JOURNAL OF APPLIED INTELLIGENT SYSTEMS & INFORMATION SCIENCES
VOL. 1, ISSUE. 1, PP. 54-63, FEBRUARY 2020.
*Available at:*www.journal.research.fanap.com
DOI: 10.22034/JAISIS.2020.219089.1006

AUTHORSHIP CLUSTERING USING HOMOGENEOUS FEATURE SPACE AND TWO-STEPPED AUTOMATIC FUZZY CMEANS CLUSTEING

Mohammad Aminian[1,*], Mahdi Eskandari[2]
[1]*Computer Engineering Department, Bu Ali Sina University, Hamedan, Iran*
[2]*Computer Engineering Department, Bu Ali Sina University, Hamedan, Iran*

ABSTRACT

Identifying the authorship of either an anonymous or a doubtful document constitutes a cornerstone for automatic forensic applications. Moreover, it is a challenging task for both humans and computers considering the complex content of documents with a variety of backgrounds. Due to its nature, this task is always considered as an unsupervised one. Clustering documents according to the linguistic style of the authors who wrote them is a less charted territory. PAN evaluation framework is the first effort to promote the development of the author clustering as an attempt to address this problem. There are different approaches to address the task and this article proposes a method based on a set of homogeneous features and two-step automatic FCM clustering. We use word Ngram, part-of-speech tagging and some other context free features, then document similarity graph (DSG) was used for estimating the number of clusters; finally we used FCM to cluster corpus. We conducted the task in a very short period of time and our performance results are comparable with leaderboard competitors in PAN CLEF 2017 challenge.

***KEYWORDS*:** Author Clustering, Homogeneous Feature Space, Word Ngram, Part-of-Speech, Fuzzy Cmeans, PAN.

1. INTRODUCTION

Stylometry is the study of distinct linguistic styles and individual writing practices with the purpose of determining the authorship of a written piece of text (Holmes, 1998). A writing style represents the linguistic choices of a writer that persist throughout one's work. Stylometric research is inspired by the hypothesis that every person has a unique and distinct writing style, referred to as "stylistic fingerprint" Holmes (1994) that can be measured and learned. Here, a stylistic fingerprint of a writer means a set of features frequently used by that author such as word length, sentence length, choice of certain words, and syntactic structure of a sentence. Stat-of-the-art perspective of stylometry research is authorship analysis (Holmes,1994; Juola, 2008; Chaski, 2001; Koppel et al., 2011; Abbasi & Chen, 2005; Stamatatos et al., 2000). In the recent past, the domain of authorship analysis has embraced new dimensions of research typically with the emergence of machine learning techniques for text mining. One of the recent and emerging trends in authorship analysis is computational extraction of stylometric features from the text of an author instead of engineering the stylometric features manually (Rosen-Zvi et al., 2012; Seroussi et al., 2011; Caliskan-Islam, 2015). The main focus of authorship identification is deciding the most probable author of a target document among a list of known authors (Juola, 2008). From machine learning perspective, authorship identification can be perceived as one label multiclass text classification problem where the role of classes is played by contestant authors (Sebastiani, 2002).

The detailed literature review in the domain of authorship identification for the last two decades revealed that it is a field of great interest and has been mainly applied to the English language (Chaski, 2011; Abbasi& Chen, 2005; Holmes et al., 2001; Koppel et al., 2009). Additionally, few solitary efforts were undertaken for

*Corresponding Author, Email: m.aminiyan@gmail.com

RECEIVED: 9 FEBRUARY, 2020; **ACCEPTED:** 30 FEBRUARY, 2020; **PUBLISHED ONLINE:** 30 FEBRUARY, 2020

other languages including Arabic (Abbasi& Chen, 2005; Altheneyan&Menai, 2014) Dutch (Juala&Baayen, 2005; Hoorn et al., 1999), Greek (Stamatatos et al., 2000; Kešelj et al., 2003), and Portuguese (Pavelecet et al., 2008; Silva et al., 2001).

In the following sections, we provide a detailed description of our approaches to solve the two subtasks of the Author Identification track of PAN 2017. The problem instance is a tuple <*K*; *U*; *L*> where *K* is a set of documents <k_1, k_2, k_3,…,k_n> authored by the different authors, *U* is the genre of the document and *L* is the enumerated value specifying the language of the documents: English, Dutch or Greek. All documents in a problem instance are in the same language and same genre.

