Information Processing and Management 53 (2017) 1103–1119

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Distance measures in author profiling

Mirco Kocher, Jacques Savoy*

University of Neuchatel, rue Emile Argand 11, 2000 Neuchatel, Switzerland

ARTICLE INFO

Article history:
Received 29 November 2016
Revised 4 April 2017
Accepted 19 April 2017

Keywords:
Distance measure
Author profiling
PAN-CLEF
Text categorization

ABSTRACT

Determining some demographics about the author of a document (e.g., gender, age) has attracted many studies during the last decade. To solve this author profiling task, various classification models have been proposed based on stylistic features (e.g., function word frequencies, *n*-gram of letters or words, POS distributions), as well as various vocabulary richness or overall stylistic measures. To determine the targeted category, different distance measures have been suggested without one approach clearly dominating all others. In this paper, 24 distance measures are studied, extracted from five general families of functions. Moreover, six theoretical properties are presented and we show that the Tanimoto or Matusita distance measures respect all proposed properties. To complement this analysis, 13 test collections extracted from the last CLEF evaluation campaigns are employed to evaluate empirically the effectiveness of these distance measures. This test set covers four languages (English, Spanish, Dutch, and Italian), four text genres (blogs, tweets, reviews, and social media) with respect to two genders and between four to five age groups. The empirical evaluations indicate that the Canberra or Clark distance measures tend to produce better effectiveness than the rest, at least in the context of an author profiling task. Moreover, our experiments indicate that having a training set closely related to the test set (e.g., the same collection) has a clear impact on the overall performance. The gender accuracy rate is decreased by 7% (19% for the age) when using the same text genre during the training compared to using the same collection (leaving-one-out methodology). Employing a different text genre in the training and in the test phases tends to hurt the overall performance, showing a decrease of the final accuracy rate of around 11% for the gender classification to 26% for the age.

1. Introduction

In our digital world, author profiling and authorship attribution are viewed as important questions from a security perspective or regarding the increased number of pseudonymous posts and messages (Olsson, 2008). In literary studies, being able to verify the gender of a given character may open new research directions (e.g., is Juliet really a female figure? (Craig & Kinney, 2009)).

To solve these questions, various approaches have been suggested based on vocabulary richness measures (Holmes, 1998), (Baayen, 2008), stylometric similarities (Burrows, 2002; Savoy, 2012), or machine learning models (Stamatatos, 2009; Jockers & Witten, 2010). In many cases, texts are represented by vectors in which the different dimensions correspond to words, characters, *n*-grams of letters or words, part-of-speech (POS) categories, or other possible stylistic measures (e.g., sentence

* Corresponding author.
E-mail addresses: Mirco.Kocher@unine.ch (M. Kocher), Jacques.Savoy@unine.ch (J. Savoy).

http://dx.doi.org/10.1016/j.ipm.2017.04.004
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

length, lexical density, etc.). These models assume usually that the corresponding dimensions are orthogonal and the number of dimensions varies widely from one model to another (e.g., 3 in (Fung, 2003), 10 in (Mosteller & Wallace, 1964), 2907 in (Jockers & Witten, 2010), and more than 73,000 in (Khonji & Iraqi, 2014)).

To define the exact demographic category of the author, several proposed approaches need to compute a distance (or similarity) measure between the query text and the representations of the different categories. The shortest distance (or the maximum similarity) determines the predicted class. The choice of the distance measure is often based on *ad hoc* considerations, tradition, or limited empirical evidence.

The objectives of this paper are the following three. First, we want to establish a set of useful properties that a distance measure must respect. Second, and based on a large number of different test collections, we want to determine a reduced set of distance measures showing the most effective performance. Third, using a relatively large number of test collections, we have the opportunity to quantify the influence of the training set on the test set. Thus, we want to estimate the possible performance variations when using the same collection during the training and test phases, when using different collections with the same text genre, or when there are different text genres in both stages.

The rest of this paper is organized as follows. The next section presents the state of the art in author profiling with the focus on the gender and age determination. The third section explains the distance measures and the properties we can expect from an effective one in the context of authorship attribution or profiling. In the fourth section, we perform a theoretical assessment of the different distance measures. The fifth section describes the test collections and the evaluation methodology used in the experiments. The evaluation of the different distance measures is exposed in the sixth section, together with the evaluation of different combinations during the training and test phase. A conclusion draws the main findings of this study.

2. Related work

The main objective of an author profiling task is to determine, as accurately as possible, some author's demographics from text (e.g., gender, age, some personality traits, social class, native language, etc. (Argamon, Koppel, Pennebaker, & Schler, 2009)). The gender distinction might be viewed as the simplest one. The classification decision can be binary and a relatively large amount of data can be collected. However, such a classification system can be effective only if the writing style between genders does differ (Eckert & McConnell-Ginet, 2013) and if such stylistic differences can be detected.

Past studies tend to demonstrate that such differences do occur when considering pervasive and frequent features such as determiners, pronouns, or part-of-speech (POS) distributions. According to Pennebaker (2011), women tend to employ more personal pronouns (especially more *I* and *we*) than men (in relative frequencies, 14.2% vs. 12.7% in blog posts). The signal does not seem to be really strong, but it exists. Looking at other lexical groups, Pennebaker (2011) indicates that men tend to employ more big words (composed of six letters or more), determiners, prepositions, nouns, numbers, and swear words. On the other hand, women use more verbs, negations (e.g., never, not), cognitive words (e.g., consider, explain, think), social words (e.g., family, folks), emotion words (e.g., fears, crying, losses) (Talbot, 2010; Rangel & Rosso, 2016), and certainty words (e.g., always, must). Of course, each individual can depict a more or less strong masculine or feminine figure.

As another way to detect the author gender, Alowibdi, Buy, and Yu (2013) suggest taking account of the first names and user names both transformed into phonemes (with the set of possible phonemes limited to 40). With other languages than English, the gender detection can be determined by considering a few words (e.g., in Portuguese, thank you is *obrigado* for a man, and *obrigada* for a woman) (Ciot, Sonderegger, & Ruths, 2013).

For most of those features, simple lists of words can be created mainly because some grammatical categories such as determiners or pronouns form a closed set. Within a given language, a new preposition cannot be created. For other POS such as nouns or verbs, new instances can occur (e.g., to google). Their identification requires however a language-dependant POS tagger. As an alternative, LIWC (Linguistic Inquiry and Word Count) (Tausczik & Pennebaker, 2010) proposes a set of word lists to measure some stylistic features (e.g., determiners, personal pronouns, modal verb forms) as well as other semantic-based categories such as positive emotions or social words.

A simple count based on a single feature cannot provide a reliable measure. The text register has an impact on those predictors, as for example, pronouns are in general less frequent in a formal context. On the other hand, political speeches delivered by US presidents contain more pronouns, even when the context is official (Savoy, 2016). Therefore, generalization based on a single experiment or using a unique text register should be viewed with caution.

As expected, some topical words are used more frequently by one of the genders (e.g., sports, job, money vs. family, shopping, friends) (Schler, Koppel, Argamon, & Pennebaker, 2006). The two genders have their preferred subjects and this aspect is reflected in their lexical choice. Based on around 100,000 blog posts (50% were written by men, 50% by women), the computer can correctly classify 72% of them based on very frequent words (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Argamon et al., 2009) (19,320 authors; mean text length: 7250 words). Including also the topical terms, the machine can reach an accuracy rate of 76%. In this case, men use more terms related to technology (e.g., game, software, Linux) while women prefer writing about friends and social relations (e.g., love, cute, mom). Those examples are however related to the weblog in which other lexical features can be used to discriminate between the two genders (e.g., emoticons (Crystal, 2006)). Changing the text source requires that the most discriminative topical words between the two genders should be redefined (e.g., selecting the 1000 words depicting the highest information gain ratio (Argamon et al., 2009)).

At the syntactical level, differences between genders can be found (Yule, 2010). For example, women tend to use higher-prestige constructions (*I saw it* vs. *I seen it*) more frequently. On the other hand, double negatives (e.g., *I don't want none*) is a structure occurring more with men than women who have a higher sensitivity to linguistics norms (Coates & Pichler, 2011). In dialogue, men are more likely to interrupt women than the opposite (Talbot, 2010).

In the author profiling task at PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse) in CLEF (Conference and Labs of the Evaluation Forum) 2014, there are distance based approaches, but some solutions used distance measures in implicit form. For instance, the best approaches to solve the task were based on machine learning approaches such as SVM (Support Vector Machines) (López-Monroy, Montes-Y-Gómez, Jair-Escalante, & Villaseñor-Pineda, 2014), logistic regression (Maharjan et al., 2014), or a mix of different classifiers (Weren, Moreira, & de Oliveira, 2014). Similarly, in 2015, the best approaches (Álvarez-Carmona, López-Monroy, Montes-Y-Gómez, Villaseñor-Pineda, & Jair-Escalante, 2015; González-Gallardo, Montes, Sierra, Núñez, Adolfo, & Ek, 2015; Grivas et al., 2015) were based on SVM models. Finally, for the cross-genre classification task in 2016, both SVM (Busger Op Vollenbroek et al., 2016) and logistic regression approaches (Modaresi, Liebeck, & Conrad, 2016; Bilan et al., 2016) have demonstrated the highest performances by correctly determining the gender in 3 out of 4 texts on average. The gender detection remains a hard task (Nguyen et al., 2014).

