Ensemble-based Author Identification Using Character N-grams

Efstathios Stamatatos[1]

Abstract. This paper deals with the problem of identifying the most likely author of a text. Several thousands of character *n*-grams, rather than lexical or syntactic information, are used to represent the style of a text. Thus, the author identification task can be viewed as a single-label multiclass classification problem of high dimensional feature space and sparse data. In order to cope with such properties, we propose a suitable learning ensemble based on feature set subspacing. Performance results on two well-tested benchmark text corpora for author identification show that this classification scheme is quite effective, significantly improving the best reported results so far. Additionally, this approach is proved to be quite stable in comparison with support vector machines when using limited number of training texts, a condition usually met in this kind of problem.

1 INTRODUCTION

Author identification is the task of predicting the most likely author of a text given a predefined set of candidate authors. This task can be seen as a single-label multi-class text categorization problem [17] where the candidate authors play the role of the classes. Early attempts to author identification focused mainly on cases of disputed authorship [13] or literary works [3] with limited number of candidate authors, sometimes providing controversial results. However, a growing number of studies indicate that the field is now mature to handle difficult cases with many candidate authors and limited number of short training texts [1, 4, 5, 8, 9, 10, 15, 18, 20].

One major subtask of the author identification problem is the extraction of the most appropriate features for representing the style of an author, the so-called *stylometry*. Several measures have been proposed, including attempts to quantify vocabulary richness, function word frequencies and part-of-speech frequencies. A good review of stylometric techniques is given by Holmes [6].

Obviously, the most straightforward approach to represent a text is by using word frequencies, a method widely applied to topic-related text categorization as well. To this end, the most appropriate words for author identification may be selected arbitrarily [13], according to their discriminatory potential on a given set of candidate authors. Burrows [3] first indicated that the most frequent words of the texts (like ‘and’, ‘to’, etc.) have the highest discriminative power for stylistic purposes. Interestingly, these words are usually excluded from topic-related text categorization systems. Additionally, this approach for selecting appropriate words is language-independent.

A recent study [9] shows that sub-word units like character *n*-grams (i.e., character sequences of length *n*) can be very effective for capturing the nuances of an author’s style. The most frequent *n*-grams of a text provide crucial information about the author’s stylistic choices on the lexical, syntactical, and structural level. For example, the most frequent 3-grams of an English corpus indicate lexical (‘the’, ‘ to’, ‘tha’, ‘con’), syntactical (‘ing’, ‘ed ’), or structural (‘. T’, ‘ “T’) information.

In this paper, we follow the language-independent stylometric approach proposed by Burrows [3] using character *n*-gram frequencies instead of word frequencies. Several thousands of the most frequent *n*-grams are used to repressent the style of a text. From a machine learning point of view, the task of author identification can, then, be viewed as a classification problem of high dimensional feature space (several thousands of valuable features). As proved by previous studies, every word (and subsequently *n*-gram) is valuable for text classification [7]. Therefore, feature selection methods that attempt to reduce the feature set seem not suitable for this task. Moreover, the longer the feature set, the more sparse the data (i.e., the less frequent an *n*-gram, the less likely to be found in a given text).

A machine learning approach able to cope with such a classification task is an ensemble of classifiers based on *feature set subspacing* [2]. That is, to avoid the curse of dimensionality problem, the feature set is divided into smaller parts, each used to train a base learner. The predictions of the base classifiers are, then, combined to provide the most likely class. In this paper, we propose a suitable ensemble-based model and apply it to character *n*-gram representations of authors’ style. Comparative performance results are provided for the ensemble-based approach and an alternative model using support vector machines, based on two benchmark text corpora previously used by author identification studies. Moreover, we focus on practical considerations of the task in question, such as limited number of training texts, a condition usually met in real-world author identification problems.

The rest of this paper is organized as follows. Section 2 presents the learning ensemble classification scheme as used in this study. The *n*-gram data sets and the other methods used for comparative purposes are described in section 3. The performance results of the examined schemes are included in section 4. Finally, section 5 summarizes the conclusions drawn and suggests future work directions.

