This is a preprint of the paper published in *Information Processing and Management*, 44(2), pp. 790-799, 2008.

Author Identification: Using Text Sampling to Handle the Class Imbalance Problem

EFSTATHIOS STAMATATOS

Dept. of Information and Communication Systems Eng.
University of the Aegean
83200 – Karlovassi, Samos, Greece
stamatatos@aegean.gr

Abstract

Authorship analysis of electronic texts assists digital forensics and anti-terror investigation. Author identification can be seen as a single-label multi-class text categorization problem. Very often, there are extremely few training texts at least for some of the candidate authors or there is a significant variation in the text-length among the available training texts of the candidate authors. Moreover, in this task usually there is no similarity between the distribution of training and test texts over the classes, that is, a basic assumption of inductive learning does not apply. In this paper, we present methods to handle imbalanced multi-class textual datasets. The main idea is to segment the training texts into text samples according to the size of the class, thus producing a fairer classification model. Hence, minority classes can be segmented into many short samples and majority classes into less and longer samples. We explore text sampling methods in order to construct a training set according to a desirable distribution over the classes. Essentially, by text sampling we provide new synthetic data that artificially increase the training size of a class. Based on two text corpora of two languages, namely, newswire stories in English and newspaper reportage in Arabic, we present a series of authorship identification experiments on various multi-class imbalanced cases that reveal the properties of the presented methods.

Keywords: Author identification, class imbalance, text categorization.

1 INTRODUCTION

Authorship analysis has a long history mainly due to research on literary works of disputed or unknown authorship. The *Federalist Papers* (some of them claimed by both Alexander Hamilton and James Madison) is a famous case (Mosteller & Wallace, 1984). However, in certain cases, the results of authorship attribution studies were considered controversial (Labbé & Labbé, 2001). In recent years, researchers have paid increasing attention to authorship analysis in the framework of practical applications, such as verifying the authorship of emails and electronic messages (de Vel et al., 2001; Argamon et al., 2003; Abbasi & Chen, 2005), plagiarism detection in student essays (van Halteren, 2004), and forensic cases (Chaski, 2001).

Authorship identification is the task of predicting the most likely author of a text given a predefined set of candidate authors and a number of text samples per author of undisputed authorship (Stamatatos et al., 2000; Peng et al., 2003). From a machine learning point of view, this task can be seen as a single-label multi-class text categorization problem (Sebastiani, 2002) where the candidate authors play the role of the classes. As concerns the text representation, various measures have been proposed in order to quantify the stylistic

choices of the authors. Among them, function word frequencies, character *n*-gram frequencies, vocabulary richness measures, word-class frequencies, and syntactic analysis measures. Holmes (1998) provides an excellent review of the different stylometric techniques while Zheng et al. (2006) emphasize on modern approaches.

Very often, a common problem in authorship identification case (at least for some of the candidate authors) is the lack of text samples of undisputed authorship to be used for training. It is not unusual, only extremely limited text samples to be available for some authors. On the other hand, a big amount of text samples may be available for other candidate authors. Note that text samples should be of comparable length. Another realistic scenario is to have (more or less) equal amount of texts of undisputed authorship for all the candidate authors, however short texts are available for some of them and long texts for others. Hence, in the procedure of normalizing the length of training text samples, few text samples will be produced for some authors and many text samples for others. From a machine learning point of view, this constitutes the *class imbalance* problem (i.e., uneven distribution of the training set over the classes) in a classification task. This problem has been studied mainly within the framework of two-class datasets (Japkowicz & Stephen, 2002). The main approaches to deal with class imbalance attempt to re-balance the training set by performing:

- Under-sampling of the majority class, and
- Over-sampling of the minority class.

In general, it is unclear which approach is more effective and there have been attempts to combine them (Estabrooks et al., 2004). Another main approach is to attempt to modify the sensitivity of the classification algorithm so that errors on minority class to be costlier than errors on majority class (Veropoulos et al., 1999). Last but not least, the SMOTE approach (Chawla et al, 2002) creates new synthetic training data for the minority class. This is achieved by adding a small random value to some of the features of original training data and producing new data which lie close to the original ones in the multi-dimensional space of the problem.