This lab report is structured as follows: In section 3.1 we propose a number of different features that characterize documents from widely different points of view: character, word, parts-of speech, sentence length and punctuation. We construct non-overlapping groups of homogeneous feature. In section 3.2 we present the two-step method for authorship clustering task by employing a graph-based approach and the standard Fuzzy C means algorithm. Then we employ a new feature space to determine links between documents. Finally, in section 4 we describe our results on the training corpus and the final evaluation corpus of PAN-2017.

2. Related Works

In the literature, a large number of works in the past had focused on computational linguistics-based methods for the identification of stylometric features in the texts and their application in order to attribute the possible author of each text. The focus of these approaches was on improving various tasks of authorship analysis of a piece of text such as authorship identification, author verification, and author profiling.

The first approach to authorship identification is the use of univariate or multivariate measures that can reflect the style of a particular author. Individual measures such as word occurrence or the frequency of specific words (Mendenhall, 1887), mean sentence length or word length (Yule, 1944), and word richness (Sebastiani, 2002) were proposed; however, none of these univariate measures prove to be adequate (Grieve, 2007). The philosophy behind the multivariate approach is to take the documents as points in vector space, and to assign the query document to the author, whose documents are closest to the query document (Burrows, 2002) by using some suitable similarity measures; furthermore, other distance-based similarity metrics such as Euclidean distance, Kullback-Leibler, and Hellinger distance were applied to various feature sets for authorship identification (Chaski, 2001; Kešelj et al., 2003; Reza et al., 2009).

The second approach includes statistical machine learning techniques. Individual author is a category value, and a classification model is built. Machine learning techniques are further separated into two subgroups, one is supervised and the other is unsupervised. In supervised learning, a classifier is built using both features and the categorical value. However, unsupervised models work on unlabelled data (Blei et al., 2003). For authorship identification, supervised techniques include support vector machine (SVM) (Koppel et al., 2009; Argamon et al., 2009; Stamatatos, 2008), decision trees (Abbasi& Chen ,2005), linear discriminant analysis (Chaski ,2005), and neural networks(Tweedieetal., 1996; Zheng et al., 2006). Support vector machine out-performed other supervised techniques such as linear discriminant analysis and neural networks in terms of accuracy. Unsupervised classification techniques include principal component analysis (PCA) (Burrows 1997; Jamaket et al., 2012), cluster analysis (Holmes, 1992), word2vec (Mikolov et al., 2013), doc2vec or distributed document representation (Le &Mikolov, 2014), and LDA (Rosen-Zvi et al. 2004; Seroussietal, 2017; Arun et al., 2009; Savoy, 2013). The work discussed by Blei et al. (2003) is the first attempt to address author identification in Urdu text and that approach is improved in this paper by using tf-idf along cosine similarity and a KNN-based classification module for more accurate results. First systematic study of authorship identification by using the enhanced version of LDA was presented by Rosen-Zvi et al. (2004). LDA model has the ability to identify all hidden topics from a large number of features and present them as LDA topics, serving for dimensionality reduction and making it attractive for text analysis problems. Anwar et al. (2019) presented an approach for authorship identification in English and Urdu texts using the LDA model with *n*-grams texts of authors and cosine similarity. The approach uses similarity metrics to identify various learned representations of stylometric features and uses them to identify the writing style of a particular author. The proposed LDA-based approach emphasizes instance-based and profile-based classifications of an author's text.

Although our work focuses specifically on PAN CLEF 2017 data (Tschuggnall et al., 2017), we briefly review other approaches as well. Kocher & Savoy (2017) proposed an effective unsupervised author clustering and authorship linking model called SPATIUM. The suggested strategy can be adapted without any difficulty to different languages (such as Dutch, English, and Greek) in different text genres (e.g., newspaper articles and reviews). As features, they suggest using the *m* most frequent terms (isolated words and punctuation symbols) or the *m* most frequent character *n*-grams of each text. Applying a simple distance measure, they determine whether there is enough indication that two texts were written by the same author or not.

Gómez-Adorno et al. (2017) performed a hierarchical clustering analysis of different document features: typed and untyped character *n*-grams, and word *n*-grams. We experimented with two feature representation methods, log-entropy model, and tf-idf; while tuning minimum frequency threshold values to reduce the dimensionality. Their system was ranked 1 in both subtasks, author clustering and authorship-link ranking.