As a second important profiling variable, the author's age was also analyzed by different studies. In this case, the age year cannot be defined precisely and the challenge is usually to predict an age range. Moreover, the target value is the chronological age range and not the psychological one. It is known that differences may occur between them (Yule, 2010). To avoid problems in the limits between two groups, test collections tend to ignore intermediate age groups. For example, the final categories correspond to teenagers ([13–17]), twenties ([23–27]), and thirties and more ([33–47]) (Argamon et al., 2009).

Limited to those three classes, Argamon et al. (2009) achieve an overall accuracy of 66.9% with stylistic features alone, 75.5% when using topical terms alone, and 77.7% when considering both sets of predictors. While prepositions and determiners can characterize the two older classes, the youngest is more associated with contractions (*im, dont, cant*), and as content words with *haha, wanna*, or *school*. Pennebaker (2011) mentions that younger people tend to use more past tense forms while older persons prefer using the future tense. To discriminate between different age ranges, we can consider the average sentence length or the mean word length. Younger people tend to write shorter sentences and use less complex words. This last aspect can be evaluated by considering the mean number of letters per word, with a small value serving as an indicator in favor of a young author.

For Rosenthal and McKeown (2011), combining both internet writing characteristics and lexical features tends to improve the overall performance for age determination. For example, the number of emoticons (e.g., ;-)) decreases with the age as well as the frequency of internet acronyms (e.g., LOL), or slang expressions (e.g., wazzup). The number of URL or links fluctuates with the author age and thus cannot be used as a pertinent feature. Based on Facebook messages, Sap et al. (2014) suggest to build a lexicon of words with their weights reflecting their usage across age and gender. Applying the generated lexicon on other sources (e.g., blogs, tweets) tends to decrease the overall performance of the prediction, indicating that there are stylistic or content differences between the different sources.

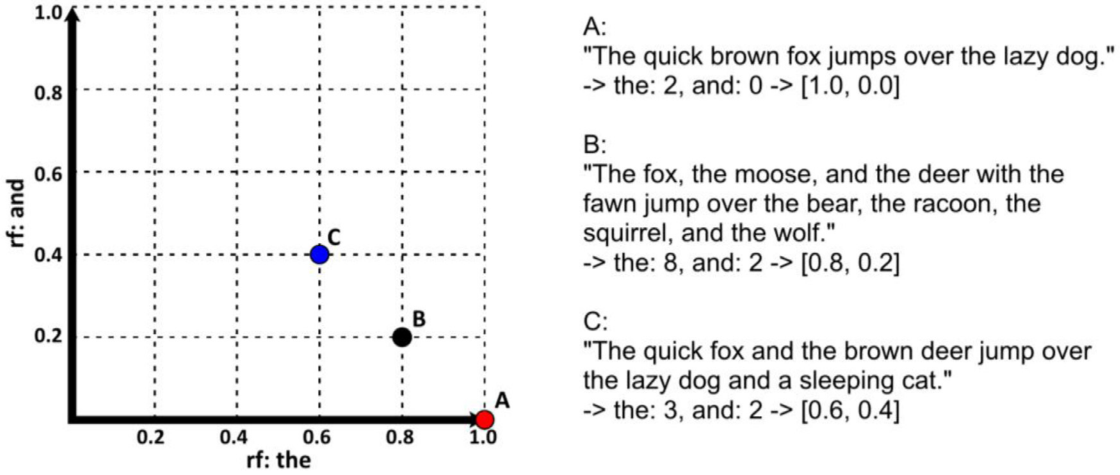
In these previous studies, the main focus is set on determining the most effective features while the choice of the distance (or similarity) measure is usually marginal. In information retrieval (IR) (Manning, Raghavan, & Schütze, 2008; Zhai & Massung, 2016), the relative effectiveness of different similarity measures has been the subject of various studies and evaluation campaigns. As for example, Zobel and Moffat (1998) indicate that the overall effectiveness of a similarity measure depends on the corpus used in the evaluation, the performance measures, and the query type. Thus, a single measure does not always perform better than the others in all contexts. Moreover, implementation details may play a significant role, such as the base for the logarithm, adding one in a formula for smoothing purposes, etc. Gronenschild, Habets, Jacobs, Mengelers, van Os, and Marcelis (2012) and Collberg and Proebsting (2016) made similar findings in evaluating the output of a given system on various platforms. Those results are not directly applicable in our context, in part because in IR the query size is rather short (e.g., composed in mean of one to two words in web search (Manning et al., 2008)) compared to the document length.

3. Distance measures

To build a text classifier, the most effective features are selected and a distance or machine learning approach is applied to determine the author's demographic category. In this paper the most frequent words have been selected as features and various distance measures can then be applied. To be able to discriminate between them, this section presents some useful properties and explains the usefulness of some families of distance measures.

3.1. Distance properties

In the context of authorship attribution or author profiling, the definition of an effective distance measure should not be based on a simple *ad hoc* consideration. A set of properties have to be clearly defined first. To achieve this, in our notation, uppercase letters will denote vectors (or points) while lowercase letters with a subscript indicate the value inside a vector. Thus, A, B, or C specify vectors, a_i indicates the element in the ith position of vector A, and m is the length of the vector.



A:
"The quick brown fox jumps over the lazy dog."
-> the: 2, and: 0 -> [1.0, 0.0]

B:
"The fox, the moose, and the deer with the fawn jump over the bear, the racoon, the squirrel, and the wolf."
-> the: 8, and: 2 -> [0.8, 0.2]

C:
"The quick fox and the brown deer jump over the lazy dog and a sleeping cat."
-> the: 3, and: 2 -> [0.6, 0.4]

Fig. 1. Example of three points and the distance between them.

First, a distance measure must be equal to zero when computing the distance from a point to itself. Nothing is more similar to a vector than the vector itself. Second, all distance measures between two points must be greater than or equal to zero. Negative values that can be useful in some cases, do not have generally a clear semantics in our context. Thus, for now, we will assume that all a_i values are non-negative ($\forall i$: $a_i \geq 0$). Moreover, the distance is zero only when computing the distance from a point to itself. Otherwise, the distance between two distinct points must be greater than zero such that we can detect a difference. Third, it is usually convenient to admit that the distance between two points is symmetric as the ordering of the texts is flexible. Going from A to B corresponds to the same distance as going from B to A. This property is not always satisfied in practice (e.g., the presence of one-way streets). Moreover, in our context, one vector may reflect an author profile or a gender category and thus may correspond to a larger text than the second vector (the query text). Thus, for some measure definitions, the symmetry property is not respected without affecting largely the effectiveness of the distance function. More formally these first three properties can be specified as follows.

P1: Property 1. *Zero distance*
When two vectors are identical, their distance must be zero. $dist(A, A) = 0$.

P2: Property 2. *Positivity*
When two vectors differ, the distance between them must be positive.
$dist(A, B) > 0$.

P3: Property 3. *Symmetry*
Computing the distance from vector A to B must return the same value as computing the distance from B to A. $dist(A, B) = dist(B, A)$.

Furthermore, a distance measure can only be a metric if it respects the previous three properties plus the triangle inequality. This last property specifies that adding a point between two points cannot decrease the distance between the first two points. For instance, when measuring the style difference in the first half of a text and the second half separately, then the sum of those two can not be smaller than when directly measuring the overall change in style. This property is usually respected by the majority of the measures used in practice.

P4: Property 4. *Triangle inequality*
For any triangle, the sum of the distances of any two sides must be greater than or equal to the distance of the remaining side.
$dist(A, C) \leq dist(A, B) + dist(B, C)$.

The next two properties are more specific to our context in which each feature included in a vector usually corresponds to a stylistic marker. The fifth property emphasizes the fact that the absence of a feature used frequently in one vector must have a bigger impact on the distance than the absence of an infrequent element. This property underlines the fact that not all dimensions have the same importance, and frequent features should have a larger influence in the distance measure. This helps to reduce the influence of noise or outliers which could obscure meaningful information.

P5: Property 5. *Frequent feature*
The absence of a frequent feature must be penalized more than the absence of an infrequent one.

Finally, we consider the case when the distance between two pairs of points is equal according to the previous properties. In this situation, the last property indicates that the distance including the presence of a feature must be smaller than when this feature is absent. For example, in Fig. 1 we show three short sample texts and represent them in a vector space that has the relative frequencies of the two words “the” and “and” as its dimensions. The point B is located in equidistance between points A and C. As shown in the figure, $dist(A, B) = dist(B, C)$. To respect this last criterion, when comparing two cases returning the same distance, the preference has to be given to the vector pair depicting the presence of more features. In our example, the distance between B and C should be viewed as smaller than the distance between B and A since A is missing one information that was important to B.

P6. *Presence of the feature*

When the distance measure returns the same value, the presence of a feature is better than the absence of it.

3.2. L^p family

The distance measures can be regrouped under different families (Cha, 2007; Duda, Hart, & Stork, 2001; Manning et al., 2008) where the most frequent one is the L^p family (or L^p norm). In this paradigm, when changing the value of the parameter p, several distance measures can be defined.

