Taylor & Francis
Taylor & Francis Group

Cybernetics and Systems
An International Journal

ISSN: 0196-9722 (Print) 1087-6553 (Online) Journal homepage: https://www.tandfonline.com/loi/ucbs20

AUTHORSHIP ATTRIBUTION OF *RESPONSA* USING CLUSTERING

Yaakov HaCohen-Kerner & Orr Margaliot

To cite this article: Yaakov HaCohen-Kerner & Orr Margaliot (2014) AUTHORSHIP ATTRIBUTION OF *RESPONSA* USING CLUSTERING, Cybernetics and Systems, 45:6, 530-545, DOI: 10.1080/01969722.2014.945311

To link to this article: https://doi.org/10.1080/01969722.2014.945311

Cybernetics and Systems: An International Journal, 45:530–545

ISSN: 0196-9722 print/1087-6553 online
DOI: 10.1080/01969722.2014.945311

Taylor & Francis
Taylor & Francis Group

Authorship Attribution of *Responsa* Using Clustering

YAAKOV HACOHEN-KERNER and ORR MARGALIOT

Department of Computer Science, Jerusalem College of Technology – Lev Academic Center, Jerusalem, Israel

Authorship attribution of text documents is a "hot" domain in research; however, almost all of its applications use supervised machine learning (ML) methods. In this research, we explore authorship attribution as a clustering problem, that is, we attempt to complete the task of authorship attribution using unsupervised machine learning methods. The application domain is responsa*, which are answers written by well-known Jewish rabbis in response to various Jewish religious questions. We have built a corpus of 6,079* responsa*, composed by five authors who lived mainly in the 20th century and containing almost 10M words. The clustering tasks that have been performed were according to two or three or four or five authors. Clustering has been performed using three kinds of word lists: most frequent words (FW) including function words (stopwords), most frequent filtered words (FFW) excluding function words, and words with the highest variance values (HVW); and two unsupervised machine learning methods: K-means and Expectation Maximization (EM). The best clustering tasks according to two or three or four authors achieved results above 98%, and the improvement rates were above 40% in comparison to the "majority" (baseline) results. The EM method has been found to be superior to K-means for the discussed tasks. FW has been found as the best word list, far superior to FFW. FW, in contrast to FFW, includes function words, which are usually regarded as words that have little lexical meaning. This might imply that normalized frequencies of function words can serve as good indicators for authorship attribution using unsupervised ML methods. This finding supports previous findings about*

Address correspondence to Yaakov HaCohen-Kerner, Department of Computer Science, Jerusalem College of Technology – Lev Academic Center, 21 Havaad Haleumi St., P.O.B. 16031, 9116001, Jerusalem, Israel. E-mail: kerner@jct.ac.il

the usefulness of function words for other tasks, such as authorship attribution, using supervised ML methods, and genre and sentiment classification.

KEYWORDS authorship attribution, Hebrew, responsa*, text clustering, unsupervised machine learning methods, word lists*

INTRODUCTION

Authorship attribution of a text document is the identification of the author of a tested document from a group of potential authors. Authorship attribution of a set of text documents is the categorization of each document according to the most appropriate author from a group of potential authors. Authorship attribution (also called authorship identification or authorship prediction) studies methods for differentiating among the styles of different authors.

The basic variant of the authorship attribution problem is as follows: given text-writing examples of several authors, there is a need to determine which of them authored a given anonymous text. This variant is equivalent to the basic paradigm of the text categorization problem (Lewis and Ringuette 1994; Sebastiani 2002). A comprehensive overview of various authorship attribution methods including history of such methods is presented by Koppel et al. (2009) and Stamatatos (2009).

Authorship attribution is a subdomain of text classification (TC). Nearly all existing authorship attribution applications that combine machine learning (ML) methods use supervised ML methods.

In the case of supervised ML, all documents in a training set are preassigned a class before the training process, after which, in most cases, the model has to assign each given document to one predefined category (Meretakis and Wuthrich 1999).

The availability of tens of supervised ML methods enables researchers to deal with many features. For instance, Madigan et al. (2005) use all words that appear at least twice in the corpus, and Stamatatos (2006) uses the 1,000 most frequent words. A comprehensive summary of various ML methods used for TC can be found in Sebastiani (2002).

However, because most text documents are not preassigned to a class, they are unsuitable for supervised ML. Moreover, the problem of usage of unsupervised methods in authorship attribution is not well studied.

A possible solution to deal with authorship attribution for such documents is to utilize unsupervised ML methods. Unsupervised ML finds similarities between text data in order to determine whether they can be characterized as forming a group (cluster).

In this research, we want to apply authorship attribution using unsupervised ML for *responsa*. *Responsa* are unlabeled answers written approximately

during the last 1,000 years. The *responsa* were authored by well-known Jewish rabbis in response to various questions submitted to them. These documents are taken from a widespread variety of Jewish domains: customs, holidays, kosher food, prayers, and synagogues. Each *responsa* is based on both ancient Jewish writings and answers given by previous rabbis over the years. The *responsa* are written mainly in Hebrew and some words are written in Aramaic.