Another competitor team in CLEF 2017, Karas et al. (2017), proposed methods for author identification task divided into author clustering and style breach detection. Their solution to the first problem consists of locality-sensitive hashing based clustering of real-valued vectors, which are mixtures of stylometric features and bag of *n*-grams. For the second problem, we propose a statistical approach based on some different tf-idf features that characterizes documents. Applying the Wilcoxon Signed Rank test to these features, we determine the style breaches.

García et al. (2017) proposed a graph-based method, specifically β compact clustering, for discovering the groups of documents written by the same author. The β-compact algorithm is based on the analysis of the similarity between documents and they belong to the same group as long as the similarity between them exceeds the threshold β and it is the maximum similarity with respect to the other documents.

A simple distance measure has been applied to the author clustering problem to determine which documents are written by the same author. This simple distance measure works with the probability distribution of character sequences of a document, making it insensitive to language differences. The most frequent feature *k*, where *k* is chosen to be 300, determines the distribution where punctuation is present. Also, the uppercase letters are transformed to lowercase symbols, while a threshold of 3:0 remains for the symmetric distance score. In addition, characters 2-grams are chosen due to their best outcomes (Alberts, 2017).

Finally Halvani&Graner (2017) proposed a simple method; In order to group the documents by their authors, they use k-Medoids, where the optimal k is determined through the computation of silhouettes. To determine links between the documents in each cluster, they apply a predefined compressor as well as a dissimilarity measure. The resultant compression-based dissimilarity scores are then used to rank all document pairs.

The proposed scheme does not require (text-) preprocessing, feature engineering or hyper-parameter optimization, which are often necessary in author clustering and/or other related fields.

3. Proposed Method

This section describes the proposed method. First preprocessing and feature extraction from documents is introduced, then the way we use this features to cluster authors is explained.

3.1. Preprocessing

The proposed method extracts a number of different features from each document. For ease of presentation, these features are grouped in homogeneous categories, as described below.

3.1.1. Features

Word ngrams (WG): all characters are converted to lowercase and then the document is transformed to a sequence of words. White spaces, punctuation characters and digits are considered as word separators. We count all word ngrams, with $n \leq 3$, and we obtain a feature for each different word ngram which occurs in the training documents set of a given language (Mansoorizadeh et al., 2016). It should be mentioned that, word unigrams and 2-gram are used in clustering task as well as the pre-processes related to it and word 3-gram only used in link computation phase.

In order to normalize these sets of features we use term frequency-inverse document frequency (tf -idf) for each set of documents (each problem) (Mansoorizadeh et al., 2016).

POS (parts-of- speech) tag ngrams (PG): We apply a parts of speech (POS) tagger on each document, which assigns words with similar syntactic proper ties to the same POS tag. We count all POS ngrams, with $n \leq 2$, and we obtain a feature for each different POS ngram which occurs in the training set documents of a given language (Mansoorizadeh et al., 2016).

Sentence lengths (SL): We transform the document to a sequence of tokens, each token being a sequence of characters separated by one or more blank spaces. Next, we transform the sequence of tokens to a sequence of sentences, each sentence being a sequence of tokens separated by one of the following characters:

., ;, :, !, ?. We count the number of sentences whose length in tokens is n, with $n \in \{1,..,15\}$: we obtain a feature for each value of n (Mansoorizadeh et al., 2016).

Punctuation ngrams (MG): We transform the document by removing all characters not included in the following set: {,, ., ;, :, !, ?, "}—the resultant document thus consists of a (possibly empty) sequence of characters in that set. We then count all character ngrams of the resultant document, with $n \geq 2$, and we obtain a feature for each different punctuation ngrams which occurs in the training documents set of a given language (Mansoorizadeh et al., 2016).

In order to preprocess documents we use python NLTK 3.4.5 package (Juola, 2008). After creating the feature space we simply separate word 2grams for authorship link task and use the rest of features for clustering. We assume that word 2grams consist of very specific relation which can have a better effect inside each cluster for determining the level of similarity between documents.

3.1.2. Data normalization

After feature extraction, we normalize the value of each feature using min-max normalization in order to remove the impact of different scale spaces:

$$X_{new} = \frac{X_{old} - Min}{Max - Min} \tag{1}$$

Where X_{old} is the old value of X and Max is the maximum value of feature X and min is the Minimum value of feature X.

