STUDENT NAME: Simisani Ndaba

STUDENT NUMBER: 200600916

COURSE NAME: Masters in Computer Information Systems

DEPARTMENT: Computer Science

SUPERVISOR: Dr. E. Thuma

SECOND SUPERVISOR: Dr. G. Mosweunyane

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

ACKNOWLEDGEMENTS

I would like to thank my dissertation supervisors Dr.E. Thuma and Dr. G Mosweunyane from the Department of Computer Science at the University of Botswana for their commitment to me during the course of my research work. I could not ask for more committed supervisors. I would also like to extend my gratitude to the Graduate Coordinator Mr. S.D. Asare, Head of Computer Science Dr. A.N. Masizana-Katongo of the Department of Computer Science and the School of Graduate Studies for allowing me to pursue my work and part time work at the University of Botswana.

I must express my very profound gratitude to my family for providing me with their unfailing assistance throughout my years of study and through the process of researching and writing this dissertation. The completion of my schooling would not have been possible without them.

ABSTRACT

This dissertation describes an evaluation of writing style features for cross-topic and cross-genre documents in Authorship Identification. The study sets out to investigate this by extracting writing style features from related works and evaluates which writing style features work best for cross-topic and cross-genre documents by using an ablation process. The ablation process demonstrates that writing style features increase or decrease performance with their removal from or addition to a classification model. This study uses the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology as it provides a structured approach. The classification techniques used include Naïve Bayes, Support Vector Machine and Random Forest, which were chosen because evidence from previous studies suggest that they generally perform well in a variety of tasks.

The study first investigates whether the writing style features used in successful related works that had single-topic and single-genre documents can be used effectively on cross-genre and cross-topic documents for Authorship Identification. The evaluation results showed that the writing style features used in single-topic and single-genre Authorship Identification can be used in cross-genre and cross-topic Authorship Identification because they performed reasonably well when used in the classification model. In addition, the study investigated which type of writing style features work ideally for cross-genre and cross-topic in Authorship Identification. The Syntactical writing style features that were identified as being ideal were; ***Parts of Speech Tag (POST) unigram, bigram, trigram*** and ***quad-gram*** and ***Punctuation Bigram***. This shows that word-based adjectives have a positive contribution in Authorship Identification performance.

Furthermore, the study continued to find out which writing style features can be combined to work best on cross-genre and cross-topic documents in Authorship Identification. It was found that the best combination of feature set that showed to be used in cross-genre and cross-topic documents for Authorship Identification with high results was the **Lexical, Syntactical, Structural** and **Content** feature combination set. This shows that a combination of adjectives (Content), layout (Structural) and character-word collocations (Lexical, Syntactical) features attributes to a successful cross-genre and cross-topic document Authorship Identification.

Finally, the study also set out to find out whether the results from this study generalise across the three different family of classifiers. The results generally showed that regardless of the classifier used, most of the highest results were generated from Syntactical set, then secondly Lexical, then Content followed by Structural set. This generalisation is the same as the initial evaluation and after the ablation process. When the feature set are combined, the Syntactical and Lexical feature set generated the highest results. The combination of features that had mostly Content features performed moderately, and the combination features that had mostly Structural sets had the lowest results across the classifiers. The study achieved its highest result score of 0.837 from the **Lexical, Syntactical, Structural** and **Content** feature set.

Table of Content

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ARFF	Attribute Related File Format
AUC	Area Under the Curve
CLEF	Conference and laboratory of the evaluation forum
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma Separated Value
IDF	Inverted document frequency
FN	False Negative
FP	False Positive
NLP	Natural Language Processing
PAN	Uncovering Plagiarism, Authorship and Social Software Misuse
POST	Parts of Speech Tag
RBF	Radical Base Function
RF	Random Forest
ROC	Receiver operating characteristic
SVM	Support Vector Machine
TN	True Negative
TF	Term Frequency
TF-IDF	Term Frequency-Inverted Document Frequency
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis
XML	Extensible Mark-up Language

1 INTRODUCTION

To determine a writer of an anonymous text has been of interest in many areas since the nineteenth century, (Stamatatos, 2009). 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 (Juola and Stamatatos., 2013). Castro et al., (2015) cites many practical examples where knowing the author of a document may be very important. For example, finding an author of a malicious mail sent from an anonymous email account, plagiarism detection and to catch paedophiles by tracing conversations through topics and lines of conversations which are sexual in nature (Inches and Crestani, 2012). Other instances include, spam filtering and linking terrorist proclamations to their writers. 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 as they are fundamental in identifying writers of texts. Authorship Identification defines a specific character of an author and finds the differences between documents (Coyotl-Morales et al., 2006). 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. For example, a Syntactical feature set is characterized by the concatenation and frequency of certain words and characters. The measure of these words and characters within a text is compared to another text. If the difference is low, the texts are likely to be written by the same person, otherwise, there are likely written by different people.

The problem is complicated by the fact that an author may consciously or unconsciously vary his or her writing style from text to text (Sari and Steven, 2015). The writing style of an author is affected by the genre in addition to the personal style of an author. It is also 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 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 (Stamatatos et al., 2015). 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.

This dissertation sets out to identify the ideal writing style features for cross-genre and cross-topic documents in Authorship Identification. Sari and Steven, (2015) explain that the genre/topic between known and unknown documents differ significantly. The dissertation plans on using the writing style features in previous successful Authorship Identification studies in a model to perform on cross-genre and cross-topic documents. Three different classifiers will be used in the empirical evaluation to see whether the results are generalised well across the different family of classifiers for cross-genre and cross-topic documents Authorship Identification tasks.

1.2 Motivation

This dissertation was inspired by the Uncovering Plagiarism, Authorship and Social Software Misuse (PAN) at the Conference and Laboratory of the Evaluation Forum (CLEF). PAN is a forum for the digital text forensics, where researchers and practitioners study technologies that analyse texts with regard to originality, authorship, and trustworthiness, (Rosso et al., 2016). It focuses on the evaluation of selected tasks from digital text forensics in order to develop and assess the latest large scale techniques. It presents three tasks through which important variations of problems are studied. The tasks are explained as follows:

Plagiarism detection, is divided into source retrieval and text alignment. Source retrieval searches for most probable sources of a suspicious document. Text alignment matches passages of reused text between a pair of documents.

Authorship identification, focuses on answering the question on whether an unknown document is written by a particular author or not. This task emulates real world problems that most forensic linguists face every day (Stamatatos et al., 2015).

Authorship profiling, is concerned with predicting an author's demographics from their writing. For example, an author's writing style may reveal the age, gender, and personality.

1.2.1 The Evolution of Authorship Identification at PAN CLEF

The advancement of the Authorship Identification task at PAN CLEF has been highlighted by Argamon and Juola (2011) who reported the two variations of the Authorship Identification task that were explored which are Authorship Attribution and Authorship Verification. Authorship Attribution refers to determining which of a known set of authors wrote a text, and Authorship Verification is determining if a specific author did or did not write a text. Stamatatos and Juola (2013) outline that in 2011, PAN focused on a dataset consisting of single genre documents extracted from the Enron dataset and the PAN 2012 dataset was made up of fictional documents from Feedbooks.com site. Thereafter, the Authorship Identification task focused on the author attribution sub-task in 2012 and in 2013, the focus changed to Authorship Verification sub-task. The 2013 dataset incorporated a substantial multilingual element, including English, Spanish and Greek natural languages (Stamatatos and Juola, 2013).

In 2014 as compared to 2013, a larger dataset was built comprising over a hundred documents in each of the four languages; English, Spanish, Greek and Dutch. Four genres; reviews, novels, essays and opinion articles were also included. Eventually, Stamatatos et al., (2015) points out that in contrast to the authorship identification task evaluation setup in 2013 and 2014, as well as previous work after 2015, it is not assumed that all documents match in genre and topic. Instead, documents were considered as cross-topic and cross-genre documents. A new dataset was built, covering the four languages Dutch, English, Greek, and Spanish and comprising a variety of genres and topics. Subsequently, in 2016, the task focused on author clustering and author diarization (also known as intrinsic plagiarism detection). Both subtasks are concerned with measuring author's writing style similarity within texts.

