Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.5, No.4, 2014

Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data

Fatma Howedi[1*] Masnizah Mohd[2]

1. School of Computer Science, Universiti Kebangsaan Malaysia, Bangi, Malaysia
2. Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

*fatma.howdy1@gmail.com

Abstract

Authorship attribution (AA) is the task of identifying authors of disputed or anonymous texts. It can be seen as a single, multi-class text classification task. It is concerned with writing style rather than topic matter. The scalability issue in traditional AA studies concerns the effect of data size, the amount of data per candidate author. This has not been probed in much depth yet, since most stylometry researches tend to focus on long texts per author or multiple short texts, because stylistic choices frequently occur less in such short texts. This paper investigates the task of authorship attribution on short historical Arabic texts written by10 different authors. Several experiments are conducted on these texts by extracting various lexical and character features of the writing style of each author, using N-grams word level (1,2,3, and 4) and character level (1,2,3, and 4) grams as a text representation. Then Naive Bayes (NB) classifier is employed in order to classify the texts to their authors. This is to show robustness of NB classifier in doing AA on very short-sized texts when compared to Support Vector Machines (SVMs). Using dataset (called AAAT) which consists of 3 short texts per author's book, it is shown our method is at least as effective as Information Gain (IG) for the selection of the most significant n-grams. Moreover, the significance of punctuation marks is explored in order to distinguish between authors, showing that an increase in the performance can be achieved. As well, the NB classifier achieved high accuracy results. Since the experiments of AA task that are done on AAAT dataset show interesting results with a classification accuracy of the best score obtained up to 96% using N-gram word level 1gram.

Keywords: Authorship attribution, Text classification, Naive Bayes classifier, Character n-grams features, Word n-grams features.

1. Introduction

Authorship attribution (AA) is a research field of stylometry, which consists of deciding for a specific text which the author has in written, usually from a predefined set of the authors. Authors have their own style of speaking and writing. The writing style can be used as distinctive features to recognize its author. Broadly speaking, writing style can be viewed as the underlying methods of sentences constructions that can be analysed by examination of a variety of elements, like the sequences of words, the frequency of characters, and the length of sentences. The accepted assumption behind AA is that every author writes in a distinct way; some writing characteristics cannot be manipulated by the writer's will, and therefore can be identified by an automated process (Zhao 2007). In general, applications of AA include plagiarism detection, deducing the author of inappropriate communications, that were sent anonymously or under a pseudonym, and resolving historical questions of unclear or disputed authorship. In recent years, practical applications for AA have grown in areas such as criminal law, civil law and computer security.

Authorship attribution is a kind of classification problem but it is different from text classification (TC), because in AA the writing style is also important besides the text content, which is the only factor used in text classification. In addition, the features in AA task are not deterministic as in text classification. Also, with different data such as books and articles the classifiers and feature sets may behave differently in AA (Bozkurt et al. 2007). Therefore, these differences make AA task more challenging. In text classification task the texts should be assigned to one or more predefined classes based on the topics, while in AA task the texts should be assigned to one or more predefined classes based on the authors (Zhao 2007). Thus the texts in AA are classified into different classes by their authors.

In recent years, the field of AA has witnessed a number of studies on very short texts in many languages. Short text authorship attribution constitutes a particular challenge to the TC approach we adopt, and by extension to any approach. Working with short texts requires robust and reliable representation of such texts as well as a Machine Learning (ML) algorithm that is able to be handled with limited data. In most studies, texts of book

length are used for training phase, while studies with short text are relatively rare. In Stamatatos (2009), it is reported that the samples of texts should be long enough therefore the text representation features can sufficiently represent their style. Luyckx, (2010) showed that reducing the length of the training samples has a direct impact on performance. Traditionally, 10,000 words per author are considered to be a reliable minimum for an authorial set (Burrows 2007). Some studies have shown promising results with short texts of 500 characters (Sanderson & Guenter 2006) or 500 words (Koppel et al. 2007). Siham and Halim (2012) stated that the longer is the text; the better is the identification accuracy. This paper uses short texts between 290 and 800 words per text. This allows us to probe the scalability of the proposed approach with limited training data and very short text documents.