Supervised classification of *responsa* has been carried out in several previous studies. Koppel et al. (2004, 2006) and HaCohen-Kerner et al. (2008) use different types of "bags of words," and HaCohen-Kerner et al. (2010a, 2010b) use stylistic feature sets.

We aim to cluster a corpus of *responsa* according to authors using unsupervised ML methods. The chosen feature set for each clustering experiment is one of the three following word lists: (1) FFW (most frequent words excluding function words), (2) FW (most frequent words including function words), and (3) HVW (words with the highest variance values, words whose frequencies vary considerably according to the documents). Our initial assumption is that FFW will perform better than FW and HVW because function words have little lexical value and words with the highest variance values are assumed to be "bad" features for clustering tasks.

The structure of this article is as follows. "Text Clustering and Its Application to Authorship Attribution of *Responsa*" gives a brief review of text clustering, its application for authorship attribution of documents in general and of *responsa* in particular, and a short overview of document clustering using "bags of words." The section following that introduces our model. "Experimental Results" describes the examined corpus, presents the experimental results, and analyzes them; this is followed by "Analysis of the Precision, Recall, and F-Measure Values." The final section summarizes and proposes future directions for research.

TEXT CLUSTERING AND ITS APPLICATION TO AUTHORSHIP ATTRIBUTION OF *RESPONSA*

Text Clustering

Text clustering is an automatic grouping of unlabeled text documents into groups that are called clusters. Clustering of text documents is the process of creating a set of clusters in such a way that documents within one cluster are similar and documents from different clusters are dissimilar.

The basic text clustering method utilizes a "bag of words," in which each word serves as a feature. The feature selection process is important for the success and efficiency of the clustering process. That is to say, relevant features and an optimal number of features are critical for the success of the clustering process and its efficiency. Furthermore, an optimal set of features

might reduce the run time and/or the amount of required training data. There are various types of text clustering applications:

1. *Document clustering*: the creation of groups of similar documents in such a way that documents within one cluster are predominantly similar and documents from different clusters are mostly dissimilar. This application is the most straightforward and probably the most frequent. Examples of document clustering are provided later in this article.
2. *Word clustering*: the creation of groups of similar words or concept hierarchies, e.g., Baker and McCallum (1998) and Bekkerman et al. (2001).
3. *Document classification*: word-clusters (see previous) and cotraining methods (Nigam et al. 1998) created by clustering methods might help to improve the classification accuracy of the supervised classification methods.
4. *Document organization and browsing*: smart hierarchical organization of documents into clear and logical categories might be very useful for systematic browsing of the document collection in general and search of a specific document in particular. An example of document organization and browsing is the Scatter/Gather method (Cutting et al. 1992).
5. *Document summarization*: document summaries can be generated by using sentence clusters (Goldstein et al. 2000; Zha 2002), and cluster-digests (Schutze and Silverstein 1997).

Clustering methods can be divided into nonhierarchical methods such as K-means and Expectation Maximization (EM), and hierarchical methods such as hierarchical clustering. Comprehensive surveys of various clustering methods and/or applications can be seen at Jain et al. (1991), Dubes and Jain (1998), Steinbach et al. (2003), Li et al. (2006), and Aggarwal and Zhai (2012).

Mingzhe and Minghu (2012) use two kinds of features, two kinds of clustering methods (K-means and hierarchical clustering), and three different kinds of distance measures. The clustering experiments were carried out for 2, 3, 4, or 5 authors.

Author Clustering of Documents in General and of *Responsa* in Particular

As mentioned, most of the researchers who explore authorship attribution use supervised ML methods. In this subsection, we shall introduce a few works that investigate authorship attribution using unsupervised ML methods.

Chan et al. (2006) use author clustering and keyword clustering for a system that visualizes relationships and clusters of authors. They found that authors tend to be correctly grouped together in the same clusters because of joint publications and co-citations. He and Hui (2002) use agglomerative hierarchical clustering for author clustering and multidimensional scaling

for displaying author cluster maps. Layton et al. (2013) proposed an automated and unsupervised methodology for clustering documents by authorship. Their methodology is called N-gram Unsupervised Automated Natural Cluster Ensemble (NUANCE). It uses the K-means algorithm.

Basic clustering of *responsa* according to authors has been described in HaCohen-Kerner and Margaliot (2013). In that study, clustering was performed for a small corpus of documents (1,370 *responsa*), using only one unsupervised ML method; and experiments were carried for only two or three authors. Moreover, analysis of the precision, recall, and f-measure values is not supplied in this article.

In the current research, clustering of a larger corpus containing 6,079 *responsa* has been done according to two, three, four, or five authors (others than those in the research mentioned in the previous paragraph) using two unsupervised ML methods: K-means and EM. Analysis of the of the precision, recall, and f-measure values is also provided.