Fixing $p = 1$, the Manhattan distance is obtained as defined in Eq. (1). The underlying assumption is that the distance is computed according to the sum of the absolute differences for all dimensions. When the number of dimensions $m = 2$, this L^1 metric corresponds to the city block distance in New York. Similarly, the Gower distance is the L^1 distance divided by the vector length m as shown in the right part of Eq. (1). With this formulation, the distance value can be decomposed into contributions made by each dimension (or feature). When only the rank of the different distances is required, both the Manhattan and Gower measure return the same ordering.

$$dist_{Manhattan}(A, B) = \sum_{i=1}^{m} |a_i - b_i| \ \propto \ dist_{Gower}(A, B) = \frac{1}{m} \sum_{i=1}^{m} |a_i - b_i| \tag{1}$$

Changing the value of p to 2, the Euclidean (L^2 norm) distance is obtained as depicted in Eq. (2). This metric corresponds to our physical concept of a distance between two points, which is the direct straight line. Ignoring the square root can shorten the computation time without changing the ordering of the distances.

$$dist_{Euclidean}(A, B) = \sqrt{\sum_{i=1}^{m} |a_i - b_i|^2} \ \propto \ \sum_{i=1}^{m} |a_i - b_i|^2 \tag{2}$$

The parameter p can take other values and this parameter can be included in the definition of the distance. This general case is known as L^p norm or Minkowski distance as depicted in Eq. (3).

$$dist_{Minkowski}(A, B, p) = \sqrt[p]{\sum_{i=1}^{m} |a_i - b_i|^p} \tag{3}$$

When p goes to infinity, the L^∞ norm or Chebyshev distance is obtained as depicted in Eq. (4). This formulation is also known as maximum metric, or minimax approximation.

$$dist_{Chebyshev}(A, B) = \sqrt[\infty]{\sum_{i=1}^{m} |a_i - b_i|^\infty} = max(|a_1 - b_1|, \ldots, |a_m - b_m|) \tag{4}$$

Some studies proposed to compute the mean between the L^1 and L^∞ distance measure to take account of the advantages of both functions. Eq. (5) shows the corresponding formulation called “Average” distance.

$$dist_{Average}(A, B) = \frac{1}{2}\left(\sum_{i=1}^{m} |a_i - b_i| + max(|a_1 - b_1|, \ldots, |a_m - b_m|)\right) \tag{5}$$

Before comparing these distance measures according to their respective properties and effectiveness (see below), the Appendix visualizes with colors how the distance decreases when the second point is moving away from a given fixed point. For example, we can see that the Euclidian distance describes circles (isodistances) around a given fixed point while the Manhattan approach is based on squares.

3.3. Variants of the L^1 family

Based on the L^1 norm (absolute difference), several variants of the Manhattan distance measure have been proposed. In fact, the value returned by the Manhattan distance is not normalized and it is sometimes difficult to figure out if a given distance is small or large. To propose a solution, the Sørensen, also called Czekanowski or Bray–Curtis distance (Eq. (6)), suggests to normalize the classical Manhattan distance by the sum of all components. As we assume that all vector values are non-negative, the Sørensen distance returns a value between 0 and 1, allowing a clearer interpretation of the distance value than the Manhattan one.

$$dist_{Sørensen}(A, B) = \frac{\sum_{i=1}^{m} |a_i - b_i|}{\sum_{i=1}^{m} (a_i + b_i)} \tag{6}$$

The Tanimoto distance (Eq. (7)), also called Soergel, and similarly the Kulczynski distance (Eq. (8)) also correspond to the Manhattan distance with a normalization factor (divided by the max (or min) of the coefficients). The Motyka distance (Eq. (9)) is based on the maximum value instead of the difference, and the normalization is the sum of all coefficient pairs.

The Canberra distance (Eq. (10)) suggests that the absolute differences of the individual terms are normalized based on the sum of them. One drawback of this last definition is its sensitivity to small changes near zero. The Lorentzian distance (Eq. (11)) is based on the natural logarithm while the Wave-Hedges distance (Eq. (12)) normalizes the difference of each pair of coefficients with its maximum.

$$dist_{\text{Tanimoto}}(A,B) = \frac{\sum_{i=1}^{m} |a_i - b_i|}{\sum_{i=1}^{m} max(a_i, b_i)} \tag{7}$$

$$dist_{Kulcynski}(A,B) = \frac{\sum_{i=1}^{m} |a_i - b_i|}{\sum_{i=1}^{m} min(a_i, b_i)} \tag{8}$$

$$dist_{Motyka}(A,B) = \frac{\sum_{i=1}^{m} max(a_i, b_i)}{\sum_{i=1}^{m} (a_i + b_i)} \tag{9}$$

$$dist_{Canberra}(A,B) = \sum_{i=1}^{m} \frac{|a_i - b_i|}{a_i + b_i} \tag{10}$$

$$dist_{Lorentzian}(A,B) = \sum_{i=1}^{m} ln(1 + |a_i - b_i|) \tag{11}$$

$$dist_{Wave-Hedges}(A,B) = \sum_{i=1}^{m} \frac{|a_i - b_i|}{max(a_i, b_i)} \tag{12}$$

3.4. Variants of the L^2 family

Based on the Euclidian distance or L^2 norm, different variants have been suggested. First, we have the Matusita distance (which imposes the presence of non-negative values for all vector elements). As other variations we have the squared χ^2 and the Clark measure. Those distance measures are variants of the squared difference as used in L^2 and they all result in almost the same visualization (see Appendix).

$$dist_{Matusita}(A,B) = \sqrt{\sum_{i=1}^{m} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2} \tag{13}$$

$$dist_{Squared\ \chi^2}(A,B) = \sum_{i=1}^{m} \frac{(a_i - b_i)^2}{a_i + b_i} \tag{14}$$

$$dist_{Clark}(A,B) = \sqrt{\sum_{i=1}^{m} \left(\frac{|a_i - b_i|}{a_i + b_i}\right)^2} \tag{15}$$

3.5. Inner product family

As another well-known family, different variants based on the inner product (or dot product, see Eq. (16)) have been suggested. The main drawback of the inner product is the absence of a normalization. It is not clear when a distance value should be interpreted as large or small. Therefore, different variants have been proposed, and the most popular is certainly the Cosine similarity (Eq. (17)) which can be transformed into a distance value between 0 and 1 (Eq. (18)) (Zobel & Moffat, 1998; Manning et al., 2008). According to this measure, two similar points indicate similar direction.

$$dist_{Inner\ Product}(A,B) = \sum_{i=1}^{m} a_i b_i \tag{16}$$

$$sim_{Cosine}(A,B) = \frac{\sum_{i=1}^{m} a_i b_i}{\sqrt{\sum_{i=1}^{m} a_i^2}\sqrt{\sum_{i=1}^{m} b_i^2}} \tag{17}$$

$$dist_{Cosine}(A,B) = \frac{1}{\pi} cos^{-1}(sim_{Cosine}(A,B)) \tag{18}$$

As other possible variants, the Jaccard (Eq. (19)) and Dice (Eq. (20)) distances are based on two different normalization approaches. Therefore, similar points have to be in the same direction but also located closely. The Appendix shows clearly the difference between the Cosine and the Jaccard distance.

$$dist_{Jaccard}(A,B) = 1 - \frac{\sum_{i=1}^{m} a_i b_i}{\sum_{i=1}^{m} a_i^2 + \sum_{i=1}^{m} b_i^2 - \sum_{i=1}^{m} a_i b_i} \tag{19}$$

$$dist_{Dice}(A,B) = 1 - \frac{2\sum_{i=1}^{m} a_i b_i}{\sum_{i=1}^{m} a_i^2 + \sum_{i=1}^{m} b_i^2} \tag{20}$$

When comparing the retrieval effectiveness of these four measures in the IR domain, Zobel and Moffat (1998) indicate that the Cosine distance usually tends to produce the best performance. This conclusion cannot be however confirmed in a clear and systematic way.

3.6. Entropy family

Shannon's concept of entropy (Manning et al., 2008) is also the main source of a family of distance measures. The Kullback–Leibler divergence (KLD), also known as relative entropy or information deviation, computes the difference between two probability distributions (see Eq. (21)). In this case, it is required that all values a_i of each vector are non-negative and that their sum is equal to 1. Moreover, the basis of the logarithm is fixed to two in Shannon's entropy measure. However, in the author profiling context, or when only the ranking of the different categories is relevant, changing the basis of the logarithm doesn't affect the ordering of the answers. As for other distance measures, a larger value indicates a larger distance between the two vectors (or points).

$$dist_{KLD}(A,B) = \sum_{i=1}^{m} a_i \; log\left(\frac{a_i}{b_i}\right) \tag{21}$$

The term *divergence* emphasizes the fact that this distance measure is not symmetric. As a variant of KLD, we can mention the Jeffrey or JDivergence defined by Eq. (22) while the KDivergence is depicted in Eq. (23).

$$dist_{JDivergence}(A,B) = \sum_{i=1}^{m} (a_i - b_i) \; log\left(\frac{a_i}{b_i}\right) \tag{22}$$

$$dist_{KDivergence}(A,B) = \sum_{i=1}^{m} a_i \; log\left(\frac{2\; a_i}{a_i + \; b_i}\right) \tag{23}$$

To obtain symmetric measure, one solution is to add to the distance from A to B the distance from B to A. Based on this technique, the KDivergence is used to define the Topsoe distance as shown in Eq. (24). When dividing the Topsoe distance by 2, we obtain the Jensen-Shannon divergence (which is also symmetric). Finally, the Jensen difference is shown in Eq. (25) representing a more complex formulation.

$$dist_{Topsoe}(A,B) = \sum_{i=1}^{m} \left(a_i \; log\left(\frac{2\; a_i}{a_i + b_i}\right) + \; b_i \; log\left(\frac{2\; a_i}{a_i + \; b_i}\right)\right) \tag{24}$$

$$dist_{Jensen}(A,B) = \sum_{i=1}^{m} \left(\frac{a_i \log a_i + \; b_i \log b_i}{2} - \frac{a_i + b_i}{2} log\left(\frac{a_i + b_i}{2}\right)\right) \tag{25}$$

3.7. Combination family

To define a more appropriate distance, different propositions suggest to combine two or more sources of distance measures. For example, Taneja proposes to take account of the arithmetic mean and the geometric mean divergence to define the distance measure given in Eq. (26).

$$dist_{Taneja}(A,B) = \sum_{i=1}^{m} \frac{a_i + b_i}{2} ln\left(\frac{a_i + b_i}{2\sqrt{a_i b_i}}\right) \tag{26}$$

In a related vein, the Kumar-Johnson distance is based on the symmetric χ^2, and both the arithmetic and geometric divergence as shown in Eq. (27).

$$dist_{Kumar-Johnson}(A,B) = \sum_{i=1}^{m} \frac{\left(a_i^2 - b_i^2\right)^2}{2\left(a_i b_i\right)^{3/2}} \tag{27}$$

4. Theoretical assessment

The previous section shows that numerous distance measures can be derived and regrouped under five large families. In this section, we verify whether those distance measures respect the six defined properties. Table 1 describes the results where the 24 distance measures are listed with an indication specifying whether or not they obey to the corresponding property. In the last column, we indicate the number of properties respected by the given measure. Overall, only the Tanimoto and Matusita distance measures fulfill all theoretical properties.