Given a text categorization task, each training text is considered as a unit for constructing the training set. Usually, the length of the training texts is fixed or defined by the source of the documents (Sebastiani, 2002). Little work has been done on how the training texts can be efficiently segmented in order to provide multiple training text samples to assist the re-balancing of the training set. Moreover, text categorization often requires several thousand of features producing sparse data. Hence, producing synthetic data based on an approach such as SMOTE does not seem to fit well.

In this paper, we present methods to segment the training texts into text samples according to the size of the class. The main idea is that textual data can be handled in a flexible way so

that to produce a variable amount of text samples of variable length. That is, minority class can be segmented into short samples and majority class into longer samples. Therefore, we transform an imbalanced multi-class textual training set into a balanced set. Moreover, we explore text re-sampling methods in order to construct a training set according to a desirable distribution over the classes. In other words, text re-sampling can be viewed as providing new synthetic data that increase the training size of a class. Based on two text corpora, namely, newswire stories in English and newspaper reportage in Arabic, we present a series of authorship identification experiments on various multi-class imbalanced cases.

A basic assumption of inductive learning is that the test set distribution over the classes is similar to the training set distribution. This is obvious in tasks such as disease detection, where the disease may appear only in a few cases. Hence, the percentage of disease cases in training and test set should be similar. However, in other tasks, such as author identification, the distribution of the training set over the classes is affected by factors irrelevant to the dataset itself. For instance, only a couple of texts of unquestioned authorship may be available for a certain author. This should not be taken as evidence that this author is unlikely to be the author of an unknown text. Therefore, where tasks as author identification are examined, the test set should not follow the distribution of the training set. Rather, the test set should be equally distributed over the classes so that the performance of the produced model to be fairly evaluated. In this paper, we follow this procedure.

The rest of this paper is organized as follows. Section 2 describes our authorship identification approach focusing on language-independent text representation. Section 3 briefly presents the text corpora and imbalanced multi-class datasets used in this study. Section 4 includes the presented methods and the evaluation experiments. Finally, section 5 summarizes the main conclusions drawn by this study and indicates future work directions.

2 AUTHOR IDENTIFICATION

2.1 Representing Style

One great challenge is to define an appropriate quantitative text representation so that the stylistic choices of the author to be revealed. Several types of features have been used so far including lexical and character features, syntactic features (part-of-speech frequencies), structural features (use of greetings, signatures) etc. (Zheng, et al., 2006) Since the focus of this study is on text sampling techniques to avoid the class imbalance problem, features on the document level (e.g., use of greetings, types of signatures, number of paragraphs etc.) are not appropriate. Moreover, the features should be stable in representing the style of very short text samples.

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 stylistic purposes may be selected in an arbitrary way (Mosteller & Wallace, 1984) according to their discriminatory potential on a given set of candidate authors. Burrows (1987) 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. Moreover, using high-frequency words as style markers is a language-independent procedure.

A simple but powerful text representation technique for stylistic purposes is a 'bag of character *n*-grams'. Character *n*-grams (contiguous characters of fixed length) are able to capture complicated stylistic information on the lexical, syntactic, or structural level. For example, the most frequent character 3-grams of an English corpus indicate lexical (|the|, |_to|, |tha|), word-class (|ing|, |ed_|), or punctuation usage (|. T|, |_"T|) information[1]. Kjell et al. (1994) used character bigrams and trigrams to visualize stylistic differences while Keselj et al., (2003) proposed a model based on character *n*-gram representation for author identification. A variation of this model achieved the best results in the ad-hoc authorship attribution contest (Juola, 2004), a competition based on a collection of 13 text corpora in various languages (English, French, Latin, Dutch, and Serbian-Slavonic). Interestingly, character n-grams proved to be a useful representation for topic-based text categorization as well (Lodhi et al., 2002). Note also that using the *n*-gram representation on the character level, the sparse data problems that arise in *n*-grams on the word level are significantly reduced.

2.2 The Bag of Character N-grams Approach

In this paper, we are based on the frequencies of occurrence of the most frequent character *n*-grams of the training corpus in order to represent a text sample. 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 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 $<f_{i1}, f_{i2}, \ldots, f_{id}>$. In this study, d=5,000 and n=3. In other words, each text is represented as a vector of 5,000 character 3-gram frequencies of occurrence.

