# Authorship Attribution using Rough Sets based Feature Selection Techniques

Ignatius Ikechukwu Ayogu
Department of Computer Science,
Joseph Ayo-Babalola University
Ikeji-Arakeji, Osun State,Nigeria.

Victor Akinbola Olutayo
Department of Computer Science,
Joseph Ayo-Babalola University
Ikeji-Arakeji, Osun State,Nigeria.

## ABSTRACT
This presents an investigation into the usefulness of rough set theory in the context of authorship attribution using writing style. The problem was setup as a standard supervised machine learning problem. The rough set based feature subset computation techniques reduced the dimensionality of the feature space from 346 conditional attributes to an average of 8 features. Experiments were performed experiment using five different subsets of the original attributes computed using rough sets techniques with the results showing that the rough set based techniques improved the performances of neural network (NN) and Support Vector Machines (SVM) models. The overall classification accuracy increased from 8.712 % for on the baseline data to 50.505 % for the NN and from 7.197 % to 28.662 % for the SVM model. The improvements in performance compared to the baseline model are evidenced across all other performance metrics used. However, the NN model performed generally better than the SVM model.

## General Terms
Authorship Attribution, Rough Sets Theory, Machine Learning.

## Keywords
Stylometry, Feature Selection, Neural Networks, Support Vector Machines, Supervised Machine Learning.

## 1. INTRODUCTION
The act of writing is a habit and accordingly, it gets mature as the writer/author increases his/her capacity to recognize and store certain characteristics that influence his learning and thereby develops/acquires specific skills of writing that not only tend to reflect him/her but distinguishes him from other writers. Habitual tendencies become irresistible over time and the urge to exhibit the promptings become very hard to control because they are autonomic and are not within the control of the individual [40; 12; 15] and as [40] puts it, "habits are activated in memory in an autonomous fashion without requiring executive control". Writing, skills mature with practise and the writer-specific traits becomes more difficult to obfuscate [42; 38]. Thus, the author or writer, like every other addict, becomes addicted to his style of writing which remain largely invariant across topics/discourses. An author's writing may vary only significantly in the richness of vocabulary as he writes across varying topics. Research has largely proven that authors' personal writing idiosyncrasies remain relatively invariant, regardless of the discourse of writing [33; 6]. This fact is the singular most important source of motivation for the subject of authorship attribution and attribution techniques.

Authorship attribution, sometimes referred to as authorship classification or authorship identification with slight modification of intent, has a major goal of ascribing authorship to an anonymous or disputed piece of writing. It uses statistics to study the linguistic and computational characteristics of written texts in order to correctly identify the author [37]. Apart from its traditional use in settling authorship disputes of literary texts, authorship attribution has been employed in a number of very important tasks: computer forensic investigation, criminology, plagiarism detection, authorship identification/verification, civil law procedures, detection of the author of malicious computer programs, fraud detection and in e-commerce.

The problem of authorship attribution is age-long [17] and has attracted a lot of research efforts over the last two centuries [18; 35]. Though an old problem, the dynamics of technology and the plethora of technological innovations and inventions have introduced more challenging dimensions to settings by which authorship attribution problems are situated, which has rendered the hitherto prolific traditional solution approaches unattractive. Nowadays, social media platforms have become an integral, indispensable, part of the society. These platforms, though seemingly indispensable and useful have been turned into a platform for profanity, denigration, mud-slinging and undue harassments that literally amounts into character assassination and even irredeemable economic losses. While earlier attribution techniques relied on longer texts for statistics, social media texts are often short, with ungrammatical sentences, high rates of abbreviations and spelling contractions. This is a challenge for authorship attribution methods.

Authorship attribution research in the last decade has focused mostly on developing methodologies that scale well on short weakly structured messages by exploring features that tend to talk more of the writer than the subject of the writing [32; 35; 37; 6]. This work explores the power of a feature extraction/dimensionality reduction technique as it applies to the authorship attribution problem. It shows that the Rough Set methodology for feature extraction/selection improves the performance of two algorithms: Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) in a multi-class authorship classification scenario respectively. A second major intent of this research was to identify a scalable set of features that improves the classification accuracies of algorithms across various genres and topics at the long run. The paper also shows that such set of features improved classification as the number of candidate authors increases as opposed to the current observation by a good number of researchers that performance deteriorates with increasing number of authorship class [35; 37; 6]. An extensive evaluation of the feature subsets computed in this work in a short-length data scenario is on-going. This paper in the meantime presents our initial attempt in the exploration of the rough set based techniques for feature selection in authorship attribution task with experiments involving SVM and NN,

using the implementations in the *rminer* [9], the *roughsets* and *discretization* packages in R using news texts written by selected Nigerian columnists.

## 2. PROBLEM STATEMENT

Given a piece of text document, d, whose author is in dispute or unknown, and a set of authors *A* consisting of $a_1, a_2, \ldots, a_n$, for which there is are sample documents $d_{ai}$, of writings by each of the suspected authors respectively, the task of authorship attribution is to analyse the given documents and correctly assign the authorship of the document *d* to one of the authors in *A* to which the characteristics of *d* is a match according to some threshold. Otherwise, no author should be assigned.

$$g(d,[A]) = \begin{cases} 1 \; iff \; d \; matches \; any \; one \; of \; da_i \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases}$$

where $g$ is the authorship attribution model.

