Journal of King Saud University – Computer and Information Sciences (2014) 26, 473–484

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Naïve Bayes classifiers for authorship attribution of Arabic texts

Alaa Saleh Altheneyan [a,1], Mohamed El Bachir Menai [b,*]

[a] Department of Information Technology, College of Computer and Information Sciences, King Saud University, P.O. Box 89638, Riyadh 11692, Saudi Arabia
[b] Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

Available online 28 September 2014

KEYWORDS

Authorship attribution;
Arabic language;
Naïve Bayes classifier;
Event model

Abstract Authorship attribution is the process of assigning an author to an anonymous text based on writing characteristics. Several authorship attribution methods were developed for natural languages, such as English, Chinese and Dutch. However, the number of related works for Arabic is limited. Naïve Bayes classifiers have been widely used for various natural language processing tasks. However, there is generally no mention of the event model used, which can have a considerable impact on the performance of the classifier. To the best of our knowledge, naïve Bayes classifiers have not yet been considered for authorship attribution in Arabic. Therefore, we propose to study their use for this problem, taking into account different event models, namely, simple naïve Bayes (NB), multinomial naïve Bayes (MNB), multi-variant Bernoulli naïve Bayes (MBNB) and multi-variant Poisson naïve Bayes (MPNB). We evaluate these models' performances on a large Arabic dataset extracted from books of 10 different authors and compare them with other existing methods. The experimental results show that MBNB provides the best results and could attribute the author of a text with an accuracy of 97.43%. Comparison results with related methods indicate that MBNB and MNB are appropriate for authorship attribution.

1. Introduction

Authorship attribution is a subfield of authorship analysis. It is the process of attributing the author of an anonymous text based on its characteristics (Juola et al., 2006). This problem has a long history; studies of authorship attribution can be traced back to the 19th century. The early traditional approaches were human expert-based, then from 1964 up until the 1990s, the non-traditional authorship attribution studies were performed. The focus of research at that time was on defining features that measure the writing style of authors. In recent

* Corresponding author. Tel.: +966 1 4670687.
E-mail addresses: atheneyan@ksu.edu.sa (A.S. Altheneyan), menai@ksu.edu.sa (M.E.B. Menai).
[1] Tel.: +966 504 882499.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.006

years, the development in fields such as information retrieval, machine learning and natural language processing has had a great impact on authorship attribution studies (Stamatatos, 2009). Authorship attribution can be used in a broad range of applications in diverse areas, including intelligence, criminal and civil law, computer forensics, and cybercrime investigation as well as in the traditional application to literary research.

A large number of methods have been developed to tackle the authorship attribution problem. These methods can be divided into three classes based on their approach: the unitary invariant approach, multivariate analysis and machine learning approach (Koppel et al., 2009). These methods rely on the linguistic devices used unconsciously by authors, such as semantic, syntactic, lexicographic, orthographic and morphological devices. Although the Arabic language is one of the official languages of the United Nations and is widely used by hundreds of millions of people, only a very small number of authorship attribution studies have been published for Arabic texts so far (Shaker and Corne, 2010).

Naïve Bayes classifiers have been used for authorship attribution in many languages, including English (Hoorn et al., 1999; Zhao and Zobel, 2005; Tan and Tsai, 2010; Pillay and Solorio, 2010), Turkish (Türkoğlu et al., 2007), and Mexican (Coyotl-Morales et al., 2006)). However, there is generally no mention of the event model used. Naïve Bayes classifiers have also been used for Arabic text classification (El Kourdi et al., 2004; Al-Salemi, 2011; Al-Shammari, 2010; Alsaleem, 2011; Noaman et al., 2010). The results provided by the classifier were encouraging.

In this paper, we propose to investigate the naïve Bayes event models for Arabic authorship attribution because they have not been considered for this problem before. Four naïve Bayes event models are examined in this study, namely, the simple naïve Bayes (NB), multinomial naïve Bayes (MNB), multi-variant Bernoulli naïve Bayes (MBNB) and multi-variant Poisson naïve Bayes (MPNB). The rest of this paper is organized as follows. Section 2 presents a general overview of authorship attribution and writing style features. In Section 3, characteristics of the Arabic language are discussed. In Section 4, an extensive study of the different authorship attribution methods is provided along with a study of the available feature selection methods. In Section 5, the naïve Bayes event models are described. In Section 6, the Arabic authorship attribution system is detailed. Section 7 presents and discusses experimental results. Finally, a general conclusion of this work is presented in Section 8.

2. Background

2.1. Authorship attribution

Authorship attribution addresses the problem of determining the author of an anonymous text from a set of candidate authors based only on internal characteristics of the text. It fits a typical text classification problem where each author represents a class (Koppel et al., 2009). The main key research topics in authorship attribution are feature selection and attribution techniques.

2.2. Writing style features

Writing style features are extracted text characteristics that assist the attribution of texts (Abbasi and Chen, 2005a). According to authorship attribution studies, taxonomies of many feature sets exist, which can be categorized into: lexical, character, syntactic, semantic, content-specific, structural and language-specific (Abbasi and Chen, 2005a,b; Stamatatos, 2009).

- *Lexical:* Lexical features are one of the earliest and most traditional features used for attributing authorship. Examples of these features are word length, sentence length, word frequencies, and vocabulary richness. One main problem with lexical features is that in some oriental languages (e.g., Chinese), there are no boundaries separating words, making it hard to apply these measures without requiring special tools.
- *Character:* Character-based measures treat texts as a sequence of characters. There are several measures, such as character type, letter frequencies and character *n*-grams. The significance of character *n*-gram measures is that they can capture lexical information and contextual information. They can be applied to any language easily without requiring any special tools. However, the dimensionality of the representation is very high compared to the lexical approach because of redundant information (e.g., |or_|, |_or|).
- *Syntactic:* Syntactic features are used by authors unconsciously, which makes them more reliable than lexical features. Different syntactic measures were used in attribution studies, including part-of-speech (POS) frequencies, rewrite rule frequencies, syntactic errors and function words. These features require accurate language dependent tools to extract them.
- *Semantic:* Current natural language processing (NLP) tools for handling semantic analysis are not sufficient. As a result, only a few attempts to exploit semantic features have been performed. These features include semantic dependencies, synonyms and the most significant systemic functional linguistics (SFL), which define functional words summed with POS features.
- *Content-specific:* Content-specific features are used when the available texts for all authors are of the same content. Content-specific words are key words of a particular topic that can be used to aid other stylistic features.
- *Structural:* These features capture the habits of an author when organizing a text. They were defined as a result of applying authorship attribution to emails and online forum messages. Examples of these measures include paragraph length, use of signature, font color and font size. The structural features are significant when attributing short texts because it is hard to capture stylistic properties of the text.
- *Language-specific:* These features are specific for a particular language. Measures for these features have to be defined manually.

Lexical, character, syntactic and semantic features can be extracted from any text independent of the application or text language by using the appropriate tools. According to attribution studies, lexical and syntactic features are the most used features for attribution (Abbasi and Chen, 2005a; Stamatatos, 2009).

3. Arabic characteristics

The complex linguistic structure of the Arabic language introduces several challenges: inflection, diacritics, word length, and elongation.

- Inflection

Arabic is a highly inflectional language. Stems are derived from roots by adding affixes (prefixes, infixes and suffixes). Words are a result of adding affixes to stems (e.g., root: كتب stem: مكتب word: المكاتب) (de Roeck and Al-Fares, 2000). Inflection increases the number of words, which might cause particular problems when extracting lexical features, e.g., some vocabulary richness measures will not be that effective (Abbasi and Chen, 2005a).

