The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Eliminating Sycophants to Improve Authorship Attribution

Ivan Petrovic,[1] Smiljana Petrovic,[2] Ileana Palesi,[2] Anthony Calise[2]

[1]Bronx Community College of CUNY, Bronx, NY 10453, USA
[2]Iona College, New Rochelle, NY 10801, USA
spetrovic@iona.edu

Abstract

Classification problem in authorship attribution consists of choosing the correct author of a document from an exhaustive list of candidates presented by the samples of their writing. A typical approach is to assign a vector representing measurements of a stylometric feature to each sample document and apply a supervised machine learning method to build a classifier. Different classifiers vary in the accuracy and attributions of the disputed documents. In our previous research, we have shown that a large number of classifiers can be combined into an effective jury via weighted voting. Such a jury is almost always more accurate than individual classifiers.

In this paper, we investigate whether it is possible to improve a jury's accuracy by eliminating some of its members. We test and compare two methods of reduction. Dynamic reduction selects a subset of original jury members by eliminating sycophants. Static reduction tests the behavior of preselected juries. Our testbed is a collection of 18th-century political writings, a fertile research ground rich with disputed works.

Authorship Attribution

Authorship attribution is the task of identifying the author of an anonymous text or a text whose authorship is in doubt (Love 2002). In this research, we consider a classification problem within authorship attribution: given an exhaustive list of possible authors and the samples of their work, how do we identify the author of a disputed text? Many modern authorship attribution methods have roots in the seminal work conducted by Mosteller and Wallace on the Federalist Papers (Mosteller and Wallace 1964). These methods rely on the premise that stylistic features, used unconsciously and consistently, can be measured to identify the author.

We combine a stylistic feature with a supervised machine learning method to create a *base-classifier*. The fifty most frequent values of a feature are identified and each sample document is assigned a normalized 50-dimensional vector of frequencies. Supervised learning is conducted, resulting in a classifier that can be applied to the attribution of unknown documents (Berton et al. 2016; Petrovic et al. 2015).

Stylistic Features

In our work, we consider seventeen different stylistic features outlined in Table 1. Among the most common "off-the-shelf" lexical features are function words, n-grams of characters and words, and Part of Speech (Stamatatos 2009).

Table 1. Features used in our analysis and their descriptions

Style Marker	Abbreviation	Description
MW Function Words	MWFW	Function words as defined by Mosteller-Wallace in their Federalist Papers study (Mosteller and Wallace 1964)
Word n-grams	WG2	The sequence of n items from a given sequence of words (in our case, n = 2)
Character n-grams	CG2, CG3	The sequence of n characters from a given sequence of characters (in our case, n is 2 or 3)
Part of Speech	POS	Uses the Maxent Tagger developed by the Stanford NLP Group (Toutanova et al. 2003)
POS n-grams	POSG2, POSG3	The sequence of n parts-of-speech tags (n is 2 or 3)
First Word in Sentence	FWIS	The first word in each sentence
Prepositions	PREP	The most common prepositions
Vowel initial Words	VIW	Words beginning with vowels
Suffices	SUF	The last three letters of every word
Coarse POS Tagger	CPOST	A simplification of the normal part-of-speech tagger, neutralizing

		minor variations such as plural inflection (singular/plural words are grouped)
Lexical Frequencies	LFREQ	Log-scaled frequencies of words from the general purpose HAL corpus as recorded in the English Lexicon Project (ELP) database (Balota et al. 2007)
Naming Reaction Times	NRT	Naming times from the ELP database; Each word is converted to the time it takes to name that word in the database (Balota et al. 2007)
Sorted Character n-grams	SCG2, SCG3	Alphabetically sorted characters in each n-gram (in our case, n is 2 or 3)
Word Stems	WS	Stems of the words obtained from Porter's stemming algorithm (Porter 1980)

Learning Methods

In our work, each feature is paired with one of the three following machine learning methods.