1.3 Dissertation Statement

The statement of this dissertation is that not all writing style features work well for cross-genre and cross-topic documents Authorship Identification. This statement will be validated through work which seeks to answer the following research questions:

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?

2. Which type of writing style features work best for cross-genre and cross-topic documents 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. Do the results from this study generalise across the three different family of classifiers?

1.4 Dissertation Outline

The rest of the dissertation is organised as follows,

Chapter 2 surveys the background of processes and the different techniques used in learning methods. It also provides an overview of evaluation measures.

Chapter 3 discusses the background and evolution of writing style features particularly used in Authorship Identification. The chapter also reviews previous works that have used writing style features in their research capacities and their findings.

Chapter 4 reviews the CRISP-DM methodology used in the study and the arrangement of the study following the CRISP-DM process.

Chapter 5 presents the experimental results generated from the evaluation experiment of the writing style features, combination writing style feature sets and comparison of previous work findings.

Chapter 6 summarises the dissertation, discusses the conclusions from the evaluation experiment and the findings from the dissertation which inspire future research work.

2 BACKGROUND

2.1 Introduction

In this section, a description of techniques that are used in the dissertation is provided, in particular, the descriptions of different classification techniques in section 2.2. In section 2.3 shows a description of the different pre-processing techniques used to prepare the data for classification. In section 2.4, evaluation method measurements used in the dissertation are explained in detail.

2.2 Classification Learning

According to Tan et al., (2006), classification learning is a systematic approach that applies a learning algorithm to create a model to accurately predict the class label (category) of an input data based on predefined features (attributes). This technique is used in different applications, for example, recognising terms in spamming email messages, feature pattern recognition in Bioinformatics and face detection. Figure 1 shows a model flow of a classification learning model. The training data is the collection of records which makes up the input data in a classification task and where the predefined features are extracted that make up a particular class. The machine learning algorithm creates a predictive model (classifier) from the training data features that is then applied to new data/undefined class data to categorise it to a predefined class.

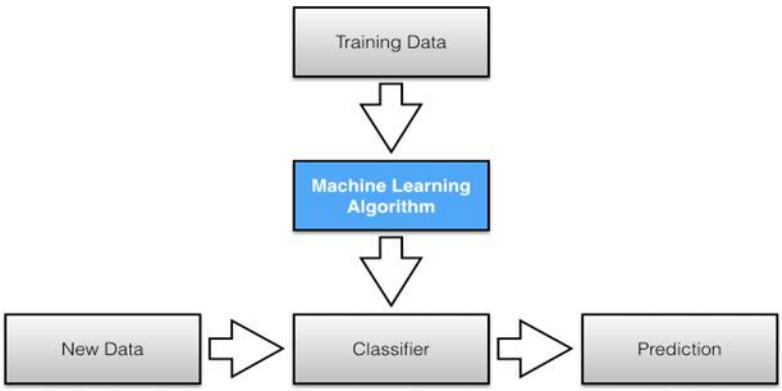


Figure 1: A model flow of a classification learning model, (Raschka, 2014).

It is important that the type of problem and type of data investigated dictate which classification technique should be chosen. The different classification techniques are compared based on particular datasets used for a task, prediction performances as well as their computational

efficiency, (Raschka, 2013). Examples of these techniques include Tree (Rule) based, Kernel based and Probabilistic based methods.

2.2.1 Tree Based Method

Tree based methods are a hierarchical way of partitioning the data which builds classification and regression trees (CARTS) for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The data is repeatedly divided into smaller regions until the end where every region is assigned with a class label. The characteristics of the data are modelled as a tree structure. Tree based methods classify instances by sorting them from the root to the node, which provides the classification of an instance. The Random Forest classifier is an example of a tree based method.

Random Forest is a scheme proposed by Breiman (2001) in the 2000's as a predictor with a set of decision trees that grow in randomly selected subspaces of data. Polamuri (2017) outlines the way Random Forest works as described in the following steps:

1. ***K*** features are randomly selected from a total ***m*** features where ***k < m,***
2. Among the ***K*** features, the node ***d*** is created from the data and is calculated using the best split feature,
3. The node is split into more **nodes** using the **best feature,**
4. Step 1 to 3 is repeated until l number of nodes has been reached
5. A forest is built by repeating steps **1 to 4** for ***n*** number times to create ***n*** number of trees.

To calculate the class prediction:

1. The method takes the **test features** and uses the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
2. The **votes** for each predicted target are calculated.
3. The **highest voted** predicted target is considered as the **final prediction** from the random forest algorithm. This concept of voting is known as **majority voting**. Figure 2 illustrates this Random Forest process.

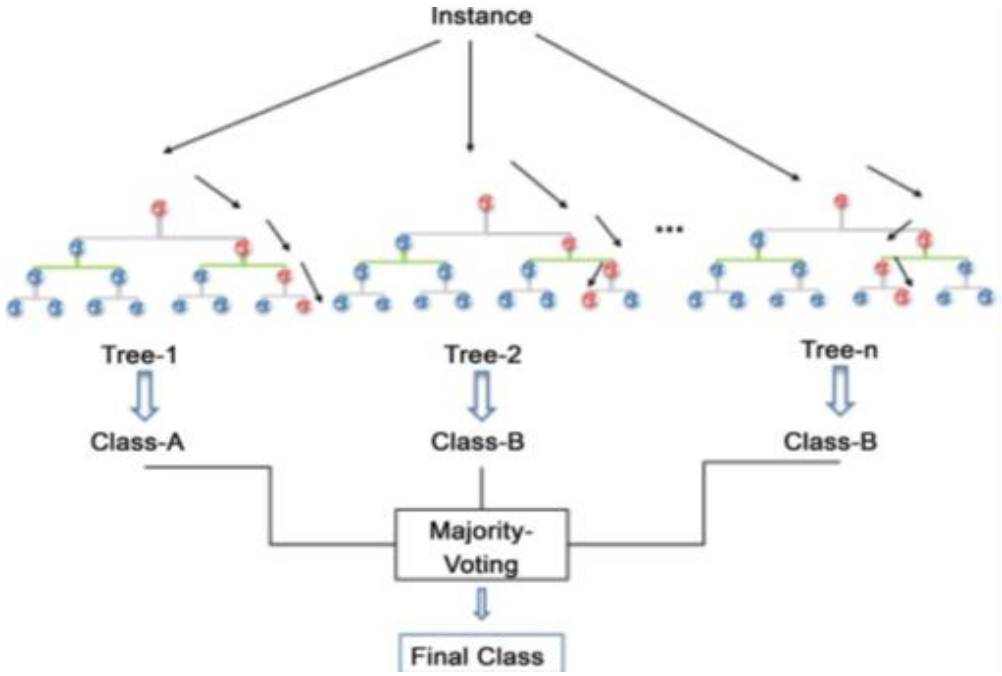


Figure 2: Demonstration of the Random Forest methodology, (Fu, 2017).

2.2.2 Kernel Based Methods

Kernel based methods are a class of pattern analysis. The goal of Kernel based methods is to map input data to feature spaces. The mapping to the new space is defined by a function called the Kernel function. Due to its effective generalization performance, kernel methods have been widely used in many applications but may not always be the most efficient technique, (Agarwal, 2007). The best known kernel based system is the Support Vector Machine (SVM).

The SVM classifier is a binary classifier where the output of learned function is either a positive or negative value ranking. Binary SVMs are classifiers which discriminate data into two labels. Each data object (data) is represented by an n-dimensional vector. Each of these data points belongs to only one of two classes. A linear classifier separates them with a hyperplane (a line that splits the input variable space). There are many linear classifiers that correctly classify the two groups of data using separating hyperplanes (L1, L2 and L3) in a dimensional space depicted in figure 3.