- Diacritics

Diacritics are special marks placed above or below letters to represent short vowels. The use of diacritics changes both the pronunciation and meaning of words. However, diacritics are rarely used in writings because the readers are expected to infer the missing short vowels using their semantic knowledge of the language. However, for computers, it is not possible for feature extraction programs to infer this knowledge.

This might reduce the effectiveness of using function words as a feature. For example, without using diacritics, the function words مَن (man) and مِن (men) are identical, and computers cannot distinguish between them (Abbasi and Chen, 2005a; Farghaly and Shaalan, 2009).

- Word length

Arabic words tend to be short. This might reduce the effectiveness of lexical features, such as word length distribution (Abbasi and Chen, 2005a).

- Elongation

Elongation is the use of a special dash between two letters in Arabic writing for purely stylistic reasons. Although elongation can be used as a significant attribution feature, it causes a problem when extracting lexical features, especially the word length feature because some word lengths double after using elongation. For example, the word مكتب is a four-letter word. After the addition of four dashes, the elongated word مـــكتب is eight letters (Abbasi and Chen, 2005a).

4. Literature review

4.1. Feature selection

One of the major issues that need to be considered when tackling an authorship attribution problem is the high dimensionality of the feature set, especially when using lexical and character features because every word and phrase represents a feature. Feature selection is essential to reduce the feature set, speed up the computation and improve the classification process (Yang and Pedersen, 1997; Forman, 2003). Feature selection methods can be classified into two main approaches: wrappers and filters.

Wrappers select feature subsets using classical search methods in artificial intelligence (e.g., hill-climbing and beam search) that explore the search space for appropriate feature subsets. Each subset is evaluated using the induction algorithm, which is a time-consuming operation. Therefore, wrappers are unpractical for large-scale problems (Forman, 2003).

Filters methods use feature scoring measures to score each feature independently. The feature subset is then formed by choosing a predefined number of the best features. A number of effective feature scoring measures are used with texts such as: chi-squared χ^2 (CHI), document frequency (DF), information gain (IG), term strength (TS), mutual information (MI), odd ratio (OR), cross entropy (CE), Weight Of Evidence (WOE), random, Ng-Goh-Low (NGL) coefficient, Galavotti–Sebastiani–Simi (GSS) coefficient, and term frequency–inverse document frequency (TF–IDF) (Forman, 2003).

4.2. Authorship attribution approaches

Authorship attribution methods fall into three main classes: unitary invariant, multivariate analysis and machine learning classes.

4.2.1. Unitary invariant

Unitary invariant is the oldest approach used to attribute the author of a text. It uses a single textual feature to discriminate between authors, such as sentence length and word length (Koppel et al., 2009). Mendenhall (1887) used curves that represent word length frequencies to attribute text to Marlowe, Bacon or Shakespeare. Yule (1939) examined the authorship of De Imitatione Christi (a published religious treatise in 1418) and Observations upon the Bills of Morality (an economic writing believed to have been written by either John Graunt or Sir William Petty) using sentence length. Brinegar (1963) also used word length frequencies for the attribution of the Quintus Curtius Snodgrass Letters (10 letters published in the New Orleans Daily Crescent in 1861). None of these methods provided reliable results, which gave way to multivariate analysis methods.

4.2.2. Multivariate analysis

The multivariate analysis method uses a set of features to statistically attribute texts. Mosteller and Wallace (1964) initiated the use of this method by proposing the use of Bayesian statistical analysis to attribute the Federalist Papers (a number of political newspaper essays written by John Jay, Alexander Hamilton, and James Madison; both Hamilton and Madison claimed that they wrote 12 of these essays). Their method based on the most frequent function words provided reliable results, which encouraged scholars to explore other types of features and techniques.

Principal component analysis (PCA) (Pearson, 1901) is a statistical analysis method that uses as few features as possible to examine the variation in texts. It was used for the authorship attribution of many disputed documents (Binongo and Smith, 1999; Holmes et al., 2001a,b; Baayen et al., 2002; Binongo, 2003). Linear discriminant analysis (LDA) (Fisher, 1936) is another statistical method used for attribution (Baayen et al., 1996; Stamatatos et al., 2000; Baayen et al., 2002; Chaski, 2005).

Distance-based methods attribute the author of an anonymous text by measuring the distance between the anonymous text and the available documents written by the candidate authors using some distance measures (Burrows, 2002; Keselj et al., 2003; Hoover, 2004; Juola, 2005; Zhao et al., 2006; Zhao and Zobel, 2007; Zhao and Vines, 2007; Koppel et al., 2010).

Other statistical techniques based on Markov chains were used for authorship attribution (Khmelev and Tweedie, 2001; Kukushkina et al., 2001). Data compression techniques were also considered, including the Lempel and Ziv (LZ77) compression method used by Benedetto et al. (2002) for authorship attribution, the prediction by partial matching (PPM) text compression scheme used by Teahan and Harper (2003) for text categorization, and the R-measure based method proposed by Khmelev and Teahan (2003) for plagiarism detection and text categorization.

4.2.3. Machine learning methods

Supervised machine learning methods are applied on training documents represented as vectors of features to build classifiers that attribute anonymous documents. Various machine learning methods have been used for authorship attribution such as naïve Bayes (Hoorn et al., 1999; Zhao and Zobel, 2005; Coyotl-Morales et al., 2006; Türkoğlu et al., 2007; Tan and Tsai, 2010; Pillay and Solorio, 2010), Bayesian classifiers (Kjell, 1994; Zhao and Zobel, 2005; Zhao et al., 2006; Pillay and Solorio, 2010), K-nearest neighbor (Hoorn et al., 1999; Zhao and Zobel, 2005; Türkoğlu et al., 2007), decision trees (Zheng et al., 2003; Zhao and Zobel, 2005; Zheng et al., 2006; Türkoğlu et al., 2007; Pillay and Solorio, 2010), neural networks (Hoorn et al., 1999; Zheng et al., 2003; Zhao and Zobel, 2005; Zheng et al., 2006; Türkoğlu et al., 2007) and support vector machines (SVM) (Diederich et al., 2003; Zheng et al., 2003; Argamon and Levitan, 2005; Sanderson and Guenter, 2006; Zhao et al., 2006; Zheng et al., 2006; Türkoğlu et al., 2007; Pavelec et al., 2007).

4.3. Authorship attribution of Arabic text

Abbasi and Chen (2005a) used support vector machine (SVM) and C4.5 decision trees on political and social Arabic web forum messages from Yahoo groups for authorship analysis. They preprocessed the texts before extracting the features to remove elongation using an elongation filter; however, the number of elongation characters and elongated words were calculated to be used later as features. The feature set used by Abbasi and Chen (2005a) consists of 410 features including lexical features such as frequent roots and sentence length, syntactic features such as function words and structural features such as the number of attachments and content-specific features. These features were partitioned into different sets for testing as follows:

set1: Lexical features
set2: Lexical + syntactic features
set3: Lexical + syntactic + structural features
set4: Lexical + syntactic + structural + content-specific features

A cluster algorithm by de Roeck and Al-Fares (2000) was used to extract the roots and to use them as features. In each experiment, five authors were selected and for each author, 20 texts were used. The best results were achieved when SVM and set4 features were used.