Document Clustering Using "Bags of Words"

Miao et al. (2005) implement three methods for document clustering: methods that use (1) words (after removing stopwords, stemming, pruning rare terms, and tf-idf weighting), (2) terms (based on their C-value, i.e., a frequency-based weight that accounts for nested terms), and (3) frequent character *n*-grams. They found that the *n*-gram-based representation provides the best performance with the lowest dimensionality.

Li and Chung (2005) and Li et al. (2008) implement text clustering based on frequent word sequences and frequent word meaning sequences instead of bags of words; "word meaning" refers to a concept expressed by synonymous word forms. Using these two sequence variants, they reduce the high dimensionality of the documents and measure the closeness between documents.

Yu (2012) uses function words[1] for the problem of Chinese authorship attribution in three different genres: novel, essay, and blog. Her system is able to distinguish three authors in each genre with various levels of success. The use of Chinese function words is most effective in distinguishing authors of novels (90%), followed by essays (85%), and blogs as the most difficult (68%).

Barring Yu (2012), all studies mentioned use bags of words as one of their features, inter alia process removal of stop words from their bags of words. Moreover, they use neither the most frequent words nor the words with the highest variance values.

[1]Function words express grammatical relationships. Function words include articles (e.g., the, a), pronouns (e.g., he, him, she, her), particles (e.g., if, then, well, however, thus), conjunctions (e.g., for, and, or, nor, but, yet, so), and auxiliary verbs (be, have, shall, will, may, and can). Function words were probably first proposed by Mosteller and Wallace (1964).

Forman (2003) presents an empirical study of 12 feature selection metrics. He claims that overly common words (stopwords), such as function words, may be removed on the grounds that they are ambiguous and occur so frequently that they are not discriminating for any particular cluster. In his research, a common word was identified if it occurred in over 50% of all documents.

THE MODEL

As mentioned above, Forman claims that stopwords should not be chosen as features. We decided to examine his claim by conducting two sets of experiments: one uses all these documents' words including stopwords, and the other excludes stopwords. In both sets, we used a normalized frequency for each feature. In addition, we examined a third word-list: words with the highest variance values.

We have defined the following terms:

1. **Common Features (CF)** – As define by Forman (2003), a word is identified as a common feature if it occurs in over 50% of all documents (at least once in each document). In these corpora, 84 CFs have been found. Following the test performed, around 80 of these are stopwords.
2. **Frequent Words (FW)** – Most frequent words including CF.
3. **Filtered Frequent Words (FFW)** – Most frequent words excluding CF.
4. **Highest Variance Words (HVW)** – Words with the highest variance values, i.e., words whose frequencies vary considerably according to the documents

We have defined the following algorithm:

- For each specific combination C of n authors ($n = 2, 3, 4, 5$)
  - Activate a clustering method (K-Means, EM) on the corpus that contains the documents composed by the authors included in the specific combination
    - For each word list: *FW, FFW*, and *HVW*
      - select: 100, 200, …, 1,000 as the number of its features

The accuracy in all experiments was measured by dividing the sum of correctly clustered documents by the total number of documents to be clustered.

The K-Means algorithm (MacQueen 1967; Lloyd 1982) is one of the simplest and oldest unsupervised ML methods that aim to solve the clustering task. This method partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The main idea is to define k centroids, one for each cluster. The method uses an iterative

refinement technique in order to calculate the new means to act as the new centroids of the new clusters. The optimal solution of this problem is NP-hard. Therefore, various heuristic polynomial variants have been defined and implemented.

The EM algorithm (Dempster et al. 1977) is another simple unsupervised ML method. It uses an iterative computation of maximum-likelihood estimates. Each algorithm's iteration consists of an expectation step followed by a maximization step.

We chose the K-Means and the EM methods for the following three reasons: (1) their simplicity, generality, and relatively quick run time, (2) the wide range of examples that fall under their definitions, and (3) the convenience and ease of their activation in WEKA (Witten and Frank 2005; Hall et al. 2009), using the default values as performed by Forman (2003).

EXPERIMENTAL RESULTS

The application domain contains *responsa* written in Hebrew and Aramaic. The *responsa* are answers written by foremost Jewish rabbis in response to various questions submitted to them. These documents are taken from a widespread variety of Jewish domains: customs, holidays, kosher food, prayers, and synagogues. Each *responsa* is based on both ancient writings and answers given by previous rabbis.

The examined *responsa* were downloaded from the Global Jewish Database (The *Responsa* Project[2]) at Bar-Ilan University. The corpus includes 6,079 *responsa*, composed by five authors, mainly in the 20th century, which contain 9,706,496 words (an average of 1597 words per document). These authors wrote *responsa* on the same general issues such as Shabbat, holidays, family life, kosher food, and so forth. Table 1 presents various details about these *responsa* and their authors.

