IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 10, 2019, accepted July 11, 2019, date of publication July 23, 2019, date of current version August 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2930536

Incorporating Topic Information in a Global Feature Selection Schema for Authorship Attribution

HAYRI VOLKAN AGUN[1] **AND OZGUR YILMAZEL**[2]
[1]Dilisim, Teknopark, Anadolu University, 26210 Eskişehir, Turkey
[2]Eskisehir Vocational School, Anadolu University, 26555 Eskişehir, Turkey
Corresponding author: Ozgur Yilmazel (ozgur@anadolu.edu.tr)

ABSTRACT Authorship attribution (AA) is a stylometric analysis task of finding the author of an anonymous/disputed text document. In AA, the performance improvement of class-based feature selection schemas, such as Chi-square, and Gini index over frequency-based feature selection schemas, such as document frequency, common n-grams, and inverted document frequency has been shown to be limited. In AA, the feature selection process is significantly affected by topic distributions. In this paper, we assess the performance of a global feature selection approach into which the document's topic category is incorporated to scale the existing feature weights. In this approach, the common features of an author among different topics indicate higher relevance for the author and thus have higher weights. On the other hand, features with biased topic distributions are assumed to have high topic relevance and lower weights. In this approach, the global topic measure and the author specific topic measure are combined in order to scale the existing selection weights of the features. The ten-fold cross-validation experiment result on a multi-topic dataset with a random topic distribution indicates that our approach improves the performance of Chi-square, modified Gini index, and common n-grams schemas significantly in the best performing configurations of the classifiers.

INDEX TERMS Authorship attribution, feature selection, text classification.

I. INTRODUCTION

The task of authorship attribution (AA) is the identification of the author of a disputed/unknown text document. Modern AA methods focus on determining the authorship through the supervised text classification methods; however, they are different from other text classification tasks in terms of feature engineering [1]–[4]. Feature sets suggested for exploiting the stylometric properties of the authors are generally assumed to be topic independent, and they indeed encode little or no information about the content of the document. In recent studies, these feature sets are addressed as vocabulary richness, readability measures, character n-grams, terms and function words [5]–[8].

In AA, the number of studies involving a feature selection process is limited, and traditional feature selection schemas, such as document frequency filtering, information gain, odds ratio and chi square have been compared on datasets with very few authors. According to these comparisons, simple document frequency (DF) based term selection has been reported to be quite competitive with other feature selection methods [9], [10]. Along with traditional class-based feature selection schemas, a frequency-based feature selection method known as local common n-grams (CNG) has been specifically proposed for feature selection in AA tasks [11]–[13]. However, CNG has not been compared with existing feature selection approaches on the same datasets.

Traditional filter-based feature selection approaches in text classification use co-occurrence frequencies of the features and the class-labels for ranking features according to their discriminative power. In this study, we propose a new feature selection schema, which assists a class-based feature selection process by using topic information of the document. Our approach works on the assumption that the uniform distribution of a feature among different topics is a good property for stylometric analysis. Therefore, such features should have

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

higher weights. On the other hand, the features correlated with certain topics should have appropriate weights based on the divergence of the author specific distributions in a given topic from their general distributions in the same topics. This approach is shaped according to the idea of both rewarding the features with function word characteristics for a given author, and punishing the features with topic bias for the same author. The proposed approach computes the appropriate scaling constants for the existing class-based weights through topical properties of the feature. In this study, we directly used general tags that the authors assign to their documents as topic categories. These tags represent the general topics, such as sports, travel, computers and technology. In order to assess the significance, we compared the performance of common n-grams (CNG), modified Gini index (GI), and Chi-square (CHI) with and without the adjustments on the feature weights made by the proposed method. Our experimental setup consists of 10-fold cross validation for multinomial naïve Bayes, multilayer perceptron and support vector machine classifiers on an English blog dataset with 100 authors. The comparisons are made on the dimensionality reduction sizes of 5000, 1000, 500, 100, and 50. The classification results indicate a significant performance increase for the scaling approaches on the baseline feature selection schema with the same set of classification parameters. Introducing the scaling approach to the existing feature selection schema improves the performance more as the feature selection dimension is reduced.

In Section II, common feature selection methods applied in AA tasks are briefly explained. In Section III, the details of three scaling techniques are given. In Section IV, the details of our datasets, evaluation strategy and evaluations for feature weighting and selection methods are given. In Section V, the results are interpreted and some conclusions are made.