Abbasi and Chen (2005b) also used SVM and C4.5 with lexical, syntactic, structural and content-specific features. They also used a filter to remove elongations and the cluster algorithm of de Roeck and Al-Fares (2000) to extract root words. They tested their method on English and Arabic web forum messages. The Arabic set was extracted from a Yahoo group forum for the Al-Aqsa Martyrs group using four different sets as in Abbasi and Chen (2005a). For each experiment, five authors with 20 texts for each were used. The best average precision obtained was 94.83% for Arabic and 97.00% for English, when all four sets of features were used with SVM.

Abbasi and Chen (2006) used SVM and writeprint, an authorship visualization, which creates patterns for different author writing styles using a number of documents written by them. They tested their method on the same data set used by Abbasi and Chen (2005b), which consists of a group of 10 messages for each author. Writeprint outperformed SVM when testing the attribution of a group of messages written by one author. However, SVM performed better when testing the attribution of a single message.

Stamatatos (2008) tested the use of SVM on Arabic newspaper reports from an Alhayat newspaper. The aim of the study was to propose a solution for the class imbalance problem: some authors have long and diverse training documents, while others have only a few short documents. He concluded that the best results are obtained when the method uses many short texts for some authors and a few long texts for the others.

Shaker and Corne (2010) used linear discriminant analysis (LDA) for the attribution of 12 Arabic books. They used function words as a feature and started with 104 words of common conjunctions and prepositions. Then, they built their data set based on the English set used by Mosteller and Wallace (1964); however, only 64 words were used because they omitted the forty most frequently used words from the set. For the selection of function words, an evolutionary search was used to choose subsets of function words. Two authors were selected for each of the experiments conducted. For each author, two books were selected: one for testing and the other for training. The books were divided into 1000-word chunks for the first experiment and 2000-word chunks for the second experiment with both 65 and 54 function words. The best performance obtained was 87.63% accuracy, when 2000-word chunks and 54 function words were used.

5. Naïve Bayes models for arabic authorship attribution

Let a, A, f, and n denote an author, the total number of authors, a feature, and the total number of features, respectively. For the naïve Bayes classifier, a set of training documents is provided for each author $a \in A$. Each document is represented by a set of features $\{f_1, f_2, \ldots, f_n\}$. A new document is described by the same set of features $\{f_1, f_2, \ldots, f_n\}$, and the learner is asked to predict the author of the new document, assuming that the occurrences of the features are mutually independent (Mitchell, 1997).

5.1. Simple naïve Bayes

The simple naïve Bayes classifier (NB) attributes a new document with a set of features $\{f_1, f_2, \ldots, f_n\}$ to the most probable target author a according to Eq. (1).

$$a = argmax_{a \in A} P(a|f_1, f_2, \ldots, f_n) \quad (1)$$

The probability $P(a|f_1,f_2,\ldots,f_n)$ needs to be calculated for each $a \in A$ using the following Bayes formula:

$$P(a|f_1,f_2,\ldots,f_n) = \frac{P(f_1,f_2,\ldots,f_n|a)\cdot P(a)}{P(f_1,f_2,\ldots,f_n)} \quad (2)$$

where $P(f_1,f_2,\ldots,f_n)\neq 0$.

Assuming the uniformity of $(f_1,f_2,\ldots,f_n)$, Eq. (2) can be simplified into Eq. (3).

$$P(a|f_1,f_2,\ldots,f_n) = P(f_1,f_2,\ldots,f_n|a)\cdot P(a) \quad (3)$$

By using the chain rule, we obtain:

$$P(f_1,f_2,\ldots,f_n|a)\cdot P(a) = P(a).\prod_{i=1}^{n} P(f_i|a) \quad (4)$$

Therefore, an author a is attributed according to Eq. (5)

$$a = argmax_{a\in A} P(a)\prod_{i=1}^{n} P(f_i|a) \quad (5)$$

where the probability $P(a)$ is estimated by the frequency of a in the training data.

$$P(a) = \frac{\text{number of documents written by } a}{\text{total number of documents}} \quad (6)$$

$P(f_i|a)$ can be estimated using a Gaussian distribution (Zhao and Zobel, 2005) or Laplacian prior (Al-Salemi, 2011):

$$P(f_i|a) = g(f_i,\mu_i,\sigma_i)$$
$$g(f_i,\mu_i,\sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i}} e^{\frac{(f_i-\mu_i)^2}{2\sigma^2}} (\sigma > 0) \quad (7)$$

where μ_i is the mean value of feature f_i in documents written by author a and σ_i is its standard deviation.

The Laplacian prior is given by Eq. (8)

$$P(f_i|a) = \frac{1 + D_{ai}}{A + D_a} \quad (8)$$

where D_{ai} is the total number of documents written by a and containing f_i and D_a is the total number of documents written by α Absent features can cause zero probabilities, which mislead the classifier. To overcome this problem, the number of documents D_{ai} is primed with a count of one using a Laplacian prior. Continuous features such as word length, vocabulary richness and sentence length can only be calculated using a Gaussian distribution.

5.2. Multinomial naïve Bayes

The multinomial model captures feature frequency information (Yang and Liu, 1999). So, instead of representing a document as a set of features $\{f_1,f_2,\ldots,f_n\}$, such as in the simple model, the document is represented as a vector $v = v_1, v_2, \ldots, v_n$, where v_i is the frequency of f_i in the document. So, the new document is attributed to the most probable target author a according to Eq. (9).

$$a = argmax_{a\in A} P(a)\prod_{i=1}^{n} P(v_i|a) \quad (9)$$

The probability $P(v_i|a)$ is calculated using Eq. (10) (Manning et al., 2008),

$$P(v_i|a) = \frac{1 + v_{ia}}{n + n_a} \quad (10)$$

where v_{ia} is the frequency of feature f_i in documents written by author a, n_a, is the total number of features in documents written by author a and a Laplacian prior is used to prime feature frequency with one to avoid the zero probability problem.

5.3. Multi-variant Bernoulli naïve Bayes

The multi-variant Bernoulli naïve Bayes model is similar to the multinomial model, but instead of representing the document as a frequency vector, it is represented as a binary vector (Al-Salemi, 2011) $b =< b_1, b_2, \ldots, b_n >$. If f_i occurs in the document, then $b_i = 1$; otherwise, $b_i = 0$. A new document is attributed to the most probable target author a according to Eq. (11).

$$a = argmax_{a\in A} P(a)\prod_{i=1}^{n} (b_i P(f_i|a) + (1 - b_i)(1 - P(f_i|a))) \quad (11)$$

The probability $P(f_i|a)$ is calculated using Eq. (8).

5.4. Multi-variant Poisson naïve Bayes

The Poisson statistical distribution is commonly used for modeling random events in a fixed unit of time. Poisson distribution has been used for text classification in English (Kim et al., 2006; Huang and Li, 2011). A document is represented as a random vector $x = x_1, x_2, \ldots, x_n$, where x_i is a Poisson random variable assigned the value v_i from within the term-frequency of feature f_i (Kim et al., 2006). The attribution of a new document to the most probable target author a is given by Eq. (12).

$$a = argmax_{a\in A} P(a)\prod_{i=1}^{n} e^{-\lambda_{ia}} \lambda_{ia}^{v_i} \quad (12)$$

When using MNB, MPNB and MBNB, some features such as word length are not suitable. Ideal features are "frequency-based" features because documents are represented as frequency and binary vectors.