The *Centroid Nearest-Neighbor* (NNCos) approach represents authors by their centroid vectors (average of vectors assigned to their sample documents). An unknown document is associated with the author whose centroid is the nearest according to the cosine distance, i.e. the normalized scalar product.

The *Support Vector Machines (SVM)* method is a linear separation algorithm that seeks a hyperplane in the *n*-dimensional input space which best separates points corresponding to different candidate authors. We use here the implementation of John Platt's Sequential Minimal Optimization Algorithm (SMO) (Platt 1999).

The *Multilayer Perceptron* (MLP) is another linear separation algorithm that implements a Backpropagation Neural Network with the sigmoid activation function and the number of hidden nodes set to the average of the number of attributes and the number of candidate authors.

Leave-one-out for Assessing the Accuracy

To evaluate the selected base-classifier (feature-learning method pair), we adopted "leave-one-out" testing: *n-1* of the available *n* documents are used for training, and testing is carried out on the single remaining document. This procedure is repeated *n* times, in such a way that every document is used for testing exactly once. As a result, for each document, each base-classifier selects an author based on its learning from the remaining (*n-1*) documents. We record the accuracy (percentage of correctly classified documents) of each base-classifier.

Choosing and Combining Classifiers

Accuracy-weighted Method

Given k candidate authors, we assume that observed accuracy p of a base-classifier m signifies that when m votes for candidate A, the probability that A is the correct author is p, and the probability for each other candidate is an equal share of the complement. We assign the base-classifier's supports for candidates accordingly (eq. 1).

$$support_m(A) = \begin{cases} p & if\ m\ selects\ A \\ \frac{1-p}{k-1} & otherwise \end{cases} \quad (1)$$

This approach allows more accurate methods to have a greater contribution in voting for their choice of author. We define overall support for a candidate A as the product of supports for A by individual classifiers (eq.2). If base-classifiers were independent, the supports for candidates would be proportional to the probabilities of each of them being the correct author.

$$support(A) = \prod support_m(A) \quad (2)$$

The accuracy-weighted method selects the author with the highest overall support. To prevent any single method from taking over the voting (0 appearing in the products eliminates all other choices), we took the position that observed accuracy of 1 corresponds to probability $p = 0.999$.

Condorcet's Jury Theorem and Independence of Voters

In 1785, the Marquis de Condorcet established in his Jury Theorem that, for n independent and equally competent voters (with fixed accuracy $p > 0.5$), the accuracy of the majority vote is an increasing function of n and approaches 1 (certainty) as n approaches infinity. The same holds for accuracy-weighted voting of independent voters (the jury's accuracy increases with n, and tends to 1 when n tends to infinity), even for voters of unequal accuracies (accuracy $\geq c > 0.5$). In particular, an independent jury is at least as accurate as the most accurate member (Boland 1989; Grofman, Owen, and Feld 1983). If the voters are not required to be independent, it is possible to construct both examples: a) where the accuracy of a dependent jury is worse than the accuracy of any single member and b) where the accuracy of a dependent jury is significantly better than the accuracy of an independent jury with the same voter accuracies.

Example. Let A, B and C represent the events that the first, second and third voters vote correctly. Let each voter have 60% accuracy, i.e. $P(A) = P(B) = P(C) = 0.6$. Since all voters have the same accuracy, accuracy-weighted voting is the same as majority voting.

Case 1. If voters are independent, the probability of an intersection is a product of probabilities, so:
$P(ABC) = (0.6)^3 = 0.216, \quad P(\bar{A}\bar{B}\bar{C}) = (0.4)^3 = 0.064$
$P(AB\bar{C}) = P(A\bar{B}C) = P(\bar{A}BC) = (0.6)^2(0.4) = 0.144$
$P(A\bar{B}\bar{C}) = P(\bar{A}B\bar{C}) = P(\bar{A}\bar{B}C) = (0.4)^2(0.6) = 0.096$
The majority chooses correctly in the first four cases, producing a jury with $0.216 + 3(0.144) = 0.643 = 64.3\%$ accuracy.