Table 1
Summary of the evaluation of the theoretical properties.

Measure	Eq.	P1	P2	P3	P4	P5	P6	Total
Manhattan	1	Yes	Yes	Yes	Yes	Yes	**No**	5
Euclidean	2	Yes	Yes	Yes	Yes	Yes	**No**	5
Chebyshev	4	Yes	Yes	Yes	Yes	Yes	**No**	5
Average	5	Yes	Yes	Yes	Yes	Yes	**No**	5
Sørensen	6	Yes	Yes	Yes	**No**	Yes	Yes	5
Tanimoto	7	Yes	Yes	Yes	Yes	Yes	Yes	6
Kulczynski	8	Yes	Yes	Yes	**No**	Yes	Yes	5
Motyka	9	**No**	Yes	Yes	Yes	Yes	Yes	5
Canberra	10	Yes	Yes	Yes	Yes	**No**	Yes	5
Lorentzian	11	Yes	Yes	Yes	Yes	Yes	**No**	5
Wave-Hedges	12	Yes	Yes	Yes	Yes	**No**	Yes	5
Matusita	13	Yes	Yes	Yes	Yes	Yes	Yes	6
Squared χ^2	14	Yes	Yes	Yes	**No**	Yes	Yes	5
Clark	15	Yes	Yes	Yes	Yes	**No**	Yes	5
Cosine	17	Yes	**No**	Yes	Yes	Yes	Yes	5
Jaccard	19	Yes	Yes	Yes	**No**	Yes	Yes	5
Dice	20	Yes	Yes	Yes	**No**	Yes	Yes	5
KLD	21	Yes	**No**	**No**	**No**	Yes	Yes	3
JDivergence	22	Yes	Yes	Yes	**No**	Yes	Yes	5
KDivergence	23	Yes	**No**	**No**	**No**	Yes	Yes	3
Topsoe	24	Yes	Yes	Yes	**No**	Yes	Yes	5
Jensen	25	Yes	Yes	Yes	**No**	Yes	Yes	5
Taneja	26	Yes	Yes	Yes	**No**	Yes	Yes	5
Kumar-Johnson	27	Yes	Yes	Yes	**No**	Yes	Yes	5

To illustrate some of the entries in this table, a few numerical examples will be given based on the following set of points in a two-dimensional space.

$$A = \begin{pmatrix} 0.30 \\ 0.70 \end{pmatrix} B = \begin{pmatrix} 0.15 \\ 0.35 \end{pmatrix} C = \begin{pmatrix} 0.50 \\ 0.50 \end{pmatrix} D = \begin{pmatrix} 0.70 \\ 0.30 \end{pmatrix} E = \begin{pmatrix} 0.30 \\ 0.00 \end{pmatrix} F = \begin{pmatrix} 0.00 \\ 0.70 \end{pmatrix} G = \begin{pmatrix} 0.15 \\ 0.70 \end{pmatrix}$$

The first property specifies that the distance from a point to itself must be zero. This feature seems evident, and usually respected by all distance measures. A closer look reveals that applying the Motyka formulation (see Eq. (9)), the distance to itself is not equal to zero but 0.5. The following numerical example illustrates this issue with vector A.

$$dist_{Motyka}(A, A) = \frac{0.3 + 0.7}{0.3 + 0.3 + 0.7 + 0.7} = 0.5$$

This property does not always imply the reverse. Thus, when the distance between two points is zero, one cannot infer the two points are identical. For example, computing the Cosine similarity (Eq. (17)) between the points A and B, the similarity value is 1.0 and therefore the distance between them, according to the Cosine distance, is zero. The second property imposes that the distance between two distinct points must be larger than zero; this is not the case here.

$$sim_{Cosine}(A, B) = \frac{0.3 * 0.15 + 0.7 * 0.35}{\sqrt{0.3^2 + 0.7^2} * \sqrt{0.15^2 + 0.35^2}} = 1.0$$

$$dist_{Cosine}(A, B) = \frac{1}{\pi} cos^{-1}(1.0) = 0.0$$

Both vectors are pointing towards the same direction, but they do not have the same length. As one can see, the vector A is twice the vector B, and therefore the angle between them is zero, resulting in a Cosine distance of 0.0.

When considering the positivity (P2) and the symmetry property (P3), the distance measure based on the Kullback–Leibler divergence (KLD) (Eq. (21)) does not respect these two characteristics. In the following computation, one can see that the resulting value is negative. When KLD is applied between two probabilistic distributions, the returned value is always non-negative. In our context, it is not imposed that the sum of the elements of a vector is 1.0. Therefore, in some cases, the returned value could be negative.

$$dist_{KLD}(B, \mathrm{A}) = 0.15\ ln\left(\frac{0.15}{0.3}\right) + 0.35\ ln\left(\frac{0.35}{0.7}\right) = -0.347$$

Using the same argument, one can verify that the KDivergence distance (Eq. (23)) can return negative values. For the symmetry, the following computation shows that with the Kullback-Leibler divergence (KLD), this property is not respected. The distance is 0.693 while the distance from B to A is −0.347.

$$dist_{KLD}(A, \mathrm{B}) = 0.3\ ln\left(\frac{0.3}{0.15}\right) + 0.7\ ln\left(\frac{0.7}{0.35}\right) = 0.693$$

Many distance measures do not respect the fourth property, the triangle inequality. When considering the triangle {A, C, D}, the distance from A to D must be smaller (or equal) to the distance from A to C plus the distance from C to D. For example, with the Dice formula (Eq. (20)), one can obtain:

$$dist_{Dice}(A, D) = 1 - \frac{2 \cdot (0.3 \cdot 0.7 + 0.7 \cdot 03)}{0.3^2 + 0.7^2 + 0.7^2 + 0.3^2} = 0.276$$

$$dist_{Dice}(A, C) = 1 - \frac{2 \cdot (0.3 \cdot 0.5 + 0.7 \cdot 0.5)}{0.3^2 + 0.7^2 + 0.5^2 + 0.5^2} = 0.074$$

$$dist_{Dice}(C, D) = 1 - \frac{2 \cdot (0.5 \cdot 0.7 + 0.5 \cdot 0.3)}{0.5^2 + 0.5^2 + 0.7^2 + 0.3^2} = 0.074$$

and $dist(A, D) = 0.276 > dist(A, C) + dist(C, D) = 0.074 + 0.074 = 0.148$. As shown in Table 1, this property is not respected by several distance measures.

Regarding the fifth property (absence of an important feature), a few distance measures do not respect it, as, for example, the Canberra (Eq. (10)) or the Clark equation (see Eq. (15)). In our example, the vector A is composed of an important second component (with a value 0.7) while the first is smaller (0.3). The vector E has a zero value for the second coordinate, an important feature in describing vector A. On the other hand, the vector F owns exactly the same value for the second coordinate than A, but does not have the first one, a less important feature. Computing the distance from A to E or A to F with the Canberra measure, the same value is obtained. To obey the fifth property, the distance (A, E) must be larger than the distance (A, F).

$$dist_{Canberra}(A, E) = \frac{|0.3 - 0.3|}{|0.3| + |0.3|} + \frac{|0.7 - 0.0|}{|0.7| + |0.0|} = 1$$

$$dist_{Canberra}(A, F) = \frac{|0.3 - 0.0|}{|0.3| + |0.0|} + \frac{|0.7 - 0.7|}{|0.7| + |0.7|} = 1$$

Concerning the last property, we specify that the presence of a feature is better than its absence given the fact that the absolute difference is the same. This property is usually satisfied by numerous formulations. However, some of them do not follow it as, for example, the Manhattan distance. With the following numerical examples, the distance between vector A and G is the same as the distance between vector F and G. But in this example, the vector F does not have the first feature, and thus to respect this sixth property, the distance (F, G) must be larger than between A and G.

$$dist_{Manhattan}(A, G) = |0.3 - 0.15| + |0.7 - 0.7| = 0.15$$

$$dist_{Manhattan}(F, G) = |0.0 - 0.15| + |0.7 - 0.7| = 0.15$$

Respecting the set of proposed properties is important from a theoretical point of view and can form a set of criteria to select the most appropriate distance formulation. However, in practice a distance measure should also be easy to compute and provide overall good effectiveness for the targeted task. To evaluate this last aspect, we will evaluate the 24 distance measures according to the 13 test collections described in the next section.