There is a long history of linguistic and stylistic investigation into AA. That is why, the present paper proposes an overall research work of AA that handles 30 different texts written by 10 ancient Arabic travelers who wrote several books describing their travels. A special Arabic dataset has been built in order to assess several features and classifiers that are usually employed in stylometry. This paper focuses on employment Naive Bayes classifier in the classification task, due to the fact that, using Naive Bayes classifier only requires a small amount of training data to estimate the parameters necessary for classification. Since independent variables are assumed, only the variances of the variables for each label need to be determined and not the entire covariance matrix (Bhargavi & Jyothi 2009). In contrast to NB, this study also implements SVM which is more suited to extremely big datasets (Elayidom et al. 2013). Therefore for attribution problems with limited data, SVMs may not superior to other learning approaches (Zhao 2007). We use the SVM for comparison to work by Siham and Halim (2012) where the SVMs did not give good results with classification failure using the word-based features (classification about 10% and 70%).

2. Dataset Description

The dataset is collected from an Arabic website. Also, all the texts that have been collected are about Travel subject. We collect the same dataset as Siham and Halim (2012) used in their work. The dataset is called AAAT dataset (*i.e Authorship attribution of Ancient Arabic Texts*). We collect the dataset from "Alwaraq library" website by choosing 10 different authors, and collecting 3 Arabic texts per author. Table 1 views the authors and the average size of texts for each author. These texts are stored in standard text files of format (.txt) in 10 folders, where each folder stores 3 texts per author.

Table 1. Size of the texts in terms of words

Author Designation	Text 1	Text 2	Text 3	Total size per word	Average size in %
Author 1	630	605	308	1543	514
Author 2	575	540	598	1713	571
Author 3	657	800	290	1747	582
Author 4	599	593	593	1785	595
Author 5	459	511	722	1692	564
Author 6	511	559	636	1706	568
Author 7	599	460	541	1600	533
Author 8	515	653	578	1746	582
Author 9	322	629	548	1499	551
Author 10	591	345	353	1289	429

3. Size of the Texts

As we can see in Table 1 each author is presented by 3 different texts. The texts are very short with the average of about 550 words; moreover some of the texts have even less than 300 words. This situation involves bad experimental conditions since it has been shown in some researches done by Eder (2010) and Signoriello et al. (2005) that the minimum number of words per text should be 2500 words in order to obtain a good attribution results. We have chosen to use short text documents in order to evaluate the Naive Bayes classifier using different features. Because we know when to use short texts, the AA performances decrease. Also to investigate

which features show more robustness to the effect of short texts size. We also compared two ML algorithms in terms of their ability to deal with limited data, using 3 texts per author. Because of testing the system in very limited data, we can estimate its viability when it applied to small collections of texts (Luyckx 2010).

4. Methodology

In this section various steps are considered. Firstly, data pre-processing, secondly feature extraction, then classification and finally author identification. In this paper, we approach AA as a classification task. Where most of TC systems apply two stages approach which first extracts features with high predictive value for the classes, then it trains an ML algorithm to classify new documents by employing the selected features in the first stage. Automatic TC labels documents according to a set of pre-defined authorship classes (Zhao 2007). The TC approach we apply to the AA task begins from a set of pre-processed data, the data is separated into train and test sets. In the first phase, predictive features are extracted from the data, after that training and test instances are created, on the basis of these features. In the second phase, an ML model is built from training data, so as to be tested on unknown test data. The training and test instances are numerical features vectors that represent term frequency of every selected feature, followed by the author label. We perform supervised classification, the situation in which labeled training data are used to train a machine learner, as it allows the evaluation of classification, and thus is the best technique for investigating the scalability of the TC approach (Luyckx 2010). Also the task of AA here is conducted as multi class AA.