Case 2. If the voters are not independent, a jury of three voters with 60% accuracies can have an accuracy as low as 40% (the worst case).
$P(ABC) = 0.4, \quad P(AB\bar{C}) = P(A\bar{B}C) = P(\bar{A}BC) = 0$
$P(A\bar{B}\bar{C}) = P(\bar{A}B\bar{C}) = P(\bar{A}\bar{B}C) = 0.2, \quad P(\bar{A}\bar{B}\bar{C}) = 0$
The majority chooses correctly in the first four cases, so the jury's accuracy is 40%.

Case 3. If the voters are not independent, a jury of three voters with 60% accuracies can have an accuracy as high as 90% (the best case).
$P(ABC) = 0, \quad P(AB\bar{C}) = P(A\bar{B}C) = P(\bar{A}BC) = 0.3$
$P(A\bar{B}\bar{C}) = P(\bar{A}B\bar{C}) = P(\bar{A}\bar{B}C) = 0, \quad P(\bar{A}\bar{B}\bar{C}) = 0.1$
The majority chooses correctly in the first four cases producing a jury with an accuracy of 90%.

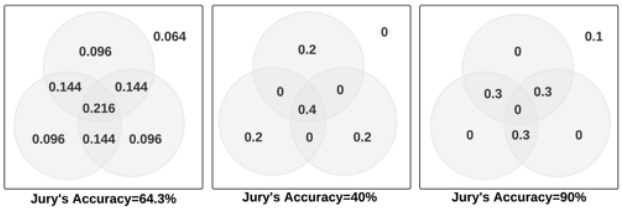

Figure 1: In each of the three cases, all voters have the same accuracy (60%), but the jury's accuracies differ greatly.

The example above indicates that low accuracies arise when correct choices win by unnecessarily large majorities (or in this case, the unanimous vote), while incorrect choices win with a minimal majority. This way, there are wasted votes for the correct choice, while every vote for an incorrect choice influences the decision. Conversely, in highly accurate juries, the pattern is reversed.

The Criterion for Eliminating Base-classifiers

The base-classifier voters we use in our experiments are not independent. A natural task is an investigation whether reducing the number of base-classifiers can improve accuracy. The task is more complicated than our example for two reasons. A large number of voters (up to 51) makes the brute force evaluation of all combinations impossible (2^{51}). The situation is made more complex by the presence of multiple incorrect alternatives; typically, we are dealing with more than ten possible authors (up to 42), with only one being the correct author. Thus, a pairwise diversity measure and a heuristic approach to reduction is indicated.

We define a *sycophant* as a voter for which there exists a more accurate voter such that they agree on a wrong choice more frequently than the specified threshold. Our strategy for reducing a jury set is to eliminate sycophants While variety of pairwise diversity measures (e.g. Yule's Q) are explored in literature (Butler et al. 2018; Kaniovski and Zaigraev 2011; Kuncheva and Whitaker 2003), we chose the frequency of "agreement-on-wrong" as a criterion for reduction for two reasons:

First, it is highly unlikely that two independent voters would agree on the wrong choice. For example, for two independent voters with an accuracy of 80% and 11 candidate values (one of which is correct), it is expected that they agree on the correct choice with a probability of 0.8^2=0.64=64% and agree on each of the incorrect values with a probability of 0.04%, hence the probability of agreeing-on-wrong is 0.4% (see Figure 2). Since independent voters are expected to have a fairly high frequency of agreeing on the right choice and a very low frequency of agreeing-on-wrong, using agreement-on-wrong as a heuristic cut-off criterion is a better choice than using total agreement.

Second, as we saw from the example in Figure 1, the independence of voters is not necessarily the goal when reducing a jury. An author receiving two votes starts with a significant advantage over the other candidates. Situations when two voters agree on a correct choice more often than the independent jurors would agree, may be beneficial or detrimental to overall accuracy (see cases 2 and 3 in the previous example). On the other hand, when two voters agree on an incorrect choice, that greatly increases the chances of their choice being selected, and this is never desirable.