A support vector machine, a supervised learning algorithm based on the *structural risk minimization* principle (Vapnik, 1995), is then applied to these multi-dimensional vectors. This is an appropriate classification algorithm for text categorization tasks since its learning ability is independent of the feature space dimensionality, because it measures the complexity

[1] The characters '|' and '_' are used to denote n-gram boundaries and a single space character, respectively.

of the hypotheses based on the margin with which they separate the data, instead of the features.

Note that this approach is language-independent. However, for achieving best results one should explore the most appropriate amount and length of *n*-grams for a particular natural language (e.g., for Arabic less and longer *n*-grams would be more effective). This is out of the purpose of this study since we focus on evaluating the performance of authorship identification under imbalance conditions.

3 TEXT CORPORA AND DATASETS

Two text corpora have been used in this study, one in English and one in Arabic. Both include texts by ten different authors (100 texts per author). In more detail, each text corpus consists of texts belonging to the same genre:

- English Corpus: newswire stories in English taken from the publicly available Reuters Corpus Volume 1 (RCV1) (Lewis et al., 2004). There are four main topic classes in RCV1: CCAT (corporate/industrial), ECAT (economics), GCAT (government/social), and MCAT (markets). Each of these main topics has many subtopics and a document may belong to a subset of these subtopics. Although, not particularly designed for evaluating author identification approaches, the RCV1 corpus contains 'by-lines' in many documents indicating authorship. The top 10 authors with respect to the amount of texts belonging to the topic class CCAT were selected.
- Arabic Corpus: newspaper reportage in Arabic downloaded from the website of *Al-Hayat*. Ten authors were selected according to the amount of available texts online.

In both corpora, only the main body of the text was considered (titles, author names, dates, etc. were excluded). No pre-processing was performed on these texts apart from removing xml and html tags irrelevant to text content,.Note that for the English corpus steps to reduce both the genre and the topic factor have been taken to be hoped that the authorship differences will be a more significant factor in differentiating the texts. On the other hand, in the Arabic corpus the texts for a certain author may cover several topics. On the other hand, some authors may share specific topics. Since the features used in this study (character *n*-grams) are able to capture stylistic as well as thematic information, the topic factor may strengthen or weaken the differences between the authors of the Arabic corpus. Moreover, the average text of the Arabic corpus (4,378 characters) is longer than the average English corpus text (3,089 characters). Each corpus was divided into 50 training and 50 test texts per author.