The probability λ_{ai} is calculated by Eq. (13),

$$\lambda_{ai} = \frac{c_1 + f_{ai}}{c_2 + D_a} \quad (13)$$

where $c_1, c_2 \in [0, 1]$.

6. Arabic authorship attribution system

In this section, we describe the main components of the system that we implemented to test the four naïve Bayes event models for Arabic authorship attribution. The four main phases of the authorship attribution process consist of preprocessing of the texts, extraction of the features, selection of a sub-set of features, and then training and attributing. Fig. 1 illustrates this process.

6.1. Preprocessing

For preprocessing, the following steps were taken:

- *Normalization:* Normalization is used to help overcome the variation in Arabic text representation. We chose the following normalization steps:
  - Use CP1256 for text encoding.
  - Replace إ,آ or أ with bare alif ا.
  - Replace the sequence ىء with ى.

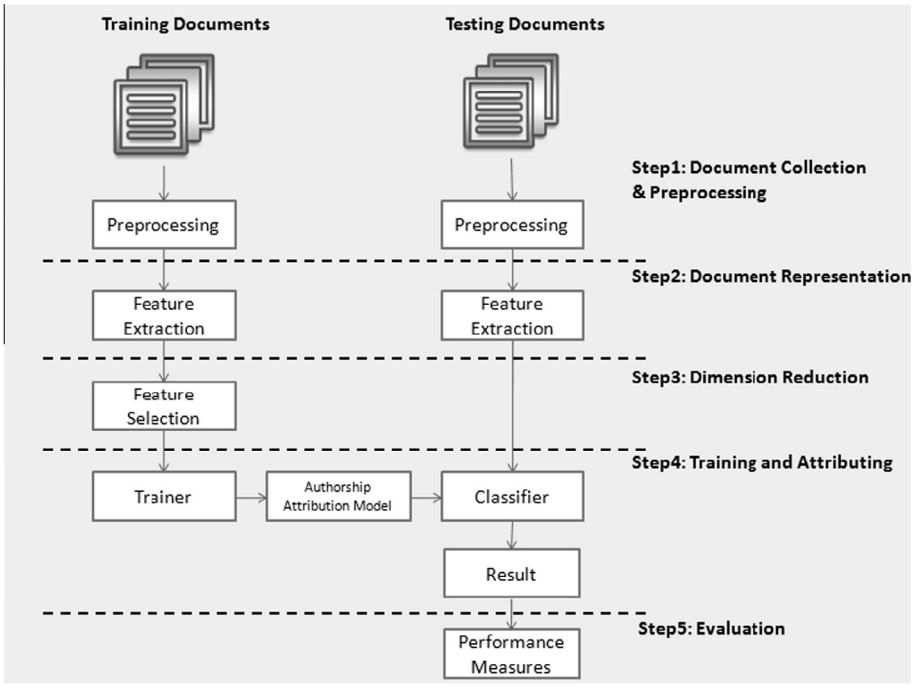

Figure 1 Authorship attribution of Arabic texts.

- Replace final ى with ي.
- Replace final ة with ه.

We implemented our own preprocessing tool. Each document in the training and test sets is preprocessed before extracting its features.

- *Function words, punctuation, diacritics and non-letter removal:* Non-letters, diacritics, punctuation and function words (stop words) are kept because they can provide authorial evidence.
- *Elongation:* Elongation can be used as a significant attribution feature, but it introduces a problem when extracting lexical features, particularly the word length. To overcome this problem, we implemented an elongation filter to extract the number of elongations and the number of elongated words. They are then used as features before removing elongation.
- *Stemming:* Stemming is the process of finding roots for Arabic words. Stemming methods are divided into root-based and stem-based classes. Abbasi and Chen, 2005a,b used the most common roots as features. They used the clustering algorithm of de Roeck and Al-Fares (2000) and a root dictionary to extract the roots, while other features were extracted from the original texts. Stamatatos (2008), Shaker and Corne (2010) did not use any stemming preprocessing. In our work, we used Khojah's stemmer (http://zeus.cs.pacificu.edu/shereen/research.htm) to extract the roots of words.

6.2. Feature extraction

Documents are represented as numerical vectors of features. We used a feature set similar to the one used by Abbasi and Chen, 2005a,b, 2006, which has a total of 408 features for the simple naïve Bayes model and 374 features for the other models. Two hundred distinct words are used as features. A complete description of the feature set is presented in Table 1. The extraction of these features is performed in two steps: first, all of the distinct words are extracted and then, the features and 200 words are selected based on some feature selection method.

6.3. Feature selection

For feature selection, we used term frequency feature selection with the NB model because the calculation of the probability for this model depends on the mean and standard deviation of the features. Chi-squared is used for MNB, MBNB, and MPNB because this measure provided good results when used in Arabic text classification. Indeed, Al-Salemi (2011) used a naïve Bayes classifier with different feature selection methods and showed that chi-squared provided the best result. Chi-Squared also provided the best result among other feature selection methods when tested by Mesleh (2008) with his SVM classifier for Arabic text. Chi-squared was also used by Al-Harbi et al. (2008) and Mesleh (2007).

6.4. Training and attributing

The four models NB, MNB, MBNB, and MPNB are trained and tested on a large Arabic corpus for authorship attribution. Their performance evaluation and comparison are detailed in the next section.

7. Experimental evaluation

The authorship attribution system was implemented using the JAVA programing language under the NetBeans IDE 6.9.1 environment on a personal computer with an Intel Core 2 Duo CPU P8700 @2.53 GHz CPU, a 4-Gbyte RAM and a 32-bit Windows Vista operating system.

Table 1 Extracted features. Note that M = total number of words, V = total number of distinct words, and C = total number of characters in a document.

Feature	Number	Feature extraction tool
Total number of characters (C)	0	Tokenizer
Total number of words (M)	1	Tokenizer
Frequency of elongation	2	Elongation filter
Frequency of elongated words	3	Elongation filter
Total number of punctuations	4	Character dictionary
Total number of whitespace characters {\t \n \f \r space}	5	Character dictionary
Total number of space characters/C	6	Character dictionary
Total number of space characters/number white-space characters	7	Character dictionary
Total number of tab spaces/C	8	Character dictionary
Total number of tab spaces/number white-space characters	9	Character dictionary
Total number of punctuations/C	10	Character dictionary
Number of blank lines/total number of lines	11	Character dictionary
Average sentence length	12–13	Sentence splitter
Frequency of punctuation (8 features) “,”,“?”,“.”, “:”, “!”,“ ’ ”,“ ” ”,”; ”	14–21	Character dictionary
Average word length (number of characters)	22	Tokenizer
Word length frequency distribution/M	23–37	Tokenizer
Total number of function words/M (from Kojah stemmer)	38	Tokenizer
Frequency of function word	39–202	Tokenizer
Total number of hapax legomena/V	203	Tokenizer
Total number of hapax legomena/M	204	Tokenizer
Vocabulary richness i.e., V/M	205–208	Tokenizer
200 distinct words	209–408	Tokenizer

Different performance metrics are used to measure the effectiveness of the attribution methods. Abbasi and Chen (2005a, 2006), Stamatatos (2008), and Shaker and Corne (2010) used the accuracy measure to evaluate the performance of their proposed methods, while Abbasi and Chen (2005b) used the precision measure. To be able to compare our results with the results of the other Arabic authorship attribution systems, we used both accuracy and precision. We also used the recall and F1-measure as additional metrics.