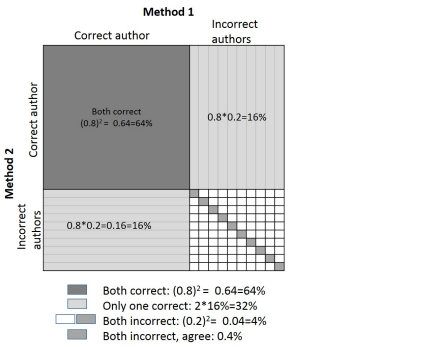

Figure 2: Example with two independent methods that are correct 80% of the time and select from eleven candidate authors

Dynamic Reduction of Jury Algorithm

In order to eliminate sycophants, we used a greedy (in classifier accuracy) algorithm that considers unassigned base-classifiers and eliminates from further consideration all that agree-on-the-wrong with the newest juror more frequently than a given benchmark. The most accurate of the remaining base-classifiers is chosen for the jury and the process is repeated until the jury-candidate pool is empty.

Pool = set of all base-methods
Selected = ∅
While Pool is not empty:
Move the most accurate method c from Pool to Selected
For all methods m remaining in Pool:
If m and c agree on wrong more often than the threshold:
Remove m from Pool

Figure 3: Pseudo code for our sycophant eliminating algorithm

Observed Correlation and Static Reduction

While running our dynamic reduction experiments, we noticed that some base-methods are consistently in high agreement over a wide range of experiments. For all three learning methods (NNCos, MLP, and SVM), there was a group of similarly-voting base-methods (MWFW, NRT, WS, and LFREQ) and several pairs: CG2 and SCG2, CG3 and SCG3, POSG2 and POSG3, and POS and CPOST. In order to reduce the size of juries formed from 17 features combined with the same learning method, we experimented with static selections, ensuring that only one feature from each of these groups was selected.

In the juries comprised of 51 base-classifiers (17 features combined with 3 learning methods), there were additional strong similarities between base-classifiers using the same features and different learning methods. SVM and MLP with the same feature were almost always highly correlated. Correlation between NNCos and these two was somewhat weaker (NNCos-SVM slightly stronger than NNCos-MLP), but still often present. Statically chosen sets were based on these observations with bigger set also including all observed exceptions to general rules.

Experimental Design and Results

Experimental Design

For different experiments, we used 12 groupings of authors selected among the pool of the following 42 authors: J.Adams, AmMerc, Barlow, Benezet, Brackenridge, Burgh, Burke, Cartwright, Cassandra, Chatham, Dickinson, Francis, Franklin, Freneau, G.Morris, Grenville, Hamilton, Hopkins, Hopkinson, Jay, Jefferson, Lafayette, Macauley, Madison, Matlack, Monroe, Moore, Paine, Peale, Price, Priestley, Rittenhouse, Rush, Sackville, S.Adams, Shelburne, Stanhope, Temple, Tooke, Wilkes, Witherspoon, Young. Some groups are based on the authors' origin (American or European), some on the period in which the authors were writing (1790s-1800s American or 1770s-1780s American), and others by political beliefs (Whigs-American). Writings were obtained from the Institute of Thomas Paine Studies, Iona College.

Instead of using the original papers, we represented each author with five documents, created by combining all the available work of a given author and separating it into five documents of approximately equal size. We found that this approach was more robust, particularly in cases when authors were originally represented by only a few writings (data not provided).

Results are obtained using the JGAAP (Java Graphical Authorship Attribution Program) open source software (Juola 2008), implemented WEKA libraries (Hall et al. 2009), and programs written by members of our research team.

As a baseline, we ran weighted sums of 17 base-classifiers using each of the three machine-learning methods (NNcos, SVM and MLP) and the weighted sum of all 51 base-classifiers. We then ran the dynamic reduction algorithm and let the reduced jury vote. We use 10% agreement-on-wrong as a threshold for eliminating sycophants in the experiments with 17 base-classifiers and both 4% and 10% in the 51 base-classifiers experiments.