3.2. Two-step unsurpervised method

After extracting features from each document, we use train data to build a model in order to cluster documents and then assign each article to the right author.

3.2.1. Step 1: Determining the number of authors

Considering the fact that the number of authors is unknown first we have to determine the number of authors for each problem, for example, we have to determine the number of clusters for the clustering algorithm (Fuzzy Cmeans is used and the number of clusters for this method should be set). The number of clusters should be set by the developer based on specifications of the problem. Assigning a proper number is a challenging task. A document similarity graph (DSG) algorithm has been used. DSG is an undirected graph showing similarity relations between documents based on their contents (Chaski, 2001). The nodes of this graph are documents and the edges between documents are defined by the similarities between them using Eq. (2):

$$Z(i,j) = \frac{X.Y}{|X||Y|} = \frac{\sum_{i=1}^{n} X_i.Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{2}$$

$$GS_{mat} = \begin{cases} 1 & Z(i,j) \geq \delta \\ 0 & Z(i,j) \prec \delta \end{cases}$$

Where x_i and y_i are features of X_i and Y_j documents respectively and δ is the threshold which defines the existence of similarity between the two documents. In this paper, the δ parameter is set to 0.5. Moreover, Z is the cosine similarity between the two documents (Koppel et al., 2011). The number of clusters has been determined using the number of sub-graphs drawn from DSG. To find the number we just count the nodes with value more than 65 percent of the number of all document for example if we have 100 documents in problem folder, we count nodes which have more than 65 incoming edges.

3.2.2. Step 2: clustering and computing links

After calculating the number of clusters, we use fuzzy *C*-means clustering Phu et al. (2017) in order to perform clustering task. When clustering is completed, we collect the results and employ simple similarity task in each cluster. We compute similarity based word 3grams features and cosine similarity Eq. (2), in order to assign each new document to the right author.

4. EVALUATION

In this section we perform evaluation on PAN 2017 data sets (train and test). First we introduce evaluation parameters, then we report our individual and comparison results.The evaluation was performed using the TIRA platform, which is an automated tool for deployment and evaluation of the software (Tschuggnallet al., 2017). The data access is restricted such that during a software run the system is encapsulated and thus ensuring that there is no data leakage back to the task participants (Tschuggnallet al., 2017). This evaluation procedure also offers a fair evaluation of the time needed to produce an answer.

During the PAN CLEF 2017 evaluation campaign, six corpora (or test collections) were built each containing 30 problems (10 for training and 20 for testing). In each problem, all the texts matching the same language, are in the same text genre, and are single-authored, but they may differ in text-length and can be cross-topic (Tschuggnall et al., 2017). The number of distinct authors is not given. In this context, the task is defined as:

Given a problem of up to 50 short documents, identify authorship links and groups of documents by the same author. Each of the six corpora is a combination of one of the three languages (English, Dutch, or Greek) and one of the two genres: newspaper articles or reviews. An overview of these corpora is depicted in Table 1. Considering the six benchmarks, we have 120 problems to test and 60 problems to train (pre-evaluate) our system.

Table 1. PAN CLEF 2017 training corpora statistics.

Corpus			Training sets			
			Text	Authors	Single	Terms
English	News paper	(EN)	20	5.6 [3-10]	1.8 [0-6]	62 [56-67]
English	Reviews	(ER)	19.4	6.1 [4-10]	1.9 [0-5]	73 [70-77]
Dutch	News paper	(DN)	20	5.3 [4-8]	2.0 [0-5]	59 [53-66]
Dutch	Reviews	(DR)	18.2	6.5 [5-8]	0.3 [0-2]	159 [143-186]
Greek	News paper	(GN)	20	6.0[4-8]	1.5 [0-5]	76 [66-88]
Greek	Reviews	(GN)	20	6.1 [4-8]	2.1 [0-6]	62 [53-70]

For each corpus, we have 10 problems in the training dataset containing the average number of texts given under the label “Texts”. The number of distinct authors on average together with the range for each corpus is indicated in the column “Authors”, and the average with the minimum and maximum number of authors with only a single document is presented under the label “Single”. Finally, the average number of the terms (isolated words and punctuation symbols) is given in the column “Terms”. For example, with the English newspaper collection (training set), 20 texts are written, on average, by 5.6 authors and we can find 1.8 authors who wrote only one single article. These metrics are not available for the test corpora because the datasets remain undisclosed because of the TIRA system. We only know that the same combinations of language and genre are present.