4.1 Text Pre-processing

In this step, the data is sent to pre-processing algorithm, for tokenization, punctuation marks removal, and normalization. This step of text pre-processing is crucial in determining the quality of the next stages, feature extraction and classification stage.

In normalization process, some of Arabic letters such as (alef) is normalized into all its forms from (آ, إ, أ) to (ا). All letters are converted to the same case of its forms to more accurately reflect the dimensionality of the vector space. For tokenization process, data is processed into word grams (and character grams) by tokenizing words (and characters) on whitespace in order to facilitate the use of n-gram feature. In terms of punctuation mark removal, all punctuation marks (e.g. \:;.,‘"!?) are removed from the texts of each document by replacing all these punctuation marks with the empty string in all experiments of word-level n-grams. However, in character-level grams we include these punctuation marks in specific experiments of character-level grams for all numbers n of n-grams. This because punctuation marks can present author style. For instance, while some authors rarely use the exclamation marks, some others use the exclamation marks in more cases, some authors use dot frequently because they like short sentences while others use comma more frequently by using long sentences in their writing. These types of details in the texts have vital importance in AA task (Takc & Ekinci 2007). Thus two experiments were done for each character n-gram feature, one includes all punctuation marks, and another excludes them.

4.2 Extracting Features

An important stage is a process of dataset to find distinctive features which exhibit the writing style of each an authorship individually. Assumption that every style of each author has particular features can be accessible to exploit these stylomatric features. In contemporary research, we can discern four main kinds of features that carry potential cues for authorship: lexical, character, syntactic, and semantic features. In this paper, we report on experiments using lexical and character features, since they are more reliable than semantic features, considering the state of the art in semantic analysis; and are most commonly applied behind syntactic features. The features we use are listed in Table 2. Character-n-grams have been demonstrated that they are able to reliably handle limited data (Luyckx 2010), which is why we tried them for short text AA. The success of character-n-grams can be explained through their ability to capture nuances in different linguistic levels (Houvardas & Stamatatos 2006), but mainly for their ability to handle limited data. While, lexical-n-grams features are by far the most widely used kind of features, the idea that they contain interesting stylistic information is rather than intuitive. Whereas researches done by Hong et al. (2010) and Luyckx (2010) stated that lexical features are good for small datasets.

Table 2. Character and lexical features used in this study

Feature	N-grams level	Description	Feature Type
Character Uni-gram	1-gram	individual characters	Character
Character Bi-gram	2-grams	two consecutive characters	
Character Tri-gram	3-grams	three consecutive characters	
Character Tetra-gram	4-grams	four consecutive characters	
Word Uni-gram	1-gram	single words	Lexical
Word Bi-gram	2-grams	two consecutive words	
Word Tri-gram	3-grams	three consecutive words	
Word Tetra-gram	4-grams	four consecutive words	
Rare Words		low frequency	

As we can see from the Table 2, n-gram based approaches can operate at either word level or character level. In using such techniques, a text document or a piece of text is regarded as a sequence of n words (or n characters), where n is the number of words (or characters), in that document. For instance, in the word unigrams a list of all individual words (where n=1 gram) that occur in the text are constructed; this represents the list of features. For each text the occurrence of a feature is accounted. If a text contains more words, then the probability of occurrence of a word in the text increases. Unigrams data are built on the character level in the same manner, where each feature indicates an individual character rather than an individual word. Rare words are the words that are repeated in a text in low frequency. Rare words in this paper are all single words that occur in less than 3 times in each document per author.

4.3 Feature Selection Methods

The aim of feature selection methods is to reduce the dimensionality of dataset by removing irrelevant features for the classification task (Ikonomakis 2005). Some types of features, such as character and lexical features can considerably increase the dimensionality of the features' set (Stamatatos 2009). In such case, feature selection methods can be used to reduce such dimensionality of the representation. In this way the classification algorithm is helped to avoid overfitting on the training sets. The most important criterion of feature selection in AA task is their frequency (Stamatatos 2009). In general, the more frequent feature is, the more stylistic variation it captures. Two feature selection methods including chi-squared (x^2) and information gain are used in this paper. What is common to these selecting features methods is that they come to conclusion by ranking the features by their independently determined scores, and then select the top scoring features using evaluation function.