7.1. Arabic corpus

Different Arabic corpora have been used by scholars to tackle the Arabic authorship attribution problem. Abbasi and Chen (2005a, 2006) used Arabic web forum messages from a Yahoo group forum for the Al-Aqsa Martyrs group. Abbasi and Chen (2005a) used political and social Arabic web forum messages from Yahoo groups. Stamatatos (2008) used Arabic newspaper reports from an Alhayat newspaper. Finally, Shaker and Corne (2010) used 12 Arabic books from the Arab Writers Union's website to test their method. We planned to use the same corpus used by Abbasi and Chen (2005a,b, 2006) to effectively compare our results with theirs, but unfortunately, we were not able to obtain it from the authors. We then built our own dataset, which consists of 30 Arabic books written by 10 different authors collected from the Alwaraq website (http://www.alwaraq.net): Alfarabi, Alghazali, Aljahedh, Almas3ody, Almeqrezi, Altabary, Altow7edy, Ibnaljawzy, Ibnrshd, and Ibnsena. Each book is partitioned into chunks of 1980 to 2020 words so that each author has 60 book chunks.

7.2. Sensitivity to stemming and/or normalization

We applied stemming and normalization to our corpus and divided it into four different datasets. The first dataset contains the original documents that have been neither stemmed nor normalized (denoted $-R-N$). The second dataset contains the documents that have not been stemmed, but normalized (denoted $-R+N$).

The third dataset contains the documents that have been stemmed, but not normalized (denoted $+R-N$).

Finally, the fourth dataset contains the documents that have been both stemmed and normalized (denoted $+R+N$). A ten-fold cross validation was used for training and testing of the four naïve Bayes models (NB, MNB, MBNB, and MPNB).

Tables 2–5 summarize the results obtained by the four classifiers on different datasets in terms of the mean and standard deviation of the recall, precision, accuracy, and F1-measure. Note that μ and σ represent the mean value and its standard deviation, respectively. Fig. 2 shows the variation of the mean accuracy for the four classifiers on different datasets. It is clear from the results that the MBNB model achieves the best results, regardless of the dataset used. The MBNB model achieves the best results with 97.43% accuracy and 86.07% F1-measure when applied on data that have been neither stemmed nor normalized ($-R-N$). The MNB model results are worse than MBNB, but it outperforms the MPNB and NB models. Its best performance is obtained when it is applied on data that have been neither stemmed nor normalized ($-R-N$) with an accuracy of 92.03% and an F1-measure of 56.26%. The NB model performance is the worst; its overall accuracy is 82.30%. Normalizing the data prior to its classification has a small impact on the results of its classification. Indeed, the overall results obtained on normalized data show that the accuracy of the different naïve Bayes models increases maximally up to 2.93% and decreases up to 1.64%, in the case of non-stemmed and stemmed data, respectively. However, stemming the data has a larger impact on the results because the accuracy of different models decreases to 7.1%.

Table 2 Results of NB, MNB, MBNB, and MPNB classifiers on the dataset −R + N.

	μ(Recall) (%)	σ(Recall) (%)	μ(Precision) (%)	σ(Precision) (%)	μ(Accuracy) (%)	σ(Accuracy) (%)	μ(F1-measure) (%)	σ(F1-measure) (%)
NB	10.50	18.53	8.93	18.11	82.10	15.24	7.12	12.43
MNB	53.50	38.87	54.29	36.60	89.10	8.31	48.10	31.07
MBNB	**87.00**	**22.34**	**89.14**	**17.49**	**97.40**	**2.66**	**85.48**	**17.32**
MPNB	36.00	37.00	35.00	35.00	87.00	8.00	31.00	30.00

The best results are shown in bold.

Table 3 Results of NB, MNB, MBNB, and MPNB classifiers on the dataset −R − N.

	μ(Recall) (%)	σ(Recall) (%)	μ(Precision) (%)	σ(Precision) (%)	μ(Accuracy) (%)	σ(Accuracy) (%)	μ(F1-measure) (%)	σ(F1-measure) (%)
NB	11.50	19.02	9.05	17.59	82.30	12.78	7.95	12.93
MNB	60.17	35.24	63.64	29.24	92.03	4.30	56.26	26.16
MBNB	**87.17**	**21.16**	**89.44**	**16.31**	**97.43**	**2.73**	**86.07**	**16.26**
MPNB	37.00	39.26	33.62	32.21	87.40	6.79	30.19	28.56

The best results are shown in bold.

Table 4 Results of NB, MNB, MBNB, and MPNB classifiers on the dataset +R + N.

	μ(Recall) (%)	σ(Recall) (%)	μ(Precision) (%)	σ(Precision) (%)	μ(Accuracy) (%)	σ(Accuracy) (%)	μ(F1-measure) (%)	σ(F1-measure) (%)
NB	2.00	5.49	0.65	1.91	80.40	19.67	0.89	2.53
MNB	30.83	39.34	29.47	35.38	86.17	10.53	22.49	24.46
MBNB	**53.67**	**41.18**	**53.11**	**36.62**	**90.73**	**5.47**	**46.60**	**31.66**
MPNB	24.67	34.68	20.16	28.38	84.93	9.14	16.99	20.93

The best results are shown in bold.

Table 5 Results of NB, MNB, MBNB, and MPNB classifiers on the dataset +R − N.

	μ(Recall) (%)	σ(Recall) (%)	μ(Precision) (%)	σ(Precision) (%)	μ(Accuracy) (%)	σ(Accuracy) (%)	μ(F1-measure) (%)	σ(F1-measure) (%)
NB	0.83	2.64	1.63	5.17	80.17	15.98	0.75	2.36
MNB	31.33	40.10	27.38	36.06	86.27	11.23	22.53	25.87
MBNB	**61.83**	**40.24**	**61.26**	**36.26**	**92.37**	**5.14**	**55.62**	**32.81**
MPNB	28.33	39.09	19.35	26.74	85.67	9.86	19.30	25.20

The best results are shown in bold.

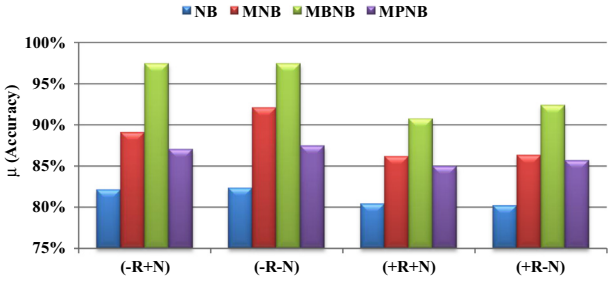

Figure 2 Variation of the mean accuracy for the four classifiers on different datasets.

7.3. Performance comparison

Figs. 3–6 show the average results of applying the different naïve Bayes models on data that have been neither stemmed nor normalized for 10 different authors. The NB model attributed texts with an accuracy of 90% and above to 5/10 authors: Alfarabi, Alghazali, Altabary, Ibnrshd, and Ibnsena. Its lowest performance (approximately 60% accuracy) was given on texts of Ibnaljawzy. The MNB model attributed texts with an accuracy of 90% and above to 7/10 authors: Alfarabi, Alghazali, Almeqrezi, Altabary, Ibnaljawzy, Ibnrshd, and Ibnsena. Its accuracy on texts of Alghazali and Almeqrezi is greater than 96%. The accuracy of the MBNB model is greater than 95% for all authors and exceeds 98% for Alfarabi, Alghazali, Almeqrezi, and Altabary. The MPNB model attributed texts with an accuracy of 90% and above to 2/10 authors: Alghazali and Almeqrezi. However, its lowest accuracy is approximately 82% (Alfarabi).