4.1. Evaluation parameters

There are multiple evaluation measures available for clustering tasks. In general, a clustering evaluation measure can be intrinsic (when the true labels of data are not available) or extrinsic (when true labels of data

are available). Given that the information about the true authors of the documents are available, our task fits the latter case. Among a variety of extrinsic clustering evaluation metrics, we opted to use BCubed Precision, Recall, and the *F*-score. The latter has been found to satisfy several formal constraints including cluster homogeneity, cluster completeness, and the rag bag criterion (where multiple unrelated items are merged into a single cluster). Let d_i be a document in a collection (i = 1,…, N). Let C (d_i) be the cluster di is put into by a clustering model and A (d_i) be the true author of d_i. Then, considering two documents of the collection d_i and d_j, a correctness function can be defined as follows:

$$correct(d_i, d_j) = \begin{cases} 1 & \text{if } A(d_i) \wedge C(d_i) = C(d_j) \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

The BCubed precision of a document d_i is the proportion of the documents in the cluster of d_i (including itself) by the same author of d_i. Moreover, BCubed recall of d_i is the proportion of documents by the author of d_i that are found in the cluster of d_i (including itself). Let C_i be the set of documents in the cluster of di and A_i be the set of documents in the collection by the author of d_i BCubed precision and recall of d_i are then defined as follows:

$$precision(d_i) = \frac{\sum_{d_j \in c_i} Correction(d_i, d_j)}{|C_i|} \tag{4}$$

$$recall(d_i) = \frac{\sum_{d_j \in c_j} Correct(d_i, d_j)}{|A_i|} \tag{5}$$

Finally, the overall BCubed precision and recall for one collection is the average of precision and recall of documents in the collection, whereas the BCubed *F*-score is the harmonic mean of BCubed precision and recall:

$$BCubedprecision = \frac{1}{N}\sum_{i=1}^{N} precision(d_i) \tag{6}$$

$$BCubedrecall = \frac{1}{N}\sum_{i=1}^{N} recall(d_i) \tag{7}$$

$$BCubed\ F = 2 \times \frac{BCubed\ \text{precision} \times \text{BCubed recall}}{BCubed\ \text{precision} + \text{BCubed recall}} \tag{8}$$

Regarding the authorship-link ranking task, we use average precision (AP) to evaluate submissions. This is a standard scalar evaluation measure for ranked retrieval results. Given a ranked list of authorship links for a document collection, average precision is the average of non-interpolated precision values at all ranks where true authorship links were found. Let L be the set of ranked links provided by a submitted system and T the set of true links for a given collection. If l_i is the authorship link at i-th position of L then a relevance function, precision at cutoff i in the ranked list, and AP are defined as follows:

$$relevent(i) = \begin{cases} 1 & \text{if } l_i \in T \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$precision(i) = \frac{\sum_{j=1}^{i} relevent(j)}{i} \tag{10}$$

$$AP = \frac{\sum_{i=1}^{|L|} precision(i) \times relevent(i)}{|T|} \tag{11}$$

It is important to note that AP does not punish verbosity, i.e. every true link counts even if it is at a very low rank. Therefore, by providing all possible authorship links one can attempt to maximize AP. In order to show how effective a system is in top-ranked predictions, we also provide *R*-precision (*RP*) and *P*@10, which are defined as follows:

$$R - precision = \frac{\sum_{i=1}^{R} relevent(i)}{R} \tag{12}$$

$$P@10 = \frac{\sum_{i=1}^{10} relevent(j)}{10} \tag{13}$$

Where *R* is the number of true authorship links. Focusing on either the top *R* or the top 10 results, these metrics ignore all other answers.

For multiple instances of author clustering problems, mean scores of all the above measures are used to evaluate the overall performance of submissions in all available collections. Finally, submissions are ranked according to Mean *F*-score (MF) and Mean Average Precision (MAP) for complete author clustering and authorship-link ranking, respectively.Table 1 shows the results of train dataset. It is obvious that our method have high Bcubed recall hence we can say the method clusters the same items almost great in each cluster but by investigating our method's Bcubed precision, we can clearly say that the number of cluster or even the way we measure similarity does not tuned perfectly.Like Table 1, Table 2 results of test dataset, also illustrates high level of Bcubed recall in most of the problem sets, in contrast with Bcubed precision which is not high as Bcubed precision. But it is obvious that results obtained from test dataset are better than train data. It shows the ability of the system to generalize the new problems. But the major defect of the system which is lower Bcubed precision than that of the recall one still exists.