The input to the model is a set of features extracted from the documents. The task of feature extraction is also an important problem since the technique must select only and all relevant features. Identifying the most important features is again another problem of this research.

The overall aim of this research is to identify a compact set of features from a combination of all the classes of authorship measures that exhibits appreciable 'universality' for the identification of the writer of any piece of text. The objectives are to identify most relevant features from amongst the 346 features we extracted from the texts and ascertain whether they improve the performance of the test models.

Authorship attribution techniques use statistical feature significance to create a model that attempts to assign one out of two or more authors to a piece of writing whose author is unknown. Theoretically, the best model should be that which combines all the possible features in the feature space. This however is not practicable because of the 'curse of dimensionality', the increased complexity of models which is often a leading cause of poor performance [27]. It is practically impossible to define a model that uses all features in the feature space, hence the need to create a 'universal' feature set, consisting of the best combination of features. Obtaining this 'universal' feature set, of the most relevant features from the seven classes/categories of stylometric measures we have selected is both a theoretical and practical possibility if intelligent feature engineering techniques are carefully employed. From a practical standpoint, this research is motivated to contribute to the development of authorship models that are not adversely domain-restricted.

## 3. AUTHORSHIP ATTRIBUTION

Early authorship identification methodologies experienced a number of significant challenges. [35] identified some of these challenges as it concerns their evaluation procedure to include: excessively long textual data which were often not stylistically homogeneous, very small candidate authors, lack of control for topic, subjective evaluation method and non-availability of suitable benchmark data.

Authorship attribution in the last two decades has witnessed a major shift from the more traditional methods and applications to literary works to a more technically sound methodologies and approaches that now earns the subject matter a potential for applications in forensic analysis, criminal investigation, fraud detection, terrorists profiling, copyright dispute or plagiarism detection, source code authorship analysis, and verification of suicide notes [32; 33; 34; 37; 6]. This change of focus is attributable to the Internet and Internet-enabled technologies that produces very large volumes of textual data in a very short time. Examples of these include blog posts, e-mail chat, social network forum posts, etc, which necessarily leads to the failure of traditional methods [25; 21; 23].

Authorship attribution tasks can be seen in various forms: authorship verification, author profiling, plagiarism detection and authorship identification [6]. In authorship attribution, the likelihood that a particular piece of text is written by a given author is determined, author profiling focuses on characterizing an author from a piece or pieces of texts written by him or her. Authorship attribution techniques utilize style-based measures to ascertain the authorship of a given text document. These style markers are quantified through a set of features that are extracted from text [43]. Stylometry is the most widely used method for authorial identification; a summary of stylometric features with computational tools and resources needed for their measurement can be found in [35].

Lexical and character-based style markers are the most widely developed and used among the stylometric features. The most significant reason for this as we find it in literature is that these features are language-independent and besides, they are easy to extract, not requiring most of the complex, not-too-easy-to-come-by tools that are often required for deeper level analysis and feature extraction and most importantly, they are relatively author-invariant [32; 33; 34; 35; 37; 6; 3; 7]. For instance, function words are known to be topic-independent and more related to the writer's style than the discourse of the document. These feature sets however are known to generally increase exponentially in size and thereby increasing model perplexity, leading to overfitting and poor generalization.

In order to ensure scalability and good performance, the selection of the most relevant features is crucial. The feature set size is a determinant of a number of important effects noted in previous authorship attribution research [3; 7]. The feature set size must be determined alongside a consideration for the available data set and the classification algorithms that would be trained on it. Certain algorithms are known to cope well with high-dimensional features set (e.g SVM) without overfitting whereas neural networks apart form overfitting problems also takes a longer time to train in the face of high dimensional features. Other algorithms like decision tree will overfit and generalize poorly when learned on a dataset with high dimensional feature space.

Word and character *n*-grams represent another variant of lexical and character features. Research however indicates that the decision for the size of *n* is purely a subjective issue [19] and are a function of the level of expertise available to the designer. It is noted that larger n values capture better lexical and contextual information but increases the dimensionality of the representation (often hundreds of thousands of features). Small n (2 or 3) is able to represent sub-word information, but represents contextual information poorly [3]. [14] proposed the extraction of n-grams of variable length as a solution to this problem.

Other stylometric features (syntactic and semantic) needs robust NLP tools which are vastly unavailable for minority, less resourced languages to successfully extract them. The work of [4] is acclaimed the first to use syntactic features based on a syntactically annotated English corpus from which they were able to extract re-write rule frequencies. [30; 31],

exploited syntactic information for authorship attribution. Their method avoided the more complex method of [4], by counting noun phrases, verb phrases, length of noun phrases, length of verb phrases and so on. These authors also discovered a class of features they referred to as analysis-level features involving an incremental fashion of analysis proceeding from simple to complex where outputs from previous steps are reused in the next higher steps.

Research has noted that deep syntactic and semantic measures are far more useful for short length texts. It is also observed from literature that deeper-level measures of syntax and semantics alone have yet to outperform the lexical and character-based measures. Thus the advantage of these methods is still being seen as too little to justify the high level of work involved in building them. This work presented in this paper did not consider the use of deep-syntactic and semantic measures; the reason is due to scalability issues resulting from language dependencies and the non-availability of the required tools for many languages. Moreover, lexical and character-based measures have been shown to perform well across language and topics.