II. FEATURE SELECTION METHODS

In text classification, feature weighting and selection have been analyzed in depth [14], [15]. In the literature of classification, the general approaches to feature selection is divided into three categories: wrappers, filters and embedded methods [16]. Wrappers use a classifier in order to access the performance in selecting subset of features. They consider feature dependencies in order to achieve better performance. However, for large feature sets, they are computationally expensive. Filter methods use information about the features, such as class frequency and document frequency to properly weight and select discriminative features. Embedded approaches act with the classifier states to classify and select features. They are more time-efficient than wrapper approaches [17], [18]. Most of the feature selection methods in text classification are filter-based methods since selecting tens of thousands of features by combinatorial elimination as in wrapper approaches is inappropriate.

Traditional feature selection approaches on text classification use the class labels to select the features in multi-class dataset. However, in a recent study, a wrapper approach has been presented to select subset of features in multi-label dataset [19]. In this study, the similarity of the samples with the same labels are maximized or preserved by the selection of a feature. The evaluations indicate significant performance improvement in the best performing baseline approaches. However, this approach is not scalable to hundreds of thousand of features as in our case.

Common filter-based feature selection approaches in text classification are document frequency (DF), odds ratio (OR), mutual information, Chi-square (CHI), improved Gini Index (GI), and Information Gain (IG). In AA, CNG approach is proved its effectiveness. Savoy experimented in the topics of sports and politics in Italian and English newspaper articles [10]. He indicated that among nine different feature selection approaches, CNG achieves more robust results than IG and CHI2.

AA approaches differ from topic classification in terms of features. In AA, function words (stop words) are not eliminated since they have stylometric properties [6], [20], [21] and are assumed to have less topic correlation [22]. In general, function word frequency in documents is much higher than topic words. In IDF, function words have lower weights but high term frequencies. With multiplication of IDF term weights and the term frequency, a balance in IDF based term weighting is maintained IDF achieves robust performance in datasets where authors write about a single topic. However, in a multi-topic dataset, the distribution of function words may differ across topics which create topic-biased weights in IDF. To handle topic bias in AA, the traditional approach is to select common features with simple document frequency filtering. In general, features with higher DF have been effective in most AA datasets [23]. However, in imbalanced topic distributions, features with high DF may have topic dependencies. Through this study, this problem is solved by introducing topic-based scaling for existing feature weights. In subsections A through C, the details of IDF, CHI2, GI and common n-gram methods are given.

A. INVERTED DOCUMENT FREQUENCY

Inverted document frequency (IDF) has been widely applied in text classification. IDF assumes that rare terms are more discriminative than common terms. Therefore it assigns higher scores to rare terms and lower scores to common terms. Although IDF is a common feature-weighting schema, it is not quite appropriate for feature selection in AA tasks since stylometric features have higher occurrence rates than other features. For instance, function words – a well-known feature set in AA – have higher document frequencies, thus when IDF selection schema is applied on arbitrary words, most function words will get lower scores and be eliminated. For this reason, the application of IDF in AA is restricted to feature weighting rather than feature selection. The common formulation of IDF is given in (1).

$$IDF\ (feature_i) = \log\left(\frac{N+1}{DF_i+1}\right) \tag{1}$$

In (1), N is the total number of documents, and DF_i is the number of documents where $feature_i$ appears. In order to eliminate infinity values, 1 is added to the nominator and the denominator before applying the logarithm. Since IDF is a global feature weighting schema, the final weight of the feature is equal to its IDF score.

B. CHI-SQUARE

Chi-square is a popular term selection method which is commonly used in classification [24], [25]. It is based on the independence assumption of two events, which are probability values computed by global occurrence of the term and the occurrence of term for a given class. If these events are dependent, then it can be inferred that the term has author bias thus is discriminative. The simplified formulation of Chi-square is given in (2).

$$ChiSquare\left(author_j, term_i\right) = \frac{N\,(AD-BC)^2}{(A+C)\,(B+D)\,(A+B)\,(C+D)} \quad (2)$$

In (2), N is total number of documents, A is the document frequency of $term_i$ for documents of $author_j$, B is the document frequency $term_i$ in the documents of other authors, C is the number of documents of $author_j$ where $term_i$ does not appear, and D is the number of documents of other authors where $term_i$ does not appear.

$$weight\,(term_i) = \frac{1}{M}\sum_{j}^{authors} ChiSquare(author_j, term_i) \quad (3)$$

The final weight of a term is computed by averaging the sum of Chi-square for all authors as given in (3), where M is the total number of authors, $author_j$ is current author and $term_i$ is the term to be weighted.