We have performed all 26 possible combinations of clustering tasks according to two, three, four, or five authors. Each combination has been checked for each word list and for each ML method. Because we have three word lists and two ML methods, we have conducted a total of 156 different experiments.

Table 2 presents the best clustering results for two, three, four, or five authors using three word lists and the EM and K-Means ML methods. Results colored with bold red represent the best among all results in the same row (experiment) for all combinations of word lists and ML methods. Results colored with red (not in bold) represent the best among the results for the ML method, which achieves inferior results.

[2]http://www.biu.ac.il/ICJI/Responsa/index.html.

TABLE 1 General Statistical Information about the Examined Corpus

#	Book's Hebrew Title (pronounced in English)	Author	Period	Location	Documents	Total words	Average words per document
1	Iggerot Moshe	R. Moses Feinstein	1895–1985	Lithuania and USA	1832	2,264,231	1235.93
2	Divrei Yatziv	R. Yekutiel Yehuda Halberstam	1904–1995	Israel	930	1,461,930	1571.97
3	Be-Tzel Ha-Chochmah	R. Betzalel Stern	1911–1989	Slovakia, Austria, Australia and Israel	668	1,016,748	1522.08
4	Yabbia Omer	R. Ovadiah Yosef	1920–2013	Israel and Egypt	841	3,503,266	4165.6
5	Shevet Ha-Levi	R. Shmuel Halevi Wosner	1914–Today	Austria and Israel	1808	1,460,321	807.7

In addition, we present the improvement rates of the best word list for each ML method for every experiment. The improvement rate is calculated by subtracting the "majority" (the baseline result) from the best result. The

TABLE 2 Clustering Results for Two, Three, Four, or Five Authors Using Three Word Lists and Both EM and K-Means

		EM				K-Means			
#	Authors	FW	FFW	HVW	Best improvement rate	FW	FFW	HVW	Best improvement rate
1	1,2	**99.78%**	98.99%	92.36%	**33.45%**	99.64%	65.89%	96.71%	33.31%
2	1,3	**99.08%**	96.56%	97.56%	**25.80%**	98.48%	98.40%	98.32%	25.20%
3	1,4	**99.59%**	99.40%	98.32%	**31.05%**	99.40%	99.10%	98.92%	30.86%
4	1,5	**92.09%**	89.92%	75.19%	**41.76%**	75.91%	65.12%	64.70%	25.58%
5	2,3	**99.69%**	98.06%	98.81%	**41.49%**	99.37%	98.81%	97.56%	41.17%
6	2,4	**99.89%**	99.83%	99.27%	**47.38%**	**99.89%**	99.55%	99.94%	**47.38%**
7	2,5	80.53%	**97.44%**	92.62%	**31.41%**	89.74%	93.83%	91.38%	27.80%
8	3,4	97.68%	**99.80%**	98.01%	**44.07%**	97.61%	57.46%	97.42%	41.88%
9	3,5	98.18%	69.18%	**98.30%**	**25.28%**	70.44%	72.94%	69.87%	−0.08%
10	4,5	96.68%	97.74%	96.90%	29.49%	97.66%	97.24%	**98.75%**	**30.50%**
11	1,2,3	**98.75%**	80.90%	85.60%	**45.34%**	98.63%	86.47%	96.38%	45.22%
12	1,2,4	92.48%	93.34%	91.17%	42.49%	96.25%	**97.45%**	95.70%	**46.60%**
13	1,2,5	64.46%	70.33%	57.57%	30.24%	**71.60%**	61.53%	60.94%	**31.51%**
14	1,3,4	**98.32%**	92.25%	94.43%	**43.49%**	98.02%	77.58%	97.99%	43.19%
15	1,3,5	80.76%	72.24%	**82.89%**	**40.36%**	72.86%	66.13%	69.22%	30.33%
16	1,4,5	**87.44%**	77.75%	78.93%	**46.56%**	60.72%	69.85%	61.08%	28.97%
17	2,3,4	**98.52%**	97.46%	97.99%	**60.39%**	98.11%	94.05%	97.95%	59.98%
18	2,3,5	75.04%	74.93%	**89.93%**	**36.85%**	74.49%	72.14%	73.08%	21.41%
19	2,4,5	69.43%	**97.51%**	90.25%	**46.99%**	63.82%	62.59%	54.29%	13.30%
20	3,4,5	**96.68%**	71.66%	93.88%	**42.17%**	96.47%	77.75%	78.63%	41.96%
21	1,2,3,4	**98.08%**	86.00%	89.37%	**55.19%**	82.42%	81.83%	76.35%	39.53%
22	1,2,3,5	68.50%	**68.79%**	65.29%	**33.81%**	66.82%	66.78%	67.03%	32.05%
23	1,2,4,5	62.47%	**70.02%**	60.00%	**36.16%**	58.21%	65.87%	57.62%	32.01%
24	1,3,4,5	**81.39%**	65.70%	76.40%	**45.81%**	48.61%	51.72%	56.50%	20.92%
25	2,3,4,5	69.11%	71.65%	**85.75%**	**43.18%**	60.28%	68.71%	54.77%	26.14%
26	1,2,3,4,5	65.41%	**66.51%**	59.65%	**36.37%**	44.28%	58.50%	65.52%	35.38%

"majority" method is our baseline classifier. It assumes that every document belongs to the larger of the n categories (n is the number of authors in each experiment).