Table 2. Evaluation for the training corpora

Corpus	Precision	Recall	F1	MAP
EN	0.3533	0..8351	0.4836	0.4029
ER	0.4098	0.8599	0.5332	0.3876
DN	0.3344	0.8905	0.4762	0.4508
DR	0.4464	0.9247	0.5988	0.3310
GN	0.3863	0.8630	0.5316	0.3409
GR	0.3515	0.8725	0.4929	0.3458
Overall	0.3803	0.873	0.521	0.3765

To put those values in perspective, we can see our results in Table 4 in comparison with the other participants using macro-averaging for the effectiveness measures and showing the total runtime sorted by the final score. Overall, the method is ranked 2 out of 7 approaches. Generally, there are only small differences in the BCubed *F*1 between the participants. Conversely, the MAP shows substantial variations and impacts the final score the most. The runtime only shows the actual time spent to classify the test set.

On TIRA, there was the possibility to first train the system using the training set which had no influence on the final runtime. Since we have an unsupervised system it did not need to train any parameters, but this possibility might have been used by other participants. Overall, we achieve excellent results using a rather simple and fast approach in comparison with the other solutions.

Table 3. Evaluation for the test corpora

Corpus	Precision	Recall	*F*1	MAP
EN	0.5244	0.7539	0.6068	0.4700
ER	0.5751	0.6609	0.5696	0.3876
DN	0.5291	0.7381	0.5860	0.4508
DR	0.5597	0.5428	0.5988	0.3310
GN	0.4865	0.7520	0.5316	0.4451
GR	0.4876	0.6162	0.5021	0.3458
Overall	0.522	0.6773	0.5572	0.4104

Table 4. Comparison results for the test corpora

Rank	Method	*F*1	MAP	RunTime
1	Gómez-Adorno et al. (2017)	0.5732	0.4552	00:02:05
2	Proposed Method	0.5572	0.4104	00:00:27
3	Kocher & Savoy (2017)	0.5517	0.3951	00:00:41
4	García et al. (2017)	0.5647	0.3800	00:15:49
5	Halvani&Graner (2017)	0.5488	0.1394	00:12:25
6	Karas (2017)	0.4663	0.1252	00:00:26
7	Alberts (2017)	0.5276	0.0416	00:01:45

As we can see in Table 4, the proposed method is the fastest method compared to other competitors, this happens because we use simple but effective features combined with fuzzy clustering algorithm and simple similarity measure. Our method works based on this theory that if we could cluster train document properly and define efficient cluster heads, a simple similarity as well as complex similarity measure could perform. Results show that the methods of ours and those of the competitors' do not illustrate perfect *F*1 measure but at least our works do not consume large amount of resources to prepare results.

5. Conclusion and Future Works

In this research we proposed a simple but effective method to cluster authors and the identified links between them. We use a set of homogenous features combined with FCM clustering. We determine a number of clusters automatically. The most important aspect of our method is that we use a number of simple features along with the automatic clustering procedure which is performed in a very short period of time (the shortest time per document in comparison with other methods). Results show that our work can compete with the work of other CLEF 2017 competitors but the low level of precision in our results informs us that, we should focus on the method to identify authors more accurately. For future works we suggest fellow researchers that they work on distance measures and use semi-supervised methods like GMM in order to complete the task with higher performance. Also based on previous works, deep neural networks would be a good approach to follow.

References

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

Altheneyan, A.S., & Menai, M.E.B. (2014). Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, 26(4), 473-484.

Alberts, H. (2017). Author clustering with the Aid of a Simple Distance Measure. *In CLEF (Working Notes)*.

Anwar, W., Bajwa, I.S., & Ramzan, S. (2019). Design and implementation of a machine learning-based authorship identification model. *Scientific Programming.*

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119-123.

Arun, R., Saradha, R., Suresh, V., Murty, M., & Madhavan, C. (2009). Stopwords and stylometry: a latent Dirichlet allocation approach. *In NIPS workshop on Applications for Topic Models*.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation, Journal of Machine Learning Research, 3(3), 993-1022.