C. IMPROVED GINI INDEX

Traditional Gini index measures the impurity (non-purity) of term among the classes. It is introduced by learning of classification and regression trees in order to give a binary decision according the term distribution among categories. In this study we used the improved version of Gini Index [15]. Improved Gini Index (GI) is computed according to (4)

$$GI\,(t) = \sum_{i}^{authors} p\,(t|c_i)^2\,p\,(c_i|t)^2 \quad (4)$$

Probabilities of $p(t\text{ę}c_i)$and $p(c_i|t)$ are computed as in (5) and (6) respectively.

$$p\,(t|c_i) = \frac{Document\ frequency\ of term\ t\ in\ author\ c_i}{Number\ of\ documents\ for\ author\ c_i} \quad (5)$$

$$p\,(c_i|t) = \frac{Document\ frequency\ of\ term\ t\ in\ author c_i}{Number\ of\ documents\ where\ term\ t\ appears} \quad (6)$$

D. COMMON N-GRAMS

CNG is a feature selection technique specifically applied in AA [4], [12], [11]. Finding representative features for authors is applied by selecting common set of features for all authors separately. A feature is discriminative when it occurs in most of the documents written by a given author. Based on this definition, a score is computed by (7).

$$weight\,(term) = \frac{1}{M}\sum_{j}^{M} \frac{Document\ frequency\ of\ term\ in\ author\ c_j}{Number\ of\ document\ of\ author\ c_j} \quad (7)$$

CNG method is appropriately formulated in (7), where M is the total number of authors applied for averaging the score. Unlike inverted document frequency, this approach does not consider common or rare occurrences of features.

III. PROPOSED METHODOLOGY

Author term choice is determined by the elements of the content and author style [6], [12], [21]. In a broad sense, the content elements are topic and genre. Filter-based term selection methods use author and term correlations to determine a global score for the term. In AA tasks, the general assumption for selecting useful features is vulnerable to author – topic correlations. For instance, consider the case where the training documents of an author belong to a single topic, then the existing feature selection procedure for this author becomes biased to topic. In the worst-case scenario, this topic is not shared by other authors. In this case, a test document of another author written about this topic is likely to have non-discriminative features. Thus, in AA, the term – topic and author – topic correlations are effective properties for selecting features.

In the proposed approach, global topic and author specific topic measures are combined to scale existing feature weights. Global topic measure computes the entropy of features according to the topics. Entropy gives high scores for features with uniform distributions among topics and lower score for topic-biased features. Along with global topic measures, the author's topical divergence from the global topical divergence is an important property especially when authors are identified by their topic correlations. To measure these correlations, we measure the symmetric divergence between author specific topic and global topic distributions. Divergence computes the distance of the author topic distribution from global topic distribution for a specific feature and author. The final measure is obtained by averaging all the distances of the authors. Higher divergence average of a feature indicates higher discriminative power. On the other hand, lower divergence average of a feature indicates that the term choices in most of the topics is highly similar to their global occurrences in respective topics. Thus, they are less likely to be discriminative. In subsection A, the global topic measure is explained, and in subsection B, the adaptation of symmetric divergence to author specific measure is given. Finally, in subsection C, the final formulation for the combination of these measures is given.

A. GLOBAL TOPIC MEASURE

Global topic measure is applied for scoring the imbalance ratio of the feature topic distribution. In AA tasks,

topic bias has a negative effect because topics are usually shared among authors, and an author cannot be identified by the topic. Although the features with high occurrence rates, such as function words are likely to be uncorrelated with the topics, they might contain some topical dependencies. Thus, they should not be assumed to be discriminative. In order to measure how much a common feature is dependent to topic, entropy is used. Entropy is a measure of uncertainty of events [26]. It gives high scores for uniform distributions and low scores for skew distributions. In our case, a feature's uniform appearance among topics is a good property. Higher entropy values satisfy this criteria. Entropy measure is given in (8).

$$Entropy = \sum_{k}^{topics} p\,(topic|feature) \log_2 p(topic|feature) \tag{8}$$

In (8), we calculate $p\,(topic|feature)$ by dividing the document frequency of feature appearing in topic to global document frequency of the feature. This formulation is given in (9).

$$p\,(topic|feature) = \frac{document\ frequency\ of\ feature\ in\ topic}{total\ document\ frequency\ of\ feature} \tag{9}$$

After the entropy is computed for a given feature, the final weight of this feature is scaled by multiplying its weight with its entropy. It is given in (10).