When considering the results in all 26 experiments in Table 2,

1. For the EM ML method, FW has been found as the superior set with 13 best results, FFW with 9 best results, and HVW with only 4 best results.
2. For the K-Means ML method, FW has been found as the superior set with 16 best results, FFW with 6 best results, and HVW with only 4 best results.
3. That is to say, for both ML methods, FW is definitely the best word list. FW includes function words, which are common to most documents of all the authors and, therefore, are usually regarded as words that have little lexical meaning. Thus, the results that FW is far superior to FFW are in contrast to Forman's opinion that common words should not be used for classification tasks. HVW, which represents words with the highest variance values, was found as less relevant for the performed clustering tasks.
4. The improvement rates of the best results in all experiments according to two, three, four, or five authors vary from 25.28% to 47.38%, 31.51% to 60.39%, 33.81% to 55.19%, and 35.38% to 36.37%, respectively.
5. The EM method was found to be superior to the K-Means method for the discussed clustering experiments. The best results in 22 out of 26 experiments were achieved using the EM method.
6. The best clustering results for two, three, or four authors were between 98% and 100%. Contrasting those, the best clustering result for five authors was rather low (66.51%).

Tables 3–6 present the clustering results for specific combinations of two, three, four, or five authors in greater detail, from the viewpoint of various numbers of words (100, 200, . . ., 1,000) contained in each word list

TABLE 3 Clustering Results for Two Authors (# 1, 2) Using Three Word Lists and Both EM and K-Means

	EM			K-means		
# of features (words)	FW	FFW	HVW	FW	FFW	HVW
100	54.6%	56.19%	92.36%	68.25%	65.82%	**96.71%**
200	86.82%	56.99%	59.45%	**98.08%**	65.89%	68.18%
300	51.27%	63.69%	58.33%	65.82%	65.86%	**96.34%**
400	60.72%	63.40%	62.64%	65.71%	**65.86%**	65.75%
500	**99.71%**	62.02%	61.91%	99.64%	65.86%	65.71%
600	**99.78%**	62.20%	62.27%	51.01%	65.86%	65.75%
700	**99.71%**	63.72%	63.25%	51.01%	65.86%	65.71%
800	**99.78%**	98.99%	63.36%	51.05%	65.71%	65.75%
900	**99.78%**	98.88%	63.76%	99.2%	65.71%	65.71%
1000	**99.75%**	98.88%	63.4%	99.2%	65.71%	65.68%

TABLE 4 Clustering Results for Three Authors (# 1, 2, 3) Using Three Word Lists and Both EM and K-Means

	EM			K-means		
# of features (words)	FW	FFW	HVW	FW	FFW	HVW
100	56.56%	56.76%	**85.6%**	70.44%	65.51%	67.87%
200	83.99%	71.52%	64.58%	63.06%	**86.47%**	52.07%
300	92.33%	62.04%	69.39%	**96.50%**	56.12%	59.04%
400	64.43%	61.40%	69.36%	**97.23%**	40.67%	94.75%
500	69.07%	61.46%	68.22%	**97.70%**	40.73%	49.85%
600	71.87%	80.90%	68.92%	**98.45%**	40.79%	74.26%
700	71.34%	78.66%	69.27%	98.51%	59.45%	59.13%
800	**98.72%**	80.52%	69.21%	98.54%	40.73%	96.38%
900	**98.75%**	78.16%	70.5%	98.57%	58.89%	96.30%
1000	**98.66%**	77.35%	70.47%	98.63%	58.8%	72.19%

TABLE 5 Clustering Results for Four Authors (# 1, 2, 3, 4) Using Three Word Lists and Both EM and K-Means

	EM			K-means		
# of features (words)	FW	FFW	HVW	FW	FFW	HVW
100	63.47%	72.54%	70.83%	51.28%	**75.81%**	56.83%
200	**91.29%**	86.00%	89.37%	71.06%	76.19%	67.83%
300	**91.95%**	75.51%	86.07%	73.05%	49.73%	75.58%
400	70.45%	75.25%	53.9%	73.82%	49.57%	**76.05%**
500	**87.82%**	76.12%	53.88%	73.89%	49.52%	73.03%
600	**86.91%**	77.99%	64.08%	82.42%	81.18%	71.69%
700	**87.15%**	70.59%	73.68%	78.62%	81.83%	76.21%
800	**97.94%**	76.21%	63.1%	79.00%	81.46%	76.24%
900	71.44%	80.05%	62.75%	75.09%	**81.71%**	76.33%
1000	**98.08%**	81.57%	63.31%	63.26%	81.48%	76.35%