5. Test collections and evaluation methodology

To provide large and reusable test collections, the CLEF was launched in 1999. In 2010, the PAN CLEF track was created to detect plagiarism, and in 2011 the authorship attribution issue was added. During the PAN CLEF 2013 campaign (Rangel Pardo, Rosso, Koppel, Stamatatos, & Inches, 2013) and in 2014 (Rangel et al., 2014), a profiling task was proposed. In this case, only the gender and age range are required to be determined based on blog posts, sequences of tweets, or reviews written in the English or Spanish language. The corresponding demographic category was extracted from the author's profile with some verifications (e.g., consulting Facebook or LinkedIn websites). The selected text register corresponds to messages written more or less spontaneously, without corrections done by an editor (as for newspaper articles) or a group of advisors/experts (as in official speeches).

In 2015, the PAN CLEF campaign (Rangel, Celli, Rosso, Potthast, Stein, & Daelemans, 2015; Stamatatos, Potthast, Rangel, Rosso, & Stein, 2015) added the Italian and Dutch languages but the text genre was limited to tweets. For the two new languages, only the gender of the author is provided. Finally, in 2016 (Rosso et al., 2016), the evaluation text genre was unknown but the training had to be performed using Twitter data in which the Dutch collection did not require an age range detection. More general information about these 13 author profiling collections extracted from the PAN CLEF campaigns are given in Table 2.

In this table, the corpus name corresponds to the concatenation of the last two digits of the year, the first letter of both the language and text genre. For example, 15ET denotes a test collection of year 2015, written in English, and containing tweets. The following columns indicate the year, language, and text genre of the corresponding corpus. Under the label "Problems", the value indicates the number of texts for which the system determines the gender, and the age range. The

Table 2
PAN CLEF 2014 to 2016 test collection statistics.

Name	Year	Language	Genre	Problems	Gender	Age Groups
14EB	2014	English	Blog	147	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
14SB	2014	Spanish	Blog	88	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
14ET	2014	English	Twitter	306	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
14ST	2014	Spanish	Twitter	178	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
14ER	2014	English	Review	4160	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
14SS	2014	Spanish	Social media	1272	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
15DT	2015	Dutch	Twitter	34	Male female	NA
15ET	2015	English	Twitter	152	Male female	18–24, 25–34, 35–49, 50-xx
15IT	2015	Italian	Twitter	38	Male female	NA
15ST	2015	Spanish	Twitter	100	Male female	18–24, 25–34, 35–49, 50-xx
16DT	2016	Dutch	Twitter	384	Male female	NA
16ET	2016	English	Twitter	436	Male female	18–24, 25–34, 35–49, 50–64, 65-xx
16SP	2016	Spanish	Twitter	250	Male female	18–24, 25–34, 35–49, 50–64, 65-xx

last two columns provide the possible value for the gender and age classes. A closer look on this table reveals that in 2015 only four age groups have been specified. For that year, the two oldest classes (50–64 and 65-xx) were merged into a single class (50-xx). When considering all test collections, one can find 7545 problems in total where 5201 are written in the English language, 1888 in Spanish, 418 in Dutch, and 38 in Italian.

As a performance measure, the accuracy rate (or success rate) has been adopted in the PAN CLEF evaluation campaign. This measure varies from 0 to 1 (or 100%), where a higher rate means a better result. This performance score can be computed for each demographic category individually, namely gender and age group. For example, if the system correctly predicts the author gender 7 times in 10 problems, the accuracy will be 0.7 for this category alone. The accuracy rate for both demographic categories will be reported in our experiments.

However, in the CLEF evaluation campaigns, the different systems are ranked according to a single value. To obtain a single overall effectiveness value, the fraction of problems where both the gender and age group are correctly predicted for the same problem is computed. Continuing with our previous example: If, for the age ranges, the classifier predicts correctly 5 times the age class over 10 problems, the accuracy is 0.5. To determine the quality of this classifier with respect to both the gender and the author age, the arithmetic mean could be used (e.g., ½ $(0.7 + 0.5) = 0.6$). In the CLEF campaigns, the evaluation corresponding to both demographic categories is based on the number of correct assignments for *both* the gender and the age class. In our example, the classifier was able to correctly determine the gender and the age group for 4 problems, giving us an accuracy rate of $4/10 = 0.4$. As one can see in the evaluation shown in the following tables, the accuracy rate for determining both categories is always lower than the simple arithmetic mean.

6. Evaluation

Before presenting the evaluation results, the first section describes the *k* nearest neighbors classifier (*k*-NN) used in all our experiments. The following section presents the results over the 13-test collection using two different learning phases. In the last section, distinct text genres are employed in the training and test phase. With this experiment, one can estimate the loss of accuracy due to the use of different text genres in the training and testing phases.

6.1. K-nearest neighbors classifier

To evaluate the different distance measures, the top *m* most frequent terms (isolated words without stemming but with the punctuation symbols) forms the feature set. For determining the value of *m*, previous studies have shown that a value between 50 to 300 tends to provide the best performance (Burrows, 2002; Savoy, 2012, 2015; Kocher & Savoy, 2016). For all our experiments, we fixed $m = 200$.

When considering the *m* most frequent terms from a query text, the terms appearing once (*hapax legomenon*) are ignored. Of course, for some short texts, the resulting representation can include less than 200 terms. The character appearing in uppercase are replaced by the corresponding lowercase letter. Thus, from the text “The cat jumps over the cat and over the table” the representation is [the: 3, cat: 2, over: 2]. The word “jumps”, “and”, or “table” occurring once are ignored. Finally, instead of directly using the occurrence frequencies of the *i*th term (denoted tf_i), the estimated probability is computed by dividing its occurrence frequency by the text length (measured in remaining tokens and denoted n). Therefore, for each vector component we have tf_i / n. In our previous example, the final representation of the query vector is [the: 3/7, cat: 2/7, over: 2/7] and for the comparison all other document vectors are built according to those three terms.

Web-based textual communication contains other forms than words. In a tweet, one can find hashtags (e.g., #nasa, #noobama) or various URLs (e.g., http://shakespeare.mit.edu, www.un.org). To take them into account, all hashtags or URLs are replaced by a common marker. The frequency of hashtags or URLs forms two additional features that have been shown to be effective, for example, to discriminate between Democrats and Republicans (Sylwester & Purver, 2015).

Table 3
Profiling results based on the same collection in training and test phase (macro-average over 13 test collections, leaving-one-out).

Measure	Training on the Same Corpus		
	Gender	Age	Both
Manhattan	0.6064	0.4535	0.3514
Euclidean	0.6131	0.4463	0.3566
Chebyshev	0.6049†	0.4313	0.3534†
Average	0.6105	0.4581	0.3643
Sørensen, Tanimoto, Kulczynski, Motyka	*0.6362†	*0.4626	***0.3865**
Canberra	*0.6290†	0.4504	0.3668†
Lorentzian	0.6281†	0.4446	0.3772†
Wave-Hedges	***0.6439**	0.4504	*0.3825†
Clark	0.6262†	0.4604	0.3682†
Matusita	0.6227†	***0.4868**	*0.3791†
Squared χ^2	*0.6293†	*0.4688	*0.3821†
Cosine	0.6114	0.4402	0.3569
Jaccard, Dice	0.6133	0.4489	0.3640
KLD	0.6102	*0.4608	0.3632†
JDivergence	0.6211†	0.4526	0.3670†
KDivergence	0.6246†	0.4476	0.3638†
Topsoe, Jensen	*0.6322†	*0.4727†	*0.3832†
Taneja	0.6162†	0.4594	0.3689†
Kumar–Johnson	0.6092	0.4558	0.3605
Mean	0.6204	0.4553	0.3682

Finally, in determining the corresponding demographic category of a query text, the distance with all other texts is computed and the five nearest neighbors (5-NN) are taken into account. From this set of the five closest neighbors, the majority determines the returned category, and in case of a tie, the closest neighbor defines the returned category. For example, if the age groups of the five nearest neighbors, in increasing distance, are 25–34, 18–24, 25–34, 18–24, and 35–49, the system assigns the label “25–34” to the query text because it is the closest group with the most members. This kind of classifier model has demonstrated overall good performance in a similar task (Kocher & Savoy, 2016). The remaining question is to know which distance measure offers the best performance.

6.2. Evaluation based on same text genre for training and test

To determine the accuracy rate for each of the 13 test collections and considering the 24 distance measures, a first set of experiments is based on the leaving-one-out (LOO) methodology (Witten, Frank, & Hall, 2011; Zhai et al, 2016). This evaluation approach guarantees an unbiased estimation of the true performance. Instead of reporting all possible combinations of each corpus according to each distance measure, only the average will be reported as shown in Table 3. To achieve this, the macro-average principle (Sebastiani, 2002) was applied, giving the same importance to each test collection. In other words, the mean is computed over all corpora instead of over each decision (micro-averaging). When considering the size of each corpus given in Table 2, the result of the micro-average method will be dominated by the 14ER corpus having 4160 problems while, for example, the 15DT (34 problems) will have an insignificant effect on the overall performance. Thus, we prefer giving the same importance to each test collection, and the macro-averaging method was adopted.

When adopting a distance measure, the returned distance is used to select the top five closest neighbors for each problem. The distance value by itself is not directly used. Therefore, even if one can see a small difference between two distance formulations, the effect in selecting the five nearest neighbors is nil. For example, applying the Jaccard (Eq. (19)) or Dice (Eq. (20)) distances, the returned value is not strictly the same but the selection of the five closest neighbors is the same (however, maybe not in the same order). Therefore, instead of presenting both measures in Table 3, both distances are merged into a single row. The same phenomenon appears with the Topsoe (Eq. (24)) and Jensen (Eq. (25)) measures, as well as with the following four distance formulas: Sørensen (Eq. (6)), Tanimoto (Eq. (7)), Kulczynski (Eq. (8)), and Motyka (Eq. (9)).