We report here results of two statically chosen sets of eight features in combination with each of the three learning methods:
Static-1={Scg2, Viw, Lfreq, Cg3, Prep, Posg2, Pos, Suf}
Static-2={Cg2, Viw, Nrt, Cg3, Prep, Posg2, Cpost, Suf}

We also included two statically chosen sets out of the 51 base-classifiers:
Static-11 = {SVM+Cg2, SVM+Cg3, MLP+Posg2, MLP+Pos, SVM+Lfreq, MLP+Mwfw, NNCos+Lfreq, SVM+Prop, MLP+Suff, SVM+Viw, NNCos+Viw}

Static-15 = {SVM+Cg2, MLP+Cg2, SVM+Cg3, MLP+Scg3, NNCos+Scg3, MLP+Posg2, MLP+Pos, SVM+Lfreq, MLP+Lfreq, NNCos+Lfreq, SVM+Prop, SVM+Suff, MLP+Viw, NNCos+Viw, SVM+Wg2}

Experimental Results and Analysis

Several observations can be made from the results in Table 2. SVM and MLP have a very similar performance while NNCos is a uniformly weaker method. Accuracy typically decreases with the number of authors. Note, however, that random selection from 42 authors would result in the expected accuracy of 1/42=2.38%, while we achieve over 80% accuracy on datasets of this size. Weighted voting usually outperforms all individual base-methods, and there is no individual method that performs consistently well in all cases (data not shown).

All four of our static selections (two each for our 17 and 51 base-classifiers) behaved very similarly to original juries in all experiments. The only benefit of reduction is an improved speed of classification. While speed can be essential in some applications, it is of secondary importance in the automated authorship attribution of historical documents. An unexpected attribution may lead to months of historical research, making accuracy much more important than execution speed.

Table 2: Accuracy of original jury, dynamic reduction with 0.1 threshold and two static reductions in experiments where the base-classifiers are obtained by combining 17 features with each of three machine-learning methods independently

	Original jury	Dynamic reduction, 0.1 threshold	Number of features	Static-1	Static-2
NNCos					
A (7)	68.6%	74.3%	5	74.3%	74.3%
D (8)	77.5%	80.0%	6	75.0%	80.0%
B (10)	62.5%	75.0%	4	68.0%	64.0%
G (10)	74.0%	74.0%	7	72.0%	74.0%
H (11)	76.4%	87.3%	8	80.0%	78.2%
C (16)	75.0%	77.5%	8	70.0%	73.8%
Whigs(16)	67.5%	71.3%	7	66.3%	70.0%
90-00 (16)	66.3%	70.0%	6	63.8%	68.8%
Euro (17)	83.5%	84.7%	10	84.7%	82.4%
70-80 (22)	71.8%	72.7%	7	69.1%	70.0%
Amer. (25)	66.4%	67.2%	7	64.0%	64.0%
All (42)	68.6%	73.3%	7	70.0%	69.5%
Averages	*71.51%*	*75.61%*	*6.8*	*71.43%*	*72.42%*
SVM					
A (7)	88.6%	91.4%	7	85.7%	85.7%
D (8)	87.5%	92.5%	12	87.5%	92.5%
B (10)	90.0%	90.0%	13	90.0%	88.0%
G (10)	94.0%	94.0%	14	92.0%	92.0%
H (11)	92.7%	96.4%	14	90.9%	90.9%
C (16)	93.8%	93.8%	11	95.0%	91.3%
Whigs(16)	88.8%	91.3%	11	88.8%	86.3%
90-00 (16)	87.5%	90.0%	12	87.5%	88.8%
Euro (17)	89.4%	90.6.%	12	91.8%	90.6%
70-80 (22)	90.0%	87.3%	11	90.0%	87.3%
Amer. (25)	87.2%	88.0%	12	87.2%	87.2%
All (42)	83.8%	83.8%	10	84.3%	82.4%
Averages	*89.44%*	*90.77%*	*11.6*	*89.23%*	*88.58%*
MLP					
A (7)	88.6%	97.1%	9	94.3%	88.6%
D (8)	90.0%	92.5%	13	85.0%	87.5%
B (10)	90.0%	94.0%	9	88.0%	90.0%
G (10)	90.0%	90.0%	12	90.0%	94.0%
H (11)	90.9%	92.7%	13	90.9%	90.9%
C (16)	90.0%	91.3%	14	91.3%	91.3%
Whigs(16)	88.8%	91.3%	14	91.3%	90.0%
90-00 (16)	91.3%	92.5%	14	93.8%	95.0%
Euro (17)	89.4%	90.6%	14	91.8%	90.6%
70-80 (22)	90.0%	90.0%	12	87.3%	86.4%
Amer. (25)	88.8%	87.2%	13	84.8%	87.2%
All (42)	87.6%	88.1%	14	86.2%	83.8%
Averages	*89.62%*	*91.44%*	*12.6*	*89.56%*	*89.61%*