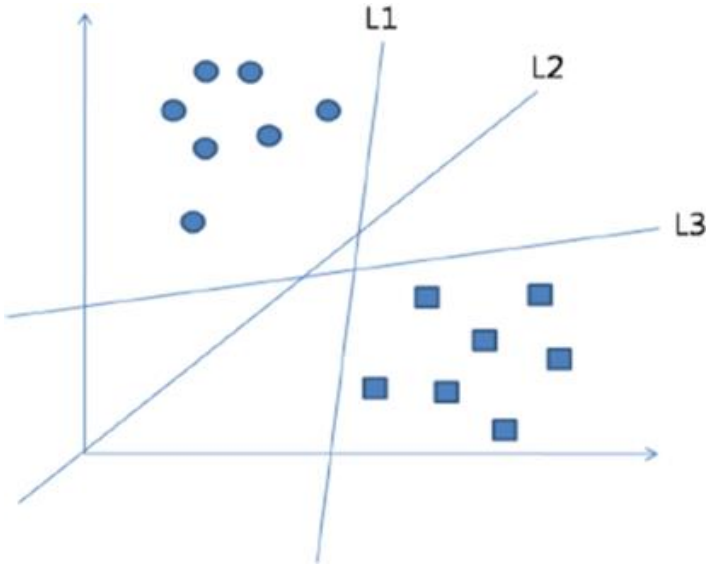


Figure 3: Support Vector Machine differentiating two groups of data, (Yu and Kim, 2012)

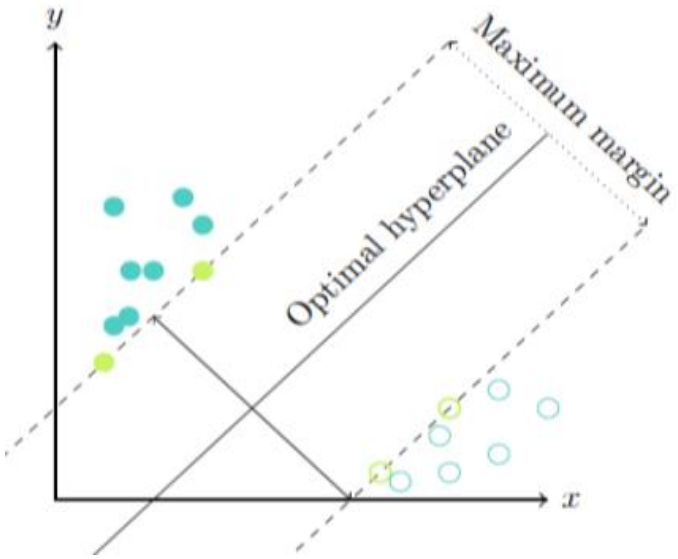


Figure 4: An example of the maximal margin of the hyperplane, (Skoglund, 2015).

According to Skoglund (2015), to achieve a maximum separation between the two categories, a hyperplane is picked by the SVM to determine which has the largest margin. A margin is the summation of the shortest distance from the separating hyperplane to the nearest data point of both labels (categories) as seen in figure 4 that shows an example of the maximum hyperplane. The hyperplane that has the largest distance is likely to generalize better, meaning that the hyperplane correctly classifies "unseen" or testing data points. SVMs do the mapping from input space to

feature space to support nonlinear classification problems. The object rearranging process is known as mapping (transformation) illustrated in figure 5.

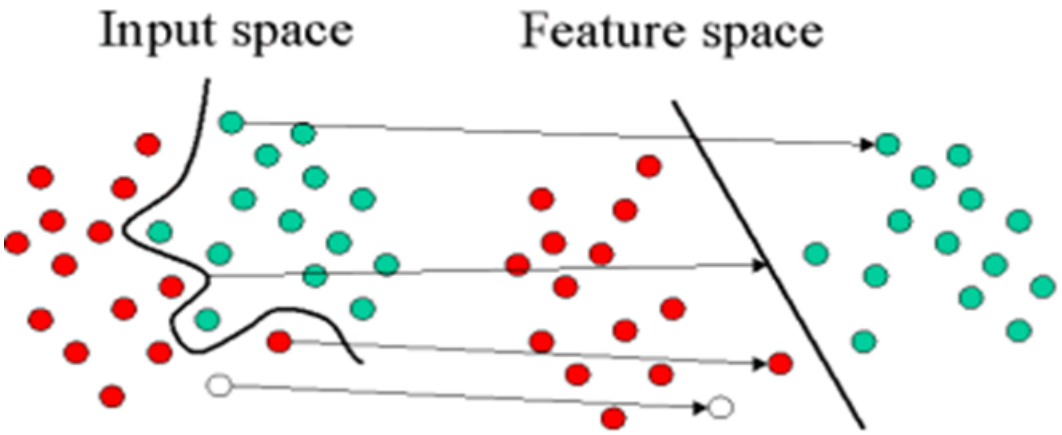


Figure 5: Support Vector Machine object rearranging process from input space to feature space, (Hill and Lewicki, 2007).

The SVM as described only separates two classes. But in many situations, like Author identification, it has to distinguish between more classes. This can be performed using pair wise classification. This classification method constructs classifiers for each pair of classes, while ignoring the data that does not belong to one of these two classes. So, for C classes C(C-1)/2, binary classifiers need to be constructed. The unseen data sample gets the class label that is predicted most by the classifiers, (Fissette, 2010).

2.2.3 Probabilistic Methods

Probabilistic methods use data from past events into unknown situations by assuming that previous existing trends will continue in future events. The most commonly used Probabilistic classifier is the Naïve Bayes.

The Naive Bayes classifier uses the Bayes Theorem. It predicts probabilities for each class such as the probability that a given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. The Bayes Theorem works on Conditional probability, which is the probability that something will happen, given that something

else has already occurred. Using the conditional probability, the probability of an event using its prior knowledge can be calculated.

2.3 Data Representation

Data representation is a critical step in order to identify features from data, they need to be represented in a way they can be processed in learning methods and categorised. Typically, for large scale applications, how to learn the structure of data and discover valuable information from data becomes continuously more urgent, important and challenging. The processes consist of the following in order;

- **Pre-processing** is the process of amending or removing data in a text that is incomplete, improperly formatted, or duplicated to prepare for further processing.

- **Feature Extraction** refers to identifying terms (features) and assigning a numeric value to them. They will be used for categorising text input different classes.

2.3.1 Pre-Processing

In order to help improve the quality of the data and results, raw data is pre-processed so as to improve the efficiency and ease of the recall process. The following pre-processing techniques used in document pre-processing are in no particular order;

2.3.1.1 Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization,

- Some characters like *punctuation marks* are discarded.

- Tokens or words are separated by whitespace, punctuation marks or line breaks.

- White space or punctuation marks may or may not be included depending on the need. All characters within contiguous strings are part of the token.

- Tokens can be made up of all alpha characters, alphanumeric characters or numeric characters only.

2.3.1.2 Stop (Function) word removal; Stop words are frequently used words in a text which are language-specific functional words that join the flow of a sentence and carry no information (i.e., prepositions, conjunctions, pronouns, prepositions). There are about 400-500 Stop words in the English language, examples of such words include 'the', 'of', 'and', 'to'. This is usually the first step in data pre-processing, depending on a particular writing style feature measure (Gaigole et al., 2013).

2.3.1.3 Stemming (Lemmatisation); A report on text classification by Elayidom et al., (2013) defines Stemming as the process of reducing words to their root or base form known as a stem. A stem may not be the same as its base form, but it is enough that related words map to the same stem, even if that stem is not a convincing base form. For example, Stemming reduces the words "fishing", "fished", "fishes", and "fisher" to the root or base form of the word that is the word "fish".

2.3.1.4 Normalization; In Normalisation, all terms in texts are converted to the same form for more accurate reflect of terms used so that matches occur despite superficial differences. For example, uppercase letters (‘A’) are changed to lowercase letters (‘a’) for text analogy (Howedi and Mohd, 2014).

2.3.1.5 Punctuation mark removal; Howedi and Mohd (2014) explain all punctuation marks (e.g. \:;.,‘"!?) are removed from the texts of each document by replacing all these punctuation marks with an empty string. However, in *character-level writing style features*, these punctuation marks are considered. This is because punctuation marks can represent an author’s writing style. For instance, while some authors rarely use exclamation marks, some use other distinct exclamation marks in more cases. Some authors may use *full stops* frequently because they like short sentences while others use *commas* more frequently by using long sentences in their writing.

2.3.2 Feature Extraction

This is the process of using features from data which exhibit the characteristic and understand the peculiarity of a class (Howedi and Mohd, 2014). Ozgür (2004) explains these features can be extracted by using Natural Language Processing (NLP) or Statistical techniques by considering

their frequency appearing in data. The features are represented using Vector space modelling which reflect their frequency in data.