In Table 3, the first column indicates the name of the distance measure and the next three columns report the accuracy rates when applying the leaving-one-out approach. The first value corresponds to the gender problem, the second to the age class determination, and the third indicates the percentage of correct answers for both the gender and age group at the same time. As one can see, the gender categorization problem is always easier than the age group. The third evaluation was clearly the most difficult, and the reported accuracy rates are always smaller than for the age detection. For example, applying the Euclidian distance, the average over 13 test collections for the gender problem, the accuracy rate is 0.6131. Under the same condition, the age group determination achieved a mean performance of 0.4463 while the proportion of correct decisions for both the gender and age group is 0.3566.

In Table 3, the highest performance per column is depicted in bold and an asterisk indicates the top best five cells per columns. When considering measures appearing in the top five, the Sørensen (and Tanimoto, Kulczynski, and Motyka), Squared χ^2, and Topsoe (and Jensen) formulas occur three times, while the Wave-Hedges and Matusita appear two times.

Table 4
Corpus used in the training and test phase.

Training	Test	Training	Test
14ET	15ET	14ET	16ET
15ET	14ET	15ET	16ET
16ET	14ET	16ET	15ET
14ST	15ST	14ST	16ST
15ST	14ST	15ST	16ST
16ST	14ST	16ST	15ST
15DT	16DT	16DT	15DT

Table 5
Profiling results based on the same text genre in training and test phase (macro-average over 14 collections).

Measure	Training on the Same Corpus Leaving-one-out			Training on a Second Corpus Same Text Genre		
	Gender	Age	Both	Gender	Age	Both
Manhattan	0.6249	0.5065	0.3657	0.5857	*0.4377	0.3131
Euclidean	0.6270	0.5151	0.3803	0.5888	0.4191	0.3053
Chebyshev	0.6091	0.4805	0.3503	0.5752	0.3881	0.2821
Average	0.6257	0.5113	0.3801	0.5914	*0.4444	0.3192
Sørensen, Tanimoto, Kulczynski, Motyka	0.6569†	*0.5294	*0.4096†	0.5926	0.4131	0.3106
Canberra	*0.6665†	0.5126	0.3930	***0.6248**	*0.4353	*0.3436†
Lorentzian	0.6475	0.5022	0.3920†	0.5927†	0.4287†	0.3054
Wave-Hedges	***0.6707**	0.5076	0.4000†	*0.6200†	0.4263	*0.3379†
Clark	*0.6631†	*0.5285†	0.3939†	*0.6190†	***0.4528**	***0.3463**
Matusita	0.6556†	***0.5497**	*0.4096†	*0.6079†	0.4219	0.3144
Squared χ^2	0.6550†	*0.5353†	*0.4089†	0.6072†	0.4212	*0.3256
Cosine	0.6227	0.5053	0.3733	0.5767	0.4119	0.2977
Jaccard, Dice	0.6240	0.5094	0.3814	0.5796	0.4014	0.3018
KLD	0.6466	0.5253	0.3931	0.6065†	0.4217†	0.3130
JDivergence	0.6596†	0.5052	0.3895†	0.6065†	0.4259†	0.3098
KDivergence	0.6438	0.4891	0.3737	0.5900	0.3679	0.3002
Topsoe, Jensen	*0.6613†	*0.5387†	***0.4101**	*0.6183†	0.4190	*0.3260
Taneja	*0.6600†	0.5222	*0.4044†	0.5987†	*0.4292†	0.3048
Kumar–Johnson	0.6275	0.5177	0.3772	0.5628	0.4050	0.2763
Mean	0.6446	0.5153	0.3887	0.5971	0.4195	0.3123
Accuracy loss				7.4%	19.1%	19.7%

Canberra and KLD each make it once in the top five. On the other hand, the Manhattan, Euclidian, Chebyshev, Average, Lorentzian, Clark, Cosine, Jaccard (and Dice), JDivergence, KDivergence, Taneja, or Kumar-Johnson never appear in the best five measures on the three tasks. To statistically determine whether or not a given distance measure is statistically worse than the best one (depicted in bold), we applied the *t*-test whereby the null hypothesis H_0 states that both distance measures result in similar performance levels. In the experiments, statistically non-significant differences are indicated by a cross (†) (paired, two-sided test, significance level $\alpha = 5\%$).

Table 3 reports the mean accuracy rate achieved when using the same collection for both the training and test phase. However, the instances used during the test phase never occur during the training (leaving-one-out methodology) (Witten et al., 2011). When considering the available corpora, the training stage can be performed using another text collection. As shown in Table 2, the different collections share some common characteristics such as the language or the text genre. Thus, instead of deriving the text representations from the same corpus as previously, these profiles can be built according to another corpus written, of course, in the same language but also having the same text genre. For example, the decisions related to corpus ET14 can be based on corpus 15ET or 16ET. Table 4 indicates the 14 different combinations that can be obtained when considering the 13 test collection. Obviously, having just one corpus in the Italian language, it was impossible to perform this kind of evaluation in Italian. Moreover, one can observe that the English collections 14ET, 15ET, and 16ET appear twice in the test stage. The same occurs with the Spanish corpora 14ST, 15ST, and 16ST. However, the two Dutch collections (15DT and 16DT) appear only once.

To have a fair comparison between the two forms of training, the accuracy rates reported in Table 3 are not appropriate. Thus, in Table 5 the left part indicates the overall performance under the leaving-one-out methodology but using the set of collection appearing in Table 4 under the column "Test". This means the performance achieved by the corpus 14ET, 15ET, and 16ET are computed twice while the accuracy of the Italian collection is ignored. Therefore, in total 14 collections will be used to estimate the accuracy rate.

The left part of Table 5 indicates the overall performance achieved with those 14 corpora, using in the training the same collection (leaving-one-out). On the right side of this table, one can find the same three accuracy rates obtained with another

Table 6
Corpus used in the training and test phase (cross-genre evaluation).

Training	Test	Training	Test	Training	Test	Training	Test
14EB	14ER	14ER	16ET	14SS	14SB	15ET	14ER
14EB	14ET	14ET	14EB	14SS	14ST	15ST	14SB
14EB	15ET	14ET	14ER	14SS	15ST	15ST	14SS
14EB	16ET	14SB	14SS	14SS	16ST	16ET	14EB
14ER	14EB	14SB	14ST	14ST	14SB	16ET	14ER
14ER	14ET	14SB	15ST	14ST	14SS	16ST	14SB
14ER	15ET	14SB	16ST	15ET	14EB	16ST	14SS

corpus for the two demographic categories and the combined evaluation. The last row reports the arithmetic average over the 24 distance measures.

The effectiveness values reported in Table 5 indicate that the Clark and Canberra appear to achieve the best overall performances when the learning is based on a distinct collection having the same text genre. Overall, when considering measures appearing in the top five, the Clark and Topsoe (and Jensen) formula occur five times, and the Canberra four times. On the other hand, the Euclidian, Chebyshev, Lorentzian, Cosine, Jaccard (and Dice), KLD, JDivergence, KDivergence, or Kumar-Johnson never appear in the best five measures over these six categorization tasks.

Using a distinct corpus with the training stage tends to hurt the overall performance of the classification system, whatever the distance measure is. As depicted in the last row of Table 5, the mean degradation goes from 7.4% for the gender classification to 19.1% for the age. When considering both the gender and age classification, the decrease reaches 19.7%.

From an efficiency point of view, the increasing complexity from the L^1 family, to the L^2 family, to the Inner Product family, and to the Entropy family is directly reflected in an increasing runtime. Therefore, computing a distance from the L^1 family is faster than any of the distances from other families while one from the Entropy family is slower than all measures from any other distance family. Based on our experiments, the Manhattan distance (Eq. (1)) takes the least computing time, while Topsoe's formulation (Eq. (24)) can be from 20% up to 80% slower.

6.3. Cross-genre evaluation

The training phase can be performed on a corpus written, obviously, in the same language as the test instances, but having a distinct text genre. For some applications, the correspondence between the training and test sets cannot be as close as one would wish. Therefore, the training has to be performed on a different text genre, which is a solution that will have an impact on the overall classification performance. But to what extent? Certainly, the distance between the training and test set should be as close as possible. In the 13 test collections described in Table 2, one can observe that they correspond to web-based communication with an emphasis on tweets. The remaining question is to estimate the loss of effectiveness when learning on one web-based text genre to test on another one. Of course, considering two very distinct text genres (such as oral vs. written formal speech (Biber & Conrad, 2009)) will produce higher accuracy rate degradation.

To evaluate this degradation in the accuracy rate, we design a set of experiments in which the training text genre differs from the test phase. Table 6 depicts the different possible configurations applied to obtain results shown in Table 7. As for the previous tables, on the left one can see the accuracy rates achieved using the same collection during the training and the test phases. On the right, the performance values are obtained using different text genres in training and testing.

As in the evaluation on the same text genre in Table 5, the Clark measure appears again to achieve the best overall performance when the learning is based on a distinct collection having a different text genre. When considering measures appearing in the top five, the KLD, JDivergence, and Taneja formula occur three times in the cross-genre evaluation but they never occur in the results when the training was on the same corpus. Conversely, the Sørensen (and Tanimoto, Kulczynski, and Motyka) appears three times in the top five on the left-hand side of the table, but are missing from the top five in the right part when the training was on a different text genre. On the other hand, the Manhattan, Euclidian, Chebyshev, Cosine, Jaccard (and Dice), or Kumar-Johnson never appear in the best five measures over these six categorization tasks.