Table 3: Three learning methods, NNCos+SVM+MLP, combined with 17 features, resulting in 51 base-classifiers

NNCos+SVM+MLP							
	Original jury	Dynamic reduction, threshold 0.04	Number of features 0.04	Dynamic reduction, threshold 0.1	Number of features 0.1	Static-11	Static-15
A	85.7%	97.1%	6	94.3%	18	82.9%	85.7%
D	85.0%	97.5%	6	92.5%	22	85.0%	88.0%
B	86.0%	96.0%	7	96.0%	11	88.0%	88.0%
G	90.0%	92.0%	11	94.0%	26	88.0%	92.0%
H	90.9%	94.5%	12	94.6%	32	90.9%	94.6%
C	90.0%	95.0%	10	92.5%	27	90.0%	88.8%
Whigs	86.3%	90.0%	9	90.0%	23	83.8%	85.0%
90-00	86.3%	93.8%	9	90.0%	22	86.3%	86.3%
Euro	91.8%	90.6%	12	89.4%	29	87.1%	90.6%
70-80	88.2%	90.9%	10	87.3%	23	88.2%	88.2%
Amer.	86.4%	89.6%	10	88.8%	21	82.4%	85.6%
All	84.8%	88.6%	11	87.1%	23	83.3%	86.2%
Avgs:	*87.6%*	*93.0%*	*9*	*91.4%*	*23*	*86.7%*	*88.2%*

Dynamically reduced sets outperformed original juries in average accuracy over all learning methods (NNCos, SVM, MLP, and the combination of the three). Furthermore, the improvement of averages is the result of a consistently better performance on almost every experiment.

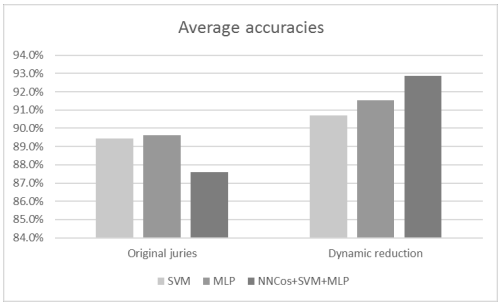

Figure 4: Comparison of average accuracies of original juries and dynamic reduction (threshold 0.1 for 17 base-classifiers pools and 0.04 for 51 base-classifiers pool)

The effect of sycophant elimination on the jury of 51 base-classifiers was especially encouraging. While adding (independent) classifiers should improve accuracy, our (dependent) jury did not behave that way. Combinations of 17 base-classifiers based on both SVM and MLP outperformed the combination of all 51 base-classifiers. The dynamic reduction of 51 base-classifiers significantly outperformed not only the original jury, but also SVM and MLP juries and their reductions.

Our experiments with different thresholds indicate that dynamic reduction is a robust method. Thresholds of 0.04 and 0.1 resulted in very different juries, averaging 9 and 23 base-classifiers selected. Both methods outperformed original jury of 51 classifiers as well as statically chosen juries of 11 and 15 classifiers. This seems to indicate that, while it could be difficult to find the optimal threshold for each problem, there is a fairly wide range of well-performing thresholds.