4.3.1 Chi-Squared

Chi-x^2 is a statistical feature selection method and it measures divergence from the expected distribution, assuming that feature occurrence is independent of class value. It is also generalized well to multi-class problems (Nicolosi 2008). It is based on the chi-squared distribution in the fields of statistics and probability. This method has been used in many studies in TC in common, and in AA specifically, one of recent studies are Grieve (2007) and Luyckx (2010).

4.3.2 Information Gain

Information Gain (IG) represents the entropy reduction given a certain feature, that is, the number of bits of IG about the class by knowing the presence or absence of a term in a document. Since IG considers each feature independent of others, it offers a ranking of the features depending on their IG score, thus a certain number of features can be selected easily (Houvardas & Stamatatos 2006). It can also generalize any number of classes (Nicolosi 2008). One of the studies that used IG as a baseline feature selection method in AA was by Stamatatos and Houvardas (2006).

4.4 Classification

In our experiments, the authorship is classified using Naive Bayes learners and 3-fold cross validation which divides the AAAT dataset into training and test data. The description of Naive Bayes classifier and 3-fold cross validation that we use is as follows:

4.4.1 Three-Fold Cross Validation

As the case of this study, when the classification problem has only a small dataset to work with, it would be difficult to provide enough data for separate training set and testing set. In this case, it is possible to apply *n*-fold cross validation technique, this technique was generally applied in TC and ML research, to provide a more meaningful result by using all the data in dataset as both training and testing data. We performed 3-fold cross validation. Three equally sized parts were randomly created from the dataset. In per fold, the authorship model is trained on two partitions, and tested on the remaining one. The procedure is then repeated thus each fold is held out for testing. This way ensures that all data is used for both training and testing, and that there is no overlap between them. Therefore, the classification task was performed 3 times, each time different partition was used as testing data and the remaining two partitions were used as training. Thus, each partition was used just once as test data. The results of these 3 classification tasks were then combined for calculating the average results for the dataset. This method reduces the variability of the classification.

4.4.2 Applying Naive Bayes classifier

Naive Bayes classifier builds a probabilistic model of each authorship class based on training data of that class. Then it calculates and multiplies the probabilities of all features to give the probability of test text. The highest probability among all authors is most likely an author of that anonymous or test text. The problem arises using the Naive Bayes in that the test data contains features in which the model has not been seen in training data (Boutwell 2011). So some probabilities yield zero result since none of the training data falls in the range. These zero counts have a zero probability, leaving the Naive Bayes classifier unable to predict a class. Thus, Laplace correction parameter is applied. Laplace correction parameter is an expert parameter. This parameter indicates whether Laplace correction should be used for preventing high influences of zero probability or not (Akthar & Hahne 2012). By using Laplace correction parameter there is a simple process to avoid zero probabilities. This is by adding a value of one "1" to every count in the dataset, that we need would make only a negligible difference in estimated probabilities. This prevents a zero probability situation to ensure that each function has a probability of occurrence based on at least a single count, even if it does not appear in the training data.

4.4.3 Description of the SVM Classifier

The different experiments of authorship attribution are also made using a SVM. The SVM is a very accurate classifier that uses bad examples to form the boundaries of the different classes. On the other hand, SVMs usually require large numbers of samples for training in order to achieve satisfactory effectiveness in TC tasks such as AA. Our classification system was trained using the 3-fold cross validation, with 2 features vectors files which were used to train the SVM classifier and the third one used to test it.

4.5 Evaluation

The performance of the classification algorithms with different selected features on the AAAT dataset is evaluated by looking at standard evaluation metrics. Accuracy is used to indicate the number of correctly classified instances over the total number of test instances by calculating the average of accuracy, as in Eq.(1).