2 CLASSIFICATION SCHEME

In the current approach, each text is represented as a vector of character *n*-gram frequencies of occurrence. Let $\mathbf{G}_d=\{g_1, g_2, \ldots, g_d\}$ be the ordered set (by decreasing frequency of occurrence) of the most frequent *n*-grams (i.e., character sequences of length *n*) of the

[1] Dept. of Information and Communication Systems Eng., University of the Aegean, 83200 – Karlovassi, Greece, email: stamatatos@aegean.gr

training set. Consider f_{ij} as the normalized frequency of occurrence of the j-th n-gram of $\mathbf{G}_d$ in the i-th text. Then, a text x_i is represented as the ordered vector $<f_{i1}, f_{i2}, \ldots, f_{id}>$.

For constructing a classifier ensemble based on feature set subspacing we follow an approach we call *exhaustive disjoint subspacing*. That is, a large feature set is divided into equally-sized disjoint feature subsets drawn at random. Each particular attribute is used exactly once. Each resulting feature subset is used to train a base classifier using a learning algorithm able to provide posterior probabilities. In this study, *linear discriminant analysis* is used. This standard technique from multivariate statistics is a well-known stable classification algorithm proven to be a good compromise between classification accuracy and training time cost [12]. The predictions of the base classifiers are, then, combined based on an appropriate combination method as described in the following subsections.

2.1 Base Classifiers

Let $G_{m:d}$ be a subset of m features drawn (without replacement) at random from the set $\mathbf{G}_d$ of the most frequent n-grams of the training corpus ($m \leq d$). Consider $C(G_{m:d})$ as a single linear discriminant classifier trained on the frequencies of these m n-grams in the training set texts. Then, $E(C(G_{m:d}), combination)$ is an ensemble of such base classifiers according to the *combination* method. When every feature is used exactly once in the framework of an ensemble, we have an exhaustive disjoint subspacing ensemble. In this case, the number of base classifiers is d/m. Preliminary experiments indicated that the lower the m, the better (and more stable) the performance of the ensemble model. In the experiments described in this study, feature subsets of minimal length are used (m=2).

Consider $\mathbf{L}$ as the set of all possible classes (authors), then the i-th classifier assigns a posterior probability $P_i(C_i(G_{m:d}), x, c)$ to an input text x for each $c \in \mathbf{L}$, so that

$$\sum_{j=1}^{|L|} P_i(C_i(G_{m:d}), x, c_j) = 1$$

where $|\mathbf{L}|$ is the size of $\mathbf{L}$. In case of learning algorithms that provide crisp predictions, the posterior probabilities can only take binary values (0 or 1).

2.2 Combination Method

Provided the posterior probabilities of the constituent classifiers, an ensemble assigns a posterior probability to an input text for each class according to the combination of the predictions of the base classifiers. Commonly, a combined decision is obtained by just averaging the estimated posterior probabilities (the *mean* rule):

$$P(E(C(G_{m:d}), mean), x, c) = \frac{1}{k}\sum_{i=1}^{k} P_i(C_i(G_{m:d}), x, c)$$

where k is the number of the base classifiers. Recall that for exhaustive disjoint subspacing $k=d/m$. Given that the base classifiers are based on different feature sets, their decisions are considered to be independent. When the Bayes theorem is adopted, an alternative combination rule can, then, be applied to the outputs of the base classifiers (geometric mean or the *product* rule):

$$P(E(C(G_{m:d}), product), x, c) = \sqrt[k]{\prod_{i=1}^{k} P_i(C_i(G_{m:d}), x, c)}$$

Comparison of these two combination rules has shown that under the assumption of independence the product rule should be used. However, in case of poor posterior probability estimates, the mean rule is proved to be more fault tolerant [19].

In this study, we use a combination of these two combination rules (henceforth called *mp*). The *mp* rule is just the average of *mean* and *product* rules. Note that the *mean* rule is affected by high values of posterior probabilities, therefore it is favorable for cases where a few base classifiers have assigned a high posterior probability to a class. On the other hand, the *product* rule is affected by low values of posterior probabilities, therefore it is favorable for cases where only a few base classifiers have assigned low posterior probability to a class. Hence, *mp* is a good compromise of these two.