[44] presents an implementation of authorship attribution using writeprints, a Karhunen-Loeve-tranforms-based technique that uses pattern disruption algorithm with individual author-level feature sets. The writeprint feature dimensions were reduced by employing information gain. His work was based in the work of [1], the originator of the writeprint idea for authorship identification. The results in a 45-author scenario indicates that writeprints performed significantly better than random-chance but poorer than SVM. The authors suggested that the result may be limited to the type of writings in the corpus for evaluation.

[8] implemented an authorship similarity detection technique that targets the challenges of e-mail authorship analysis as outlined in [10]. In their model, 150 stylistic measures (indicators) were used to measure the similarity among short-length, topic-free e-mail messages using frequent pattern matching and machine learning techniques. Their method attained an accuracy of 84% and 89% when a given author had 10 and 15 short email messages respectively. This performance according to the researchers were significantly better than those of PCA and k-means clustering.

[45] identified critical issues for authorship identification analysis to include the selection of features (descriptors) that characterise texts and authors, and analytical techniques and algorithms applied to the task. In their work, they employed a committee machine approach which built neural networks models constructed using a three-point filtering methodology. These authors used paragraph as the basic sampling unit. The results obtained from their experiments indicated a performance of 73% for the first author and 74.5% for the second author. The number of authors investigated in this work is rather too few. It however further lent credence to the suitability of the neural network model approaches to the authorship analysis problem.

[26] compared the performances of various feature selection, reduction, and classification techniques on the task of authorship attribution. Chi-square, correlation-based methods, Principal Components Analysis (PCA), and Latent Semantic Analysis are among the feature selection methods investigated using eight different classifiers including Hyper-Pipes, Naive Bayes, Lazy LWL and LAD tree classifiers. LAD Tree classifier was reported to have 12% better than others. A significant limitation of this work is that it never accounted for the effect of each of the feature reduction/selection algorithm on the performance of the individual classifiers.

[41] proposed two new rough set based heuristics for feature selection: the Average Support Heuristics and the Parameterized Average Support Heuristics. These methods were shown to have two main advantages over the significance-oriented and support-oriented rough set heuristics – they produce a set of rules with balanced support distribution over all the decision classes and secondly, they consider the predictive instances that are excluded by the significance-oriented and support-oriented methods. The authors were motivated by the need to consider the contributions of information provided by inconsistent instances that are ignored by the existing methods.

[36] demonstrates the use of rough set for feature selection with application to pattern recognition. The authors subjected the results of PCA to rough set treatments, reducing the PCA projected pattern of 60 elements to only 8 element, with an accuracy of 75%. This is another indication of the usefulness of rough set in dimensionality reduction.

# 4. FEATURE SELECTION

## 4.1 Feature Selection

Enumerating all the candidate subset of a complete feature set extracted from data is difficult, if not impossible, hence the need to select a subset of the original feature set that satisfy certain optimality criteria. Feature selection techniques are inspired by various principles as could be found in literature. [36], citing the works of [11; 13; 5; 20], (1994), outlined two main streams of feature selection methods: open-loop and closed-loop methods. The open-loop methods uses what is termed between-class separability criterion, ignoring predictor quality considerations. The closed-loop method uses predictor performance as a criterion for feature subset selection.

[41], having underscored the impossibility of exhaustive search for the best subset of features in real world applications, outlined two methods: random search method and heuristic search method. The random search method proceeds by generating a random subset which is subjected against a measure criterion to see whether it satisfies it. If not, the process is repeated until either a predefined time has elapsed or a predefined number of subsets have been tested. Heuristic search method uses a heuristic function to guide the direction of the search; it is chosen to maximize the overall value of the feature subset returned by the search. The random and heuristic search methods are not guaranteed to always provide the optimal results.

The feature selection process removes irrelevant and redundant features (according to some threshold conditions) from the original feature set. The target is both to make the classifier perform better as such feature selection methods must ensure minimal loss of information content and only remove truly superfluous features. A good feature subset is that which consist only of features with high feature-to-class correlation and with zero co-linearity [16]. Despite the notable weakness of the heuristic-based feature selection method, it is notably the most widely used in the research community. [41] pointed out some salient issues in heuristic-based feature selection arising from the imposition of a partial order in the search space. The issues that arise include deciding the start state, the methodology for the search, and the stopping criterion. Two classes of measures applied are the filter and wrapper measures [41; 36].

## 4.2 Rough Sets Feature Selection

Rough set [28], is a non-statistical approach applied to the analysis of data whose methodology is concerned with the classification and analysis of imprecise, uncertain or incomplete information and knowledge [29]. A very important concept behind rough set is the approximation of lower and upper spaces of a non-crisp set. According to [39], the rough set method is based on the premise that reducing the degree of precision in the data increases the visibility of the patterns in the data. Hence the rough set approach represents a framework for discovering facts from imperfect data [29; 39].

[24], clearly showed that rough set finds use in a number of phases of the knowledge discovery process: feature selection, feature extraction, dimensionality reduction, decision rule generation and extraction of patterns from data. Rough set uses some forms of heuristics to judge the relevance of features to be included in a subset to be returned. The core consists of all the most relevant features that cannot be dropped. i.e. it consists of all attributes contained by all reducts. The heuristic functions defined for rough set include significance-oriented, support-oriented, Average Support Heuristics and Parameterized Average Support Heuristics [41].