TABLE 6 Clustering Results for Five Authors (# 1, 2, 3, 4, 5) Using Three Word Lists and Both EM and K-Means

	EM			K-means		
# of features (words)	FW	FFW	HVW	FW	FFW	HVW
100	50.78%	50.16%	**54.01%**	43.26%	52.02%	33.92%
200	44.78%	46.85%	**55.75%**	40.75%	55.19%	49.68%
300	**57.31%**	51.72%	52.53%	42.75%	50.06%	49.84%
400	50.88%	**66.51%**	53.94%	43.94%	37.87%	42.44%
500	54.84%	**64.88%**	55.93%	43.34%	37.9%	58.76%
600	**58.23%**	49.73%	51.27%	44.2%	34.92%	56.08%
700	57.59%	49.63%	59.65%	44.22%	34.92%	**65.52%**
800	**64.94%**	48.51%	49.07%	44.22%	54.04%	56.11%
900	**65.41%**	54.10%	47.87%	44.27%	48.15%	50.57%
1000	**64.25%**	52.67%	46.88%	44.28%	58.5%	50.3%

for each experiment. In three tables (out of four) FW was the best ML method. In all four tables, EM was the best ML method. At least 400 words (in some cases 900 or 1,000 words) were needed in order to achieve the best results. As mentioned before, results colored with bold red represent the best among all results in the same row (experiment) for all combinations of word lists and ML methods. Results colored with red (not in bold) represent the best among the results for the ML method, which achieves inferior results.

The best clustering result (99.78%) for authors # 1–2 (see Table 3) was achieved with the first 600 words of the FW word-list using the EM method. Therefore, the best improvement in this Table is 33.45%.

The best clustering result (98.75%) for authors # 1–3 (see Table 4) was achieved with the first 900 words of the FW word list using the EM method. Therefore, the best improvement in this Table is 45.34%.

The best clustering result (98.08%) for authors # 1–4 (see Table 5) was achieved with the first 1000 words of the FW word list using the EM method. Therefore, the best improvement in this Table is 55.19%!

The best clustering result (66.51%) for all five authors (Table 6) was achieved with the first 400 words of the FFW word list, using the EM method. Therefore, the best improvement in this Table is 36.37%. The drastic decline in the clustering results from four authors to five authors might be because author 5's documents are rather similar to many documents written by the other authors.

ANALYSIS OF THE PRECISION, RECALL, AND F-MEASURE VALUES

Analysis of the precision, recall, and F-measure values of the clustering experiments could lead to interesting findings as well as to insights that will enable definition of new features, which might improve the clustering results.

We calculated three fundamental measures: precision, recall, and F-measure.

$$\text{Precision}\,[\text{author}\,\#\,i] = a[i] : (a[i] + b[i]) \times 100\%, \text{ and}$$
$$\text{Recall}\,[\text{author}\,\#\,i] = a[i] : (a[i] + c[i]) \times 100\%,$$

where a[i] = number of documents composed by author # i and clustered to him, b[i] = number of documents not composed by author # i but clustered to him (termed "noise"), and c[i] = number of documents composed by author # i but not clustered to him (termed "silence").

Because of space limitations, we have taken into account the results presented in Table 2 for all 26 experiments using only the EM, which was found to be the best ML method for our tasks. Table 7 presents the averages of the precision and recall values for each author, for the 26 experiments, using the

TABLE 7 Average Precision, Recall, and F-Measure Values for the Five Authors, Using Three Word Lists and Both EM and K-Means.

		Authors				
Measure	Words method	1	2	3	4	5
Recall	FW	74.39%	94.00%	95.55%	78.34%	86.95%
	FFW	70.49%	97.84%	79.97%	85.67%	83.06%
	VW	79.38%	96.89%	95.39%	74.72%	78.50%
Precision	FW	86.99%	80.91%	92.89%	72.49%	95.07%
	FFW	89.62%	83.73%	60.51%	83.21%	93.01%
	VW	84.01%	85.16%	90.01%	75.20%	85.59%
F-Measure	FW	80.20%	86.97%	94.20%	75.30%	90.83%
	FFW	78.91%	90.24%	68.89%	84.42%	87.75%
	VW	81.63%	90.65%	92.62%	74.96%	81.89%

three word lists and the EM ML method. In addition, we calculated the F-measure (the weighted harmonic mean of precision and recall) values. Results colored with red represent values above 90%.