The confusion matrix, shown in Table 6 for the MBNB model on a single run, demonstrates its high performance in

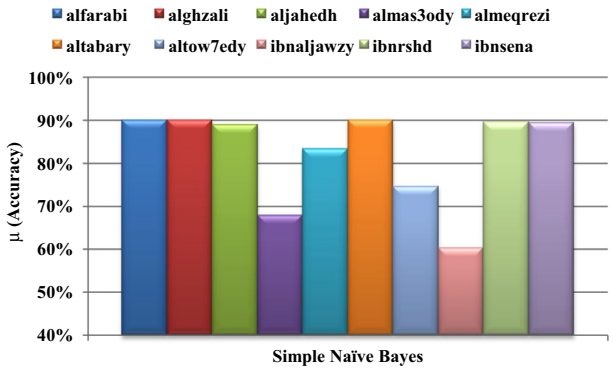

Figure 3 Variation of the mean accuracy for the NB model for different authors.

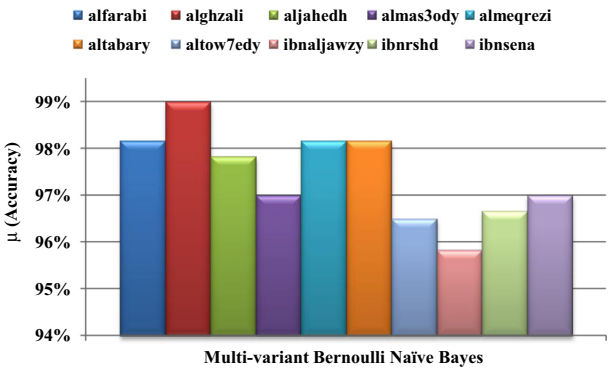

Figure 5 Variation of the mean accuracy for the MBNB model for different authors.

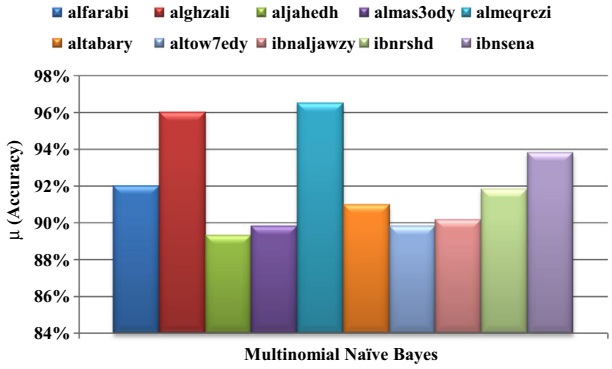

Figure 4 Variation of the mean accuracy for the MNB model for different authors.

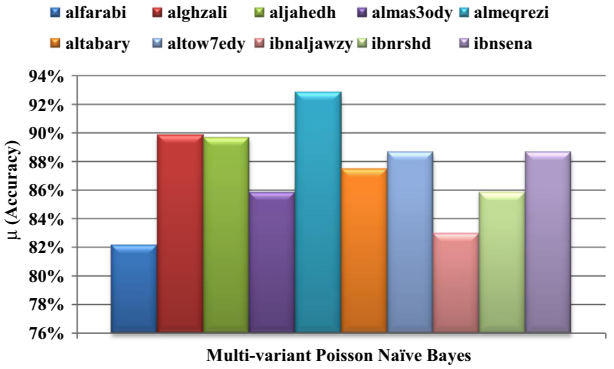

Figure 6 Variation of the mean accuracy for the MPNB model for different authors.

attributing particular texts of Alfarabi, Alghazali, Almas3ody, and Ibnaljawzy. Additionally, it can be used to underline some similarities between authors' styles. For example, it shows that nine texts of Almeqrezi are attributed to Almas3ody because of some common features, including the average word length and frequency of function words. The following is a fragment of a misclassified text:

7.4. Comparison with other methods

For comparative purposes, we considered all of the works conducted to tackle the Arabic authorship attribution problem to the best of our knowledge. Table 7 presents our results and those reported in other references in terms of recall, precision, and accuracy. The results are not in fact directly comparable

"وأوصى عديم الملك ابنه شداب بن عديم أن ينصب في كل حيز من أحياز ولايته مناراً، ويزبر عليه اسمه فانحدر إلى الأشمونين، وعمل مناراتها وزبر عليها اسمه، وعمل بها ملاعب وعمل في صحرائها مناراً أقام عليه صنماً برأسين على اسم كوكبين كانا مقترنين في الوقت الذي خرج فيه إلى اتريب وبنى فيها قبة عظيمة مرتفعة على عمد وأساطين بعضها فوق بعض، وعلى رأسها صنماً صغيراً من ذهب، وعمل هيكلاً للكواكب، ومضى إلى حيز صا فعمل فيه مناراً على رأسه مرآة من أخلاط تورى الأقاليم، ورجع وعمل شداب بن عديم هيكل ارمنت."

An example of Almas3ody's text is:

"وأرسل الله هوداً إلى عاد وهم باحقاف الرمل وملكهم الخلجان بن الوهم، وكانوا يعبدون ثلاثة أصنام وكذبوه، فدعا عليهم فأمسك الله عنهم المطر ثلاث سنين فأجهدهم ذلك فوجهوا الى مكة رجالا يستسقون لهم في الحر. فانتبه القوم لما سمعوا الشعر ونهضوا يستسقون، فلما استسقوا نشأت لهم ثلاث سحائب بيضاء وسوداء وحمراء، ونودي قيل منها اختر لقومك قال البيضاء جهام قد فرغت ماءها، والحمراء ريح والسوداء غيث فاختارها فقيل قد اخترت رماداً رمدداً لا يبقي من عاد أحداً، لا والداً ولا ولداً.فدخلت الريح على عاد من واديهم، فأقامت سبع ليال وثمانية أيام حسوماً، والحسوم الدائمة حتى هلكوا عن آخرهم، وتهدمت ديارهم ولم يمنعهم جدار ولاجبل حتى هلكوا عن آخرهم، ولم يبق إلا رسمهم."

Table 6 Confusion matrix for the MBNB classifier on the dataset −R − N.

	Alfarabi	Alghzali	Aljahedh	Almas3ody	Almeqrezi	Altabary	Altow7edy	Ibnaljawzy	Ibnrshd	Ibnsena
Alfarabi	57	0	0	0	0	0	0	0	2	1
Alghzali	0	56	0	0	0	0	0	0	1	3
Aljahedh	0	0	50	0	0	0	1	9	0	0
Almas3ody	0	0	0	60	0	0	0	0	0	0
Almeqrezi	1	0	0	9	49	0	0	1	0	0
Altabary	0	0	0	7	0	52	0	1	0	0
Altow7edy	1	0	2	2	0	1	40	11	0	3
Ibnaljawzy	0	0	1	0	0	2	0	57	0	0
Ibnrshd	6	1	0	0	0	0	0	0	47	6
Ibnsena	0	1	0	0	0	0	0	0	4	55

Table 7 Comparison of the four naïve Bayes models with other methods. Note that 'NR' means 'Not Reported'.