In applying rough set to feature selection, various strategies can be employed but as [36] notes, "the simplest approach is based on calculation of a core for discrete attribute dataset containing strongly relevant features, and reducts, containing a core plus additional weakly relevant features such that each reduct is satisfactory to determine concepts in the data set". This view is held by many researchers in rough set theory. A measure of what is strong or weak relevance is defined by [20] in terms of how strongly the occurrence of the target concept is tied to the feature. [41] and [36] are among the papers providing a good details of what constitutes relevance of features with respect to the rough set approach.

# 5. METHODOLOGY AND EXPERIMENTAL SETUP

The task of authorship attribution involves majority of the well-established data mining processes and procedures since, in itself, it can be seen as an application of data mining and machine learning. Our approach utilizes in the first place, the Cross Industry Standard Process for Data Mining (CRISP-DM) processes, followed by the use of machine learning techniques to achieve our desired goal. In the subsequent, a brief on the basic requirements of authorship attribution as a data mining-inspired process using the CRISP-DM process model is presented.

Since the task is defined for this work as a standard supervised learning (classification) problem, and following the instance based authorship attribution model described in [35], the data sets and its parameters must be defined in such a way that makes for features extraction in a similar way across the available texts for all the authors being considered. Next, a brief description of the data set that was created and used in this work is presented.

## 5.1 Data Set

Although textual data exist in such a large quantity on the web, there are no standard data sets for authorship attribution tasks. Besides, this work targets Nigerian writers. We therefore selected thirteen (13) regular columnists from a Nigerian National daily, The Nation. We harvested texts published in their columnists' column over a period spanning 2014 to early 2016 – a period that is characterised by a lot of politicking and growing reality of economic misadventure in the country – hence the subject of the authors' writing espoused politics and economic realities mostly; that is not to say these qualifies as the topic of most these writings. For this work, we collected a total of 20 articles per author, giving a total of 260 articles.

## 5.2 Feature Set

In order to keep to the objective of obtaining a universal author-specific, topic-independent set of features, this work used four classic stylometric feature categories: Lexical, Character, Syntactic and Structural. The use of syntactic and structural features were however limited to punctuations, average words/sentences for the sake of simplicity and language-dependent applications. The authors adopted some write-print signature features of [44] with modifications. The lists of all the features we have extracted are presented in Table 1.

## 5.3 Experimental Setup

Following the instance-based approach to authorship attribution, described in [35], each of the 20 articles by each author into a sample unit. From each of these, features were extracted, giving a total of 20 instances for each of the 13-author set. Experiments were performed in phases, following different rough set based feature subset computation and discretization techniques. A constant sample set which have been prepared and preprocessed to which we then apply the same model settings and parameters was maintained. A control experiment – the one in which we have not applied rough set based treatment for each of the selected classifiers and the test – the one in which we the rough set method of feature selection has been applied were performed. The target was to determine the effect of rough set based feature reduction on the set of features being investigated. The models were rated using the two strands of experiments using the most basic measures of performance of machine learning algorithms including Receiver Operating Characteristics (ROC) area, Area Under the Curve (AUC), Precision and F-measure.

**Table 1. Feature groups and their descriptions**

| Group | Category | Quantity | Description |
|---|---|---|---|
| Lexical | Word-level | 4 | Total words, average word length (1-36), frequency of large words (> 4 letters), unique words count |
| Lexical | Character-level | 3 | Total characters, % of uppercase letters, characters/word |
| Lexical | Word-length distribution | Corpus size dependent | Frequency of different word-lengths |
| Lexical | Special characters | 5 | Special characters {$, -, , =, +, &} |

| | | | |
|---|---|---|---|
| Lexical | Vocabulary richness | 1 | Hapax legomena |
| Syntactic | Function words | 200 | Frequency of function words. e.g. |
| Syntactic | Punctuation | 8 | Frequency of punctuation occurrences |

## 5.4 Implementation Details

### *5.4.1 Implementation of the NN Model*

The Neural Networks (NN) model was implemented using the *rminer* [9] package in R tool. The target attribute values were encoded with the common 1-of-$N_c$ transform, leading to $N_c$ binary classes. The *rminer* package adopts the *nnet* package implementation of the multilayer perceptron model. The multilayer perceptron architecture which includes hidden layers with logistic threshold function was used. The entire model according to [9] is given as:

$$y_i = f_i\left(w_{i,0} + \sum_{j=I+1}^{I+H} f_j\left(\sum_{n=1}^{I} x_n w_{m,n} + w_{m,0}\right) w_{i,n}\right) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

where $y_i$ is the output of the network for node $i$, $w_{ij}$, is the weight of the connection from node $j$ to $i$ and $f_j$ is the activation function for node j.

Since the problem setting is of $N_c > 2$ type, there are $N_c$ linear output neurons. The *rminer* implementation uses the softmax function to transform the output values into class probabilities:

$$P(i) = \frac{\exp(y_i)}{\sum_{i=1}^{Nc} \exp(y_i)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

where $p_i$ is the predicted probability and $y_i$ is the model's output for class $i$.

The multilayer perceptron option was chosen because it gives a better performance [9] than the simple perceptron. With this option, we are able to set the number of hidden layer for the target NN model effectively.