To complete the classification model, provided that *label*(*classifier*, *instance*) is the class assigned by a classifier to a test instance, then, a classifier ensemble chooses the class that maximizes the posterior probability for an input text x, that is:

$$label(ensemble, x) = \arg\max_{c \in L}(P(ensemble, x, c))$$

2.3 Effectiveness Measures

The performance of a classifier ensemble is directly measured by the classification accuracy on the test set. Moreover, the effectiveness of an ensemble is indirectly indicated by the diversity among the predictions of the base classifiers as well as the accuracy of the individual base classifiers. In particular, many measures have been proposed to represent the diversity of an ensemble [11]. In this study, the *entropy* measure is used, that is:

$$entropy = \frac{1}{|T|}\sum_{i=1}^{|T|}\sum_{c=1}^{|L|} -\frac{N_c^i}{k}\log_{|L|}(\frac{N_c^i}{k})$$

where k is the number of base classifiers, $|\mathbf{T}|$ is the total number of test texts and N_{ic} is the number of base classifiers that assign text i to class c. Notice that log is taken in base $|\mathbf{L}|$ to keep the entropy within the range [0,1]. The higher the entropy of an ensemble, the more diverse the predictions of the individual constituent classifiers.

3 EXPERIMENTAL SETTINGS

3.1 Data Sets

The text corpora used in this study are two well-tested benchmarks for authorship identification. In particular, the texts were published within 1998 in the Modern Greek weekly newspaper *TO BHMA* (the tribune), and were downloaded from the WWW site of the newspaper. The texts are divided into two groups of authors:

- **Group A** (hereafter GA): It consists of ten randomly selected authors whose writings are frequently found in the section A of the newspaper. This section comprises texts written mainly by journalists on a variety of current affairs. Moreover, for a certain author there may be texts from different text genres (e.g., editorial, reportage, etc.). Note that in many cases such texts are highly edited in order to conform to a predefined style, thus washing out specific characteristics of the authors which complicate the task of attributing authorship.

- **Group B** (hereafter GB): It consists of ten randomly selected authors whose writings are frequently found in the section B of the newspaper. This supplement comprises essays on science, culture, history, etc. in other words, texts in which the idiosyncratic style of the author is not overshadowed by functional objectives. In general, the texts included in the

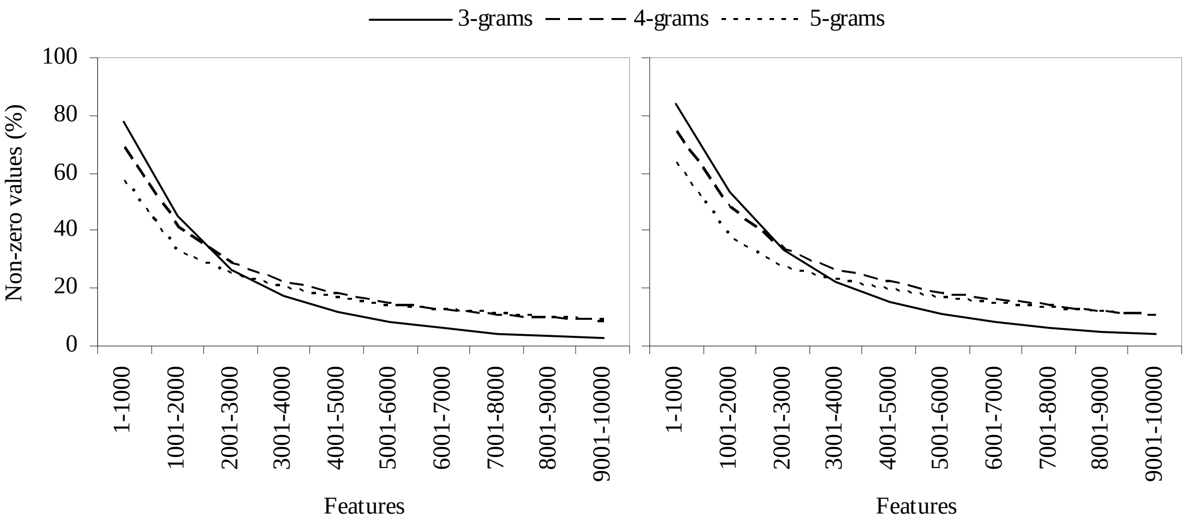

Figure 1. Average amount of non-zero attribute values per thousand of features for the training set of GA (left) and GB (right). Data sets of 3-grams, 4-grams and 5-grams are depicted.