5. Results And Discussion

$$AA.accuracy = \frac{Number of documents that are well attributed}{Total Number of documents} \quad (1)$$

Different experiments of AA are presented on an old Arabic set of texts that were written by ten Arabic travellers. Several features from character and lexical features are tested: characters, characters n-grams, words, words n-grams, and rare words. Table 4.1 displays the accuracy in % of good attribution that is obtained by implementing NB compared to SVM classifier. The classifiers are confronted with an equal number of instances per author, avoiding the case where none of the authorship classes are better represented than others, a situation that could attract misclassifications (Luyckx 2010). Furthermore, NB and SVM classifiers are employed (by using the Rapidminer tool kit) for the task of authorship classification (Akthar & Hahne 2012). We remember that every author has three different texts. We perform 3-fold cross validation. In per fold, two texts were used for the training step and the third one is used for testing. The results of AA experiments were shown in the accuracy percentage of good attribution that were obtained from different values of classification results for each feature

condition by applying chi-squared (x^2) and information gain features selection methods using different features size of 100, 500, and 1000. The features size has an impact on their frequency which is the most important criterion of feature selection in AA task and therefore has an impact on the classification performance (Stamatatos 2009).

5.1 Overall Results

Table 3. The accuracy percentage of good attribution obtained the different features by applying the NB compared to SVM

Feature	Accuracy of good attribution using NB Classifier	Accuracy of good attribution using SVM Classifier
Character Uni-gram	53.33%	50.00%
Character Bi-gram	60.00%	63.33%
Character Tri-gram	93.33%	86.67%
Character Tetra-gram	93.33%	93.33%
Word uni-gram	**96.67%**	76.67%
Word Bi-gram	76.67%	56.67%
Word Tri-gram	40.00%	**20.00%**
Word Tetra-gram	40.00%	26.67%
Rare Words	93.33%	93.33%
Average of the all used features	71.85%	62.96%

As we can observe, Table 3 shows the best classification score that obtained in using both NB and SVM, 96.67%. This number is achieved by applying NB on single word features (word unigram). This is the best score for all features that have been employed in this experiment. The lowest classification result is 20.00% which obtained by applying SVM on word tri-gram. The 20.00% cannot be used to be certain that a text was written by the author or not. This percentage was predicted by the SVM in short texts. This is probably because SVM was designed to handle sparse data with many instances and features, even more than that which was used in the experiments of this paper. In addition, the overall average of accuracy percentage which was obtained by applying NB separately for all features is 71.85%, while by applying SVM, this percentage decreases to 62.96%. It means that the method of this work shows better classification results when compared to Seham and Halim work (2012). Also, we obtained a score of good attribution of 93.33% by using one of the following three features: the character Tri-gram, the character Tetra-gram and rare words. It is important to mention that an accuracy of 93.33% with short text documents is relatively high, since several previous researches done by Eder (2010) and Signoriello et al. (2005) showed that the minimum amount of text should not be less than 2500 words in order to obtain a good attribution results.

Table 4. Percentage Accuracy of Good Attribution Obtained into Some Classes of Features by Using the NB Classifier

Feature	Accuracy of good attribution using NB Classifier
Average of the good features	85.55%
Average of the character n-grams	74.99%
Average of the word n-grams	63.33%

Table 4 summarizes the overall results into some classes of features: the average accuracy for character-grams, average accuracy for word-grams, and average accuracy for the good features, which refers to the features that obtained an accuracy up to 50%. It is clear to see that the average accuracy of the character based features are

better than the average accuracy of the word based features. However, rare words and single words feature (word unigram) seem to be pertinent in this experiment. In overall, the average accuracy for all good features is about 85.55%.

5.2 The Effect of Data Size to Features of Characters and Different Characters n-grams Including vs. Excluding Punctuation Marks

The effect of short-sized texts is also tested to the different character n-grams including vs. excluding punctuation marks. This experiment examines a pre-processing procedure for not removing punctuation marks found in texts for using them as character features. It is shown that this procedure improves the proposed method performance.