- **The Term Document-Inverse Document Frequency (TD-IDF)** is an example of a statistical method that is intended to reflect how important a term (feature) is to a document as well as in a dataset. The term frequency (*td*) of a term (*t*) in a document (d) is defined as the number of times that *t* occurs in *d*. The document frequency (df_t), the number of documents that *t* occurs in. The df_t is an inverse measure of the in formativeness of a term *t*. The *tf-idf* weight of a term is the product of its *tf* weight and its *idf* weight shown in a formula below;

$$w_{t,d} = (1 + \log tf_{t,d}) * \log (N/ df_t)$$

2.4 Evaluation Methods

2.4.1 Cross Validation

Cross validation is a partitioning technique on datasets used for predicting statistical analysis. Multiple classification studies such as Howedi and Mohd (2014) explain that applying a K-fold cross validation technique provides a more meaningful result by dividing the data into training and testing data.

In cross validation, a *k* number of equally sized parts, such as 3 or 10 which are randomly created from the dataset. In a 'Leave-out-one validation' which is often used, most of the data is used as training data but 'one-fold' is left as testing data. The procedure is repeated until each fold is held out for testing. This process ensures that all data is used for both training, testing and to ensure that there is no overlapping between them. Therefore, the classification task is performed *n* times, each time, a different partition is used as testing data. The remaining two partitions are used as a training set. Therefore, each partition is used once as test data. The results of these *k* classification tasks are then combined for calculating the average results for the dataset. This method reduces the variability of the classification.

Figure 6 demonstrates an example of a 10-fold cross-validation setup. 90% of the data is used for training and the rest of the 10% is the testing set for one-fold. This operation is repeated 10 times

with mutually exclusive training sets from other folds. In the 10-fold cross-validation setting, there are 10 different models based on the 10 different folds.

Figure 6: A 10-fold cross validation process, (Raschka, 2015).

Figure 7 demonstrates that subsequently, *n* iterations of training and validation are performed such that within an iteration, a different fold of the data is held-out for validation while the remaining k-1 folds are used for learning.

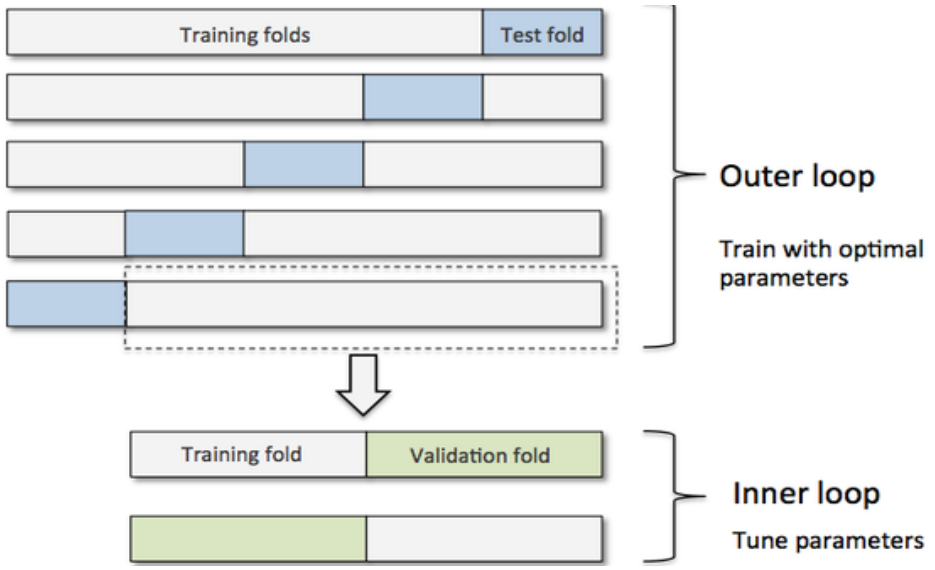


Figure 7: An Illustration of how iteration is used for training and validation in cross validation, (Raschka, 2015).

The measures of the quality of classification are built from a confusion matrix which is used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known (Sokolova et al., 2006). Table 1 shows a confusion matrix for classification followed by definitions that make up the confusion matrix below.

- True positive (TP) is the number of classifications that the classifier classified correctly as the answer.
- False positive (FP) happens when an incorrect instance is classified as correct.
- True negative (TN) is the number of classifications that the classifier correctly predicted not to be the answer.
- False negative (FN) occurs when a correct instance is classified as incorrect.

Table *1*: A Confusion Matrix, (Sokolova et al., 2006).

	Actual Class		
		True	**False**
Predicted Class	Positive	True Positives (TP)	False Positives (FP)
	Negative	True Negatives (TN)	False Negatives (FN)

2.4.2 Sensitivity (True Positive)

Sensitivity (Recall) is defined as the proportion of actual positives that are correctly identified. Sensitivity is calculated as follows;

$$\text{Sensitivity} = (TP/TP + FN)$$

2.4.3 Specificity (True Negative)

Specificity measures the proportion of actual negatives that are correctly identified. Specificity is calculated as follows;

$$\text{Specificity} = (TN/\ TN + FP)$$

2.4.4 Accuracy

Sokolova et al., (2006) say Accuracy approximates how effective a classifier is by showing the probability of the true value of the class label. In other words, it assesses the overall effectiveness of the classifier. The Accuracy formula is shown as follows;

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

2.4.5 F_1-score

According to Yedidia (2016), the F_1-score is a statistical method for determining accuracy average accounting for both precision and recall. It considers the precision and the recall of a test to compute the score. The F_1-score is the weighted average of the precision and recall, where an F_1-score reaches its best value at 1 (perfect precision and recall) and worst at 0. The F_1-score formula is shown below as follows;

$$F_1\text{-score} = (2 * (\text{Precision} * \text{Recall})) / (\text{Precision} + \text{Recall})$$

2.4.6 Kappa Coefficient

Xier (2010) explains the Kappa coefficient is a statistic used in assessing categorical agreement between two raters or two methods. It can also be extended to more than two methods. Equation below shows how to calculate the Kappa value;

- P(a) is the agreement between the classifier (predicted class) and the actual class.
- P(e) is the chance agreement.

$$K = Pr(a) - Pr(e)/ 1 - Pr(e)$$

A score between -1 and 1 is achievable when calculating Kappa, -1 equals perfect disagreement, 0 indicates that instances are classified by chance and 1 equals perfect agreement. A Kappa score between 0 and 0.20 is considered a poor agreement. A score between 0.20 and 0.40 is considered a fair agreement. A score between 0.40 and 0.60 is considered a moderate agreement. A score between 0.60 and 0.80 is a good agreement and between 0.8 and 1 is an excellent agreement.

2.4.7 Area under the Receiver Operating Characteristic Curve ROC (AUC)

The ROC (AUC) tests the ability of classification methods to rank scores appropriately, assigning low values ($0 < 0.5$) to negative answers and high values ($0.5 < 1$) to positive answers. The ROC AUC calculates a score threshold based on the proportion of false positives (FP) or true positives (TP) generated by a model. The use of the ROC (AUC) graphical tool is supported by Bradley (1997) who maintains that it is a good way of visualising a classifier's performance in order to select an acceptable decision threshold, or an operating point.

2.4.8 C@1

C@1 is referred to a question answering task measure which explicitly extends accuracy based on the number of problems left unanswered, (Stamatatos et al., 2014). A score greater than 0.5 is considered a positive answer and a score lower than 0.5 is considered as a negative answer. While the scores equal to 0.5 correspond to unanswered problems. The c@1 formula is defined as follows:

$$C@1 = 1/n\ (nc + (nc/n) * nu),$$

Where *nu* is the number of problems left unanswered,
nc is the number of correct answers and
n is the number of problems.

2.5 Summary

In this chapter, descriptions of the different classification techniques used in this dissertation were looked at, namely, Tree methods (Random Forest), Kernel methods (SVM) and Probabilistic methods (Naïve Bayes). These classifiers will be deployed in chapter 4 and 5. This chapter also covered the data representation stage which involves data pre-processing such as Stemming, Tokenising, and feature extraction with each having processes needed to qualify data, which this dissertation will be using. Lastly, the chapter covered different evaluation performance measures of classifiers such as Accuracy, ROC (AUC), Kappa coefficient, Sensitivity and Specificity. The evaluation measures will be deployed in chapter 5.