### *5.4.2 Implementation of the SVM model*

The *rminer* uses the sequential minimal optimization organization (SMO) learning algorithm to learn SVM. This work, adopts the Guassian kernel approach which is noted to present less parameters than other kernels. For the SVM, separate setups for the two hyperparameters γ, the kernel parameter and *C*, the penalty parameter were considered. The probabilistic output of the SVM is given by (Cortez, 2010).

$$f(x_i) = \sum_{j=1}^{m} y_j \propto_j K(x_j, x_i) + b \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

$$P(i) = \frac{1}{(1 + \exp(A_f(x_i) + B))} \dots\dots\dots\dots\dots\dots\dots\dots (4)$$

where $m$ is the number of support vectors, $y_i \in \{-1,1\}$ is the output in the binary case; but since our problem is a multiclass type, the one-versus-rest approach which gives $N_c(N_c - 1)/2$ binary classification is used, obtaining the output using pair-wise coupling as described in [9].

### *5.4.3 Feature Extraction and Data set Generation*

The feature extraction step was accomplished using the JStylo tool. To achieve our desired objective, all the documents for each author in our author mix were loaded and the parameters set as appropriate for the target output. The output of this stage is a file containing the raw counts/occurrences of each class/group of features as outlined in Table 1. We obtained a total of 260 data instances with 346 condition attributes. This is the dataset upon which all the experiments carried out in this work were based.

### *5.4.4 Discretization and Feature Subset Computation*

Rough set requires that instance attribute values be discretized. In order to compute the feature subset using the *Roughsets* package in the R programming tool, the dataset was discretized using the implementations for discretization within the R tool. This is so because the discretization is meant to work as part of the experimental pipeline. There was however a need had to idealise a walk around a seeming challenge: we are interested in obtaining a dataset with integer-valued features rather than an interval-specific output that obtains from the *RoughSets* package. To achieve our aim, the intermediate output given by the *RoughSets* package was in-turn discretized using the chi-square discretization algorithm. This gives us the desired integer-valued features which can be used in our experiments.

*"Optimal"* feature subsets were computed using four different roughest based feature subset computation/selection algorithms: *Quickreduct (qr), Dynamically adjusted approximate heuristics reduct (dhr), Greedy heuristics with superreduct (gs), Permutation heuristics reduct* (pr) and *Greedy heuristics with reduct (gr)*. For each of the four experimental runs, the effectiveness of each feature selection methodology was investigated with respect to the respective discretization method.

## 5.5 Experimentals

Experiments were conducted, covering three discretization techniques: *Unsupervised quantiles, Equal intervals* and *Local discernibility* and five feature subset computation algorithms that are based on the rough set theory as stated in the previous section. Having extracted the features from text documents written by the experimental subjects and pre-processed the data, six separate experimental runs were conducted for each of the three discretization techniques: one baseline in which the complete set of 346 condition attributes was used and five others; one for each of the five rough set based feature selection algorithm we are investigating. Three experimental runs were conducted separately for each of the NN and SVM models using a 10-fold and a 3-fold cross-validation settings respectively. The datasets for the experimentation were generated by setting up the parameters for each discretization method over the five feature subset computation methods.

## 6. RESULTS

The results obtained from the experiments are presented in this section. The outcome of the feature dimensionality reduction step is presented in Table 2. It was observed that the most significant stylometric feature categories are the character level, functions words, punctuations and word level features. This also suggests that the other three categories: special characters, vocabulary richness and word length

distribution features does not play a significant role in our experimental settings. This outcome is consistent with the outcome of other researches, particularly as it pertains to the usefulness of function words in the task of authorship attribution [22; 2].

The results of the experimental investigations based on these reduced sets are presented in Table 3 which shows the performances of each of the five subset computation/selection methods (along with the baseline) across all the selected metrics with respect to each of the discretization technique over all the performance measures that were examined. All the results are averages of three experimental runs. The ROC curves in Figures 2 (a-d), compares the performances of the algorithms (NN, SVM) on the baseline as against their performances on the computed most relevant subset features using the dynamically adjusted approximate heuristic reduct (*dhr*) method for a given author based on the unsupervised quantiles discretization.

In general, the results obtained across all the experiments conducted indicates that both the NN and SVM models performed significantly better on the rough set-computed relevance features subsets of the original data set itself. The results (Table 2) shows that that the overall classification accuracy increased from 8.712% for on the baseline data to 50.505% for the NN and from 7.197 % to 28.662% for the SVM model. This positive performance trend is demonstrated across all metrics. The SVM model however performed less impressively when compared to the NN model; this may be the effect of the drastic reduction in the feature set size, knowing that SVM does well naturally on high-dimensional data.

**Table 2. Relevance feature set obtained by RST feature subset computation methods across the three discretization techniques**

| | **Unsupervised quantiles discretization** | **Equal interval discretization** | **Local discernibility discretization** |
|---|---|---|---|
| dhr | Total characters; Function words: *i, about, would;* Punctuation: . | Word length: 3, ; Function words: *of, that* | Characters per word; Total characters, Function words: *would;* Punctuation: ?, ; |
| gr | Total characters, Function words: *are, i, about, would;* Punctuation: . | Word length: 3, 7, 10 Function words: *are, of, that, as, with, been* | Characters per word; Total characters, Word length: 5; Function words: *would, very, while;* Punctuation: ., ?, ; |
| gs | Total characters, Total words, Function words: *to, by, will, must, more* | Word length: 5, 6, 10 Function words: *of, that, all, with, been, if* | Word length: 13; Function words: *that, he, say, there at, them;* Punctuation: : |
| pr | Function words: *the, with, have, no, his, over, both, finally* | Word length: 9; Function word: *they, of, also, as, these, one, therefore, another, before, through, often;* Punctuation: : | Word length: 7; Function word: *are, he, these, whatever, those, my, getting, an through, ought;* Punctuation: ‘ |
| qr | Word length: 9, 5, 13 Function word: *from, will, those, itself* Punctuation: !, ; | Characters per word Word length: 5 Function word: *are, to, of, as, in, one, but, although, here, maybe* Punctuation: !, ; | Word length: 4, 5, 12 Function word: *were, not, many, never, but, me* Punctuation: !, ; |