Table 5. Percentage accuracy of good attribution obtained with the different character level features including vs. excluding punctuation marks

Feature	Exclude punctuation	Include punctuation
Character	26.67%	53.33%
Character Bi-gram	56.67%	60.00%
Character Tri-gram	93.33%	93.33%
Character Tetra-gram	93.33%	93.33%
Average accuracy of all character features	67.50%	74.99%

In Table 5 it is clear that summarizing the results of the experiments reveal that the use of punctuation marks in the construction of character features improves classification performance. Apparently, punctuation marks provide additional information about the author's writing style. This is possible because some authors use a lot of punctuation marks while others use punctuation marks rarely. The average accuracy of all character features obtained to 74.99% by using punctuation marks as character features, while excluding them have average accuracy of 67.5%. However, character tri-gram and character tetra-gram seem to be in the same average accuracy in this experiment. Also in this table, it can be seen that the most significant improved performance is the best when punctuation marks are added to the character unigrams (or single character) condition which showed the best classification increased from about 26% to 53%, compared to adding punctuation marks to bi-grams condition and to both tri-grams and tetra-grams conditions. This is because when the punctuation marks are combined with the bi-grams they become a part of the character of bi-gram, the bi-grams do not solely exist in characters but may consist of a character and a punctuation mark, which means that the punctuation mark occurs in different bi-grams. In one bi-gram it might occur with one character while in another it occurs with another character. Thus the frequency of these bi-grams does not necessarily increase because of the regular use of the punctuation mark. So the regular use of that punctuation mark may not become evident. Thus the influence of punctuation marks is less when they are added to the construction of bi-grams, tri-grams and tetra-grams than when the punctuation marks are added to the un-grams which have construction of single character. Also, when punctuation marks are added as features, the feature size of AAAT dataset is increased. Thus, this means that the increasing in feature size provided more information about the author of the text. Therefore the punctuation marks are indicative of the author.

6. Conclusion

In this work, an AA task has been experimented on an Arabic set of texts that where written by ten Arabic travellers, each author is presented by 3 different texts. Several state of the art features have been tested for Arabic language and particularly for very short texts (which make the main originality of this research work). The classifier that has been implemented is a Naive Bayes (NB) classifier, by using a 3-fold cross validation. Per fold, 20 texts are used for training the NB classifier and 10 are used for the testing. Experiments of AA of short texts, which have been done separately for each feature on the AAAT dataset using an NB classifier shows the following remarkable points:

- The character-based features are better than the word-based features, depending on the average accuracy of all character-level features compared to the average accuracy of all word-level features of the AAAT dataset.

- A good score of an accuracy of good attribution are achieved by using one of the following features: the character Tri-gram, the character Tetra-gram and rare words (score of 93.33% for every of those features).
- Although, in this investigation, some of the word-based features did not give high results, the word uni-gram features gave the best score for this classifier obtained up to 96% classification accuracy.
- The NB classifier shows good performance in this experiment of AA. This is because the average accuracy percentage obtained the score of 71.85% (score of all used features). When compared to SVM which obtained average accuracy percentage of 62.96%.
- The punctuation marks showed significance improves to distinguish between authors. They achieved an increase in the performance, where the average accuracy for all character features increased from 67.5% to 74.99% including punctuation marks.
- Although the size of the texts was too small (between 209 words and 800 words per text), the performance of the classifier are really interesting (96% an accuracy of the best score attribution).
- This work of AA, which is one of the rare works done on Arabic text, shows a real motivation and interest for this language.

Future work may investigate the robustness of different types of ML algorithms for tasks with many authors and small dataset of texts. It may also expand the scope of the study to investigate additional (combinations of) features.

References

Akthar, F., & Hahne, C (2012), "RapidMiner5 Operator Reference book". By Rapid-I GmbH: www. rapid-i.com. Stockumer Str.475, D-44227 Dortmund: +49(0) 23142578690, p: 971.

Bhargavi, P., & S.Jyothi (2009), "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", *IJCSNS International Journal of Computer Science and Network Security*, Vol.9, No.8.