Various conclusions can be made from Table 7:

1. Authors #3 and #5 have the highest Precision values. The very high precision value means that the EM algorithm barely clustered irrelevant documents (i.e., documents that were not written by the discussed authors) to authors #3 and #5. In other words, clusters #3 and #5 hardly contained any noise (i.e., documents that were not written by authors #3 and #5, respectively).
2. Authors #2 and #3 have the highest recall values. A very high recall value means that the EM algorithm clustered almost all the documents written by authors #2 and #3, respectively.
3. Authors #2, #3 and #5 have rather high F-measure values. Often, there is an inverse relationship between the precision value and the recall value. The high F-measure value suggests that both recall and precision values are rather high.
4. Authors #1 and #4 have relatively low values of precision, recall and F-measure. These low precision values indicate that authors #1 and #4 tend to "include" a relatively high rate of noise (i.e., documents that were not written by them). Low recall values indicate that rather high numbers of documents composed by authors #1 and #4 were clustered to other authors. A possible explanation might be that their documents are rather similar to some of the other documents.
5. Author #3 has the highest values of precision, recall and F-measure. Possible explanations might be that his documents are rather unique when compared to other documents, and perhaps this is evident by the fact that author #3 has the smallest number of documents and words when compared to other authors.

SUMMARY AND FUTURE WORK

In this research, we investigate the authorship attribution task as a clustering problem. In addition, we compare two ML methods and examine the use of three word lists as features for clustering of documents to five authors from the 20th century.

In the best clustering tasks two, three, or four authors achieved results above 98% and the improvement rates were above 40% when compared to the "majority" (the baseline) results. The EM method has been found to be superior to K-means for the discussed tasks. FW has been found as the best word list, far better than FFW. FW, in contrast to FFW, includes function words, which are usually regarded as words that have little lexical meaning. This might imply that normalized frequencies of function words can serve as good indicators for authorship attribution using unsupervised ML methods, in contrast to Forman's claim that function words are useless. This finding supports previous findings about the usefulness of function words for other tasks such as authorship attribution using supervised ML methods (Koppel et al. 2009), and genre and sentiment classification (Pang et al. 2002).

Future directions for research are (1) defining and applying additional types of features such as: subgroups of function words, special content word lists, *n*-grams, syntactic *n*-grams (*n*-grams that are constructed using paths in syntactic trees, see Sidorov et al. 2014), morphological features (e.g., nouns, verbs, adjectives, and adverbs), syntactic features (frequencies and distribution of parts of speech tags, such as adjective, adverb, conjunction, noun, numeral, preposition, pronoun, particle, verb, and punctuation), multiword expressions and references unique to each class; (2) applying various kinds of models into other domains (especially those that are important for historians or other humanities researchers), applications and languages; and (3) comparison of the unsupervised ML methods for the given tasks with supervised ML methods.

REFERENCES

Aggarwal, C. C. and C. X. Zhai. "A Survey of Text Clustering Algorithms." In *Mining Text Data*, 77–128. Springer, US, 2012.

Baker, L. and A. McCallum. "Distributional Clustering of Words for Text Classification." In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 96–103. New York, NY: ACM, 1998.

Bekkerman, R., R. El-Yaniv, Y. Winter, and N. Tishby. "On Feature Distributional Clustering for Text Categorization." In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 146–153. New York, NY: ACM, 2001.

Chan, S., R. Pon, and A. Cardenas. "Visualization and Clustering of Author Social Networks." Paper presented at Distributed Multimedia Systems Conference, 174–180. Grand Canyon, Arizona, August 30–September 1, 2006.

Cutting, D. R., J. O. Pedersen, D. R. Karger, and J. W. Tukey. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections." In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 318–329. New York, NY: ACM, 1992.

Dempster, A. P., N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B* 39, no. 1 (1977): 1–38.

Dubes, R. C. and A. K. Jain. "Algorithms for Clustering Data." Englewood Cliffs, NJ: Prentice Hall, 1998.

Forman, G. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification." *Journal of Machine Learning Research* 3 (2003): 1289–1305.

Goldstein, J., V. Mittal, J. Carbonell, and M. Kantrowitz. "Multi-Document Summarization by Sentence Extraction." In *Proceedings of the NAACL-ANLP 2000 Workshop on Automatic Summarization*, 40–47. Stroudsburg, PA: Association for Computational Linguistics, 2000.

HaCohen-Kerner, Y., H. Beck, E. Yehudai, M. Rosenstein, and D. Mughaz. "Cuisine: Classification Using Stylistic Feature Sets and/or Name-Based Feature Sets." *JASIST* 61, no. 8 (2010a): 1644–1657.

HaCohen-Kerner, Y., H. Beck, E. Yehudai, and D. Mughaz. "Stylistic Feature Sets as Classifiers of Documents According to their Historical Period and Ethnic Origin." *Applied Artificial Intelligence* 24, no. 9 (2010b): 847–862.

HaCohen-Kerner, Y. and O. Margaliot. "Various Document Clustering Tasks Using Word Lists." In *Proceedings of the 9th Asia Information Retrieval Societies Conference*, edited by R. E. Banchs et al., LNCS 8281, 156–169. Berlin, Heidelberg: Springer, 2013.