$$\text{weight} = \text{weight}_\text{i} * \text{Entropy (feature)} \tag{10}$$

B. AUTHOR SPECIFIC TOPIC MEASURE

Author specific measure is applied to score the difference of term distribution from its global distribution according to the topics. The divergence of the author's vocabulary in a given topic is a discriminative property. In this approach, a symmetric divergence (J-Divergence) is used. It is a distance measure for determining how much a given probability distribution differs from another probability distribution [27]. The average of the distance for all authors and topics are used to score the feature. In (11), the formulation of symmetric divergence is given.

$$\text{J} - \text{Divergence} = \sum_{\text{k}}^{\text{topics}} (\text{p1} - \text{p2}) * \log_2 (\frac{\text{p1}}{\text{p2}}) \tag{11}$$

In (11), p1 is the joint probability of term_i and topic_k and p2 is the joint probability of feature_i for topic_k and author_j. Topics represent the topics of the documents written by the author_j. Formulation for these probabilities are given in (12) and (13).

$$p1 = \frac{DF\ of\ feature\ in\ topic_k}{number\ of\ documents\ in\ topic_k} \tag{12}$$

$$p2 = \frac{DF\ of\ feature\ for\ author_j\ and\ topic_k}{number\ of\ documents\ for\ author_j\ and\ topic_k} \tag{13}$$

In (12), DF of feature in topic is the number of documents where feature appears in the topic_k. In (13), the DF of feature for author_j and topic_k represents the number of documents written about topic_k by author_j. Lower topical divergence of features is a good property for generalization. It implies that the author's vocabulary is similar to the general vocabulary of the given topic. On the other hand, higher divergence indicates topical bias for the author and implies a discriminative function of the feature for the author. The final average of J-Divergence for all authors is given in (14), where m represents the number of authors, and the final feature weight is calculated as in (15).

$$Avg - Divergence = \sum_{j}^{authors} \frac{J - Divergence}{m} \tag{14}$$

$$weight = weight_i * AVE - Divergence \tag{15}$$

C. BALANCED MEASURE

Both global and author specific distribution of topics for each feature are important factors in feature selection. However, when the author writes in single and distinct topics, the topic bias increases, and the author discriminative power for the feature decreases. Thus, the topic bias in the feature weight should be canceled according to the degree of the topic and author correlation. Through combining the symmetric divergence and entropy, we neutralize the topic bias on the feature. We use symmetric divergence in order to detect author topic dependency. For example, if the given feature is distinctive for the topic yet its global topic distribution is not similar to the author topic distribution, then the symmetric divergence will balance the entropy score. In this approach, the final weight of a feature is computed through entropy and symmetric divergence as given in (16).

$$weight = weight_i * (1 + Entropy * Avg - Divergence) \tag{16}$$

Equation (16) should be interpreted as follows; when the entropy is low for a given feature, then it implies that the feature has topic correlations. In this case, if the symmetric divergence score is low, then the feature encodes much of the information about topics rather than authors. However, if the divergence score is high, then there are author specific dependencies in feature topic correlations, thus the feature weight should be increased. On the other hand, when the entropy measure is high, the feature is less correlated with topics, and the feature weight should be increased according to the topic correlations.

IV. EXPERIMENTS

In this section, the general pipeline for the preparation of the dataset, the details of the evaluation strategy and the results are given. In subsection A, the dataset preparation is introduced by giving the details of partitioning algorithm and filtering parameters that control topical density of the dataset. In subsection B, the evaluation strategy for authorship attribution is presented. In subsection C, the classification performance of each feature selection schema and the proposed scaling technique are compared. It is noted that the experimental setup created many results for classifiers,

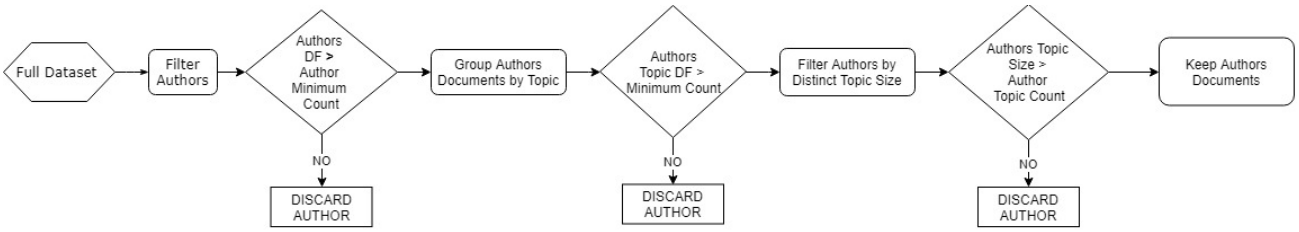

FIGURE 1. Author filtering by controlled topic distributions.