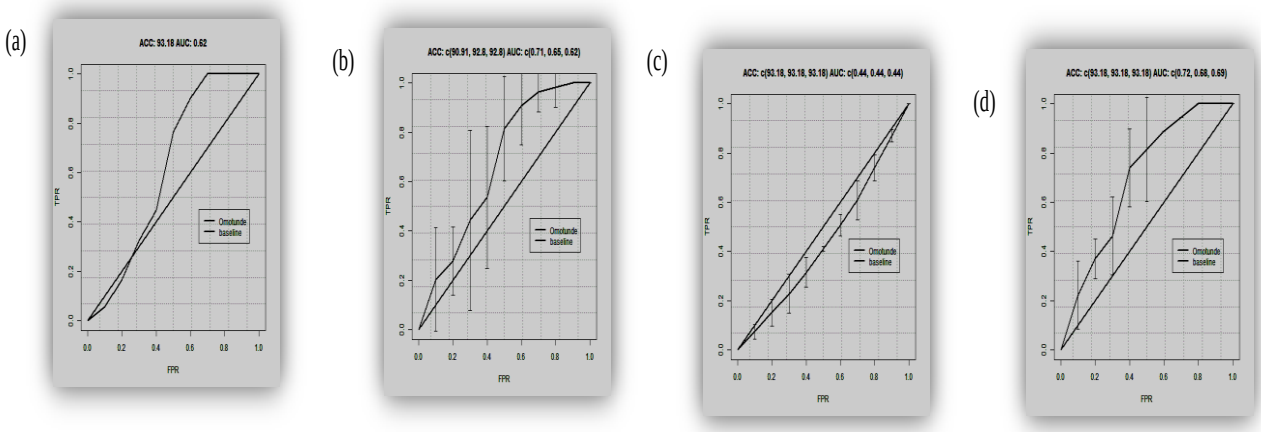


**Fig 1: ROC curves showing performance comparisons: (a) Full feature, NN, (b) Reduced feature, NN, (c) Full feature, SVM and (d) Reduced feature, SVM**