HaCohen-Kerner, Y., D. Mughaz, H. Beck, and E. Yehudai. "Words as Classifiers of Documents According to their Historical Period and the Ethnic Origin of their Authors." *Cybernetics and Systems* 39, no. 3 (2008): 213–228.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA Data Mining Software: An Update." *ACM SIGKDD Explorations Newsletter* 11, no. 1 (2009): 10–18.

He, Y. and S. C. Hui. "Mining a Web Citation Database for Author Co-Citation Analysis." *Information Processing & Management* 38, no. 4 (2002): 491–508.

Jain, K., M. N. Murty, and P. J. Flynn. "Data Clustering: A Review." *ACM Computing Surveys* 31, no. 3 (1991): 264–323.

Koppel, J., D. Mughaz, and N. Akiva. *New Methods for Attribution of Rabbinic Literature, Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics*. Bar-Ilan University Press, 2006.

Koppel, M., J. Schler, and S. Argamon. "Computational Methods in Authorship Attribution." *JASIST* 60, no. 1 (2009): 9–26.

Koppel, M., J. Schler, and D. Mughaz. "Text Categorization for Authorship Verification." Paper presented at the 8th Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, 2004.

Layton, R., P. A. Watters, and R. Dazeley. "Automated Unsupervised Authorship Analysis Using Evidence Accumulation Clustering." *Natural Language Engineering* 19, no. 1 (2013): 95–120.

Lewis, D. D. and M. Ringuette. "Comparison of Two Learning Algorithms for Text Categorization." Paper presented at Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR 94), Las Vegas, NV, 1994.

Li, Y. J. and S. M. Chung. "Document Clustering Based on Frequent Word Sequences." In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM)*, 293–294. New York, NY: ACM, 2005.

Li, Y. J., S. M. Chung, and J. Holt. "Text Document Clustering Based on Frequent Word Meaning Sequences." *Data & Knowledge Engineering* 64, no. 1 (2008): 381–404.

Li, X., O. Zaiane, and Z. Li. "A Comparative Study on Text Clustering Methods." In *Proceedings of Advanced Data Mining and Applications* (ADMA-2006), LNAI 4093: 644–651, Berlin, Heidelberg: Springer, 2006.

Lloyd, S. P. "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory* 28, no. 2 (1982): 129–137. doi:10.1109/TIT.1982.1056489. Retrieved 2009-04-15.

MacQueen, J. B. "Some Methods for Classification and Analysis of Multivariate Observations." In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1: 281–297. Berkeley, University of California Press, 1967.

Madigan, D., A. Genkin, D. Lewis, S. Argamon, D. Fradkin, and L. Ye. "Author Identification on the Large Scale." In *Proceedings of the Meeting of the Classification Society of North America CSNA-05*, (St. Louis, MI, June 2005).

Meretakis, D. and B. Wuthrich. "Extending Naive Bayes Classifiers Using Long Itemsets." In *Proceedings of the 5th ACM-SIGKDD International Conference Knowledge Discovery, Data Mining (KDD'99)*, 165–174. San Diego, CA, USA: Springer, 1999.

Miao, Y., V. Keselj, and E. Milios. "Document Clustering using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering." In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 357–358. New York, NY: ACM, 2005.

Mingzhe, J. and J. Minghu. "Text Clustering on Authorship Attribution Based on the Features of Punctuations Usage, Signal Processing (ICSP)." In *Proceedings of the 11th International IEEE Conference on Digital Object Identifiers* 3: 217–2178. IEEE, 2012.

Mosteller, F. and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley, 1964.

Nigam, K., A. McCallum, S. Thrun, and T. Mitchell. "Learning to Classify Text from Labeled and Unlabeled Documents." In *Proceedings of the AAAI Conference*. AAAI, 1998.

Pang, B., L. Lee, and S. Vaithyanathan. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 79–86. Stroudsburg, PA: ACL, 2002.

Schutze, H. and C. Silverstein. "Projections for Efficient Document Clustering." In *Proceedings of the 20th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, 74–81. New York, NY: ACM, 1997.

Sebastiani, F. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys* 34, no. 1 (2002): 1–47.

Sidorov, G., F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández. "Syntactic N-Grams as Machine Learning Features for Natural Language Processing." *Expert Systems with Applications* 41, no. 3 (2014): 853–860.

Stamatatos, E. "Authorship Attribution Based on Feature Set Subspacing Ensembles." *International Journal on Artificial Intelligence Tools* 15, no. 5 (2006): 823–838, World Scientific.

Stamatatos, E. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538–556.

Steinbach, M., L. Ertoz, and V. Kumar. "Challenges of Clustering High Dimensional Data." In *New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition*, edited by L. T. Wille. Berlin: Springer-Verlag, 2003.

Witten, I. H. and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann, 2005.

Yu, B. "Function Words for Chinese Authorship Attribution." In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 45–53. Association for Computational Linguistics, 2012.

Zha, H. Y. "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering." In *Proceedings of SIGIR 2002*, 113–120. New York, NY: ACM, 2002.