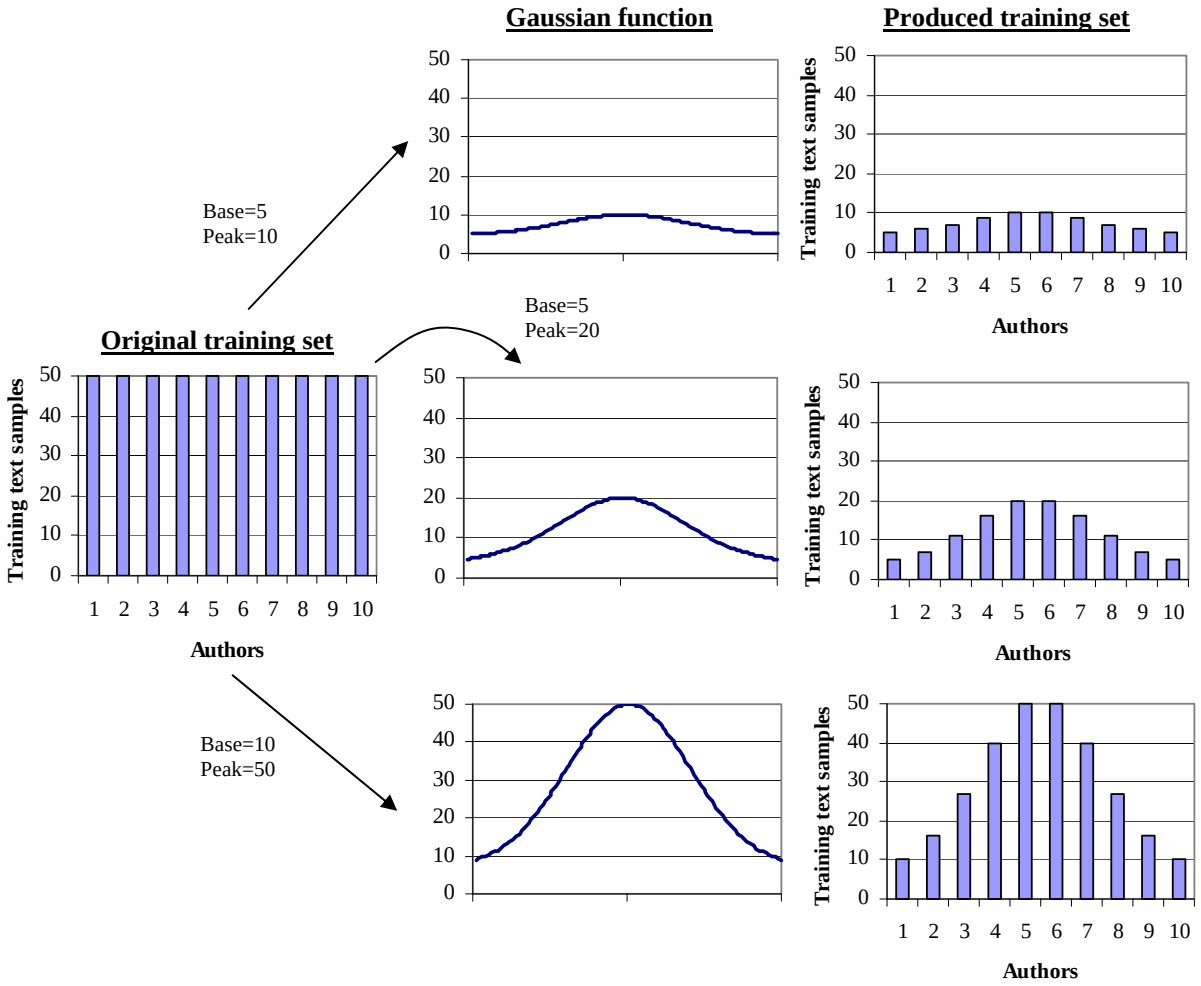

Figure 1. Artificially imbalanced training sets produced using a original balanced training set of 10 authors and Gaussian functions.

3.1 Multi-class imbalanced datasets

In the case of two-class classification problems, the class imbalance can be easily defined as the imbalance ratio of the majority class size to the minority class size. Although, any multi-class problem can be reduced to a series of two-class classifications we need a simpler way of expressing the degree of imbalance of a multi-class problem.

In order to simulate the imbalance conditions of a multi-class real-world authorship identification case, we assume a Gaussian distribution of training texts over the candidate authors. Given this setting, the multi-class imbalance ratio of the problem can be defined as:

$$\textit{Multi-class imbalance ratio} = \textit{peak}/\textit{base} \qquad (1)$$

where *peak* is the size (in training texts) of the biggest class and *base* is the size of the smallest class. Figure 1 shows examples of producing artificial imbalanced distributions of

the training set over 10 classes. Note that all authors have at least *base* training texts while each author has a separate imbalance ratio raging from 1 to *peak*/*base*. Authors near the center of the distribution are considered majority classes while the authors at both sides of the distribution are considered minority classes. By modifying *base* and *peak* values it is possible to construct multi-class imbalanced datasets that resemble a real-world scenario.

Based on the English and Arabic corpora we formed a number of datasets representing different multi-class imbalance conditions. In particular, we applied a Gaussian distribution to the entire training set using different combination of *base* (2, 5, and 10) and *peak* (10 and 50) values giving imbalance ratios from 2 to 25. The produced combinations (cases) can be seen in the first columns of Table 2. For comparative purposes, the balanced cases of equal values of *peak* and *base* (10 or 50) are also considered.