**Table 3. Performance measures for the feature subset computation techniques**

| Disc. Alg | Metric | Baseline | | Rough Set Feature Subset Computation Technique | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | dhr | | gr | | gs | | ps | | qr | |
| | | NN | SVM | NN | SVM | NN | SVM | NN | SVM | NN | SVM | NN | SVM |
| Unsupervised Quantiles | ACC | 7.576 | 7.197 | 44.066 | 28.662 | 42.424 | 27.778 | 29.672 | 16.793 | 17.298 | 14.773 | 36.111 | 19.192 |
| | ACCCL | 85.781 | 0.455 | 91.395 | 89.025 | 91.143 | 88.889 | 89.181 | 87.199 | 87.277 | 86.889 | 90.171 | 87.568 |
| | AUC | 0.662 | 0.441 | 0.825 | 0.751 | 0.807 | 0.734 | 0.742 | 0.702 | 0.641 | 0.613 | 0.775 | 0.711 |
| | BRIER | 0.070 | 0.072 | 0.058 | 0.065 | 0.059 | 0.066 | 0.071 | 0.068 | 0.08 | 0.071 | 0.067 | 0.068 |
| | F1 | 6.772 | 2.675 | 41.82 | 23.412 | 40.48 | 22.365 | 28.61 | 13.255 | 16.168 | 11.801 | 35.32 | 14.432 |
| | KAPPA | -0.413 | -0.962 | 39.376 | 22.703 | 37.573 | 21.746 | 23.796 | 9.807 | 10.397 | 7.648 | 30.751 | 12.437 |
| | PRECISION | 14.144 | 1.661 | 40.778 | 24.413 | 39.78 | 22.05 | 28.559 | 21.656 | 15.84 | 12.635 | 35.27 | 14.612 |
| | TNR | 92.276 | 92.234 | 95.339 | 94.057 | 95.2 | 93.984 | 94.141 | 93.065 | 93.109 | 92.897 | 94.675 | 93.265 |
| | TPR | 7.318 | 6.860 | 43.553 | 28.247 | 41.803 | 27.315 | 29.359 | 16.386 | 17.216 | 14.665 | 35.819 | 19.151 |
| Equal Intervals | ACC | 8.712 | 6.818 | 26.641 | 23.106 | 36.742 | 27.778 | 39.646 | 25.000 | 36.869 | 26.768 | 43.813 | 25.884 |
| | ACCCL | 85.956 | 85.665 | 88.714 | 88.171 | 90.269 | 88.889 | 90.715 | 88.462 | 90.288 | 88.734 | 91.356 | 88.598 |
| | AUC | 0.717 | 0.441 | 0.752 | 0.701 | 0.794 | 0.731 | 0.784 | 0.738 | 0.793 | 0.705 | 0.822 | 0.743 |
| | BRIER | 0.068 | 0.072 | 0.069 | 0.067 | 0.067 | 0.066 | 0.066 | 0.066 | 0.07 | 0.067 | 0.062 | 0.066 |
| | F1 | 6.271 | 1.574 | 25.487 | 16.409 | 36.504 | 21.058 | 38.824 | 18.883 | 36.781 | 21.365 | 43.331 | 18.095 |
| | KAPPA | 0.743 | -1.235 | 20.468 | 16.693 | 31.458 | 21.704 | 34.575 | 18.738 | 31.577 | 20.634 | 39.109 | 19.61 |
| | PRECISION | 6.017 | 1.574 | 25.318 | 14.652 | 36.695 | 19.112 | 38.996 | 17.635 | 36.96 | 25.363 | 43.874 | 20.173 |
| | TNR | 92.366 | 92.213 | 93.884 | 93.593 | 94.728 | 93.976 | 94.969 | 93.751 | 94.735 | 93.892 | 95.317 | 93.816 |
| | TPR | 8.317 | 6.594 | 26.235 | 22.989 | 36.83 | 27.935 | 39.298 | 24.852 | 37.088 | 27.224 | 43.669 | 25.647 |
| Local Discernibility | ACC | 6.061 | 6.818 | 46.97 | 23.737 | 50.505 | 26.768 | 30.808 | 21.338 | 25.126 | 19.318 | 37.5 | 24.116 |
| | ACCCL | 85.548 | 85.665 | 91.842 | 88.268 | 92.386 | 88.734 | 89.355 | 87.899 | 88.481 | 87.588 | 90.385 | 88.326 |
| | AUC | 0.660 | 0.441 | 0.839 | 0.761 | 0.832 | 0.784 | 0.758 | 0.656 | 0.719 | 0.657 | 0.778 | 0.769 |
| | BRIER | 0.070 | 0.072 | 0.057 | 0.065 | 0.055 | 0.064 | 0.07 | 0.069 | 0.077 | 0.069 | 0.067 | 0.067 |
| | F1 | 4.423 | 2.541 | 46.077 | 15.667 | 49.73 | 17.868 | 29.66 | 15.734 | 24.834 | 15.845 | 36.34 | 17.691 |
| | KAPPA | -2.051 | -1.235 | 42.528 | 17.367 | 46.346 | 20.748 | 25.005 | 14.7 | 18.842 | 12.543 | 32.255 | 17.738 |
| | PRECISION | 3.839 | 1.574 | 46.027 | 16.44 | 50.409 | 18.356 | 29.681 | 16.032 | 24.904 | 17.128 | 36.325 | 18.429 |
| | TNR | 92.151 | 92.213 | 95.581 | 93.645 | 95.874 | 93.906 | 94.232 | 93.44 | 93.757 | 93.271 | 94.791 | 93.674 |
| | TPR | 5.846 | 6.594 | 46.607 | 23.597 | 50.199 | 27.034 | 30.635 | 20.95 | 25.095 | 19.489 | 37.058 | 23.796 |

# 7. CONCLUSION AND FUTURE WORK

This paper presented a study of the usefulness of rough set based feature selection techniques to the authorship attribution problem. The results indicate positive showed that on multi-class authorship attribution problems, rough sets feature selection methods appreciably improves the performance of both neural network and the SVM models. The next task is to further investigate the performance of resulting set of features using short length texts across a variety of topics to ascertain their limits of applicability.

# 8. ACKNOWLEDGEMENT

The researchers wish to acknowledge the publishers of The Nation Newspapers and the Columnists whose writings have been used for this research. We wish to state that the texts/write-ups we have used from their online sources remain their property in entirety and that their use here was for the purposes of research only and it remains so.

# 9. REFERENCES

[1] Abbasi, A. and Chen, H., 2008: Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. ACM Transactions on Information Systems. 26(2):1-29

[2] Argamon, S., & Levitan, S., 2005:. Measuring the usefulness of function words for authorship attribution. In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, Association for Computing and the Humanities, Victoria, BC.

[3] Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., & Levitan, S., 2007: Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58(6), 802–822.

[4] Baayen, R., van Halteren, H., & Tweedie, F., 1996: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3), 121–131.

[5] Bishop, C. M. (1995) Neural Networks for Pattern Recognition. Oxford University Press.

[6] Buoanani, S. M. & Kassou, I., 2014: Authorship Analysis Studies: A Survey. International Journal of Computer Applications 86 (12): 22-29.

[7] Burrows, J.F.,1992: Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing, 7(2), 91–109.

[8] Chen, H. Li, J. and Zheng, R., 2006: From Fingerprint to Writeprint. Communication of the ACM, 49(4)

[9] Cortez, P., 2010: Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner (Ed.), Advances in Data Mining - Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining, LNAI 6171, Springer, pp. 572-583, Berlin, Germany, July, 2010.

[10] Devel, O., 2000: Mining e-mail Authorship. In Proceedings of the Workshop on the Text Mining in ACM International Conference on Knowledge Discovery and Data Mining.

[11] Duda, R. O. and Hart, R. E., 1973: Pattern Recognition and Scene Analysis. Wiley, New York.