One can notice that the performance drop between same-genre and cross-genre gender predictions is more than 11.2%, the estimation for the age groups decreases about 26.1%, and the loss of accuracy when classifying both attributes is 35.1% in cross-genre compared to the same-genre evaluations. This means that not all style markers are transferred from one genre to another, which leads to misclassifications.

7. Conclusion

From a practical point of view, this paper investigates the problem of the selection of the best performing distance measure when designing a classifier to solve the author profiling question. In this perspective, the gender (two categories) and the age group (four to five classes) of the author are required to be determined as accurately as possible. This problem is characterized by a relatively large number of possible features, without having some dominating all the others. In this context, 24 distance measures have been briefly described reflecting five main families of functions (L^1, L^2, inner product, entropy-based, and combination approaches).

Table 7
Profiling results based on different text genres in training and test phase (macro-average over 28 collections).

Measure	Training on the Same Corpus Leaving-one-out			Training on a Second Corpus Different Text Genre		
	Gender	Age	Both	Gender	Age	Both
Manhattan	0.6094	0.4308	0.2655	0.5287	0.3177†	0.1729†
Euclidean	0.6075	0.4169	0.2544	0.5245	0.2999	0.1635
Chebyshev	0.5867	0.4102	0.2489	0.5289	0.2926	0.1571
Average	0.6076	*0.4353	0.2731†	0.5256	0.3133†	0.1698
Sørensen, Tanimoto, Kulczynski, Motyka	*0.6216†	*0.4341	*__0.2815__	0.5373	0.3043	0.1683
Canberra	*0.6225†	0.4238	0.2725†	*0.5629†	0.3211†	*0.1877†
Lorentzian	0.6166	0.4199	*0.2782†	0.5382	0.3174	0.1734
Wave-Hedges	*__0.6296__	0.4259	*0.2795†	0.5473	0.3176	0.1754
Clark	*0.6195†	0.4312	0.2751†	*__0.5669__	0.3245†	*__0.1923__
Matusita	0.6099	*__0.4599__	*0.2805†	0.5490	*0.3266†	0.1837†
Squared χ^2	0.6132	*0.4404	0.2772†	0.5486†	0.3159	0.1798†
Cosine	0.6112	0.4122	0.2632	0.5398	0.3212†	0.1792†
Jaccard, Dice	0.6105	0.4229	0.2688	0.5289	0.3090	0.1639
KLD	0.6012	0.4331	0.2694	*0.5511†	*__0.3461__	*0.1920†
JDivergence	0.6099	0.4301	0.2731†	*0.5499†	*0.3413†	*0.1861†
KDivergence	*0.6218†	0.4298	0.2724†	0.5348	0.2852	0.1518
Topsoe, Jensen	0.6181	*0.4444	*0.2813†	0.5498†	*0.3272†	0.1839†
Taneja	0.6055	0.4325	0.2739	*0.5548†	*0.3384†	*0.1893†
Kumar–Johnson	0.5963	0.4293	0.2595	0.5455	0.3148	0.1685
Mean	0.6115	0.4296	0.2710	0.5428	0.3176	0.1757
Accuracy loss				11.2%	26.1%	35.1%

From a theoretical point of view, a set of six theoretical properties have been presented. Only two formulations (Tanimoto and Matusita) respect all these requirements, with the other 20 respecting five of these properties. Looking at their definition, the difference between these 24 distance measures is usually rather small.

From an empirical point of view, an evaluation has been performed. To achieve this, 13 test collections extracted from PAN CLEF evaluation campaigns have been selected. These corpora cover four text genres (tweets, blogs, reviews, and social media) and four languages (English, Spanish, Dutch, and Italian). To evaluate the different distance measure, the k-NN classifier is used, and the top five closest neighbors are employed to determine the demographic category. Based on the leaving-one-out methodology, the Sørensen (and Tanimoto, Kulczynski, and Motyka), Wave-Hedges, and Matusita distance measures tend to show the best performance. The performance differences are however usually not significant between the best 5 distance measures. On the other hand, the Cosine distance measure, well-known in various distributed language models (Bengio, 2009), tend to produce rather lower performance levels. In the IR domain (Manning et al., 2008), the Dice and Jaccard distance measures are also recommended but both depict a lower accuracy rate than the best performing measures. From an efficiency perspective, the Manhattan measure presents a clear advantage, having usually the smallest computation time.

In this paper, we also evaluate the differences in accuracy rates when the training corpus is not the same as the test one. Compared to having the same corpus in both the training and test phases, the overall performance decreases from around 7.4% (gender classification only) to 19.7% (both gender and age classification). As an additional evaluation, the training was performed on a distinct text genre than the test one. However, both reflect the general web-based writing style. In this case, the total accuracy rate tends to decrease from 11.1% (limited to gender classification) to 35.0% (classification of both the gender and age group).

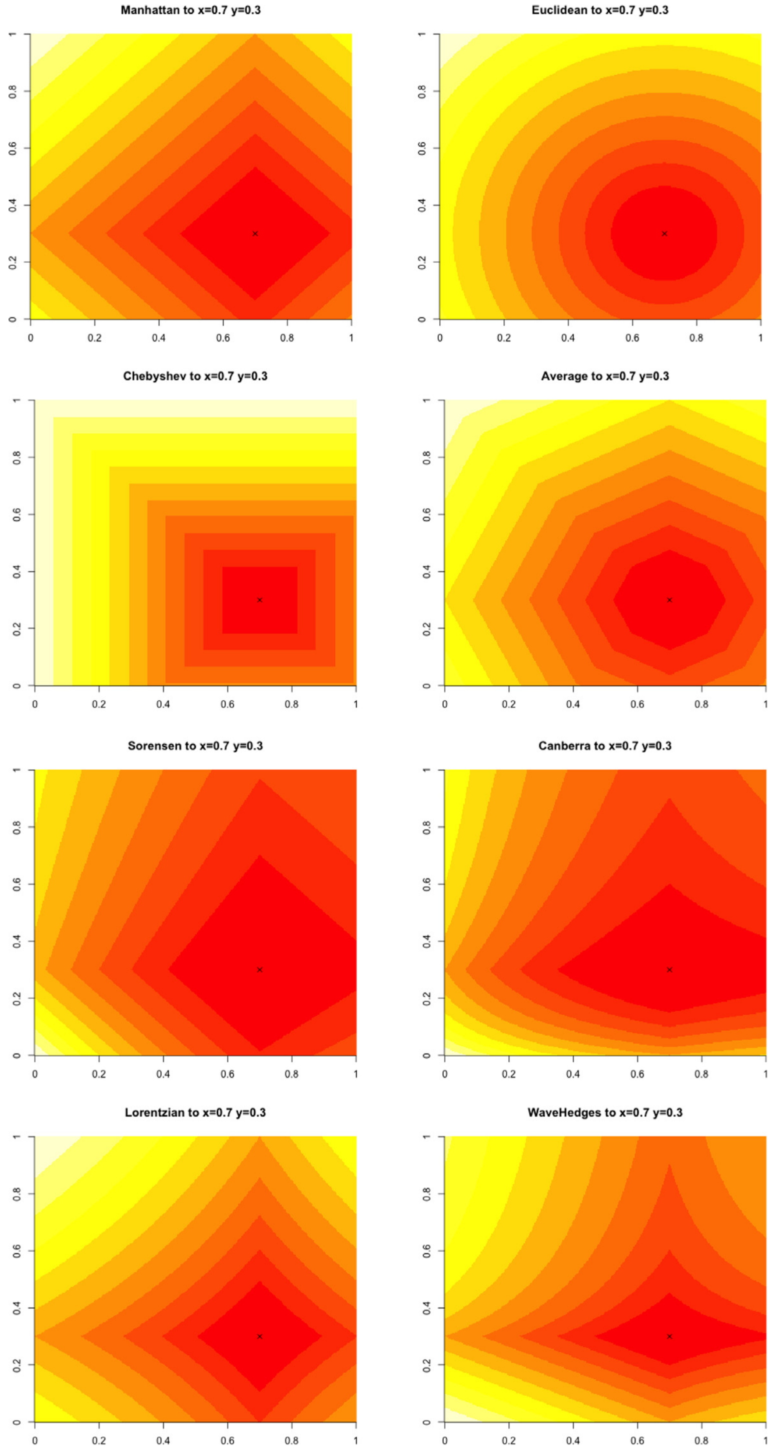
The current study has its own limitations. The focus is placed on a specific text categorization problem: that of author profiling. In this case, the number of features are relatively large and many of them tend to have similar frequencies. We do not have one or a few features dominating the others that can by themselves discriminate between the different targeted categories. The experiments were based only on web-based mediated texts and additional evaluations performed on other text genres should be done to confirm our main findings. The results are also not directly applicable to neural networks and word embedding approaches where the distance measures are not used in explicit form.