	GA	GB
Avg. words per text	866.8	1,148.2
Authors	10	10
Texts per author	20	20
Texts per author in training set	10	10
Texts per author in test set	10	10
Reported Results (accuracy %)		
Stamatatos, *et al*., 2000 [18]	72	70
Peng, *et al*., 2003 [15]	74	90
Keselj, *et al*., 2003 [9]	**85**	**97**
Peng, *et al*., 2004 [16]	-	96

Table 1. The text corpora used in this study and reported accuracy results so far.

supplement B are written by scholars, writers, etc., rather than journalists.

Each corpus is divided into disjoint training and test parts of equal size in terms of texts per author (i.e., ten texts per author in the training set and ten texts per author in the test set for each group). Some brief information about these text corpora is summarized in Table 1. More detailed information can be found in [18]. Intuitively, for the GB it is easier to discriminate between the authors since the texts are more stylistically homogenous. In addition, GB's texts are significantly longer than GA's texts.

No linguistic preprocessing of the corpora is required for constructing the data sets for the current approach. The set of the d most frequent character *n*-grams (ordered by decreasing frequency of occurrence) of the training set is extracted, for a given character sequence length n. In the following experiments, character 3-grams, 4-grams, and 5-grams are examined while the feature set size (*d*) varies from 1,000 to 10,000. Then, each text is represented by the ordered vector of *d* *n*-gram frequencies, normalized over the total amount of text characters.

To illustrate the characteristics of these data sets, figure 1 depicts the average amount of non-zero attribute values per thousand of features for both GA and GB. As can be seen, the larger the feature set size, the sparser the resulting data. Moreover, shorter *n*-grams (i.e., 3-grams) tend to be less sparse for relatively low dimensional feature spaces (until 3,000 features). Of course, this can be explained by the fact that the complete set of 3-grams is much smaller than the complete set of 5-grams and the most frequent 3-grams are more likely to be found in every text in comparison to the most frequent 5-grams. On the other hand, beyond a certain level (around 3,000 most frequent *n*-grams) 3-grams are less likely to be found in a text in comparison to the corresponding 5-grams. Notice also that GA data sets are sparser in comparison to the corresponding GB data sets.

3.2 Setting the Baseline

The GA and GB corpora provide a reliable testing ground for author identification experiments since they comprise an adequate number of candidate authors, adequate number of test texts, and the authorship of each text is undisputed. For this reason, they were used to test several author identification approaches [9, 15, 16, 18] and the best reported results so far are shown in Table 1. Notice that the considerations about the difficulty of the two text corpora are reflected in the reported results since the classification accuracy for GB is much higher in comparison to GA.

As mentioned earlier, the approach described in [9] is also based on mere character *n*-grams, thus the comparison with the presented method is straightforward. Additionally, in order to test the proposed classification algorithm, a *Support Vector Machine* (SVM) model [21] was also built, since SVMs provide one of the best available solutions when dealing with high dimensional data.

4 RESULTS

The SVM and learning ensemble classification schemes were applied to both GA and GB. In particular, common kernel options that optimize the average performance of the models were selected (linear kernel, *C*=1). In particular, the exhaustive disjoint subspacing approach with minimal feature subset length (*m*=2) was followed. The base learner combination rule *mp* was used. For each text corpus, three different data sets were examined (3-grams, 4 grams, and 5-grams) with feature set size varying from 1,000 to 10,000 with a step of 1,000 *n*-grams. Table 2 shows the performance for both classification approaches on the test set of GA

Feat.	GA						GB					
set	3-grams		4-grams		5-grams		3-grams		4-grams		5-grams	
size	SVM	Ens.	SVM	Ens.	SVM	Ens.	SVM	Ens.	SVM	Ens.	SVM	Ens.
1,000	81	80	80	77	68	68	96	96	93	96	94	94
2,000	83	79	77	76	73	73	98	96	95	95	96	94
3,000	86	86	82	79	83	81	98	99	98	96	97	97
4,000	90	95	86	83	85	85	99	99	**100**	99	**100**	**100**
5,000	89	94	87	87	85	85	99	**100**	**100**	**100**	98	**100**
6,000	92	**96**	91	93	87	89	98	99	**100**	**100**	99	**100**
7,000	92	**96**	92	93	89	92	99	**100**	99	**100**	99	99
8,000	92	**96**	92	92	92	90	99	**100**	98	**100**	98	99
9,000	92	**96**	93	93	91	92	98	**100**	97	**100**	97	99
10,000	92	**96**	94	94	91	93	98	**100**	96	**100**	97	99