4 EXPERIMENTS

4.1 Tested methods

In order to handle the class imbalance problem, several methods were tested. In more detail:

- Method-1: Under-sampling of the majority classes based on training texts. For all authors, an amount of training texts equal to the base were used. The length of each text is not modified.
- Method-2: Under-sampling of the majority classes based on training text lines. All the available training texts per author were concatenated in one big text. Let x_{min} be the size (in text lines) of the shortest big file. Then, the first x_{min} text lines of each big file were segmented into text samples of length a (in text lines). Note that, in both corpora, each text line comprised at least one full sentence. It was observed that small values of a (2 or 3) tend to provide better results. The results presented in Table 2 correspond to a=3.
- Method-3: Re-balancing the dataset by text samples of variable length. As previously mentioned, one big file is produced per author. Then, each big file is segmented into text samples according to the length of the file. That is, the text samples are of length x_i/k (k is a predefined parameter). Majority authors have long text samples and minority authors have short text samples. Thus, a balanced dataset is produced having k text samples per class. Experiments for k=10, 20, and 50 were performed. Table 2 presents results for k=50 which was the best in the majority of the cases. Note that each text line of the training corpus is used exactly once in the text samples.
- Method-4: Re-balancing the dataset by text re-sampling. Again, one big file is produced per author. Let x_i and x_{max} be the text length (in text lines) of the i-th author and the

longest file, respectively. Then, $k+x_{max}/x_i$ text samples each having x_i/k lines are produced for each author (k is a predefined parameter). Hence, a variable number of text samples is produced per author according to the length of the big file. However, the relation is now inversed. Many short text samples are produced for the minority classes and less but longer text samples are produced for the majority classes. In addition, the text lines included in a text sample are selected randomly. A text line may be included in more than one text sample. 50 runs of this method were performed and the average accuracy is presented in Table 2 (for k=50).

Table 1. Details about the text samples per author produced by the examined methods (a and k are predefined parameters).

	Amount of text samples per author	Length of text samples (in text lines)	Re-sampling
Method-1	*Base*	Defined by source	No
Method-2	x_{min}/a	a	No
Method-3	k	x_i/k	No
Method-4	$k+x_{max}/x_i$	x_i/k	Yes

Table 1 summarizes information about the amount of text samples per author and the text-length (in text lines) of each text sample produced by each examined method. The last column of this table indicates whether different text samples of the same author may share some text lines or not. Note that although method-1, method-2, and method-3 produce balanced training sets, method-4 produces an imbalanced training set. However, the originally minority classes are now represented by more samples in comparison to the originally majority classes. This is clarified in Figure 2. Figure 2 (left) depicts the distribution of the original training set, Figure 2 (right) shows the training set distribution produced by method-1, method-2, method-3 and method-4, respectively. Recall that the length of the text samples of the original training set is determined by the source of the texts. On the other hand, the length of the training text samples produced by methods 3 and 4 depends on the size of the class. Majority classes have longer text samples while minority classes have shorter text samples. Moreover, method-4 attempts to compensate this disadvantage of minority classes by adding more text samples into the training set. The smaller a class is in the original training set, the more training samples are produced for it.

4.2 Evaluation results

In order to evaluate the performance of a method handling the class imbalance problem we need a baseline. For each case, the baseline accuracy is provided when no special technique is used to re-balance the training set (each training text is considered as unit and all training texts are used by the classification model). Moreover, the balanced cases of having 10 or 50 training texts per author were also examined for comparative purposes. The latter case (i.e.,