3 RELATED WORK

3.1 Introduction

Authorship Identification goes as far back as the nineteenth century with the preliminary study of Mendenhall (1887) on the plays of Shakespeare. This was followed by statistical studies in the first half of the twentieth century by Yule (1938; 1944) and Zipf (1932). Subsequently, a detailed study by Mosteller and Wallace (1964) on the authorship of the 'Federalist Papers' was the most influential work in Authorship Identification. The Federalist Papers consisted of a series of 146 political essays written by John Jay, Alexander Hamilton, and James Madison, twelve of which claimed by both Hamilton and Madison. Mosteller and Wallaces' (1964) 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 between the candidate authors.

Stamatatos (2009) explains that in the late 1990s, research in Authorship Identification was dominated by attempts to define writing style features in the line of research known as Stylometry. Bozkurt et al., (2007) refers to Stylometry as the statistical analysis of a style and is based on the assumption 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. Rudman (1998) estimated that nearly a thousand different features had been proposed since the 1990s.

Nirkhi and Dharaskar (2013) report that to extract unique writing style from data, writing style features such as *Lexical, Syntactic, Structure* and *Content-specific* feature sets are needed to be considered. The number and types of features sets used in an Authorship Identification task produces different results. Table 2 shows examples of writing style features used in previous related work in their respective writing style feature sets and description categorised by Abbasi and Chen (2008).

Table *2*: A description of writing style features, (Abbasi and Chen, 2008).

Feature Set	**Writing Style Feature**	**Description**
Lexical	Word level	Total words; average word length; and number of short words.
	Character count (level)	Total characters; percentage of digits; percentage of uppercase letters
	n-character gram	For example, count all character *n*-grams, with n either being (2, 3,4,5,7 grams). For example, (a, ab, abc, abcde.,e.t.c)
	Letter level	Letter frequency, count of letters (i.e., a, b, c)
	prefixes	The prefix is the first character or word of a sentence. For example, Token k-prefixes, token k-prefix n-grams and word prefixes of size 2
	suffixes	The suffix is the last character or word of a sentence. For example; Token k-suffix n-grams, token k-suffixes and word suffixes of size 2.
	Word Length	frequency of 1–20 letter word
	Special Characters	i.e., (%$@#^&*)
	Vocabulary Richness	Richness (e.g., hapax legomena (words that occur only once), Dis Legomenon (number of words that appear twice) Yule K and Lexical density)
Syntactical	Function Words	frequency of function words (e.g., of, for)
	Parts of Speech Tags (POST)	Frequency of Parts of speech tag (e g. noun, adjectives, adverbs, verbs). POST n-grams (3,5,7 grams).
	Punctuation	Frequency of colon, semicolon, question mark, period, exclamation and comma. Punctuation n-grams.
	Common words	Frequency of words most used within a document.
Structural	Sentence length	Frequency of sentences in a document

	Paragraph frequency	Frequency of paragraphs in a document
Content	Word Trigrams	Word Trigrams — varies word trigrams (e.g., "editor
	Word (unigram)	varies bag-of-words (e.g., "senior", "editor")
	Word Bigrams	Word Bigrams — varies word bigrams (e.g. "senior
Ensemble Feature sets	Vocabulary richness, Common words, sentence length and misspelling words	Combination of Lexical, Structural, Syntactical and Content and Idiosyntractic feature sets

The initial Authorship Identification studies used datasets only comprising of a **single genre/topic**. The comparison of single genre/topic data was efficiently due to the attachment that words and expressions belong to a particular domain. Several datasets are available for Authorship Identification tasks which makes it easy to compare writers' writing styles from their sample texts. According to Skoglund (2015), in previous studies, due to the increasing number of text published online, popular genre for Authorship Identification have been emails, newsgroup messages and forum messages. We will see in section 3.2, a review of the Authorship Identification studies that have used same topic and same genre documents for their experiments.

Cross-topic and Cross-genre datasets contain documents from a number of authors from different domains (different topics, different genres). A typical dataset contains text samples from a variety of authors in a variety of genres such as Emails, Essays, Discussions and various topics such as, Catholic Church, Marriage, and Discrimination, (Sapkota et al., 2014). Many topics are needed to be able to test cross-topic performance and cross-genres to ensure that study findings are robust across different styles of text. The lack of datasets that should consist of texts in multiple topics and genres by the same authors is one of the biggest challenges in the field of Authorship Identification. Few attempts have been made addressing this problem, however, the most prominent attempt has been made by PAN, (Stamatatos et al., 2015). PAN does not only study the three tasks recall from section 1.2, but also provides publicly accessible cross-topic and cross-

genre datasets, (Halvani et al., 2015). A review of the Authorship Identification studies that have used cross-topic and cross-genre documents for their experiments will be seen in section 3.3.

3.2 Same topic and Same Genre

3.2.1 Lexical Features (Token-based)

3.2.1.1 Vocabulary richness

Raju et al., (2017) performed their study by using mostly *vocabulary richness measures* such as ***type token ratio***, ***hapax legomena Yule's K, Simpson's D, Sichel's S*** and ***Honore's R*** were. The other minor writing style features used include, ***number of letters***, ***number of uppercase characters***, ***digits, number of white spaces***, ***character n-grams (2 to 4) and function words***. These writing style features were considered from related studies and their dataset consisted of English editorial documents. All writing styles features, except for the vocabulary richness measures, from each document were represented by a vector using the td-idf technique. Three different classifiers namely, Naive Bayes, Support Vector Machines and Multilayer Perceptron were used to build different classification models using their default parameters. The highest average accuracy they achieved was 97.22 with SVM.

In another approach, Lou et al., (2017) used ***number of unique words***, ***Hapax Legomenon*** and ***Dis Legomenon*** in their study. They also used minor stylometric writing style features namely, ***word n-grams***, ***sentence length, word length***, and ***frequencies of punctuation.*** The single genre dataset used consisted of novels from four writers from roughly the same era. The top 100 most frequently used individual words frequency were calculated from each document. The ***word n-grams*** vectors were used in Multinomial Logistic Regression, Naive Bayes, SVM and Decision Tree to apply to multi-categorising because of the categorising of 4 writers instead of binary classification. The results showed that the ***n-grams*** didn't perform well with an accuracy of 50% and the model improved with the addition of ***vocabulary richness*** measures**.** The other minor writing style features added boosted the final result, among all the models and the SVM performed the best with an accuracy rate of about 85%. The addition of the other writing style features from other sets proves that more writing style features produce a far better performance and better Authorship Identification.

3.2.1.2 Character n-grams

Character n-grams were used in a study by Jankowska et al., (2013) where they used a Common N-Gram (CNG) dissimilarity measure applied in a k-Nearest Neighbour method (unsupervised learning method). The CNG dissimilarity measure in k-nearest neighbour is based on the differences of the frequencies of writing style features that are most common in a document. The method compared the dissimilarity of ***frequency of character n-grams (2 to 4)*** between a sample document and each document from the dataset of documents of known authorship. The dataset used comprised of Computer Science related subjects and had an accuracy of 73.3%. The ***character n-gram*** produced satisfactory results using the dissimilarity measure.

The use of character n-grams is also shown by Kešelj et al., (2003) who developed a method for computer-assisted Authorship Identification based on ***character n-grams*** author profiles. They used a dissimilarity measure between the documents to measure the average frequency for a given *n-gram* in each document. Their single genre dataset was used made up of novels from different writers from different time periods. Their approach achieved a dissimilarity measure at 83%. The dissimilarity measure is quite accurate due to the variation of writers in the dataset.

In a similar study which used character n-grams by Witten et al., (1999), they proposed to identify tokens in a single newsletter. They believed ***character n-grams*** provide a good way to recognize *lexical* tokens. In order to evaluate the effect of the context, all tokens were replaced by a symbol that was treated by Prediction by Partial Matching (PPM) as a single character. Prediction by Partial Matching (PPM) is a data compression algorithm by encoding English text in as little as 2.2 bits/character, (Cleary and Witten, 1984). The training data used for the plain-text model was transformed in this way and the text article was compressed by this model to give a dataset in form of token bits. A token frequency in a document was compared in the dataset and the similarity was measured. The model produced a low error rate and had an average accuracy of 80%.