Table 2. Performance results on test set of both GA and GB for the support vector classifier and the learning ensemble. Classification accuracy (%) is indicated for different feature set size (amount of character *n*-grams) and types of features (3-grams, 4-grams, and 5-grams). Best achieved results are in boldface.

and GB. It is obvious that for GA it is more difficult to discriminate between the authors as compared with GB. Moreover, the best results for both approaches are much better than the best reported results for the same text corpora (see table 1). In more detail, in the best case, SVM achieves 94% and 100% classification accuracy for GA and GB, respectively, while the learning ensemble achieves 96% and 100% classification accuracy for GA and GB, respectively.

Notice that the performance of both approaches increases as the feature set size increases. Beyond a certain level (around 6,000 *n*-grams) the performance is either stabilized or slightly decreased (especially in the SVM models for the GB data sets). The ensemble model is superior of the SVM model in most cases with feature set size greater than 3,000. Therefore, it seems that the ensemble model is better able to handle high dimensional feature spaces. Additionally, in most cases 3-grams are better able to discriminate between the classes for both GA and GB. Recall that the 3-gram data sets are sparser beyond 6,000 features in comparison to 4-grams or 5-grams (see figure 1). Again, the ensemble model is superior for the 3-gram data sets and large feature set sizes. This indicates that the ensemble model can cope more effectively with sparse data.

4.1 Ensemble Diversity

A more detailed insight will illustrate why the ensemble model is so successful. The base classifiers that constitute the ensemble perform quite poorly when examined as individuals. Figure 2 depicts the base learner classification accuracy on the test data of GA and GB for the 3-gram data set. Random guess accuracy is indicated as well. As can be seen, the base classifiers are very poor predictors. Moreover, the predictions for GB are constantly more accurate than that of GA.

The key-factor for the success of the ensemble model is the extremely high diversity among the predictions of the base classifiers. Figure 3 shows the diversity, in terms of entropy, among the predictions of base classifiers on the test set of GB. Note that since the base classifiers are based on disjoint feature sets, the diversity is expected to be high. However, the level of entropy depicted in Figure 3 reaches 1.0, which means random error among the predictions. In words, the wrong predictions of the base classifiers are mutually cancelled.

Moreover, the diversity of the ensemble reaches its peak value at different size of feature set (and subsequently different amount of base classifiers), according to the data set. Thus, for the 3-gram data set, the diversity reaches its peak value at 5,000 features, while for the 4-gram and 5-gram data sets the diversity reaches its peak value at 7,000 and 8,000 features, respectively. Similar diversity curves can be obtained for the GA data sets. Notice that this decrease in diversity for the 3-gram data set of GA reflects in the performance of the corresponding ensemble models. Hence, the accuracy of the GA 3-gram ensemble model, shown in table 2, is not further improved for feature spaces greater than 5,000 features. However, despite this decrease in diversity, the classification accuracy does not drop (neither for GA nor GB).

4.2 Limited Training Texts

The training set size is a crucial factor in author identification since, in real world problems there is only a limited number of texts of undisputed authorship for each candidate author to be used as training data. For that reason, it is of vital importance for the classification method to require as limited training data as possible while maintaining a high level of accurate predictions on unseen cases.

To test the degree in which the SVM and the ensemble models are affected by the training set size, the experiment of the previous section was repeated based on reduced training sets. The SVM and the ensemble models were applied to both GA and GB using 50% (i.e., 5 texts per author) and 20% (i.e., 2 texts per author) of the original training sets. Data sets of 3-grams, 4-grams, and 5-grams of 10,000 features were examined. Table 3 shows the results of this experiment. Note that the test sets remain the same, thus, the results of Table 3 can be directly compared to Table 2. To illustrate further, the last line of table 3 indicates the performance of the models using the corresponding full-sized training sets (taken from Table 2).