[12] Evans J, Stanovich K. E., 2013:. Dual-process theories of higher cognition advancing the debate. Perspect. Psychol. Sci. 8:223–41

[13] Fukunaga, K. (1990) Introduction to Statistical Pattern Recognition. Academic Press New York

[14] Forsyth, R., and Holmes, D., 1996: Feature-finding for text classification. Literary and Linguistic Computing, 11(4), 163–174

[15] Gardner, B., Lally, P. and Wardle, J., 2012: Making health habitual: the psychology of habit-formation and general practice. Br J Gen Pract 62:664–666.

[16] Hall, M. A., 1999: Correlation-Based Feature Selection for Machine Learning. PhD Thesis, Waikato University, New Zealand.

[17] Holmes, D.I., 1994: Authorship attribution. Computers and the Humanities, 28, 87–106.

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

[19] Houvards, J., & Stamatatos, E., 2006: N-gram feature selection for authorship identification. In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (pp. 77–86). Berlin, Germany: Springer.

[20] John, G. H., Kohavi, R. and Pfleger, K., 1994: Irrelevant Features and the Subset Selection Problem. Proceedings of the 11th International Conference in Machine Learning. 121-129.

[21] Juola, P., 2007: Future trends in authorship attribution. In P. Craiger & S. Shenoi (Eds.), Advances in digital forensics III (pp. 119–132). Boston: Springer.

[22] Kestemont, M., 2014: Function Words in Authorship Attribution From Black Magic to Theory? Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) at EACL 2014, pages 59–66, Gothenburg, Sweden, April 27, 2014. Association for Computational Linguistics

[23] Kukushkina, O.V., Polikarpov, A.A., and Khmelev, D.V., 2001: Using literal and grammatical statistics for authorship attribution. Problems of Information Transmission, 37(2), 172–184.

[24] Komorowski J., Pawlak Z., Polkowski L. and Skowron A., 1999: Rough sets: A tutorial, In: Rough Fuzzy Hybridization: A New Trend in Decision Making (S.K. Pal and A. Skowron, Eds.). — Singapore: Springer, pp.3–98.

[25] Luyckx, K., and Daelemans, W., 2005: Shallow text analysis and machine learning for authorship attribution. In Proceedings of the 15th meeting of Computational Linguistics in the Netherlands (pp. 149–160). Utrecht, Netherlands.

[26] Mahor, U. and Das, S., 2015: Performance Evaluation of Various Feature Extraction and Classification Techniques for Authorship Attribution. International Journal of Innovation and Scientific Research. 1(16):252-259.

[27] Mitchell, T., 1997: Machine learning. New York: McGraw-Hill.

[28] Pawlak, Z., 1982: Rough Sets. International Journal of Information and Computer Science 2:341-356.

[29] Rissino, S. and Lambert-Torres, G., 2009: Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications, Data Mining and Knowledge Discovery in Real Life Applications. 35-60.

[30] Stamatatos, E. Fakotakis, N. and Kokkinakis, G., 2000: Computer-Based Authorship Attribution without Lexical Measures. Computer and Humanities. pp 193-214.

[31] Stamatatos, E. Fakotakis, N. and Kokkinakis, G., 2001: Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 26(4).

[32] Stamatatos, E., 2006: Ensemble-based author identification using character n-grams. In Proceedings of the 3rd International Workshop onText-Based Information Retrieval (pp. 41–46).

[33] Stamatatos, E., 2007: Author identification using imbalanced and limited training texts. In Proceedings of the 4th International Workshop on Text-Based Information Retrieval (pp. 237–241).

[34] Stamatatos, E., 2008: Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management, 44(2), 790–799.

[35] Stamatatos, E., 2009: A Survey of Modern Authorship Attribution Methods, JASIST

[36] Swiniarski, R. W. and Skowron, A., 2003: Rough Set Methods in Feature Selection and Recognition. Pattern Recognition Letters, Elsevier. 24:833-849

[37] Tamboli, M. S. & Prasad, R. S., 2013: Authorship Analysis and Identification Techniques: A Review.

International Journal of Computer Applications 77(16): 11-15.

[38] Tweedie, F., Singh, S., and Holmes, D., 1996: Neural network applications in stylometry: The Federalist Papers. Computers and the Humanities, 30(1), 1–10.

[39] Walczak, B. and Massart, D. L., 1999: Rough Sets Theory. Tutorial. Chemometrics and Intelligence Laboratory Systems. 47. 1-6

[40] Wood, W., Rünger, D., 2016: Psychology of habit. Annual Review of Psychololology http://dx.doi.org/10.1146/annurev-psych-122414-033417.

[41] Zhang, M., Yao, J., 2004: A rough sets based approach to feature selection. In: Proc. 23rd Internat. Conf. of NAFIPS, pp. 434–439

[42] Zhao, Y., and Zobel, J., 2007: Searching with style: Authorship attribution in classic literature. In Proceedings of the 30th Australasian Computer Science Conference (pp. 59–68). New York: ACM Press.

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

[44] Stolerman, A., 2012: Authorship Attribution Using Writeprints. Machine Learning Final Project. Drexel University. http://www.stolerman.net/studies/cs613/cs613_Writeprints_Ariel_Stolerman_paper.pdf

[45] Can, M, Jamak, A, Savatic, A., 2012: Teaching Neural Networks to Detect the Authors of Texts Using Lexical Descriptors, Southeast Europe Journal of Soft Computing, 1, (1), pp. 57-72.