3.2.2 Syntactical Features

3.2.2.1 Function word n-gram

Coyotl-Morales et al., (2006) proposed a method for proper Authorship Identification that measured the different ***function word uni-gram to quad-gram*** sequences to classify the unknown from known authorship documents in the view that they express the most significant collocations used by an author. The ***function word uni-gram to quad-gram*** sequences were combined to create more features to be measured and for an effective performance. This dataset was gathered from the Web and consisted of 353 poems written by five different authors. The Naïve Bayes classifier was used including a 10-fold-cross validation because Coyotl-Morales et al., (2006) believe that it has proved to be competitive for most text processing tasks. The overall accuracy resulted with an accuracy of 83%. The method of calculating the frequency of ***function word n-gram*** sequences proved to be a good strategy with the recall rate at 83%. The recall rate may be the result of the many writing style features created from the combinations of ***function word n-grams.*** However, the models' goal was to identify authorship from topics thus using word based writing style features and no other writing style features that make up a personal writing style.

Another model that used ***function words n-grams*** was by Seidman (2013) who applied the General Impostor method which compares the similarity between documents and a number of external documents (generated from the web). Other writing style features included ***character n-grams.*** Seidman (2013) evaluated the writing style features using different frequency representations such as the term-document inverse-document frequency (TD-IDF) per document to see the similarities amongst the documents. The data used consisted of extracts from published textbooks on computer science and related disciplines. The overall performance achieved an accuracy rate of 79.2%.

3.2.2.2 Parts of speech tag (POST)

In the use of Parts of speech tag, Pavelec et al., (2009) used ***conjunctions*** and ***adverbs*** because of the way conjunctions are used as a characteristic of each author. The single genre dataset consisted of 30 articles with polemic subjects from different authors. All texts were pre-processed to eliminate numbers, punctuation and words were normalised. Spaces and end-of-line characters were not considered and all hyphenated words were considered as two words. The frequency of a writing style feature in a document was represented as a vector to train the SVM classifier with a

cross-validation for classification. The average performance of the strategy on the testing set was 83.2%.

A notable example of the use of **POST** by Solorio et al. (2011) proposed an approach for Authorship Identification on web forum data that generates informative meta (directory) features that can help discriminate the posts from different authors. Therefore, for their evaluation, they downloaded posts from the Chronicle of Higher Education (CHE) online forum and generated 5 data sets with a different number of authors each. They extracted ***Parts of Speech Tags*** from the dataset used which consisted of posts and minor writing style features such as ***percentage of all caps words, percentage of non-alphanumeric characters, sentence initial words with first letter capitalized, digits*** and ***word n-grams.*** The writing style features are extracted to generate a feature vector representation for each instance. However, in their model instead of having a single feature vector for a writing style, they generate smaller vectors that contain complementary types of features, or views, describing the instances. The generation of meta features uses the different vectors to produce clustering solutions for the training data with a number of clusters each ending up with different arrangements of the training instances into clusters. From each cluster in each of the clustering solutions, a centroid is computed by averaging all the feature vectors in that cluster. A similarity measure is applied on each instance to these centroids using the cosine function. These similarity values are then used as the meta features and computed them for training and testing instances, (Solorio et al., 2011). The results showed that ***word n-grams*** were not sufficient for the task due to the poor performance rate. Their approach achieved an impressive 77.38% accuracy. Their future work look into increasing the dataset by adding more authors and study the effect of having more than one topic in the data set

3.2.3 Content features (Topic features)

3.2.3.1 Word Frequencies

The evidence of word frequencies can be clearly seen in a study by Viswanathan and Mooney (2014) detected useful Business reviews using stylometric ***word frequencies*** and other minor with minor ***character*** and ***word bigram-trigram*** to capture the writing style and structural patterns of reviews and show how they can be used to distinguish useful reviews from non-useful reviews. They used extracts from the Yelp website, to demonstrate that useful reviews have a distinct style of writing that can be utilized in detecting them. The data processing involved matching analysis eliminating some writing style features in the process such as, ***Word bigram*** was found to be matching with ***character bigram***. Each review was considered as a separate document and all terms (writing style feature) frequencies across all reviews were computed to get inverse document frequency (IDF) and used in the classification models. Different classification models namely, Decision tree, Naïve Bayes, SVM, K-Nearest Neighbour and Logistic regression were used to compare performance on the dataset and applied 10-fold cross validation to tune their respective parameters. Their experiment achieved with SVM performed the best with 95.6.

A framework formulated by McDonald et al., (2012) mostly used ***word frequencies*** writing style feature and other ***Parts of Speech tags*** (***POST), sentence length, letter*** and ***character bi-trigrams*** features. They used two stylometry techniques, JStylo and Anonymouth tools for testing the consistency of anonymized writing style. JStylo is a standalone platform for Authorship Identification and Anonymouth is a writing style anonymization platform. JStylo extracted the writing style features and used SVM classifier with 10-folds cross-validation to classify the documents based on the extracted features. Thereafter, the Anonymouth performed clustering, preferential ordering, modification and document reclassification.

In clustering, using the k-nearest neighbor method, documents are clustered based on their writing style features which assists Anonymouth in selecting the target class. This forms a base that allows cluster groups to be ordered by a secondary preference calculation. The secondary preference calculation weighs writing style features with respect to their information gain ranking from feature extraction. It ensures that cluster groups that appear with a high frequency take precedence over those that appear less often. In modification and document reclassification, there is an option to modify a writing style feature once the targets are selected. Once the writing style features in the document have been changed it is considered reclassified. However, if the document has reached a sufficiently low classification, the document is considered anonymized (McDonald et

al., 2012). The dataset used consisted of samples of six authors from the Brennan-Greenstadt adversarial dataset. The features that showed to anonymize their writing style were ***sentence length*** and ***POST.*** The experiment conducted had an accuracy rate between 80% and 92.03%.

A study by Ramyaa and Rasheed (2004) also used ***word frequency*** writing style features for their study where they sourced their writing style features from Hanlein's empirical research (1999). Although Hanlein's research (1998) yielded a set of writing style features, his research only looked at a single magazine dataset and no other genres where styles of writing differ according to genre. Ramyaa and Rasheed's (2004) texts for the experiment were chosen from five Victorian era authors to ensure that the success of the model would entail that texts can be classified on the writing style of authors alone. The Decision tree classifier was chosen for the experiment because it is easy to read and understand. The writing style features used in the decision tree classifier were chosen based on whether they produced high information gain from the data, the others were not considered in the model. Neural Networks was used for its pattern recognition technique within texts. The model achieved an 82.4% accuracy using Decision trees and 88.2% accuracy using Neural Networks. In their conclusion, they state that different sets of features may be tried to see if there exists a set of feature which makes different learning techniques give the same results. Feature extraction could be done in some care to train these learners with the most relevant features. The study could also be extended to any number of authors instead of a finite number.

3.2.4 Structural features

3.2.4.1 Sentence length

Feng and Hirst (2013) applied the unmasking approach where documents written by the same author has features that detect little discrimination in authorship. Otherwise, if the texts were written by different authors, then many more features will support the discrimination. The ***average sentence length*** writing style feature was used in their model to identify discrimination between documents. Other minor stylometric writing style features include the ***average word length, word length distribution, frequencies of function words, vocabulary richness*** and ***frequencies of part-of-speech bigram***. Their experiment divided documents into not less than five pieces, then the SVM classifier with the writing style features trains their model to label each piece as coming from a particular document. The sampling creation is repeated five times, and the averaged leave-one-

out cross-validation accuracy is reported to represent the accuracy performance. The approach used a data set compiled of Computer Science subjects that performed at an evaluation accuracy of 75%. Their model probably achieved the recall rate due to the number of writing style features they had which was not alot. Their study shows that *Content* writing style features alone cannot identify authorship without addition writing style features.

Another study that used ***sentence length*** was by Ruseti and Rebedea (2012) who used it as their main writing style feature and different classification methods for Authorship Identification and used a data set collected from the free fiction collection published by Feedbooks.com. ***Character, Word length trigrams, POS bigrams*** and ***trigrams*** were also used as additional writing features which were sourced from previous studies in Authorship Identification in order to detect correct author. Text normalization was part of their data pre-processing so that the lengths of the texts do not interfere with the results. The training documents were split into pieces producing 100-200 samples for each author so a better generalization could be made by the classifier. The test documents were also split into pieces of the same size as the training data and the most common result was used as the output of the classifier for each document.