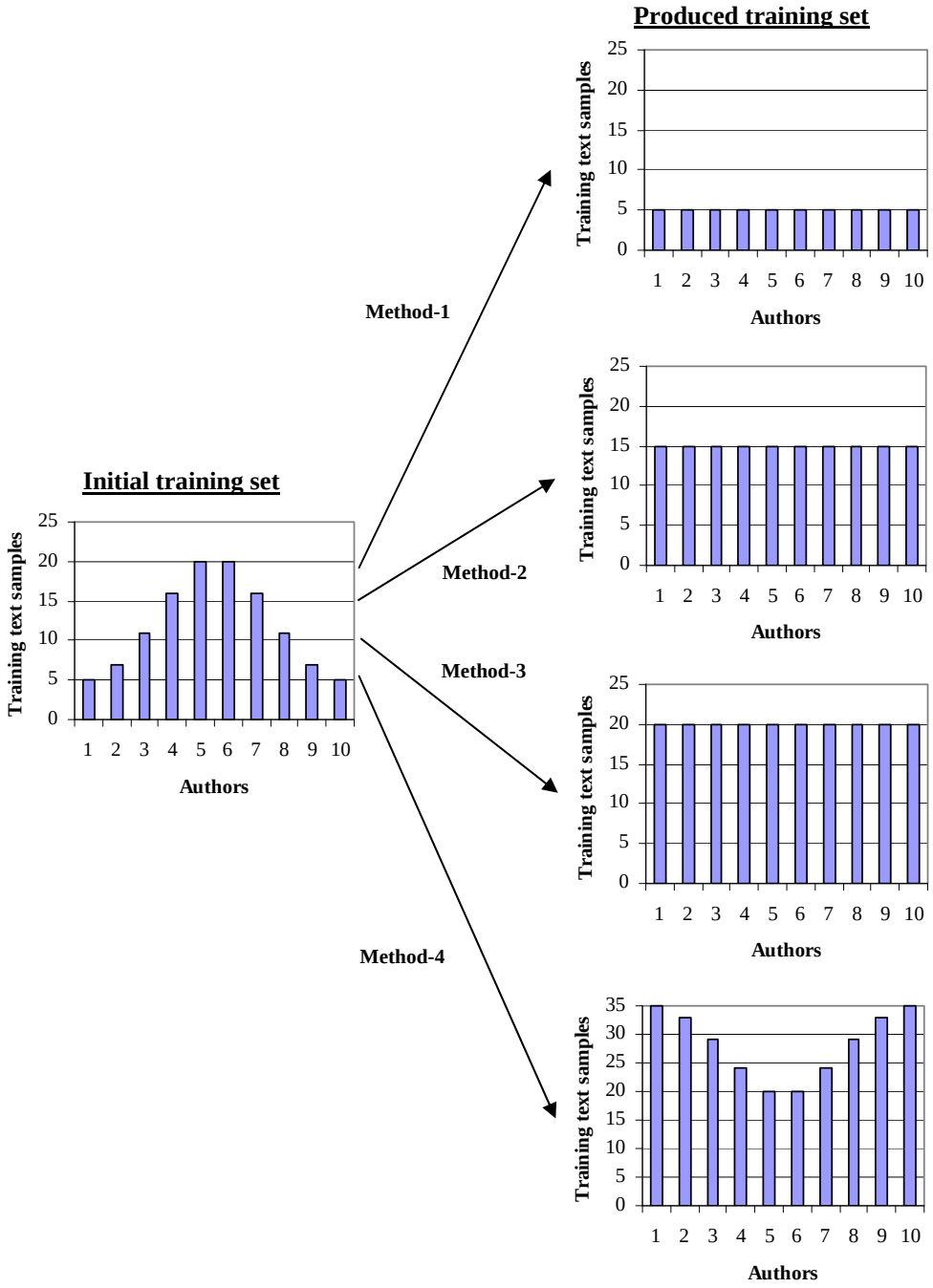

Figure 2. An initial training set distribution over 10 authors and the training sets produced by the methods examined in this study.

base=*peak*=50) could be viewed as an indicative upper bound for the performance of the remaining imbalanced cases since it is based on a significantly larger training set.

Table 2 shows the performance of the aforementioned methods on the English and the Arabic corpus. Note that in all cases the test set is the same (50 texts per author) and not overlapping with the training set. So, for a given corpus, the accuracy results obtained by different methods are directly comparable. For the balanced cases (first two lines of each corpus) the method-1 is exactly the same with the baseline approach. Note that this is an indication of the

difficulty of the two corpora. Hence, the English corpus is more difficult (recall that in average the English corpus texts are shorter than the Arabic corpus texts). The first important point is that some of the examined methods achieve to improve the performance of the baseline approach in three out of four balanced cases (ratio=1). This can be explained by the fact that Method-3 and Method-4 take into account the differences in text-length among the candidate authors. That is, despite the fact that we have equal amount of training texts per author, the variation in text-length produces another type of imbalanced training set.

Table 2. Microaverage accuracy results for the presented methods on different balanced and imbalanced cases of English and Arabic datasets. Best results for each case in boldface.