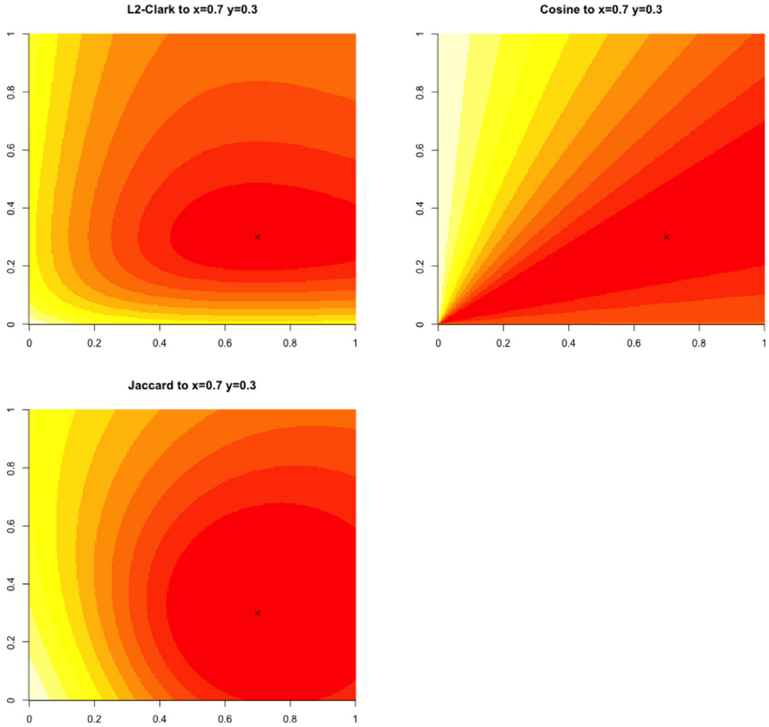
Acknowledgments

This research was supported, in part, by the NSF under Grant #200021_149665/1. The authors want to thank the anonymous reviewers for their helpful suggestions and remarks.

Appendix

To visualize the different distance measures, we consider a reduced vector space with only two dimensions. We assume that the coefficients of a vector represent the underlying probabilities. Therefore, the axis are ranging from 0 to +1. A point at position $x = 0.7$ and $y = 0.3$ is selected and the distance to all other points in $[0, 1] \times [0, 1]$ is calculated according to various measures. The plots show dark red areas where the distance to the point is small (or the corresponding similarity is high), orange zones for more distant fields, and bright yellow regions for the farthest (least similar) groups.





References

Alowibdi, J. S., Buy, U. A., & Yu, P. (2013). Empirical evaluation of profile characteristics for gender classification on twitter. In *Proceedings of the international conference on machine learning and applications* (pp. 365–369).

Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-Y-Gómez, M., Villaseñor-Pineda, L., & Jair-Escalante, H. (2015). INAOE's participation at PAN'15: Author profiling task. In L. Cappellato, N. Ferro, G. J. F. Jones, & E. SanJuan (Eds.), *Proceeding CLEF-2015, working notes*. CEUR.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical predictors of personality type. In *Proceedings of the 2005 joint annual meeting of the interface and the classification society* (pp. 1–16).

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling of the author of an anonymous text. *Commun. ACM, 52*(3), 119–123.

Baayen, H. R. (2008). *Analysis linguistic data: A practical introduction to statistics using r*. Cambridge: Cambridge University Press.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1–127.

Biber, C., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Bilan, I., & Zhekova, D. (2016). Caps: A cross-genre author profiling system. In K. Balog, L. Cappellato, N. Ferro, & C. Macdonals (Eds.), *Proceeding CLEF-2016, Working Notes* (pp. 824–835). CEUR.

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287.

Busger Op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., et al. (2016). In K. Balog, L. Cappellato, N. Ferro, & C. Macdonals (Eds.), *Proceeding CLEF-2016* (pp. 846–857). CEUR. Working Notes.

Ciot, M., Sonderegger, M., & Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. In *Proceedings of conference on empirical methods in natural language processing* (pp. 1136–1145).

Cha, S-H. (2007). Comprehensive survey on distance similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences, 1*(4), 300–307.

Coates, J., & Pichler, P. (2011). *Language and gender*. Chichester: Wiley-Blackwell.

Collberg, C., & Proebsting, T. A. (2016). Repeatability in computer systems research. *Communications of the ACM, 59*(3), 62–69.

Craig, H., & Kinney, A. F. (2009). *Shakespeare, computers, and the mystery of authorship*. Cambridge: Cambridge University Press.

Crystal, D. (2006). *Language and the internet*. Cambridge: Cambridge University Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Addison-Wesley.

Eckert, P., & McConnell-Ginet, S. (2013). *Language and gender*. Cambridge: Cambridge Univeristy Press.

Fung, G. (2003). The disputed *Federalist Papers*: SVM features selection via concave minimization. In Proceeding ACM-TAPIA Conference (pp. 42–46).

González-Gallardo, C. E., Montes, A., Sierra, G., Núñez, A., Adolfo, S., & Ek, J. (2015). In L. Cappellato, N. Ferro, G. J. F. Jones, & E. SanJuan (Eds.), *Tweets classification using corpus dependent tags, character and pos n-grams*. CEUR Working Notes.

Grivas, A., Krithara, A., & Giannakopoulos, G. (2015). Author profiling using stylometric and structural feature groupings. In L. Cappellato, N. Ferro, G. J. F. Jones, & E. SanJuan (Eds.), Proceeding CLEF-2015. CEUR *Working Notes*.

Gronenschild, B. M., Habets, P., Jacobs, I. L., Mengelers, N., van Os, J., & Marcelis, M. (2012). The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PLOS*.

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

Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing, 25*(2), 215–223.

Khonji, M., & Iraqi, Y. (2014). A slightly-modified GI-based author-verifier with lots of features (ASGALF). In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Proceeding CLEF-2014* (pp. 977–983). CEUR. *Working Notes*.

Kocher, M., & Savoy, J. (2016). A simple and efficient algorithm for authorship verification. *Journal of the American Society for Information Science and Technology, 68*(1), 259–269.

López-Monroy, A. P., Montes-Y-Gómez, M., Jair-Escalante, H., & Villaseñor-Pineda, L. (2014). Using intra-profile information for author profiling. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Proceeding CLEF-2014* (pp. 1116–1120). CEUR. *Working Notes*.

Maharjan, S., Shrestha, P., & Solorio, T. (2014). A simple approach to author profiling in mapreduce. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Proceeding CLEF-2014* (pp. 1121–1128). CEUR. *Working Notes*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Modaresi, P., Liebeck, M., & Conrad, S. (2016). Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. In K. Balog, L. Cappellato, N. Ferro, & C. Macdonals (Eds.), *Proceeding CLEF-2016* (pp. 970–977). CEUR. *Working Notes*.

Mosteller, F., & Wallace, D. L. (1964). *Applied bayesian and classical inference: The case of the federalist papers*. Reading: Addison-Wesley.

Nguyen, D., Trieschnigg, D., Seza Doğruöz, A., Gravel, R., Theune, M., Meder, T., et al. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of 25th international conference on computational linguistics* (pp. 1950–1961).

Olsson, J. (2008). *Forensic linguistics*. London: Continuum.

Pennebaker, J. W. (2011). *The secret life of pronouns. What our words say about us*. New York: Bloomsbury Press.

Rangel, F., & Rosso, P. (2016). On the impact of emotions on author profiling. *Information Processing & Management, 52*(1), 73–92.

Rangel Pardo, F. M., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at PAN 2013. *Working Notes for CLEF 2013 Conference*. Valencia, Spain.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., et al. (2014). Overview of the 2nd author profiling task at PAN 2014. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.). In *Notebook papers of CLEF 2014 labs and workshop: 1180* (pp. 827–898). Aachen.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.). *Notebook papers of CLEF 2015 labs and workshop*: 1391. Aachen.

Rosenthal, S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 763–772).

Rosso, P., Rangel, F., Potthast, M., Stein, B., Stamatatos, E., Tschuggnall, M., et al. (2016). *Experimental IR meets multilinguality, multimodality, and interaction* (pp. 332–350). Heidelberg: Springer.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., et al. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of conference on empirical methods in natural language processing* (pp. 1146–1151).

Savoy, J. (2012). Authorship attribution based on specific vocabulary. *ACM – Transactions on Information Systems, 30*(2), 170–199.

Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities, 30*(2), 246–261.

Savoy, J. (2016). Text representation strategies: An example with the *State of the Union* addresses. *Journal of the American Society for Information Science and Technology, 67*(8), 1858–1870.

Schler, J., Koppel, A., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings AAAI spring symposium on computational approaches for analyzing weblogs* (pp. 191–197).

Sebastiani, F. (2002). Machine learning in automatic text categorization. *ACM Computing Survey, 34*(1), 1–27.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 214–433.

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., & Stein, B. (2015). Overview of the PAN/CLEF 2015 evaluation lab. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Proceedings of the notebook papers of CLEF 2015 labs and workshop: 1391*. CEUR.

Sylwester, K., & Purver, M. (2015). Twitter language use to reflect psychological differences between Democrats and Republicans. *PLoS One, 10*(9). doi:10.1371/journal. pone.0137422.

Talbot, M. (2010). *Language and gender*. Malden: Polity Press.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Weren, E. R. D., Moreira, V. P., & de Oliveira, J. P. (2014). Exploring information retrieval features for author profiling. In L. Cappellato, N. Ferro, M. Halvey, & W. Kraaij (Eds.), *Proceeding CLEF-2014* (pp. 1164–1171). CEUR. Working Notes.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining. Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann.

Yule, G. (2010). *The study of language*. (4th Ed.) Cambridge: Cambridge University Press.

Zhai, C. X., & Massung, S. (2016). *Text data management and analysis*. New York: The ACM Press.

Zobel, J., & Moffat, A. (1998). Exploring the similarity space. *ACM-SIGIR Forum, 32*(1), 18–34.