Reference	Attribution method	Recall	Precision	Accuracy	Data
This paper	NB	11.50%	9.05%	82.30%	Arabic books collected from Alwaraq website
	MNB	60.17%	63.64%	92.03%	
	MBNB	87.17%	89.44%	**97.43%**	
	MPNB	37.00%	33.62%	87.40%	
Abbasi and Chen (2005a)	Decision trees (C4.5)	NR	NR	81.03%	Arabic web forum messages from Yahoo groups
	SVM	NR	NR	85.43%	
Abbasi and Chen (2005b)	Decision trees (C4.5)	NR	71.93%	NR	Arabic web forum messages from Yahoo group forum for
	SVM	NR	**94.83%**	NR	Al-Aqsa martyrs
Abbasi and Chen (2006)	SVM	NR	NR	87.00%	Arabic web forum messages from Yahoo group forum for
	Writeprint	NR	NR	68.92%	Al-Aqsa martyrs
Stamatatos (2008)	SVM	NR	NR	93.60%	Arabic newspaper report of Alhayat
Shaker and Corne (2010)	LDA	NR	NR	87.63%	Arabic books obtained from the website of the Arab Writers Union

The best results are shown in bold.

because they were not obtained on the same datasets. Moreover, the granularities of the tasks vary. However, they can give an indication of the performance of the different methods. It shows that MBNB achieved the best accuracy (97.43%), while the second best accuracy was obtained by an SVM method used by Stamatatos (2008) on Arabic newspaper reports of Alhayat (93.60%). The best precision was obtained by another SVM method used by Abbasi and Chen (2005b) on Arabic web forum messages from a Yahoo group forum for Al-Aqsa martyrs (94.83%), while MBNB achieved the second best accuracy (89.44%).

8. Conclusions and future work

We investigated the applicability of naïve Bayes classifiers and their influence on event models for authorship attribution of Arabic texts. We implemented an authorship attribution system to test and compare four different models of naïve Bayes: NB, MNB, MBNB, and MPNB. MBNB probability estimation depends on the existence or absence of a feature, while MPNB and MNB probability estimations depend on the feature frequency. Probability estimation in the NB model is based on the mean and standard deviation of the features. We evaluated their performance on a large corpus of four different datasets and examined the effect of stemming and normalization on the attribution process. The overall results show that the MBNB model provides the best results among all naïve Bayes models; it was able to attribute the author of a text with an average accuracy of 97.43%. They also show that normalization does not have a large impact on the attribution results, while stemming decreases the efficiency of the classifier because roots provide less authorial evidence than words. The results were compared with those of available methods for Arabic authorship attribution to give an indication on the performance of the naïve Bayes models. These results indicate that MBNB outperforms all of the other methods in terms of accuracy.

As future work, we intend to extend the experiments to larger datasets of more than ten authors. We also plan to investigate the impact of other feature selection methods on the performance of the naïve Bayes models.

References

Abbasi, A., Chen, H., 2005a. Applying authorship analysis to Arabic web content. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H., Merkle, R.C. (Eds.), Intelligence and Security Informatics, vol. 3495. Springer-Verlag, Berlin, Heidelberg, pp. 183–197.

Abbasi, A., Chen, H., 2005b. Applying authorship analysis to extremist-group Web forum messages. IEEE Intell Syst 20 (5), 67–75. http://dx.doi.org/10.1109/MIS.2005.81.

Abbasi, A., Chen, H., 2006. Visualizing Authorship for Identification. IN ISI, pp. 60–71.

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., Al-Rajeh, A., 2008. Automatic Arabic Text Classification. 9th International journal of statistical analysis of textual data, pp. 77–83.

Alsaleem, S., 2011. Automated Arabic Text categorization Using SVM and NB. Int. Arab J. e-Technol. 2 (2), 124–128.

Al-Salemi, 2011. Statistical Bayesian learning for automatic Arabic text categorization. J. Comput. Sci. 7 (1), 39–45. http://dx.doi.org/10.3844/jcssp.2011.39.45.

Al-Shammari, E.T., 2010. Improving Arabic document categorization: Introducing local stem. 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE, pp. 385–390. doi:10.1109/ISDA.2010.5687235.

Argamon, S., Levitan, S., 2005. Measuring the usefulness of function words for authorship attribution. Proc. Joint Conf. Assoc. Comput. Humanities Assoc. Literary Linguist. Comput., 1–3.

Baayen, H., van Halteren, H., Tweedie, F., 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Literary Linguist. Comput. 11 (3), 121–132. http://dx.doi.org/10.1093/llc/11.3.121.

Baayen, H., Halteren, H.V., Neijt, A., Tweedie, F., 2002. An experiment in authorship attribution. 6th JADT, pp. 69–75.

Benedetto, D., Caglioti, E., Loreto, V., 2002. Language trees and zipping. Phys. Rev. Lett. 88 (4), 048702. http://dx.doi.org/10.1103/PhysRevLett. 88.048702.

Binongo, J., Smith, M., 1999. The application of principal component analysis to stylometry. Literary Linguist. Comput. 14 (4), 445–466. http://dx.doi.org/10.1093/llc/14.4.445.

Binongo, J., 2003. Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. Chance 16 (2), 9–17.

Brinegar, C.S., 1963. Mark Twain and the Quintus Curtius Snodgrass letters: a statistical test of authorship. J. Am. Stat. Assoc. 58 (301), 85–96. http://dx.doi.org/10.2307/2282956.

Burrows, J., 2002. ''Delta'': a measure of stylistic difference and a guide to likely authorship. Literary Linguist. Comput. 17 (3), 267–287. http://dx.doi.org/10.1093/llc/17.3.267.

Chaski, C.E., 2005. Who's at the keyboard: authorship attribution in digital evidence investigations. Int. J. Digital Evidence 4 (1).

Coyotl-Morales, R.M., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P., 2006. Authorship attribution using word sequences. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (Eds.), Progress in Pattern Recognition, Image Analysis and Applications, Berlin, Heidelberg, vol. 4225, pp. 844–853.

de Roeck, A.N., Al-Fares, W., 2000. A morphologically sensitive clustering algorithm for identifying Arabic roots. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL'00, Stroudsburg, PA, USA, pp. 199–206. Doi: http://dx.doi.org/10.3115/1075218.1075244.

Diederich, J., Kindermann, J., Leopold, E., Paass, G., Informationstechnik, G.F., Augustin, D.-S., 2003. Authorship attribution with support vector machines. Appl. Intell. 19, 109–123.

El Kourdi, M., Bensaid, A., Rachidi, T., 2004. Automatic Arabic document categorization based on the naïve Bayes algorithm. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Semitic'04, Stroudsburg, PA, USA, pp. 51–58.

Farghaly, A., Shaalan, K., 2009. Arabic natural language processing: challenges and solutions, 8(4), 14:1–14:22. doi:10.1145/1644879.1644881.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics 7 (2), 179–188.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289–1305.

Holmes, D.I., Gordon, L.J., Wilson, C., 2001a. A widow and her soldier: stylometry and the American Civil War. Literary Linguist. Comput. 16 (4), 403–420. http://dx.doi.org/10.1093/llc/16.4.403.

Holmes, D., Robertson, M., Paez, R., 2001b. Stephen crane and the New-York Tribune: a case study in traditional and non-traditional authorship attribution. Comput. Humanities 35 (3), 315–331. http://dx.doi.org/10.1023/A:1017549100097.

Hoorn, J., Frank, S., Kowalczyk, W., van der Ham, F., 1999. Neural network identification of poets using letter sequences. Literary Linguist. Comput. 14 (3), 311–338. http://dx.doi.org/10.1093/llc/14.3.311.