Case				Accuracy (%)				
Corpus	Base	Peak	Ratio	Baseline	Method-1	Method-2	Method-3	Method-4
English	50	50	1	79.4	**79.4**	78	78.8	79.1
	10	10	1	61.4	61.4	60	65	**65.4**
	5	10	2	60	49.8	59.8	**65.2**	58.52
	10	20	2	62.8	61.4	**68.8**	65.6	66.18
	5	20	4	56.4	49.8	59.8	59	**61.44**
	2	10	5	51.2	52.4	56.8	56.6	**57.34**
	10	50	5	65	61.4	68.8	68	**69.18**
	2	20	10	47.4	52.4	**56.8**	51.8	55.38
	5	50	10	59.6	49.8	59.8	62.4	**65.76**
	2	50	25	53.6	52.4	56.8	54.6	**59.02**
Arabic	50	50	1	93.4	93.4	93	**93.6**	**93.6**
	10	10	1	74.8	74.8	39	**76.4**	76.3
	5	10	2	54.2	42.2	38.2	59	**60.2**
	10	20	2	74.6	74.8	57.4	77.6	**78**
	5	20	4	57.4	42.2	38.2	**64.6**	62.4
	2	10	5	44	33	**50.2**	48.4	49
	10	50	5	67.6	74.8	57.4	**83.2**	80
	2	20	10	54	33	50.2	58.2	**59.4**
	5	50	10	66.2	42.2	38.2	**72.6**	72.2
	2	50	25	61.6	33	50.2	66.2	**66.6**

Considering the imbalanced cases (ratio>1), method-1 fails to outperform the baseline in most of the cases. Recall that this method does not take into account the full training set available per author. Method-2 is better than the baseline for the English datasets but worse on the Arabic datasets. Recall that this method depends on the number of available text lines per author. Although Arabic texts were longer than English in average, x_{min} was roughly half of the corresponding value of the English datasets. However, this method was the best in three cases. It has also to be noted that low *a* values (many short training text samples per class) were found to perform better in most of the cases. The performance of method-3 was really competitive, especially for the Arabic datasets. However, method-4 was superior in the majority of the cases.

As concerns the imbalance ratio of the cases, it does not seem to be strongly relevant with the success of a particular method. In the Arabic datasets, for a given *base*, the best accuracy results are significantly improved as *peak* increases. This means that the extra training texts that become available as *peak* increases significantly contribute to the classification model. This happens to a lower extent in the English datasets. However, recall, that the English datasets have lower baseline accuracy.

Table 3. Performance of the presented methods on the English corpus (*base*=5, *peak*=20). Identification accuracy (%) per author is indicated for the lower bound method. The identification accuracy per author for the rest of the methods is expressed as deviation from the lower bound.

Author	Training Set	Baseline	Method-1	Method-2	Method-3	Method-4
A1	5	36	+20	+36	+10	+18
A2	7	26	+22	+28	-16	+12
A3	11	66	-22	-30	-4	-18
A4	16	38	-16	+34	+6	-4
A5	20	100	-12	-8	0	0
A6	20	100	-18	-4	0	-8
A7	16	98	-74	-36	0	-10
A8	11	56	-26	-38	+12	-2
A9	7	6	+12	+10	0	+12
A10	5	38	+48	+42	+18	+46
Accuracy		56.4	-6.6	+3.4	+2.6	+4.6

A closer look to the identification results will reveal significant properties of these methods. Table 3 shows the identification accuracy per author for the English corpus imbalanced by *base*=5 and *peak* =20. The second column indicates the number of training texts available for each author while the third column shows the results for each author using the baseline approach. The performance of methods 1 to 4 is indicated as deviation from the baseline. As can be seen, the performance of the baseline method roughly resembles the distribution of training texts over the authors. That is, the more training texts available for one author, the better the identification accuracy. Method-1 improves the accuracy for the minority authors (A1, A2, A9, and A10) but fails to keep the accuracy of the majority authors on high level. Method-2 achieves better results. It improves the identification for the minority authors (more or less the same with method-1) without a dramatic loss in majority authors. Method-3 achieves to retain the identification accuracy for the majority authors on very high level (it even improves some of them) but it fails to significantly improve the minority authors. On the other hand, method-4 considerably improves minority authors with the cost of a slight reduction on accuracy for the majority authors.