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

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

Caliskan-Islam, A. (2015). *Stylometric Fingerprints and Privacy Behavior in Textual Data.* Philadelphia:Drexel University.

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

Chaski, C.E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1), 1-13.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251-270.

García-Mondeja, Y., Castro-Castro, D., Lavielle-Castro, V., & Muñoz, R. (2017). Discovering Author Groups using a β-compact graph-based clustering. *In CLEF (Working Notes), CEUR Workshop Proceedings*.

Gómez-Adorno, H., Aleman, Y., Ayala, D.V., Sanchez-Perez, M.A., Pinto, D., & Sidorov, G. (2017). Author Clustering using Hierarchical Clustering Analysis. *In CLEF (Working Notes)*.

Halvani, O., & Graner, L. (2017). Author Clustering based on Compression-based Dissimilarity Scores. *In CLEF (Working Notes)*.

Holmes, D.I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 155(1), 91-120.

Holmes, D.I. (1998). The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3), 111-117.

Holmes, D.I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2), 87-106.

Holmes, D.I., Robertson, M., & Paez, R. (2001). Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3), 315-331.

Hoorn, J.F., Frank, S.L., Kowalczyk, W., & van Der Ham, F. (1999). Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3), 311-338.

Juola, P. (2008). Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3), 233-334.

Jamak, A., Savatić, A., & Can, M. (2012). Principal component analysis for authorship attribution. *Business Systems Research Journal: International Journal of the Society for Advancing Business & Information Technology*, 3(2), 49-56.

Juola, P. & Baayen, R.H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(1), 59-67.

Karas, D., Spiewak, M., & Sobecki, P. (2017). OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection. *In CLEF (Working Notes)*.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). *N*-gram-based author profiles for authorship attribution. *In Proceedings of the Conference Pacific Association for Computational Linguistics*, Halifax, NS, Canada.

Kocher, M., & Savoy, J. (2017). UniNE at CLEF 2017: Author Profiling Reasoning. *In CLEF (Working Notes)*.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9-26.

Koppel, M., Schler, J., & Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1), 83-94.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *In International Conference on Machine Learning*, Beijing, China.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *In Proceedings of International Conference on Machine Learning*, Atlanta, GA, USA.

Mansoorizadeh, M., Aminian, M., Rahgooy, T., & Eskandari, M. (2016). Multi Feature Space Combination for Authorship Clustering. *In CLEF (Working Notes)*.

Mendenhall, T.C. (1887). The characteristic curves of composition. *Science*, 9(214), 237-249.

Pavelec, D., Oliveira, L.S., Justino, E.J., & Batista, L.V. (2008). Using conjunctions and adverbs for author verification, *Journal of Universal Computer Science*, 14(18), 2967-2981.

Phu, V.N., Dat, N.D., Tran, V.T.N., Chau, V.T.N., & Nguyen, T.A. (2017). Fuzzy *C*-means for english sentiment classification in a distributed system. *Applied Intelligence*, 46(3), 717-738.

Raza, A.A., Athar, A., & Nadeem, S. (2009). *N*-gram based authorship attribution in Urdu poetry. *In Proceedings of the Conference on Language & Technology*, Poznan, Poland.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Bandff, Canada.

Savoy, J. (2013). Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, 49(1), 341-354.

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

Seroussi, Y., Zukerman, I., & Bohnert, F. (2011). Authorship attribution with latent Dirichlet allocation. *In Proceedings of the fifteenth Conference on Computational Natural Language Learning*,OR, USA.

Silva, R.S., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., & Maia, B. (2011). 'twazn me!!!;('automatic authorship analysis of micro-blogging messages. *In International Conference on Application of Natural Language to Information Systems*, Berlin, Heidelberg.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2), 790-799.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Text genre detection using common word frequencies. *In Proceedings of the 18th Conference on Computational linguistics*, Stroudburg, USA.

Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2017). Overview of the author identification task at PAN-2017: style breach detection and author clustering. *In Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al.* (pp. 1-22).

Tweedie, F.J., Singh, S., & Holmes, D.I. (1996). Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, 30(1), 1-10.

Yule, G. (1944). statistical study of literary vocabulary, *Modern Language Review*, 39(3), 291–293.

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