Hoover, D.L., 2004. Testing Burrows's delta. Literary Linguist. Comput. 19 (4), 453–475. http://dx.doi.org/10.1093/llc/19.4.453.

Huang, Y., Li, L., 2011. Naive Bayes classification algorithm based on small sample set. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 34–39. doi:10.1109/CCIS.2011.6045027.

Juola, P., 2005. A controlled-corpus experiment in authorship identification by cross-entropy. Literary Linguist. Comput. 20 (Suppl. 1), 59–67. http://dx.doi.org/10.1093/llc/fqi024.

Juola, P., Sofko, J., Brennan, P., 2006. A Prototype for authorship attribution studies. Literary Linguist. Comput. 21 (2), 169–178. http://dx.doi.org/10.1093/llc/fql019.

Keselj, V., Peng, F., Cercone, N., Thomas, C., 2003. N-gram-based author profiles for authorship attribution. Computat. Linguist. 3, 255–264, Doi: 10.1.1.9.7388.

Khmelev, D.V., Tweedie, F.J., 2001. Using Markov Chains for identification of writer. Literary Linguist. Comput. 16 (3), 299–307. http://dx.doi.org/10.1093/llc/16.3.299.

Khmelev, D.V., Teahan, W.J., 2003. A repetition based measure for verification of text collections and for text categorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 104–110.

Kim, S.-B., Han, K.-S., Rim, H.-C., Myaeng, S.-H., 2006. Some effective techniques for naive Bayes text classification. IEEE Trans. Knowledge Data Eng. 18 (11), 1457–1466. http://dx.doi.org/10.1109/TKDE.2006.180.

Kjell, B., 1994. Authorship attribution of text samples using neural networks and Bayesian classifiers. In: 1994 IEEE International Conference on Systems, Man, and Cybernetics, 1994. Humans, Information and Technology, vol. 2, pp. 1660–1664. Doi:10.1109/ICSMC.1994.400086.

Koppel, M., Schler, J., Argamon, S., 2009. Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. 60 (1), 9–26. http://dx.doi.org/10.1002/asi.v60:1.

Koppel, M., Schler, J., Argamon, S., 2010. Authorship attribution in the wild. Lang. Resour Evaluat. 45 (1), 83–94. http://dx.doi.org/10.1007/s10579-009-9111-2.

Kukushkina, O.V., Polikarpov, A.A., Khmelev, D.V., 2001. Using literal and grammatical statistics for authorship attribution. Probl. Inf. Transm. 37 (2), 172–184. http://dx.doi.org/10.1023/A:1010478226705.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval, 1st ed. Cambridge University Press.

Mendenhall, T.C., 1887. The characteristic curves of composition. Science ns-9, 237–246. http://dx.doi.org/10.1126/science.ns-9.214S.237.

Mesleh, A., 2007. Chi square feature extraction based SVMS Arabic language text categorization system. J. Comput. Sci. 3 (6), 430–435.

Mesleh, A.M., 2008. Support vector machines based Arabic language text classification system: feature selection comparative study. In: Sobh, T. (Ed.), Advances in Computer and Information Sciences and Engineering. Springer, Dordrecht, Netherlands, pp. 11–16.

Mitchell, T.M., 1997. Machine Learning, 1st ed. McGraw-Hill.

Mosteller, F., Wallace, D.L., 1964. Inference and disputed authorship: the federalist. The David Hume series of philosophy and cognitive science reissues. Addison-Wesley.

Noaman, H. M., Elmougy, S., Ghoneim, A., Hamza, T. 2010. Naive Bayes Classifier based Arabic document categorization. In: 2010 The 7th International Conference on Informatics and Systems (INFOS), IEEE, pp. 1–5.

Pavelec, D., Justino, E., Oliveira, L.S., 2007. Author identification using stylometric features. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial 11 (36), 59–66.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2 (11), 559–572.

Pillay, S.R., Solorio, T., 2010. Authorship attribution of web forum posts. eCrime Researchers Summit (eCrime), IEEE, pp. 1–7. doi:10.1109/ecrime.2010.5706693.

Sanderson, C., Guenter, S., 2006. On authorship attribution via markov chains and sequence kernels. In: Pattern Recognition 2006 ICPR 2006 18th International Conference on, 3(X), pp. 437–440.

Shaker, K., Corne, D., 2010. Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. In: 2010 UK Workshop on Computational Intelligence (UKCI), pp. 1–6. Doi:10.1109/UKCI.2010.5625580.

Stamatatos, E., 2008. Author identification: using text sampling to handle the class imbalance problem. Inf. Process. Manage. 44 (2), 790–799. http://dx.doi.org/10.1016/j.ipm.2007.05.012.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60 (3), 538–556. http://dx.doi.org/10.1002/asi.21001.

Stamatatos, E., Kokkinakis, G., Fakotakis, N., 2000. Automatic text categorization in terms of genre and author. Comput. Linguist. 26 (4), 471–495. http://dx.doi.org/10.1162/089120100750105920.

Tan, R.H.R., Tsai, F.S., 2010. Authorship identification for online text. In: Proceedings of the 2010 International Conference on Cyberworlds, CW'10, Washington, DC, USA, pp. 155–162. Doi:10.1109/CW.2010.50.

Teahan, W.J., Harper, D.J., 2003. Using compression-based language models for text categorization. In: Croft, W.B., Lafferty, J. (Eds.), Language modeling for information retrieval. Springer, Netherlands, pp. 141–165.

Türkoğlu, F., Diri, B., Amasyal ±, M.F., 2007. Author attribution of Turkish texts by feature mining. Corpus 1086–1093. http://dx.doi.org/10.1007/978-3-540-74171-8_110.

Yang, Y., Liu, X., 1999. A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'99, New York, NY, USA, pp. 42–49. doi:10.1145/312624.312647.

Yang, Y., Pedersen, J.O., 1997. A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, ICML'97, San Francisco, CA, USA, pp. 412–420.

Yule, G.U., 1939. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. Biometrika 30 (3/4), 363–390. http://dx.doi.org/10.2307/2332655.

Zhao, Y., Vines, P., 2007. Authorship Attribution Via Combination of Evidence. In: Amati, G., Carpineto, C., Romano, G. (Eds.), . In: Advances in Information Retrieval, LNCS, Vol. 4425. Springer, Berlin, Heidelberg, pp. 661–669.

Zhao, Y., Zobel, J., 2005. Effective and scalable authorship attribution using function words. Proc. Second Asian Inf. Retr. Symp. 3689, 174–189.

Zhao, Y., Zobel, J., 2007. Searching with style: authorship attribution in classic literature. In: Proceedings of the thirtieth Australasian conference on Computer science –ACSC'07, Darlinghurst, Australia, Australia, vol. 62, pp. 59–68.

Zhao, Y., Zobel, J., Vines, P., 2006. Using relative entropy for authorship attribution. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (Eds.), . In: Information Retrieval Technology, vol. 4182. Springer, Berlin, Heidelberg, pp. 92–105.

Zheng, R., Li, J., Chen, H., Huang, Z., 2006. A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. 57 (3), 378–393. http://dx.doi.org/10.1002/asi.v57:3.

Zheng, R., Qin, Y., Huang, Z., Chen, H., 2003. Authorship analysis in cybercrime investigation. In: Chen, H., Miranda, R., Zeng, D., Demchak, C., Schroeder, J., Madhusudan, T. (Eds.), . In: Intelligence and Security Informatics, LNCS, vol. 2665. Springer, Berlin, Heidelberg, p. 959.