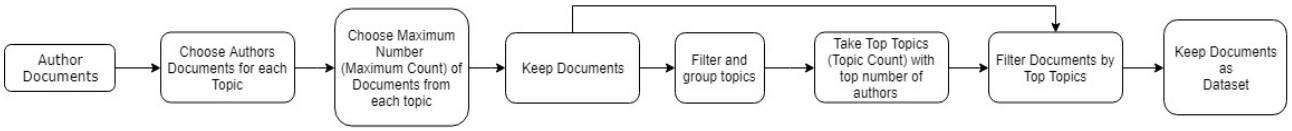

FIGURE 2. Topic and document filtering by controlled authors.

weighting and selection. Section V is dedicated to analyzing these results in detail.

A. DATASET PREPERATION

In order to build a multi-topic dataset, we crawled English blogs, namely Boing Boing, Engadet and Gizmodo. The collected documents contain a diverse set of topics written by over thousand authors. Documents are collected in plain xml format with tags of author name, title, topic keyword and paragraphs. Before dataset preparation, we applied data cleaning, where author relevant information is removed from the document content. Most of the author related content is placed in the ending paragraphs.

Such information consists of urls, emails and names. In order to remove this content, we scored the symbol density of the paragraphs where each symbol ('@', '/', '.', and ':') is counted and averaged by the number of tokens in the given paragraph. The paragraphs with the average symbol density score above 0.9 are removed from the document content.

In order to build a multi topic AA dataset, we used the topic keywords assigned by the authors. We filtered the documents according to the number of topics of each author and the number of documents in each topic. We used several parameters during filtering to maintain the balance of the number of the topics of authors. These parameters are *maximum document count*, *maximum number of topics*, *author minimum count* and *author topic count*. By changing these parameters, the topic and the document count per author are controlled. The flow diagram for document filtering based on the topic distributions is given in Fig. 1 and Fig. 2.

Fig. 1 is dedicated to choosing the authors who write at least about a certain number of distinct topics. Author filtering can be explained in three stages. In the first stage, we filter the authors with document frequency (DF) above *author minimum count*. In the second stage, we filter the topics for these authors. Topics with document frequency (DF) above *minimum count* are selected and used to filter the authors in the next stage. In the third stage, we filter authors based on the number of distinct topics selected in stage two. In this stage, the authors who write in at least minimum number of topics are filtered by using *author topic count*. Finally, all the authors and corresponding documents are stored as documents in order to be processed in the next stage of dataset generation as demonstrated in Fig. 2.

In Fig. 2, the documents and authors which are filtered in the previous filtering stage are used for filtering the topics. Firstly, we select the documents from each topic for each author separately so that each author has a certain number of documents in each topic. Secondly, we use these documents to create top topic set. A topic in top topic set has the top number of distinct authors. Later, we filter the documents generated by the top topic set. Finally, we keep or shuffle the results to feed into k-fold cross validation experiments. It is noted that the topic density of the dataset is determined via *minimum count*, *maximum count* and *topic count*. Other parameters, such as *author minimum count* are used to fix the number of documents per author. The generated dataset will have at least *author topic count* multiplied by *the minimum count* number of documents per author, and for each author, the distinct topic count will be above *author topic count*.

For creating the dataset we fixed the parameters of the first stage parameters the *author minimum count*, *minimum count, author topic count* as 120, 10, and 2 and the parameters of the second stage parameters *maximum count*, and *topic count* as 240 and 20 respectively. By using these parameters and the cross-validation, we created 10 different training/testing collections. The final collection contains 100 authors and a total of 18823 documents. In this dataset, the authors write at most in 4 topics and at least in 2 topics, and the total number of topics in the dataset is 20. Each author has at least 100 documents and at most 200 documents. The average number of distinct terms in the dataset is 53368.

B. EVALUATION

In evaluation, macro averaged f-measure along with cross validation is used. F-measure is a popular scoring technique for evaluating multi-class classification tasks. It is equally weighted harmonic mean of precision and recall. Precision and recall are the fundamental evaluation measures for text classification. Precision measures the positive prediction performance, and recall measures the positive prediction rate. Precision and recall are formulated over true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP is the correct prediction count for a given class. FP is the incorrect prediction count for a given class. TN is the number of correct predictions for the samples that belong to other classes (negative cases). FN is the number of incorrect predictions of a given class for the samples that belong to other classes. Precision, recall and f-measure are given in (17), (18), and (19) respectively.

$$precision = \frac{TP}{TP + FP} \tag{17}$$

$$recall = \frac{TP}{TP + FN} \tag{18}$$

$$F - measure = \frac{precision * recall}{precision + recall} \tag{19}$$

In the experimental setup, we applied 10-fold cross validation where equal percentage of samples from each class are selected for training and testing. Before applying 10-fold cross-validation, first the authors are sorted according to their document size, and the authors who have the most documents in the collection are selected for evaluation.

C. RESULTS

In this section, the details of evaluation of the dataset prepared by the steps given in the previous section are given. In all the experiments, the common text classifiers known as multinomial naïve Bayes (NB), multilayer neural network (NN) and C-SVM (SVM), are used. In multinomial naïve Bayes classifier, the default settings of Spark MLLIB Framework [28] is used. In multilayer neural network, single hidden layer with the 1/3 of the size of the input layer is used. On the other hand, for the C-SVM classifier, the LIBSVM implementation is used by selecting the linear classifier with shrinking [29] .

In Table I, Table II and Table III, the macro F-measure scores for CNG, Chi-square (CHI), and modified Gini index (GI) methods and their scaled versions in settings of feature selection with dimensionality sizes of 5000, 1000, 500, 100 and 50 terms are compared. The terms are extracted by a simple tokenizer without any stemming and processing. For comparability, the selected terms are weighted by TF-IDF schema in all the experiments. All the scaling approaches are separated by a dash symbol from the name of the feature selection schema. Entropy (ENT), divergence (DIV) and the balanced schema (BAL) are combined with common n-grams (CNG), Gini index (GI), and Chi-square (CHI). In Tables I, II and III, for each dimensionality, the top performances are marked with asterisk symbol (*). Moreover without using feature selection, the scores for NB, NN, and C-SVM classifiers are obtained as 54.496, 66.305, and 66.9.

TABLE 1. F-Measures for multinomial naïve bayes classifier.

Dimension / Method	5000	1000	500	100	50
CNG	61.859	53.792	47.347	25.342	18.478
CNG-ENT	61.894	54.011	47.491	26.442	18.641
CNG-DIV	62.793	54.074	48.721	29.977	22.396
CNG-BAL	62.686	53.797	47.975	28.002	19.941
GI	54.658	24.217	16.816	4.519	4.086
GI-ENT	55.204	24.739	22.888	6.915	7.206
GI-DIV	61.668	43.437	28.702	17.099	11.708
GI-BAL	56.343	29.208	23.420	13.292	8.086
CHI	60.032	35.620	23.203	6.992	6.288
CHI-ENT	54.084	37.050	23.440	16.165	10.864
CHI-DIV	62.968*	57.485*	49.356*	32.832*	24.064*
CHI-BAL	61.593	44.692	34.148	19.042	19.203

TABLE 2. F-Measures for multilayer neural network classifier.

Dimension / Method	5000	1000	500	100	50
CNG	66.442	64.770	64.310	51.651	44.378
CNG-ENT	67.690	64.981	64.804	51.549	44.929
CNG-DIV	68.134	66.827	66.959	52.485	47.019
CNG-BAL	66.634	65.775	67.061	52.444	45.469
GI	64.960	34.252	22.439	4.851	4.349
GI-ENT	66.292	34.494	22.627	7.777	7.710
GI-DIV	69.238	57.947	45.624	28.522	17.190
GI-BAL	67.931	35.735	24.582	7.890	5.842
CHI	68.489	54.592	42.544	12.182	11.248
CHI-ENT	66.738	49.754	38.972	28.342	17.801
CHI-DIV	68.528	67.757*	67.080*	52.579*	47.656*
CHI-BAL	69.157*	64.854	58.995	40.936	39.865

According to the results given in Tables I, II and III, ENT, DIV and BAL scalers improve the performance of the baseline feature selection schema for naïve Bayes, multilayer neural network and C-SVM classifiers. At least one of the three proposed scaling methods achieved the top performance in all term selection experiments.

In general, the top feature selection performances are obtained with the dimensionality size of 5000. In dimensionalities of 1000, 500, 100, and 50, the performance of CNG is significantly higher than the performance of CHI and GI schemas. As the dimensionality reduces, the top scores decrease, the performance of CNG becomes higher than the

TABLE 3. F-Measures for C-SVM classifier.

Dimension / Method	5000	1000	500	100	50
CNG	69.600	67.400	65.800	53.300	42.600
CNG-ENT	70.033	67.900	66.100	55.300	43.100
CNG-DIV	68.700	67.500	66.800	57.200	46.860*
CNG-BAL	69.990	68.100	67.100*	57.500*	44.166
GI	40.190	33.200	20.300	5.750	5.166
GI-ENT	59.700	34.200	20.700	8.200	7.933
GI-DIV	70.700	57.800	46.400	25.600	16.033
GI-BAL	68.500	50.700	22.300	7.800	9.433
CHI	70.700	53.700	33.800	10.200	8.466
CHI-ENT	68.100	63.700	35.600	24.100	15.100
CHI-DIV	72.800*	70.190*	66.900	57.300	46.800
CHI-BAL	71.100	65.600	60.500	43.300	35.830

performance of CHI and GI, and the improvement of the scalers of ENT, DIV and BAL becomes more significant. In terms of classifier comparisons, C-SVM achieves the best performances; however, as the dimensionality decreases, the performance of NN classifier becomes higher than C-SVM classifier. The performances of the proposed scaling methods vary with the baseline feature selection method and the classification algorithm, but in general, the scalers improve the best performance for all classifiers in all dimension sizes.

In all the classifiers, the best performing scaler is DIV and the performance improvements of DIV over the best classification performances in dimensions of 5000, 1000, 500, 100 and 50 are approximately 2, 2.5, 3, 4 and 5 points. On the other hand, the improvement of ENT and BAL scalers is significantly dependent on the classifier and the dimension. All the scalers improve the baseline performance significantly.

V. CONCLUSIONS

Modern feature selection schemas on text classification tasks have been experimented in content dependent tasks where the document content and target label are directly related. For instance, in topic classification, the content of the document contains terms related with the topic. However, in AA, the content of a document is not directly related with the label. Given the topic, the document content of two different authors is very likely to be similar.

In most real scenarios, the authors write in multiple topics. In large datasets with multiple topics and large number of authors, the performance of authorship attribution is mostly determined by the discriminative power of terms. Feature selection approaches on these conditions become a necessity for extracting discriminative features. In AA tasks, the frequency-based feature selection schemas, such as common n-grams and document frequency have been shown to be quite competitive with class-based feature selection schemas. In this study, n-grams, modified Gini index, Chi square in a large multi-topic AA datasets with 18.823 documents written in 20 topics are compared, and a novel term weight scaling approach to improving the performance of these feature selection schemas is proposed. Proposed scaling approaches use the distribution properties of a term in different topics and count in function word like behavior in feature selection. Our approach does not eliminate function words and other discriminative features strictly as in other feature selection schemas.

In the experimental evaluations, it is shown that using the proposed scaling approaches to scale the existing weights achieves the top performance in multinomial naïve Bayes, multilayer neural network, and C-SVM classifiers. When the best performing feature selection performance is chosen as baseline, the performance improvements of the scalers for different dimensionality reduction sizes are in a range of $2-25\%$, which indicates a significant performance improvement on macro F-measure scores obtained by 10-fold cross validation on a large dataset. However there are also cases where the scaler does not improve the performance. For example, for Chi-square (CHI) the entropy scaler (ENT) does not improve the classification results. In such cases the scaler and the baseline term selection are not behaving compatibly and the scaler may be reducing the weights of discriminative terms. This generally occurs in higher dimensionality reduction sizes and is specific to the scaler and term selection method.

In AA studies, genre and topic are such important factors that they improve the feature selection performance. On the other hand, a similar relationship between topic and author style exists in multi-domain and multi-genre topic classification tasks where the genre and author style have a negative impact on the prediction performance of the topic. In such cases, the information of the genre and author style can be employed similarly for feature selection.

REFERENCES

[1] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 9–26, 2008.
[2] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Lang. Resour. Eval.*, vol. 45, no. 1, pp. 83–94, 2011.
[3] M. Koppel, J. Schler, S. Argamon, and Y. Winter, "The 'fundamental problem' of authorship attribution," *English Stud.*, vol. 93, no. 3, pp. 284–291, 2012.
[4] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, 2009.
[5] D. L. Hoover, "Another perspective on vocabulary richness," *Comput. Humanities*, vol. 37, no. 2, pp. 151–178, 2003.
[6] M. Kestemont, "Function words in authorship attribution. From black magic to theory?" in *Proc. 3rd Workshop Comput. Linguistics Literature (CLFL)*, 2014, pp. 59–66.
[7] H. Somers and F. Tweedie, "Authorship attribution and pastiche," *Comput. Humanities*, vol. 37, no. 4, pp. 407–429, 2003.
[8] Y. Zhao and J. Zobel, "Effective and scalable authorship attribution using function words," in *Proc. Asia Inf. Retr. Symp.*, 2005, pp. 174–189.
[9] J. Savoy, "Feature selections for authorship attribution," in *Proc. 28th Annu. ACM Symp. Appl. Comput.*, 2013, pp. 939–941.