The SVM classifier was implemented along with the Logistic Regression method because they needed more exact probability estimation for each author in the training set. The Naive Bayes classifier was also tried but the results were not as good as the SVM when using cross-validation on the training set. This approach obtained an overall 77% accuracy with regard to the total number of correctly classified documents. The study shows that using only a reduced set of stylometric features has proven to offer good results for the Author Identification task. Moreover, splitting the training texts proved to be a good solution for training, evaluation and scoring the test documents. These results might have improved by adding more application specific features.

3.2.5 Ensemble Feature sets

14 ***n-gram*** patterns from *Lexical* ***tokens unigrams***, ***bigrams***, ***characters 4-grams*** and *Syntactical* ***POST unigrams*** to ***trigrams*** were used by Moreau and Vogel (2013) in their model. The model used distance measures such as Euclidean and Cosine to reflect how close the writing style features are from one document to another based on frequency value differences. A single topic dataset comprising of Computer Science related subjects was used for training their model. The SVM,

logistic regression, decision trees and Naive Bayes methods were used with their tuned parameters and evaluated on the single topic data training set 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 76.7% was achieved by their model. A reason for their result could be the configuring of their classification method parameters for an optimum performance. Moreau and Vogel (2013) mention that the ***POST n-gram*** as a syntactical feature set would have increased the model performance if it had been tagged better, this means adding more tags to terms identified.

A proposed design of feature set combination of, *lexical, content, structural* and *syntactical* by Tanguy et al., (2011) used ***word bi-tri grams frequencies***) with a large number of ad hoc features such as ***sentence length, vocabulary richness, POST n-grams*** and ***common words*** addressing different feature category performance. Their approach proposed to identify authorship using a wide range of writing style features from email messages extracted from Enron. Their data pre-processing involved the tokenisation of terms and categorising words into Part-Of-Speech categories such as nouns, verbs, etc. The Decision tree technique was adopted to split the training dataset in different subsets according to the writing style features used, i.e. the writing style features that produced the highest score of information from the data was used. During the testing phase, the ***word/character bi-trigram frequencies*** visibly increased the performance of the model more than the other writing style features. The experiment achieved an average recall score of 73.7%. The success of ***word/character bi-trigram*** frequencies investigates the number of information added by other individual features or feature sets.

Zheng et al., (2006) used *Lexical, Syntactical, Structural and Content* writing style features in their experiment on online messages to address the identity-tracing problem. The online-newsgroup messages in English and Chinese was used as their single genre, however, this dissertation will only be looking at the English dataset results that Zheng et al., (2006) used. Examples of the writing style feature in each set include in the *Lexical set,* ***total number of characters, frequency of letters and special characters, average word length*** and ***Hapax legomena***. *Syntactical* set had ***frequency of punctuations*** and ***frequency of function words***. *Content set* only had ***frequency of content specific keywords*** Structural set had ***total number of lines, total number of sentences, total number of paragraphs*** and ***number of sentences per paragraph*** to name a few.

The extracted writing style features were represented as vectors (by TD-IDF) and were trained with the Neural Networks, SVM and Decision tree classifiers used to build a feature-based classification model to identify authorship in online messages. The three classification techniques comparison results showed that *lexical* feature set performed well with an accuracy of 89% with SVM and the model improved with added feature sets achieving a high accuracy of 97.69% with SVM. The study shows that a single feature set on its own isn't as good as a combination of feature sets. The *Structural features* and *content-specific* features showed particular discriminating capabilities for Authorship Identification on online messages. ***Word*** and ***topic based*** features may be a contribution for an efficient Authorship Identification.

The feature set combination is commonly used as demonstrated in another example where Suh (2016) examined an automatic approach that adopts combination feature sets to estimate user reputations in social media. The study used a Korean Web forum for data because it is believed to be a major type of social media. Suh (2016) used writing style features that were used in previous social media studies for categorising Good or Bad user reputation based on user feedback. Some of the examples from the feature sets include, *Lexical* set with ***frequency of digit characters, frequency of white space characters*** and ***frequency of alphabetic characters.*** *Syntactical* set with f***requency 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, i.e, *F1, F1+F2, F1+F2+F3,* etc. The classification techniques used namely, 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 results with the best accuracy at 94.50 %.

Eder (2011) proposed a study that aimed to examine the effectiveness of several writing style features exclusively on a non-stemming dataset of English novels. The writing styles features chosen include ***most frequent words***, ***word n-gram (1 to 5)*** and ***letter n-gram (3 to 8).*** A few combinations of the above features included, ***words*** and ***word bi-grams***, ***words*** and ***letter 5-grams*** to make more writing style features. The procedure involved splitting the input texts into ***word***

and/or ***letter n-grams,*** and then replacing all the punctuation marks with spaces. Next, the obtained writing style features were counted and converted to frequency values in a matrix.

The testing of the writing style features effectiveness was done with Burrows's Delta which is a distance measure because it is an easily-applicable and reliable platform in stylometry. The models' performance increased with the number of the most ***frequent bi-tri-grams*** tested, following by ***word unigrams***. The frequencies ***word n-grams*** turned to be the most effective which is probably due to the combinations of ***word n-grams (1-4)*** which added up to 7500 vector writing style features that were analysed by the model. ***Letter n-grams*** were slightly less accurate than single words. The best results were obtained a combination of ***words*** and ***word bi-grams*** and the model achieved an average accuracy of 95%. A reason for the high recall rate could be due to the lack of text processing on the texts. The probability of the same unprocessed word appearing in documents is high than the base word.

3.3 Cross-topic and Cross-genre

3.3.1 Lexical Features

3.3.1.1 Character N-grams

An example of character n-grams is the study carried out by Castro et al., (2015) which compares the average similarity of documents of unknown authorship with all the documents of an author. The concept assumes that a text that is not written by an author would not exceed the average of similarity with known texts of an author. For each ***character uni- quad-gram*** and other ***word/POST n-gram*** writing style feature used from the documents, a feature that represents a document the most is obtained by statistical computation, such as Mean and Standard deviation. For the final total, Castro et al., (2015) divided the frequency value for each feature that is considered as written by the author and the unknown text, by the total of features analysed, (Castro et al., 2015). The method used a collection of dialog lines from plays which had a score of 75%. Castro et al., (2015) consider in their future work to work on documents that have high average similarity to see the efficiency of their model. They also want to evaluate the overall different genres of documents if all the writing style features contribute to the task.

The effectiveness of the character n-grams has been exemplified in a study by Bagnall (2015) 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. Figure 8 is an example of the flow feed neural network used for text formation prediction. Neural Networks are generally a good learning technique due to its precision in writing style. The pre-processing involved mapping unknown and known documents into smaller characters, i.e. capital letters decomposed into an uppercase maker followed by the corresponding lowercase letter to reduce computational complexity. The text is first converted into normal form, which decomposes accented letters into its alphabetic equivalent. For example, the character 'à' will be changed to 'a' (Bagnall, 2015). The method used a collection of dialog lines from plays which received a result of 81%.