In all cases the ensemble model performs better in comparison to SVM. In particular, for very limited training sets (20% of the original ones) the SVM model fails to maintain the previous classification accuracy. Interestingly, the performance of the ensemble model is not dramatically affected by reducing the training size. Actually, for the 3-gram data set of GB the classification accuracy remains at the top level using only 20% of the original training set, while for the corresponding GA data set the accuracy is competitive to the best reported results (see Table 1). In general, it seems that *n*-grams of short length (i.e., 3-grams) are better able to deal with limited training sets.

5 CONCLUSIONS

In this paper, an ensemble-based approach to the task of author identification was presented. Each text is represented as a vector of

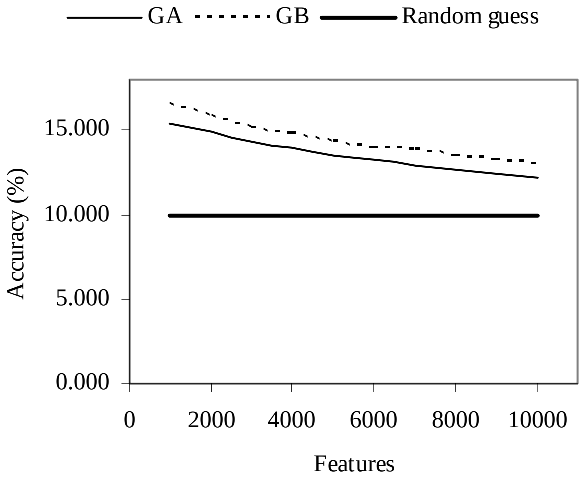

Figure 2. Classification accuracy (%) of the base classifiers for the 3-gram data set on GA and GB. Random guess accuracy is indicated as well.

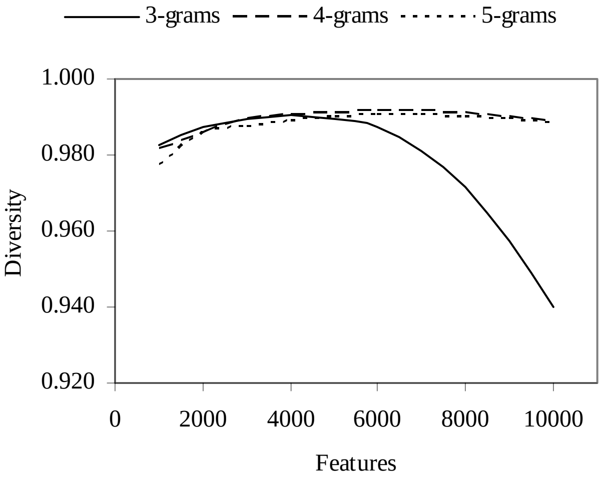

Figure 3. Diversity (in terms of entropy) of the base classifiers of the ensemble model for GB.

Train.	GA						GB					
set	3-grams		4-grams		5-grams		3-grams		4-grams		5-grams	
size	SVM	Ens.	SVM	Ens.	SVM	Ens.	SVM	Ens.	SVM	Ens.	SVM	Ens.
2	73	80	55	77	46	69	81	**100**	78	96	75	89
5	83	89	81	85	76	80	97	**100**	92	98	93	98
10	92	**96**	94	94	91	93	98	**100**	96	**100**	97	99

Table 3. Performance results on test set of both GA and GB for the support vector classifier and the learning ensemble. Classification accuracy (%) is indicated for different training set size (in texts per author) and types of features (3-grams, 4-grams, and 5-grams). In all cases, feature set size is 10,000. Best achieved results are in boldface.

frequencies of character *n*-grams. Such features require minimal text preprocessing and their extraction is a language-independent procedure. The ensemble-based approach of exhaustive disjoint subspacing was followed in order to handle such highly dimensional and sparse data. The application of this technique to two benchmark text corpora for author identification yields classification models with high accuracy, significantly higher than the best reported results for the same text corpora. First, this proves that character *n*-grams can successfully represent an author's style. Second, it demonstrates that the examined classification model can effectively cope with the author identification task.

The ensemble model proves to be significantly reliable when dealing with limited training set, a condition usually met in real-world author identification problems. Note also that the proposed technique does not require the use of a validation set for parameter tuning, minimizing the need for extra training texts. The success of the ensemble model is explained by the extremely high diversity among the predictions of the base classifiers. Previous studies have also shown that diversity alone can be used as a guide for constructing good ensembles [22]. The approach followed in this study ensures an extremely high level of diversity.