Boutwell, S. R. (2011), "Authorship attribution of short messages using multimodal features", *Master Thesis of Science in Computer Science, United State Navy B.S. Johns Hopkins University.*

Bozkurt, D., Baglıoglu, O., & Uyar, E. (2007), "Authorship Attribution: Performance of Various Features and Classification Methods" Computer and Information Sciences.

Burrows, J .(2007), "All the way through: testing for authorship in different frequency strata", *Literary and Linguistic Computing 21, 27-47.*

Eder, M. (2010), "Does Size Matter?Authorship Attribution,Small Samples, Big Problem", *Digital humanities conference, London, 132-135*

Elayidom, M. S., Jose, C., Puthussery, A., & Sasi, N. K. (2013), "Text Classification for Authorship Attribution Analysis", Advanced Computing: *An International Journal (ACIJ), Vol.4, No.5, 1-9.*

Grieve (2007), "Quantiative Authorship Attribution: An Evaluation of Techniques", *Literary and Linguistic Computing, 251-270.*

Hong, R., Tan, R., & Tsai, F, S. (2010), "Authorship Identification for Online Text", *International Conference on Cyberworlds, 155-162.*

Houvardas, J., & Stamatatos, E. (2006), "N-gram feature selection for authorship identification". J. Euzennat and J. Domingue (Eds): *Proceeding of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), 77-86.*

Ikonomakis, M., Kotsiantis, & Tampakas, V. (2005), "Text Classification Using Machine Learning Techniques", *Wseas Transactions on Computers, Vol 4 (8), 966-974.*

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007), "Measuring differentiability: Unmasking pseudonymous Authors", *Journal of Machine Learning Research, 8, 1261-1276.*

Luyckx, K. (2010), "Scalability Issues in Authorship Attribution", *PhD Thesis, Faculty of Arts and Philosophy, Dutch UPA University.*

Nicolosi, N. (2008), "Feature Selection Methods for Text Classification".

Sanderson,C. & Guenter, S. (2006), "Short text authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking", *An Investigation. Proceeding of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), 482-491.*

Signoriello, D, J., Jain, S., Berryman, M, J., & Abbott, D. (2005), "Advanced text authorship detection methods and their application to biblical texts", *Proceedings of SPIE, Vol: 6039, Publisher: Spie, 163-175.*

Siham, O. & Halim, S. (2012), "Authorship Attribution of Ancient Texts Written by Ten Arabic Travelers Using a SMO-SVM Classifier", *The 2nd International Conference on Communications and Information Technology (ICCIT): Digital Information Management, Hammamey, 44-47.*

Stamatatos, E. (2009), "A survey of Modern authorship attribution methods", *Journal of the American Society for Information Science and Technology, 538-556.*

Takc, H. & Ekinci, E. (2012), "Character Level Authorship Attribution for Turkish Text Documents", *TOJSAT: The Online Journal of Science and Technology, 12-16.*

Zhao, Y. (2007), "Effective authorship attribution in Large Document Collections", *PhD Thesis, School of Computer Science and Information Technology, RMIT University , Melbourne, Victoria, Australia.*

Masnizah Mohd is a lecturer at the faculty of information science and technology at the Universiti Kebangsaan Malaysia (UKM). She currently serves as the Head of Master programme. She received her PhD in Computer and information science from the University of Strathclyde, UK in 2010. Her PhD thesis was titled "Design and Evaluation of an Interactive Topic Detection and Tracking Interface". She holds M.IT (2002) and B.IT (1999) degrees in information science from the Universiti Kebangsaan Malaysia.Her main research interests are in the areas of Information Retrieval, Topic Detection and Tracking, and Natural Language Processing; with particular focus on aspects such as user interaction, user interface, named entity recognition, user tasks and evaluation. Currently she is a member of the Center for Artificial Intelligence Technology (CAIT) and she is in the Knowledge Technology (KT) research group.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Recent conferences: http://www.iiste.org/conference/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar