AN EMPIRICAL EVALUATION OF WRITING STYLE FEATURES IN CROSS-TOPIC AND CROSS-GENRE DOCUMENTS IN AUTHORSHIP IDENTIFICATION

Simisani Ndaba[1] Edwin Thuma[2] Gontlafetse Mosweunyane[3]

Department of Computer Science, University of Botswana, Gaborone, Botswana

ABSTRACT

In this paper, an investigation was done to identify writing style features that can be used for cross-topic and cross-genre documents in the Authorship Identification task from 2003 to 2015. Different writing style features were empirically evaluated that were previously used in single topic and single genre documents for Authorship Identification to determine whether they can be used effectively for cross-topic and cross-genre Authorship Identification using an ablation process. The dataset used was taken from the 2015 PAN CLEF Forum English collection consisting of 100 sets. Furthermore, it was investigated whether combining some of these feature sets can help improve the authorship identification task. Three different classifiers were used: Naïve Bayes, Support Vector Machine, and Random Forest. The results suggest that a combination of a lexical, syntactical, structural, and content feature set can be used effectively for cross topic and cross genre authorship identification, as it achieved an AUC result of 0.837.

KEYWORDS

Authorship Identification, Cross-topic and Cross-genre, Single-topic and Single-genre, Writing style feature

1. INTRODUCTION

To determine a writer of an anonymous text has been of interest in many domains since the nineteenth century [1]. These areas include Information Retrieval, Investigative Journalism and in Law where identifying the writer of a document such as a ransom note may be crucial in saving lives. [2] and [3] gave many practical examples where knowing the author of a document may be very important. For example, finding the author of a malicious mail sent from an anonymous email account, plagiarism detection and to catch paedophiles. Authorship identification is used to solve these problems by determining whether a known author based on his or her text samples has written an unknown text. Authorship identification uses an author's writing style in identifying writers of texts. An author's word choice, sentence structure, figurative language, and sentence arrangement are extracted from a text and categorised into writing style features for measuring an author's personal writing style.

The problem is complicated by the fact that an author may consciously or unconsciously vary their writing style from text to text [4]. This is because the writing style of an author may be affected by the genre in addition to the personal style of an author. It may also be heavily affected by topic nuances. The writing style trend of a topic for a particular author may be the same in a genre and vice versa. Thus, when some documents match in genre and topic, the personal writing style of an author would be the major discriminating factor between texts. However, it is no longer assumed that all texts within an authorship identification problem match in genre and

David C. Wyld et al. (Eds): CSML, NET, BDHI, SIPO, SOEA- 2023
pp. 39-51, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130104

topic. The assumption has been updated to a cross-genre and cross-topic idea in the authorship identification task which corresponds to a more realistic view of the problem [5]. In many applications, it is not possible to obtain text samples of known authors in specific genres and topics. For example, the author of an anonymously published crime fiction novel may be a child fiction author who has never published a crime fiction novel before.

2. OBJECTIVE

This paper set out to identify the ideal writing style features for cross-genre and cross-topic documents in the PAN CLEF 2015 Authorship Identification task from 2003 to 2015. This study plans on using the writing style features that were previously used in single topic and single genre documents for authorship identification to determine whether they can be used effectively for cross-topic and cross-genre documents for authorship identification using an ablation process. To the best of the authors' knowledge, this review and experiment set up had not been worked on from 2003 to 2015 and contributes to the task area to identify which features are best used in cross-genre and cross-topic documents. Three different classifiers were used in the empirical evaluation to see whether the results generalise well across the different family of classifiers.

The rest of the paper is organised as follows, Section 3 reviews the background, evolution of writing style features in authorship identification and related works. Section 4 describes the methodology, outlining the dataset, data preparation and experimental setup. Section 5 analysed the results, limitation and recommendations of the study and Section 6 provides the conclusion.

3. RELATED WORKS

Authorship Identification has been dated since the nineteenth century with the preliminary study of Mendenhall [6] on the plays of Shakespeare. This was followed by statistical studies in the first half of the twentieth century by [7] and [8]. Subsequently, [9] method was based on Bayesian statistical analysis of the frequencies of a small set of common words (e.g. 'and', 'to', etc.) and produced significant discrimination results. According to [1], in the late 1990s, research in Authorship Identification was dominated by attempts to define writing style features known as Stylometry. [10] referred to Stylometry as the statistical analysis of a style and assumes that every author's writing style has certain features that are unique. Hence, a great variety of writing style features including word frequencies, character frequencies, vocabulary richness, sentence length and word length had been proposed. [11] reported that to extract unique writing style from data, writing style features such as Lexical, Syntactic, Structure and Content-specific feature sets need to be considered.

Initial Authorship Identification studies used available datasets that had single-genre and single-topic documents and their composition made for ease of comparison because words and expressions belonged to a similar domain. Due to the increasing variety of text topics and genres published in different media over the years, cross-topic and cross-genre documents have formed text samples from a variety of authors in a variety of genres such as Emails, Essays, Discussions and various topics such as, Religion, Marriage, and Discrimination [12]. The following works were considered from 2003 to 2015 for the Authorship Identification task for the cross-topic and cross-genre documents used during that time.

Previous works that used single-topic and single-genre documents include [13] who used English editorial documents using Vocabulary Richness writing style features. Other writing style features used were number of letters, number of uppercase characters, digits and number of white spaces. All writing styles features, except for the vocabulary richness measures were represented by a vector using the td-idf technique. Naive Bayes, Support Vector Machines (SVM) and Multilayer

Perceptron were used to build different classification models using their default parameters. The highest result achieved was an AUC of 0.972 with SVM.

[14] used a single-genre dataset consisting of novels from four writers and used number of unique words, Vocabulary richness, word n-grams, sentence length, word length, and frequencies of punctuation writing style features. Multinomial Logistic Regression, Naive Bayes, SVM and Decision Tree were used to apply multi-categorising and their results showed that non-vocabulary richness writing style features boosted the result using SVM with an AUC of 0.85. Similarly, [15] used a single-genre dataset made up of novels from different writers from different time periods to develop a method for computer-assisted Authorship Identification using character n-grams. They used a dissimilarity measure between the documents to measure the average frequency for a given n-gram in each document and achieved an AUC of 0.83.

A single-topic dataset comprising of Computer Science related subjects was used in a model by [16]. An ensemble of 14 n-gram patterns from Lexical tokens unigrams, bigrams, characters 4-grams and Syntactical POST unigrams to trigrams writing style features were used. Euclidean and Cosine distance measures reflected how close the writing style features are from one document to another based on frequency value differences. The SVM, logistic regression, decision trees and Naive Bayes classifiers were used with their tuned parameters and evaluated on the dataset using a cross-validation method. The POST n-grams did not perform as well as lexical n-grams probably due to terms not being tagged well. An average result of 0.767 was achieved by the model.

In another writing style feature ensemble, [17] obtained writing style features in a Korean web forum for categorising Good or Bad user reputation based on user feedback. Features from Lexical set include frequency of digit characters, frequency of white space characters and frequency of alphabetic characters. Syntactical set with frequency of punctuations, frequency of stop words and frequency of POS n-grams (n = uni, bi, tri). The Structural set had measures quoted content including news, e-mail as signature and telephone number as signature and Content set had Word n-grams (n = uni, bi, tri). The feature sets were added on for evaluation in an incremental order, that is, F1, F1+F2, F1+F2+F3, etc. Naïve Bayes, SVM, Decision Tree and Neural Network with their tuned parameters applied the 10 fold cross validation and had their performances compared along with the feature sets used. The lexical, syntactical, structural and content feature set combination and SVM gave the best result with 0.945.

Other studies that used cross-genre and cross-topic documents such as [18] used a cross-genre dataset composed of essays and novels of 100 documents. A distance measure was used to calculate how close texts are to one another compared to a set of external documents to determine whether a disputed text was written by a proposed author. The writing style feature extracted was the common word frequency. The pre-processing involved tokenising the text and stemming while keeping punctuation symbols. The evaluation achieved an AUC of 0.738.

The effectiveness of character n-grams is exemplified in a study by [19] who used a cross-topic collection of dialog lines from plays for a single recurrent neural network trained to predict the flow of text by many authors while sharing a collective model of a complete language. The pre-processing involved mapping unknown and known documents into smaller characters, i.e. capital letters to lowercase letters and stemming which received a result of 0.81.

[20] measured Character n-grams, Word n-grams and POST n-grams from lexical, content and syntactical feature sets using the order of the writing style features sequences to model the writing style of an author. The experiment was conducted on a cross-topic dataset made up of a single newspaper. The td-idf was used to represent the writing style features as well as the use of the Logistic regression classifier to create their model. The combination of POST, word and character n-grams achieved a high result of 0.90.

4. METHODOLOGY

In this paper, it was theorised that not all writing style features work well for cross-genre and cross-topic documents in Authorship Identification. This hypothesis was validated by answering the following research questions:

4.1 Research Questions

1. Can writing style features used in single-genre and single-topic Authorship Identification be used effectively on cross-genre and cross-topic Authorship Identification?

2. Which type of writing style features that were effectively used on cross-genre and cross-topic Authorship Identification work best and which cannot be best used?

3. Which writing style features can be combined to work best on cross-genre and cross-topic documents in Authorship Identification?

4.2 Data Set

This paper used a corpus from the Uncovering Plagiarism, Authorship and Social Software Misuse (PAN) at the Conference and Laboratory of the Evaluation Forum (CLEF). PAN CLEF is a forum for digital text research to analyse texts on originality, authorship, and trustworthiness [21]. The dataset used was taken from the 2015 PAN CLEF English collection which consists of 100 sets. Each set contains a known-author document and an unknown-document. The dataset consists of the documents covering different topics and genres such poems, dialog lines from plays and passages from books. The documents comprise of short texts having on average 350 words per document.

4.3 Data Preparation

The texts in the dataset needed to be represented in a way they can be processed to be categorised into writing style features. The experiment followed text pre-processing techniques such as tokenising, normalising and stemming. In the process of Tokenization, some characters like white spaces are discarded. In Normalisation, characters uppercase letters (‘A’) are changed to lowercase letters (‘a’) for text analogy [22]. The Stemming process reduced words to their base form such as "fishing" to "fish".

All the documents in the dataset were processed based on writing style features. A writing style features a numeric value difference between the known and unknown documents indicating whether the documents were written by the same author or not. A known and unknown text most likely written by the same author is represented as a positive value which is over 0.5, otherwise, most likely written by different authors is represented as a negative value which is under 0.5. For instance, in a known text, if the frequency of parts of speech writing style feature count is 50 over the total number of words count is 300, the calculation would be:

$$\frac{\text{frequency of writing style feature}}{\text{total number}} = \frac{\text{frequency of parts of speech}}{\text{total number of words}} = \frac{50}{300} = 0.167$$

In an unknown file, if the calculation for the frequency of parts of speech writing style feature over the total number of words count is 0.78.

$$\frac{\text{frequency of writing style feature}}{\text{total number}} = \frac{\text{frequency of parts of speech}}{\text{total number of words}} = \frac{234}{300} = 0.78$$

The parts of speech writing style feature numeric value difference between the known and unknown text would be:

Highest value – lowest value = 0.78-0.167 = 0.613.

0.613 indicates that the known and unknown documents were most likely written by the same author because 0.613 is over 0.5 which is a positive value.

All the writing style feature calculations between known and unknown text is stored in an excel spreadsheet. The data is then converted from an excel spreadsheet into a Comma-Separated Value (CSV) file. After the data file is converted, the CSV file is then loaded into WEKA. Once the data is loaded, WEKA recognizes the writing style features as attributes. An ablation process was conducted in the experiment where the writing style features were removed from the model to see how their absence affects experimental performance and then put back into the model to see how their presence affects performance. If the removal of a feature increases performance, then it is not good for a model/set. Otherwise, if its removal decreases the performance, it is good for the experiment. Once the feature is measured, it is returned to the model/set so that another feature is modelled in the same way.

4.4 Experimental Setup

The evaluation experiment used WEKA which has classifiers for data mining tasks. The classifiers selected for the experiment are Random Forest, Decision Tree and the Naïve Bayes which builds a probabilistic model. The reason these classifiers were chosen were due to studies such as [23] who used Decision trees (Tree) and Random Forests (RF) in their evaluation experiment for comparing their results. Random Forest classifiers were chosen because they are well-known and popular supervised learning algorithms. The classifier parameters have to be changed to obtain optimal classification accuracy performance.

A cross validation technique was used to find out how well a classifier uses the training data to accurately categorise unknown data. The dataset was divided into a training set made up of 66% of the data while the remaining 34% of the test dataset.

A grid search was used for selecting the values for the parameters that maximize the accuracy of the model. The procedure of a grid search as indicated by [24] was used on Cost and Loss parameters and using the 10-fold cross validation method. The training set and test set are used to find a pair of optimal parameters C and γ (cost and loss) of the RBF Kernel function. The pairs of parameters were tested in intervals step by step as part of the Grid search. The pair is chosen when the error of cross validation is minimal and with a high accuracy cross validation. The ideal Cost and Loss pairs were found to be 1.0 and 1.0 respectively. In the Random Forest classifier, the number of trees and number of randomly parameter pairs that were used were also tried and tested to find an optimum parameter pair. It was found that the different number of trees and number of randomly pairs used in the experiment include {0,1}, {0,2},{1,1},{1,2}and {4,3}.

5. RESULTS AND ANALYSIS

In the empirical evaluation, the Sensitivity (TP), Specificity (TN), Accuracy, ROC (AUC) and Kappa coefficient evaluation measures were used. The Sensitivity (TP) measures the proportion

of actual positives that are correctly identified and Specificity (TN) measures the proportion of actual negatives that are correctly identified. The Accuracy measure approximates how effective a method is by the probability of the true value of a class label. The Kappa coefficient assesses the proportion of agreement between two or more methods for categorical items. The ROC (AUC) determines the ability of a classifier to rank scores appropriately, that is, the proportion of Sensitivity and Specificity.

Table 1 shows all the writing style features identified from related works in their respectable individual feature sets used for the empirical evaluation on the PAN CLEF 2015 English dataset. These initial evaluation results are used as a reference for further experiments to see which writing style features improve performance and which do not.

Table 1. The individual feature sets with all their writing style features.

Feature Set	**Writing Style Features used in the Feature Set**
Lexical	Uppercase frequency Character count Character{Unigram, Bigram, Trigram and Quad-gram} Word length Hapax Legomena Type token ratio
Syntactical	Parts of Speech Tag{Unigram, Bigram, Trigram and Quad-gram} Punctuation {Unigram, Bigram} Function word
Structural	Paragraph frequency Sentence Length
Content	Common words Word {Unigram, Bigram, Trigram}

5.1 Discussion of Research Question 1

The individual feature sets with all the writing style features shown in table 1 were used to generate the initial evaluation results in table 2. Based on the fact that more than 0.5 is positive (likely same author) and less than 0.5 is negative (likely different authors), the experimental results in table 2 show that the writing style features identified from the previous related works used in the experiment produced mostly positive results. This answers research question 1 (Can writing style features used in single genre and single topic documents be used effectively on cross-genre and cross-topic documents for Authorship Identification?). The results show in table 2 the writing style feature sets on the cross-genre and cross-topic dataset showed that the writing style features can be used for a successful Authorship Identification for cross-genre and cross-topic documents.

Table 2: The initial evaluation results of the individual feature sets.

Group	**Classifier**	**Sensitivity**	**Specificity**	**Accuracy %**	**AUC**	**Kappa**
Lexical	Naïve Bayes	0.560	0.620	59	0.666	0.18
	Random Forest	0.389	0.688	53	0.583	0.07
Syntactical	Naïve Bayes	0.780	0.580	68	0.738	0.36
	Random Forest	0.660	0.580	62	0.669	0.24
Structural	Naïve Bayes	0.333	0.625	47	0.483	-0.04
	Random Forest	0.278	0.750	50	0.528	0.03
Content	Naïve Bayes	0.500	0.688	58	0.583	0.18
	Random Forest	0.500	0.620	56	0.538	0.12

5.2 Discussion of Research Question 2

In order to answer research question 2 (Which type of writing style features work best for cross-genre and cross-topic documents and which cannot be best used?), the process of identifying the writing style features for best performance needs an ablation analysis. Recall from section 3.2 that the Data Preparation phase explains the experiment implemented an ablation process that removes and adds back a writing style feature to monitor how it would increase or decrease performance. The ablation process started with the full feature sets with all their writing style features from table 1.

The writing style features that were removed showed to increase performance by their removal meaning that their presence in a feature set brings down the performance result. The writing style features that were removed from Lexical feature set include Type token ratio, Word length, Hapax Legomena and Character Unigram in the Lexical features set. From the Syntactical set include Parts of Speech Tag, Punctuation and Function word. The Word Unigram feature was removed from the Structural set.

Table 3 shows the results of the feature sets with the writing style features that were kept which generated high results after the ablation process. The Syntactical set shows to have the highest results with an AUC of 0.75 answering research question 2 (Which type of writing style features work best for cross-genre and cross-topic documents and which cannot be best used?). The Syntactical writing style features are verified to be ideal for cross-genre and cross-topic document Authorship Identification because of its impressive results. The Syntactical writing style features identified as being ideal are Parts of Speech Tag (unigram, bigram, trigram and quadgram) and Punctuation Bigram. This shows that word-based adjectives help with Authorship Identification because of the number of POST writing style features used in the experiment.

Table 3: The evaluation results of the feature sets after the ablation process.

Group	Classifier	Sensitivity	Specificity	Accuracy %	AUC	Kappa
Lexical	Naïve Bayes	0.857	0.500	66	0.714	0.35
	Random Forest	0.556	0.750	64.7	0.635	0.3
Syntactical	Naïve Bayes	0.800	0.580	69	0.745	0.38
	Random Forest	0.660	0.740	70	**0.750**	0.4
Structural	Naïve Bayes	0.857	0.438	63	0.554	0.29
	Random Forest	0.500	0.625	55	0.646	0.12
Content	Naïve Bayes	0.520	0.740	63	0.634	0.26
	Random Forest	0.560	0.620	59	0.623	0.18

5.3 Discussion of Research Question 3

To answer research question 3 (Which writing style features can be combined to work best for cross-genre and cross-topic document in Authorship Identification?) writing style feature sets were combined to see how they affect the experiment performance, the process is as follows. The feature set with its remaining writing style features that were found to work the best in the experiment performance were merged with another feature set to make a combination feature set, then with another one to make another feature set combination pair. For example, a Lexical set combined with a Syntactical set, then a Lexical set combined with a Structural set. An addition of another feature set was then added to a combination feature set pair until all the feature sets were combined with one another. For example, the Structural set is added to the Lexical and

Syntactical set to make a Lexical, Syntactical and Structural set, the Syntactical set is added to the Lexical and Content set to make a Lexical, Content and Syntactical set, etc.

The combination feature sets that had the highest results had an average AUC of over 0.700. The results show that the Lexical and Syntactical and content set had the highest results with an AUC of 0.762. The other sets that had higher results include Lexical and Syntactical set with 0.751, Lexical, Syntactical and Structural set with 0.760, as well as Syntactical and Content with 0.740. The Lexical writing style features are common in the combination feature sets that performed well in the initial results. The combination feature set had had the lowest results was the Structural and Content with an AUC of 0.519.

Table 4: The initial evaluation results of the combination feature sets.

Feature Group	**Classifier**	**Sensitivity**	**Specificity**	**Accuracy %**	**AUC**	**Kappa**
Lexical and Syntactical	Naïve Bayes	0.780	0.620	70	**0.751**	0.31
	Random Forest	0.700	0.660	68	**0.751**	0.36
Lexical and Structural	Naïve Bayes	0.786	0.625	70	0.710	0.40
	Random Forest	0.500	0.688	58	0.594	0.18
Lexical and Content	Naïve Bayes	0.500	0.760	63	0.700	0.26
	Random Forest	0.520	0.620	56	0.625	0.34
Lexical, Syntactical and Structural	Naïve Bayes	0.760	0.640	70	0.744	0.4
	Random Forest	0.611	0.688	64	0.686	0.29
Lexical, Syntactical and Content	Naïve Bayes	0.780	0.620	70	**0.762**	0.4
	Random Forest	0.611	0.688	64	0.726	0.29
Lexical, Structural and Content	Naïve Bayes	0.667	0.688	67	0.646	0.35
	Random Forest	0.540	0.600	57	0.606	0.14
Lexical, Syntactical, Structural and Content	Naïve Bayes	0.722	0.623	67	0.688	0.35
	Random Forest	0.500	0.813	64	0.722	0.30
Syntactical and Structural	Naïve Bayes	0.833	0.500	67	0.733	0.3
	Random Forest	0.556	0.750	64	0.641	0.3
Syntactical and Content	Naïve Bayes	0.833	0.500	67	0.698	0.34
	Random Forest	0.680	0.720	70	0.712	0.4
Syntactical, Structural and content	Naïve Bayes	0.833	0.563	70	0.712	0.4
	Random Forest	0.571	0.625	60	0.639	0.19
Structural and Content	Naïve Bayes	0.429	0.688	56	0.545	0.12
	Random Forest	0.540	0.560	55	0.519	0.1

An ablation process was also performed on the combination feature sets to see which writing style features work best together to generate higher results in order to answer research question 3 just as it was done for the individual feature sets. The common writing style features that were removed and increased performance results kept performance low with their presence within a feature set. These writing style features include Type token, Hapax legomena, Character unigram, Parts of Speech Tag unigram, and Word unigram and bigram. The common writing style features that were removed from the combination feature sets that showed decreased results include Uppercase frequency, Character trigram and bigram, Punctuation bigram, Parts of Speech Tag

Bigram, Trigram, and quad-gram. These writing style features generate high results because of their presence and were kept in the combination feature sets and were identified as ideal for performance.

Table 5 shows the results of the feature set combination after the ablation process. The combination feature sets that had the highest results was the Lexical, Syntactical, Structural and Content set with an AUC of 0.837. Another feature set that also achieved general high results is the Syntactical and Content set with an AUC of 0.818. These feature set combinations answer research question 3. A combination of writing style features that are character and word based such as Character n-grams, Parts of Speech Tag n-grams, Common word, sentence length and Word n-grams seem to work well in Authorship Identification and generate high performance. All combination feature sets that generated high results had Syntactical writing style features as the individual feature sets had in the evaluation experiment.

Other feature sets include Lexical and Syntactical with an AUC of 0.821 and Lexical, Syntactical and Content set with 0.809. Even though these feature sets that performed well with Syntactical writing style features had Lexical writing style features, they did not have a general overall high result from True Positives, Accuracy and Kappa measures. This demonstrates that the Syntactical feature set is robust in the cross-genre and cross-topic document Authorship Identification process. This result is supported by [25] who found that in Authorship Identification, combining syntax-based (Syntactical) and token-level (Content) features performs almost equally well or even better than only using a Lexical feature set. The combination feature sets that did not have Syntactical writing style features had moderate results such as the Structural and Content set had an AUC of 0.701 and Lexical, Structural and Content set with 0.795. In comparison with previous works, the study's results are among the top five in AUC recall rates as shown in figure 1 which the PAN CLEF AUC results in 2015 [5].

Table 5: The evaluation results of combination features sets after the ablation process.

Feature Group	**Classifier**	**Sensitivity**	**Specificity**	**Accuracy %**	**AUC**	**Kappa**
Lexical and Syntactical	Naïve Bayes	0.778	0.688	74	**0.792**	0.47
	Random Forest	0.833	0.688	76	**0.821**	0.52
Lexical and Structural	Naïve Bayes	0.389	0.875	61	**0.795**	0.25
	Random Forest	0.500	0.750	61	0.625	0.25
Lexical and Content	Naïve Bayes	0.571	0.875	73	**0.799**	0.43
	Random Forest	0.600	0.680	64	0.692	0.28
Lexical, Syntactical and Structural	Naïve Bayes	0.833	0.625	73	0.774	0.46
	Random Forest	0.720	0.740	73	0.759	0.46
Lexical, Syntactical and Content	Naïve Bayes	0.889	0.625	76	**0.809**	0.52
	Random Forest	0.556	0.658	61	**0.781**	0.24
Lexical, Structural and Content	Naïve Bayes	0.556	0.875	70	**0.795**	0.42
	Random Forest	0.786	0.625	70	0.728	0.4
Lexical, Syntactical, Structural and Content	Naïve Bayes	0.889	0.878	88	**0.837**	0.76
	Random Forest	0.556	0.813	67	**0.795**	0.36
Syntactical and Structural	Naïve Bayes	0.833	70	0.778	0.46	0.563
	Random Forest	0.611	64	0.769	0.29	0.688

Syntactical and Content	Naïve Bayes	0.889	0.750	82	**0.792**	0.64
	Random Forest	0.760	0.680	72	**0.818**	0.44
Syntactical, Structural and content	Naïve Bayes	0.840	0.580	71	0.758	0.42
	Random Forest	0.389	0.813	58	0.774	0.19
Structural and Content	Naïve Bayes	0.571	0.813	70	0.701	0.39
	Random Forest	0.643	0.563	60	0.670	0.20

(b) English

Team	FS	AUC	c@1	UP	Runtime
Bagnall [2]	**0.614**	**0.811**	**0.757**	3	21:44:03
Castro-Castro et al. [5]	0.520	0.750	0.694	0	02:07:20
Gutierrez et al. [11]	0.513	0.739	0.694	39	00:37:06
Kocher and Savoy [21]	0.508	0.738	0.689	94	00:00:24
PAN15-ENSEMBLE	0.468	0.786	0.596	0	–
Halvani [13]	0.458	0.762	0.601	25	00:00:21
Moreau et al. [30]	0.453	0.709	0.638	0	24:39:22
Pacheco et al. [33]	0.438	0.763	0.574	2	00:15:01
Hürlimann et al. [14]	0.412	0.648	0.636	5	00:01:46
PAN14-BASELINE-2	0.409	0.639	0.640	0	00:26:19
PAN13-BASELINE	0.404	0.654	0.618	0	00:02:44
Posadas-Durán et al. [36]	0.400	0.680	0.588	0	01:41:50
Maitra et al. [28]	0.347	0.602	0.577	10	15:19:13
Bartoli et al. [3]	0.323	0.578	0.559	3	00:20:33
Gómez-Adorno et al. [10]	0.281	0.530	0.530	0	07:36:58

Figure 1: Evaluation results for authorship identification at PAN 2015 [5]

5.4. Limitations

This paper worked on the PAN CLEF English text collection from the PAN CLEF 2015 collection that consisted of a varying size, diversity, and featured languages, such as Chinese, Persian, and Urdu. The English texts were most appropriate to work on because it was the language the authors understood. The sample size of the PAN CLEF 2015 English collection was small and inadequate for the experiment to generate satisfactory results. Most of the references were within the range 2003 and 2015 because the Authorship Identification task was specific to the PAN CLEF 2015 Authorship Identification task that used the specific PAN CLEF 2015 dataset which was worked on for experimentation of the Authorship Identification task in 2014. Therefore, contemporary work was considered in this paper, however, the experiments may be updated with the contemporary related work for future experimentation. The writing style features used were not described in this paper in detail due to the large number of features that were identified. WEKA was used for the experimental procedure and therefore could not produce set of pseudocodes of the algorithm of the data mining and could not draw a flowchart.

5.5. Recommendations

The questions raised from this study which when answered could produce a greater degree of accuracy on Authorship Identification. For future work, to understand the reason why certain writing style features performed better, a table will need to be drawn showing the description of each feature, how and where they have been used. For future experiments, the ensemble Lexical, Syntactical, Content and Structural group that generated the best results could be used on the PAN CLEF Authorship Identification cross-genre and cross-topic English collection sample size. If the English collection is not enough, then it will need to be made larger by adding more English documents to the dataset. A larger corpus of more than 1000 documents could possibly generate the same prime results in this study. In addition, the corpus used in this study comprised of only

one known document per author and one unknown sample, which speculates whether it would be more effective if the number of known documents per author could bring on a thorough exploration for authorship identification. It would be beneficial to assess the effects of the study findings on other datasets which could be generated from the internet and not exclusively from forums. The writing style features should distinguish authors from one another no matter the topic or genres because how writers express themselves is always unique.

Possible improvements in the experimental setup include a finer grid search in parameter pairing value selection for Cost and Loss for SVM classifier as well as gamma parameters for Random Forest which could acquire better classification accuracy. Experimental setups for future work will use Google Colab for coding the algorithms using Python instead of WEKA to provide pseudocodes for data mining clarity. Future work includes the use of a clustered approach to compare against the classification recall rate this study was using to see which method is more effective. In the clustered approach, the unknown and known text samples will be used as one dataset to find similarity comparisons between texts. As another possible improvement could be an addition of features from related literature to cover over all writing style features as being ideal for authorship identification. Features such as digit frequency, occurrence of special character, feature category such as Idiosyncratic, misspellings as a feature (e.g., “beleive”, “thougth”) can represent an author’s common spelling mistake that they make.

6. CONCLUSION

The paper proposed to identify and evaluate the writing style features to be used in Authorship Identification for cross-topic and cross-genre documents. To the best of the authors’ knowledge, there have been few related works that have evaluated writing style features for Authorship Identification for cross-topic and cross-genre documents. The related works between 2003 and 2015 show that although they used writing style features, there is very little or lack of writing style feature evaluation particularly for cross-topic and cross-genre documents which are novel datasets in the Authorship Identification.

The successful related works were reviewed and extracted the writing style features were extracted to evaluate which writing style features work best for cross-topic and cross-genre Authorship Identification. The writing style features that were commonly used individually and in a combined feature set include Hapax legomena, Uppercase, Character n-grams (1 to 8), word n-grams (1 to 5), sentence length, punctuation, function words, POST, Digit, sentence count, paragraph, common word count, Alphabetic count and type token ratio. The experimental results showed that the Syntactical writing style features had the most successful results as an individual set before and after evaluation process. The Lexical, Structural, Content and Syntactical feature combination set as well as the Syntactical and Content combination feature set produced the highest AUC results with 0.837 using and 0.818 compared to other combination feature sets. The results also showed that all combination feature sets that had general high results had Syntactical writing style features such as Lexical and Syntactical with an AUC of 0.821 and Lexical, Syntactical and Content set with 0.809.

REFERENCES

[1] Stamatatos, Efstathios. "A survey of modern authorship attribution methods." Journal of the American Society for information Science and Technology 60, no. 3 (2009): 538-556.
[2] Juola, Patrick, and Efstathios Stamatatos. "Overview of the Author Identification Task at PAN 2013." CLEF (Working Notes) 1179 (2013).
[3] Castro, Daniel, Yaritza Adame, María Pelaez, and Rafael Muñoz. "Authorship verification, combining linguistic features and different similarity functions." CLEF (Working Notes) (2015).
[4] Sari, Yunita, and Mark Stevenson. "A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification." In CLEF (Working Notes). 2015.

[5] Stamatatos, Efstathios, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. "Overview of the pan/clef 2015 evaluation lab." In International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 518-538. Springer, Cham, 2015.
[6] Mendenhall, Thomas Corwin. "The characteristic curves of composition." Science 214s (1887): 237-246.
[7] Yule, G. Udny. "Reginald Hawthorn Hooker, MA." (1944): 74-77.
[8] Zipf, G.K. Selected Studies of the Principle of Relative Frequency in Language. Cambridge, MA.: Harvard University Press. 1932.
[9] Mosteller, Frederick, and David L. Wallace. Inference and disputed authorship: The Federalist. Stanford Univ Center for the Study, 2007.
[10] Bozkurt, Ilker Nadi, Ozgur Baghoglu, and Erkan Uyar. "Authorship attribution." In 2007 22nd international symposium on computer and information sciences, pp. 1-5. IEEE, 2007.
[11] Nirkhi, Smita, and Rajiv V. Dharaskar. "Comparative study of authorship identification techniques for cyber forensics analysis." arXiv preprint arXiv:1401.6118 (2013).
[12] Sapkota, Upendra, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. "Cross-topic authorship attribution: Will out-of-topic data help?." In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1228-1237. 2014.
[13] Raju, NV Ganapathi, Ch Sadhvi, P. Tejaswini, and Y. Mounica. "Style based authorship attribution on english editorial documents." International Journal of Computer Applications 159, no. 4 (2017): 5-8.
[14] Lou et al. Which Author Authored Which: Predicting Authorship from Text Excerpts. University of Stanford. Los Angeles. 2017
[15] Kešelj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. "N-gram-based author profiles for authorship attribution." In Proceedings of the conference pacific association for computational linguistics, PACLING, vol. 3, pp. 255-264. 2003.
[16] Moreau, Erwan, and Carl Vogel. "Style-based Distance Features for Author Verification-Notebook for PAN at CLEF 2013." In CLEF 2013 Evaluation Labs and Workshop-Working Notes Papers, pp. Online-proceedings. 2013.
[17] Suh, Jong Hwan. "Comparing writing style feature-based classification methods for estimating user reputations in social media." SpringerPlus 5, no. 1 (2016): 1-27.
[18] Kocher, Mirco, and Jacques Savoy. "UniNE at CLEF 2015: author identification." Working notes papers of the CLEF (2015).
[19] Bagnall, Douglas. "Author identification using multi-headed recurrent neural networks." arXiv preprint arXiv:1506.04891 (2015).
[20] Gómez-Adorno, Helena, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. "Document embeddings learned on various types of n-grams for cross-topic authorship attribution." Computing 100, no. 7 (2018): 741-756.
[21] Rosso, Paolo, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. "Overview of PAN'16." In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 332-350. Springer, Cham, 2016.
[22] Howedi, Fatma, and Masnizah Mohd. "Text classification for authorship attribution using Naive Bayes classifier with limited training data." Computer Engineering and Intelligent Systems 5, no. 4 (2014): 48- 56.
[23] Bartoli, Alberto, Alex Dagri, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. "An author verification approach based on differential features." In Conference and Labs of the Evaluation forum, vol. 1391. CEUR, 2015.
[24] Li, J., K. Hsu, A. AghaKouchak, and S. Sorooshian. "An object-based approach for verification of precipitation estimation." *International Journal of Remote Sensing* 36, no. 2 (2015): 513-529.
[25] Luyckx, Kim, and Walter Daelemans. "Shallow text analysis and machine learning for authorship attribtion." LOT Occasional Series 4 (2005): 149-160.

AUTHORS

Simisani Ndaba graduated with her Masters of Science in Computer Information Systems where her research work was based on Machine Learning. She also holds a Bachelor's degree in Business Information Systems and a Post Graduate Diploma in Education in Computer Science. Her research interests are in Data Science and Machine Learning.

Dr Edwin Thuma has a broad background in Computing Science with specific expertise in Information Retrieval (the science of search engines) and Big Data Systems. In particular, his research has been focused primarily on the development of search engines tailored to support health professionals and laypeople when searching for health content on the web. Recently he has started working on search engines that are tailored to support legal professionals when searching for precedent cases or statutes that support the current case.

Gontlafetse Mosweunyane is a lecturer at the Department of Computer Science in the University of Botswana. She received a PhD in Knowledge Organisation and Access from the University of South Hampton. She obtained her Master of Science in Computing Science from The University of Manchester and her Bachelors in Computer Science from University of Botswana. Her research interests are in Information Retrieval and Databases.