikit-learn provides skillful text vectorizers, which are utilities to build feature vectors from text documents. A vectorizer converts a collection of text documents to a matrix of intended features; within this context *tf-idf* (product of term frequency and inverse document frequency statistics) vectorizer gives a matrix of *tf-idf* features. All experiments have been done with default parameters of scikit-learn Support Vector Machines (SVM) [14] algorithm. Here an example of linear support vector classification function with default parameters:

LinearSVC(*penalty='l2',loss='squared_hinge',dual=True,tol=0.0001,C=1.0,multi_class='ovr',fit_intercept=True,intercept_scaling=1,class_weight=None,verbose=0,random_state=None,max_iter=1000*)

### A. Different Kinds of Word Structures

The first experiment consists in using different kinds of word structures. Authors often write on various subjects, in this case the word richness could not be distinctive. The authors follow the same syntactic way on their all articles without noticing it. Therefore, finding syntactic features can give more successful results on the author detection process. By this point, we developed the hypothesis that the root of the word would be better than other parts. On the other hand, if the authors use derived words in the same pattern, suffixes are valued in terms of syntactic approach. Turkish has similar features to all other agglutinative languages; such as derivational suffixes usually change the part of speech or the meaning of the word to which they are affixed.

We have designed a process, which is cutting words with a blunt knife; the first 5 letters of the words were marked as the root, the later letters were marked as the suffix part. Then the original versions of the words were marked as full word. Lastly, using the Turkish stemmer of the snowball [15] library was marked as stemmed. Thus, the dataset contains root, suffix, full word and stem.

TABLE I. F1-SCORES OF THE DATASETS VIA SVM ALGORITHM

| ***N-gram (1,3)*** | ***Full word*** | ***Root*** | ***Suffix*** | ***Stemmed*** |
|---|---|---|---|---|
| Radikal | 0.9885 | 0.9886 | 0.9792 | 0.9817 |
| Milliyet | 0.9566 | 0.9256 | 0.8768 | 0.8845 |
| Kıbrıs | 0.9621 | 0.9490 | 0.9042 | 0.9174 |

In datasets we used four different forms as mentioned above. SVM algorithm has produced average F1-scores which using *tf-idf* values of word unigram, bigram and trigram as features for each dataset can be seen in Table I. On the other hand, full word is more successful than root form and also full word is more successful than stem in all datasets. There is an important feature of agglutinative languages: derived words can be very different in terms of type and meaning from the root of the word, so the last form of the word has significant role in the Turkish language analysis. So it seems interesting to work with occurrences rather than with stems. Another comment about the results, suffix results are worse than results of other forms for Turkish. However, the gap between suffix and others is highly close. We can say that these results show promise and we could extract a syntactic clue with usage of suffixes.

### B. Gerunds Frequency

The other experiment takes advantages of Turkish characteristic features by using frequencies of gerunds. Gerunds are derived from the verbs but used as nouns in a sentence. Adding derivational suffixes to verbs in Turkish language creates gerunds. According to derivational suffix, the gerunds can be used as nouns, adjectives or adverbs in the sentence.

- **Noun:** Kardeşim oku**ma**yı öğrendi. (My sister learned to **read**.)
- **Adjective:** Ge**lecek** yıl işe başlayacak. (She will start to job **next** year.)
- **Adverb:** Yemeğimi biti**r bitirmez** gelirim. (I will come **as soon as I finish** my meal.)

We collected 590 infinitives, 587 participles and 916 verbal adverbs.

 ISBN: 978-1-61208-513-5 

On the implementation side, we use the frequencies of the gerunds as features for the SVM algorithm. The program produced 2662 features on Radikal dataset.

