Author Attribution of Turkish Texts by Feature Mining

Filiz Türkoğlu, Banu Diri, and M. Fatih Amasyalı

Narrators: Mehmetcan Güleşçi, Furkan Karakoyunlu

May 13, 2017

What is the problem?

- 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
- In some 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

The aim of the article

- They focused an author attribution of Turkish texts by extracting various feature vectors and applying different classifiers
- They studied the comparative performance of classifier algorithms using the Naive Bayes, Support Vector Machine, Random Forest, Multilayer Perceptron, and k-Nearest Neighbour
- To conclude they calculated the effectiveness of the methods by using 10-fold cross validation

Early researches before this article

- Early researchers in authorship attribution used a variety of statistical methods but it tends to vary from author to author
  - Mosteller and Wallace - Federalist Papers, by using set of function words
  - Yule, by using complexity-based features (average sentence length, average word length, type/token ratio ..)
  - Recent technical advances in automated parsing and POS tagging, by using syntactic features such as POS n-grams
  - Peng, they modeled each author by a vector of the most frequent n-grams in the text
  - Fung - Federalist Papers, by using Support Vector Machine classifier

AUTHORSHIP ATTRIBUTION

- In this article, they formed feature vectors from several categories of statistics
- Stylistic analysis in order to compare the efficacy of each
- They examined features in five main categories, which are statistical, vocabulary richness, grammatical, lexical and n-grams model
- They obtained five different feature vectors from the mentioned categories
- Then, they created five feature subsets by using the feature selector to reduce the dimension of vectors

Corpus

- In this article, they used a corpus, 630 documents written by a single author are obtained from 35 texts per 18 different authors
- On different subjects like sport, popular interest and economics
- From Turkish daily newspaper, “Hürriyet” and ”Vatan”
- In order to determine the authorship attribution performance in homogeneous and heterogeneous documents, and different dataset sizes, this corpus is divided into 3 parts: Dataset I, Dataset II, Dataset III.

General Feature Vector (gfv)

- **Statistical features:** counting features in a text and applied this to word lengths and sentence lengths
- **Vocabulary richness features:** there are different statistics to determine the richness of an author's vocabulary. These features shows author's creativity.
- **Grammatical features:** based on Turkish Word Database (TWD) 35,000 words is used.
- **Function word features:** words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with order words within a sentence.

N-Gram Model

- They extracted bi-grams and tri-grams
- While forming the bi-grams and the tri-grams of the corpus, they 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, they used a threshold value (greater than 75) to reduce the number of features.
- After removed infrequent features, the dimensions of the bi-gram *bgfv*, and tri-gram *tgfv* feature vectors are 470 and 1037 respectively.
- After that they combined *bgfv* and *tgfv*, they had a new feature vector(*btgfv*) which has a dimension of 1507.
- Finally they put together *gfv* and *btgfv* and obtained a 2148-dimensioned new vector, which is called *gbtgfv*.

Feature Selection

- 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, they used CfsSubsetEval function which is in WEKA.
- They reduced features of general feature vector, *gfv* and obtained a new vector, *rgfv*.
  - 24 dimension for Dataset I
  - 17 dimension for Dataset II
  - 40 dimension for Dataset III

Feature Selection

- Same process was applied for Bi-gram feature vector, *bgfv*, and was formed *rbgfv*
  - 25 features for Dataset I
  - 20 features for Dataset II
  - 63 features for Dataset III
- They 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
  - 101 for Dataset III
- When features decreased from btgfv, 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.

Feature Selection

- They 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)

EXPERIMENTAL RESULTS

- In this work, they used WEKA's 5 classification algorithms, which are:
  - Naive Bayes (NB)
  - Support Vector Machine (SVM)
  - Random Forest (RF)
  - k-Nearest Neighbor (k-NN)
  - Multilayer Perceptron (MLP)
- On all our dataset, five classification algorithms are applied to various combinations of feature types.

EXPERIMENTAL RESULTS

- **Dataset I:**
  - On 630 documents, 18 different authors, 35 different topic.
  - The best performance in Dataset I, 92.5%, is obtained from gbtgfv with SVM.
  - According to avg *'rgbtgfv'* gives highest accuracy rate and the most successful classifier is *'SVM'*.
  - At the same time NB(*85.6%*), RF(*82.0%*) and k-NN(*79.0%*) are achieved best performance with *rgbtgfv*

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

EXPERIMENTAL RESULTS

- **Dataset II:**
  - On 315 documents, 9 different authors, 35 same topic.
  - The best performance in Dataset II, 95.4%, is obtained from gbtgfv with MLP.
  - According to avg *'rbtgfv'* gives highest accuracy rate and the most successful classifier is *'MLP'*.
  - NB(*rbgfv-90.8%*), SVM(*gbtgfv-94.6%*), RF(*rgbtgfv-91.7%*) and k-NN(*rbtgfv-89.5%*) are achieved best performance.

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

EXPERIMENTAL RESULTS

- **Dataset III:**
  - On 315 documents, 9 different authors, 35 different topic.
  - The best performance in Dataset III, 96.9%, is obtained from btgfv with MLP.
  - According to avg *'rgbtgfv'* gives highest accuracy rate and the most successful classifier is *'MLP'*.
  - NB(*91.1%*), RF(*87.9%*) and k-NN(*90.5%*) are achieved best performance with *rgbtgfv*

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

CONCLUSION

- The final average performance: Dataset I (72.4%), Dataset II (80.0%) and Dataset III (82.9%).
- When increasing the class count, classification performance decrease.
- Dataset II and III gave better results than Dataset I.
- We compare the results of Dataset II and Dataset III to determine capability of identifying authorship for heterogenous documents.

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%	-	-

CONCLUSION

- In general, MLP and SVM give good performance with *rgbtgfv* and *rbtgfv*.
- 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.

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%	-	-

CONCLUSION

- 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.

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%	-	-

The End