[10] J. Savoy, "Comparative evaluation of term selection functions for authorship attribution," *Literary Linguistic Comput.*, vol. 30, no. 2, pp. 246–261, 2015.
[11] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proc. Pacific Assoc. Comput. Linguistics*, 2003, pp. 1–10.
[12] M. van Dam and C. Hauff, "Large-scale author verification: Temporal and Topical Influence," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, vol. 14, 2014, pp. 1039–1042.
[13] M. Jankowska, V. Kešelj, and E. Milios, "CNG text classification for authorship profiling task," in *Proc. CLEF*, 2013, vol. 1, no. 1, pp. 1–3.
[14] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, "Feature selection methods for text classification," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 230–239.
[15] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 1–5, 2007.
[16] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[17] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[18] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowl.-Based Syst.*, vol. 36, pp. 226–235, Dec. 2012.
[19] Z. Cai and W. Zhu, "Feature selection for multi-label classification using neighborhood preservation," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 320–330, Jan. 2018.
[20] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan, "Stylistic text classification using functional lexical features," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 6, pp. 802–822, 2007.
[21] K. Luyckx, "Syntax-based features and machine learning techniques for authorship attribution," M.S. thsis, Univ. Antwerp, Antwerpen, Belgium, 2004.
[22] G. K. Mikros and E. K. Argiri, "Investigating topic influence in authorship attribution," in *Proc. CEUR Workshop*, vol. 276, 2007, pp. 29–35.
[23] T.-Y. Qian, B. Liu, Q. Li, and J. Si, "Review authorship attribution in a similarity space," *J. Comput. Sci. Technol.*, vol. 30, no. 1, pp. 200–213, 2015.
[24] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression ," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.
[25] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 80–89, 2004.
[26] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[27] J. K. Chung, P. L. Kannappan, C. T. Ng, and P. K. Sahoo, "Measures of distance between probability distributions," *J. Math. Anal. Appl.*, vol. 138, no. 1, pp. 280–292, 1989.
[28] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. B. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar, "Mllib: Machine learning in apache spark," *J. Mach. Learn. Res.*, vol. 17, pp. 1–7, Jan. 2016.
[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

HAYRI VOLKAN AGUN was born in Erzurum, Turkey, in 1982. He received the B.S. degree from Anadolu University, Eskişehir, in 2001, the M.S. degree from Trakya University, Edirne, in 2008, and the Ph.D. degree from Eskisehir Technical University, Eskisehir, in 2019, all in computer engineering.

From 2005 to 2008, he was a Research Assistant with the Computer Engineering Department, Trakya University, Edirne, Turkey. From 2006 to 2008, he has collaborated in the projects of the Trakya Cognitive Science Society. He has published several scientific articles and contributed to natural language and information processing research, from 2010 to 2018. He continues his research in information processing at Dilisim, a big data and search services company, Eskişehir. His research interests include natural language processing, text classification, and information extraction.

OZGUR YILMAZEL was born in Kutahya, Turkey, in 1974. He received the B.S. degree in electrical and electronics engineering from Osmangazi University, Eskişehir, Turkey, and the M.Sc. and Ph.D. degrees in electrical engineering from Syracuse University, Syracuse, NY, USA, in 2002 and 2006, respectively.

From 1999 to 2006, he was the Chief Software Engineer with the Center for Natural Language Processing, Syracuse University. After completing his Ph.D. degree, he continued as an Assistant Research Professor at the Information School, Syracuse University, for two years. He was the CIO of Anadolu University, from 2010 to 2014, where he was the Department Chair of the Informatics Department, from 2010 to 2016. He joined the Computer Engineering Department, Anadolu University, as an Assistant Professor. He is currently an Associate Professor of information and records management with Anadolu University, Eskişehir. He is also the CEO and the Co-Founder of Dilisim, big data and search services company, where he applies his academic work into commercial applications. His research interests include big data analytics, evidence extraction, machine learning, information visualization technologies, and text classification.