Conclusion

While statically chosen reduced sets of base-classifiers performed about the same as the original jury, the dynamic reduction by eliminating sycophants proved very promising. Our results indicate that dynamic reduction is a good way to add potentially dependent classifiers to the jury, reaping the benefits and avoiding performance drop: make the biggest jury available and then eliminate sycophants. This is a robust method which does not require the optimal threshold to be viable.

Future Work

We plan on further investigating dynamic reduction, by branching in several directions. In addition to accuracy, the confidence of choice can be measured. We wish to investigate differences between correct and incorrect choices in voting support and how dynamic reduction affects this difference. Also, some base-classifiers can be set to rank their choices instead of simply reporting their top choice. We wish to explore the ranked voting of original and reduced mixtures. Automatic searching for the optimal threshold is another worthwhile project. Finally, meta-learning can be used as an alternative to weighted voting (Petrovic et al. 2018). The effect of dynamic reduction on meta-learning is another subject which we are interested in exploring further.

Acknowledgments

This research was partially supported by the generous grant from the Robert David Lion Gardiner Foundation to Iona College's Institute for Thomas Paine Studies (ITPS).

References

Balota, David A et al. 2007. "The English Lexicon Project." Behavior research methods 39(3): 445–59.

Berton, Gary, Smiljana Petrovic, Lubomir Ivanov, and Robert Schiaffino. 2016. "Examining the Thomas Paine Corpus: Automated Computer Authorship Attribution Methodology Applied to Thomas Paine's Writings." In New Directions in Thomas Paine Studies, New York: Palgrave Macmillan US, 31–47.

Boland, Philip J. 1989. "Majority Systems and the Condorcet Jury Theorem." The Statistician 38(3): 181.

Butler, Harris K, Mark A Friend, Kenneth W Bauer, and Trevor J Bihl. 2018. "The Effectiveness of Using Diversity to Select Multiple Classifier Systems with Varying Classification Thresholds." Journal of Algorithms & Computational Technology 12(3): 187–99.

Grofman, Bernard, Guillermo Owen, and Scott L. Feld. 1983. "Thirteen Theorems in Search of the Truth." Theory and Decision 15(3): 261–78.

Hall, Mark et al. 2009. "The WEKA Data Mining Software." ACM SIGKDD Explorations Newsletter 11(1): 10.

Juola, Patrick. 2008. "Authorship Attribution." Foundations and Trends® in Information Retrieval 1(3): 233–334.

Kaniovski, Serguei, and Alexander Zaigraev. 2011. "Optimal Jury Design for Homogeneous Juries with Correlated Votes." Theory and Decision 71(4): 439–59.

Kuncheva, Ludmila I., and Christopher J. Whitaker. 2003. "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy." Machine Learning 51(2): 181–207.

Love, Harold. 2002. Attributing Authorship : An Introduction. Cambridge: Cambridge University Press.

Mosteller, Frederick, and David L Wallace. 1964. Addison-Wesley series in behavioral science. Quantitative methods. Inference and Disputed Authorship: The Federalist. Center for the Study of Language and Information.

Petrovic, Smiljana, Gary Berton, Sean Campbell, and Lubomir Ivanov. 2015. "Attribution of 18th Century Political Writings Using Machine Learning." Journal of Technologies in Society 11(3): 1–13.

Petrovic, Smiljana, Ivan Petrovic, Ileana Palesi, and Anthony Calise. 2018. "Weighted Voting and Meta-Learning for Combining Authorship Attribution Methods." In International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2018), Madrid, Spain.

Platt, John C. 1999. "Fast Training of Support Vector Machines Using Sequential Minimal Optimization." In Advances in Kernel Methods, , 185–208.

Porter, M.F. 1980. "An Algorithm for Suffix Stripping." Program 14(3): 130–37.

Stamatatos, Efstathios. 2009. "A Survey of Modern Authorship Attribution Methods." Journal of the American Society for Information Science and Technology 60(3): 538–56.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In NAACL '03, Morristown, NJ, USA, 173–80.