5 CONCLUSIONS

Many text categorization tasks, including authorship identification, suffer from the class imbalance problem. Extremely few training texts are available for some authors while plenty of training texts are available for other authors. We presented an approach to handle multi-class imbalanced textual data effectively in order to re-balance the training set in favor of the minority classes. To this end, various text sampling and re-sampling methods were examined. The main idea of the most successful method was to produce many short text samples for the minority classes and less but longer text samples for the majority classes. Since textual data can be easily segmented in small pieces, they can be handled more flexibly in comparison to other kinds of data.

A character *n*-gram representation was used in order to quantify the stylistic choices of the authors. Although it requires higher dimensionality, the sparseness of the data is significantly reduced in comparison with word-based approaches. This enables efficient representation for short text samples (e.g., each comprising 1 to 5 text lines).

By following method-4, it is easy to construct synthetic data by concatenating text lines selected randomly from the available training texts. Recall that in the used corpora each text line comprised at least one full sentence. A basic assumption of this method is that such a text sample will still resemble the style of the author. To this end, the bag of *n*-grams representation is quite suitable since it is practically independent of the context of words or even sentences. It remains to be tested whether this method can be applied to topic-related text categorization tasks as well. On the other hand, a more sophisticated approach could be followed in order to select the most suitable text lines to form as good training text samples as possible.

The basic methods presented here can be combined in order to further improve the results. For instance, method-3 and method-4 can be applied together in order to train an enhanced classification model. Alternatively, they could be used to train different classification models which, then, can be combined in an ensemble of classifiers. Recall from Table 3 that the classification errors made by these methods are to a great extent uncorrelated, a crucial condition to build effective ensembles.

An important factor, not considered in this paper, is the amount of candidate authors. Both corpora were based on ten different authors. Although this number seems sufficient for many real-world author identification cases, it should be tested whether the presented methods are affected by scaling into more/less classes. Another interesting direction is the examination of different text representations for authorship identification. To this end, both word based schemas and variable-length character *n*-grams could be tested.

REFERENCES

Abbasi, A. & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 20(5), 67-75.

Argamon, S., Saric, M., & Stein, S. (2003). Style mining of electronic messages for multiple authorship discrimination: First results. In Proc. of the 9th ACM SIGKDD (pp. 475-480).

Burrows, J.F. (1987). Word patterns and story shapes: The statistical analysis of narrative style. Literary and Linguistic Computing, 2, 61-70.

Chaski, C. (2001). Empirical evaluations of language-based author identification techniques. Forensic Linguistics, 8(1), 1-65.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

de Vel, O., Anderson, A., Corney, M., & Mohay, G.M. (2001). Mining e-mail content for author identification forensics. SIGMOD Record, 30(4), 55-64.

Estabrooks, A., Jo, T., Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 20(1), 18-36.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing, 13(3), 111-117.

Japkowicz, N. & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-450.

Juola, P. (2004). Ad-hoc authorship attribution competition. In Proc. of the Joint ALLC/ACH Conference (pp. 175-176).

Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In Proc. of the Conference Pacific Association for Computational Linguistics (pp. 255-264).

Kjell, B, Woods, W.A., & Frieder, O. (1994). Discrimination of authorship using visualization. Information Processing and Management, 30(1), 141-150.

Labbé, C. & Labbé, D. (2001). Inter-textual distance and authorship attribution: Corneille and Molière. Journal of Quantitative Statistics, 8, 213-31.

Lewis, D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, 5, 361-397.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2, 419 – 444.

Mosteller, F. & Wallace, D. (1984). Applied bayesian and classical inference: The case of the Federalist papers. Springer-Verlag, New York.

Peng, F., Shuurmans, D., Keselj, V., & Wang, S.(2003). Language independent authorship attribution using character level language models. In Proc. of the 10th European Chapter of the Association for Computational Linguistics (pp. 267-274).

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471-495.

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (pp. 199-206).

Vapnik, V. (1995). The nature of statistical learning theory. Springer, New York.

Veropoulos, K., Campbell, C., & Christianini, N. (1999). Controlling the sensitivity of support vector machines. In Proc. of the 16th Int. Joint Conference on Artificial Intelligence (pp. 55-60).

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: writing style features and classification techniques. Journal of the American Society of Information Science and Technology, 57(3), 378-393.