Special attention was paid on the combination of the predictions provided by the base classifiers. A scheme that combines the arithmetic and geometric mean is proposed. This scheme chooses the most likely class based on a compromise between high scores and low scores assigned to a class. The examined ensemble model is based on feature subsets of minimal length (*m*=2). This approach yields the highest number of base classifiers and provides the best experimental results. Moreover, it minimizes the effort to group features together in order to form feature subsets. Note that preliminary experiments with different subset lengths (*m*>2) indicated that the lower the feature subset length, the better (and more stable) the performance of the ensemble model.

In this study, features are paired at random. It has to be noted that repeated experiments with randomly paired features showed that the difference in performance is not statistically significant for feature sets including at least 3,000 features. However, a more sophisticated approach involving a search through the space of all the possible feature combinations [14] can also be examined. On the other hand, such an approach would require a validation set and a considerably greater training time cost.

As concerns the task of author identification, there are still open questions. In particular, limited text-length and imbalanced training set (i.e., unequal distribution of training texts over the authors) can affect the performance of the model. Moreover, open-class problems (i.e., the true author is not included in the candidate authors), another situation usually met in real-world problems, should be thoroughly examined as well.

REFERENCES

[1] Argamon, S., M. Saric, and S. Stein. 2003. Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. In Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining: 475-480

[2] Bay S. 1998. Combining Nearest Neighbor Classifiers Through Multiple Feature Subsets. In Proc. of the 15th International Conference on Machine Learning: 37-45

[3] Burrows, J.F. 1987. Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style. Literary and Linguistic Computing, 2: 61-70.

[4] de Vel, O., A. Anderson, M. Corney, and G.M. Mohay. 2001. Mining E-mail Content for Author Identification Forensics. SIGMOD Record, 30(4): 55-64.

[5] Diederich, J., J. Kindermann, E. Leopold, and G. Paass. 2003. Authorship Attribution with Support Vector Machines. Applied Intelligence, 19(1/2): 109-123.

[6] Holmes, D. 1998. The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13(3): 111-117.

[7] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proc. of the 10th European Conference on Machine Learning.

[8] Juola, P. 2004. Ad-hoc Authorship Attribution Competition. In Proc. of the Joint Int. Conference ALLC/ACH 2004: 175-176.

[9] Keselj, V., F. Peng, N. Cercone, and C. Thomas. 2003. N-gram-based Author Profiles for Authorship Attribution. In Proc. of the Conference of the Pacific Association for Computational Linguistics.

[10] Koppel, M., and J. Schler. 2004. Authorship Verification as a One-Class Classification Problem. In Proc. of the Twenty-first Int. Conf. on Machine Learning.

[11] Kuncheva, L. & C. Whitaker. 2003. Measures of Diversity in Classifier Ensembles. Machine Learning, 51: 181-207.

[12] Lim, T., W. Loh, and Y. Shih. 2000. A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning, 40(3): 203-228.

[13] Mosteller, F. and D. Wallace. 1984. Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Springer-Verlag, New York.

[14] Opitz, D., and J. Shavlik. 1999. A Genetic Algorithm Approach for Creating Neural Network Ensembles. In A. Sharkley (ed.) Combining Artificial Neural Nets: 79-99.

[15] Peng, F., D. Shuurmans, V. Keselj, and S. Wang. 2003. Language Independent Authorship Attribution Using Character Level Language Models. In Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics.

[16] Peng, F., D. Shuurmans, and S. Wang. 2004. Augmenting Naive Bayes Classifiers with Statistical Language Models. Information Retrieval Journal, 7(1): 317-345.

[17] Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1): 1-47.

[18] Stamatatos, E., N. Fakotakis, and G. Kokkinakis. 2000. Automatic Text Categorisation in Terms of Genre and Author. Computational Linguistics, 26(4): 471-495.

[19] Tax, D., M. van Breukelen, R. Duin, and J. Kittler. 2000. Combining Multiple Classifiers by Averaging or by Multiplying? Pattern Recognition, 33: 1475-1485.

[20] van Halteren H. 2004. Linguistic Profiling for Author Recognition and Verification. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics: 199-206.

[21] Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer, New York.

[22] Zenobi, G., and P. Cunningham. 2001. Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error. In Proc. of 12th European Conference on Machine Learning: 576-587.