TABLE II. F1-SCORES OF GERUNDS AS FEATURED ON RADIKAL VIA SVM ALGORITHM

| *Author Name* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|
| AH | 0.87 | 0.67 | 0.76 |
| AO | 0.76 | 0.76 | 0.76 |
| BO | 0.72 | 0.78 | 0.75 |
| EB | 0.66 | 0.70 | 0.68 |
| FT | 0.79 | 0.84 | 0.82 |
| OC | 0.71 | 0.80 | 0.75 |
| TE | 0.83 | 0.76 | 0.79 |
| **Average** | **0.76** | **0.76** | **0.76** |

In according with Table II, the first practice implementation of gerund gives F1-score between 0.68 and 0.82. The first results are compared with reviewed Turkish studies; we can say that these results are promising. Because, the average F1-score is 0.76 and it was resulted from only gerunds frequency. Using Turkish characteristic features brings to a successful conclusion, hence the next step of the experiment will be tried to use other characteristic points of Turkish such as optative mood, synonym and free order.

## V. CONCLUSION

The expeditious increase in the number of electronic text and the development of techniques, such as machine learning and natural language processing tools have enabled the Turkish authorship attribution studies over the last two decades. These important works have been guided to develop the author detection methods that give successful results for the Turkish language. This paper includes two parts; first part is a review of Turkish authorship attribution studies. Focus of the review is the stylometric features that provide distinguishing between authors. This is the first review of Turkish authorship attribution studies, the main point of the study that obtains more successful stylometric features for Turkish language. The result of our review can show that word length, character n-gram and word n-gram models are the most important characteristics for Turkish author detection.

The second part consists of important stylometric features for Turkish and our experiments. The first one of experiments is built with n-gram and word structure by using Support Vector Machines algorithm. The average F1-score of the first experiments are 0.98, 0.90 and 0.92 for Radikal, Milliyet and Kıbrıs datasets respectively. The second experiment consisted of frequencies of gerunds by using SVM. The first practice implementation of gerund gives F1-score between 0.68 and 0.82.

Regarding the first promising results, we will continue experiments on especially n-gram, word structure and Turkish characteristic features such as optative mood, synonym and free order. Thus, we will try to provide successful solutions to the Turkish author detection problems.

## REFERENCES


[1] F. Mosteller and D. Wallace, Inference and disputed authorship: The Federalist. Addison-Wesley , 1964.

[2] E. Stamatatos, “A survey of modern authorship attribution methods”, Journal of the American Society for information Science and Technology, vol. 60.3, pp. 538-556, 2009.

[3] T. C. Mendenhall, “The characteristic curves of composition”, Science, pp. 237-249, 1887.

[4] E. Malone, A dissertation on part one, two and three of Henry IV tending to show that those plays where not written originally by Shakespeare. Henry Baldwin, 1787.

[5] C. U. Yule, The statistical study of literary vocabulary. Archon Books, 1968.

[6] A. Ellegard, “A Statistical method for determining authorship: the Junius Letters”, Gothenburg studies in English vol. 13, pp.1769-1772, 1962.

[7] F. Sebastiani, “Machine learning in automated text categorization”, ACM computing surveys (CSUR) vol. 34.1, pp. 1-47, 2002.

[8] K. Oflazer, “Turkish and its Challenges for Natural Language Processing,” Language Resources and Evaluation, vol. 48, pp. 639-653, Dec. 2014.

[9] F. C. Ekmekcioglu, M. F. Lynch and P. Willett, “Stemming and N-gram matching for term conflation in Turkish texts”, Information Research, 1(1). [Online] Available at: http://informationr.net/ir/2-2/paper13.html, 1996. [Accessed 12 July 2016]

[10] H. Takci and E. Ekinci, “Character Level Authorship Attribution for Turkish Text Documents”, The Online Journal of Science and Technology vol. 2.3, pp.12-16, 2012.

[11] F. Turkoglu, B. Diri and M. F. Amasyali, “Author attribution of turkish texts by feature mining”, Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, pp.1086-1093, 2007.

[12] Y. Bay and E. Celebi, “Feature Selection for Enhanced Author Identification of Turkish Text”, Information Sciences and Systems 2015, pp. 371-379, 2016.

[13] Scikit-learn Machine Learning in Python, http://scikit-learn.org/stable/ [Accessed 12 July 2016]

[14] C. Cortes and V. Vapnik, “Support-vector networks”, Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[15] Snowball,http://snowballstem.org/ [Accessed 12 July 2016]


 ISBN: 978-1-61208-513-5 