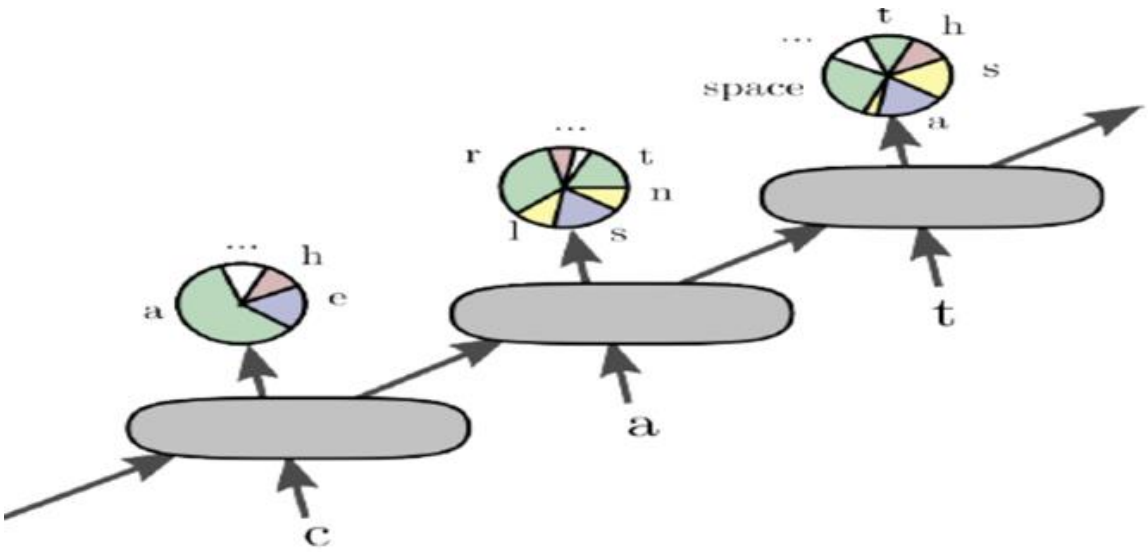


Figure *8*: Character Base Neural Network Model by Bagnall (2015).

3.3.1.2 Vocabulary Richness

A decision tree model developed by Fréry et al., (2014) used the term-frequency-inverse document frequency weighting scheme (tf-idf) to measure ***the variety of vocabulary*** writing style feature in each document in a dataset for document similarity measuring. The Classification and Regression Trees (CART) algorithm was then used to construct binary trees and thresholds that yield the largest information gain at each instance (Fréry et al., 2014). CART allows identifying good predictive writing style features. The binary trees were built by using each document of the training set separately to obtain a tree per document. The model was made of a document's writing style feature counts, averages and a label for the given unknown documents (Fréry et al., 2014).

The writing style feature that best splits the set of unknown documents into the two classes was chosen. The experimentation had been made on essays and novels containing 696 problems. For experimentation, they made a 10 cross validation for each group of problems in order to evaluate the quality of the decision trees on the training set (Fréry et al., 2014). Their method ranked an overall rate of 70.7% with an accuracy rate of 72.3%. Fréry et al., (2014) observed that the most difficult part was to gather writing style features that matched and building efficient features, like with the count method, highly improves the accuracy of CART for the dataset.

3.3.2 Syntactical Features (Syntax-based)

3.3.2.1 Punctuation

Halvani and Winter (2015) used a two-step training phase in their experiment. The first phase learns individual ***punctuation uni-gram to quad-gram*** parameters as well as decision thresholds based on equal error rates. The second phase builds feature category ensembles which are used for majority vote decisions. The pre-processing involved removing newlines, tabs and multiple spaces for better extraction of writing style features. Their data set used consisted of a variety of essays and novels with the evaluation of their model achieved an Accuracy of 76.2%.

3.3.2.2 Parts of Speech Tag n-grams

Moreau et al., (2014) used a combination of predefined value parameters to compare them to the ***Part-Of-Speech (POS) tags unigrams to 4-grams*** writing style features that were extracted from the documents they used. The frequency of these writing style features which fulfil a particular parameter value is stored for various statistical computations, such as Mean, Standard deviation and Median. The overall statistical computation for a document is compared to every other document within a data set. The dataset used contained documents from two genres, essays and reviews. The approach using these writing style features had an overall performance of an accuracy rate of 70.3%.

A similar study also used ***part-of-speech tag n-gram*** by Khonji and Iraqi (2014) using the General Impostor method which compares a set of external documents with the documents under investigation. Other minor lexical writing style features were used such as, ***letter-level, word-level,***

function word-level and ***word shape-level.*** This technique shows how close feature vectors of a document are to each other, relative to other external document vectors in the same topic (Khonji and Iraqi, 2014). Their data set had essays, articles, reviews and novels achieving an overall accuracy of 75%

3.3.3 Content features (Topic features)

3.3.3.1 Word uni-gram to tri-grams

The writing style features Gutierrez et al., (2015) used for their method were the ***frequency of word bi*** to ***tri-gram***. They used the General Impostor method to compare external populated documents with the documents being investigated. The study used the Homotopy-based Classification, a pattern recognition concept for data similarity to measure a known authorship document's writing style to an unknown document. Using a data set had a variety of essays, articles, reviews and novels, the approach performed well with an accuracy of 74%.

3.3.3.2 Common word

Kocher and Savoy, (2015) applied SPATIUM-L1, a distance measure which calculates how close texts are to one another compared to a set of external documents to determine whether or not a disputed text was written by a proposed author. The writing style feature extracted was the ***common word frequency***. To determine the value of the top most frequent terms, the effective number of terms was set to at most 200 terms, but in most cases the figure was well below it. The pre-processing involved removing stemming words and keeping punctuation symbols. The evaluation was performed on a dataset which composed of essays and novels of a 100 documents with a result of 73.8%. Kosher and Savoy (2012) have considered for improving their model to use a simpler distance measure and maintain a reduced number of writing style features. For a better feature selection scheme, they can consider the text genre, for example, the most frequent use of *personal pronouns* in narrative texts. Another possible improvement can be ignoring specific ***topic terms*** appearing frequently in an authors' writing style feature. Terms that can be selected in the feature set without being useful in discriminating between authors.

3.3.4 Structural features

3.3.4.1 Paragraph

A modified unsupervised learning method, K-Nearest Neighbour classifier by Ghaeini (2013) believed that it could have a good performance for a small dataset to make predictive decisions and used ***paragraph count*** and ***average of paragraph length*** writing style features**.** As part of the pre-processing, they used stemming to reach more general words in their approach and used a data set consisted of excerpts from newspaper editorials and short fiction, newspaper, articles and Computer Science subjects. The comparison setup of all writing style features was used to compare the known authorship documents used against unknown documents using the cosine similarity measure. The cosine similarity is a popular vector based on similarity measure in text mining and information retrieval. The weighted K Nearest Neighbour is used to balance the effect of each feature measure. The method produced a recall rate of 85.71% with 30 problems of 35 problems identified correctly and 5 problems incorrectly identified. To avoid an application that incorrectly identifies problems, 5 decision makers with different weights were used to obtain an average result to reach better and more robust result. Their overall performance had a result of 83.7%. Ghaeini (2013) concluded on increasing the dataset for an effective comparison model.

The ***number of paragraphs*** writing style features was used by Pacheco et al., (2015) in their model which used the Random Forest classifier. They used feature vector modelling to represent writing style features per document and had one to five documents per author. The Random Forest classier was applied to build their model to hierarchically determine the importance of each feature in the identification of authorship. The feature vector is valued with a number between 0 to 1, if the writing style features computed for the unknown document is closer to a known authorship document than any other in the dataset. Otherwise, it would be valued with a number between 1 to 1+. Their approach performed very well on a dataset consisting of dialog lines from plays, articles on Politics, Economy and Science, reached an accuracy score of 76.3%. Possible improvements for this approach include studying the separation of documents into paragraphs for more writing style features like ***paragraph length.*** Another possible improvement is to analyze the information gained from proposed features and include texts from other sources to broaden the dataset.

3.3.5 Ensemble Feature sets

Gómez-Adorno et al., (2018) measured ***Character n-grams, Word n-grams*** and ***POST n-grams*** from *lexical, content and syntactical* feature sets using a document embedding method which is based on context modelling. This means that the order of the writing style features sequences is used to model the writing style of an author. Documents containing similar n-gram sequences are believed to be written by the same author. The td-idf was used to represent the writing style features as well as the use of the Logistic regression classifier chosen to be used to create their model. Gómez-Adorno et al., (2018) used the Logistic Regression classifier because of its simplicity, speed and has been reported to have good results in Authorship Identification tasks.

The experiment showed that ***POST n-grams*** produced the most efficient results when added to the model. However, the best results were produced from the combination of ***POST, word*** and ***character n-grams***. The experiment was conducted on a cross-topic dataset made up of a single newspaper. Their experimental results achieved a high average 90% score. Gomez-Adorno et al., (2018) concluded to conduct experiments on other writing style features such as ***syntactic n-grams***. They considered evaluating different composition methods for the document embedding such as Neural Networks learned on various ***n-gram*** types.

The Writeprints technique for identification and similarity detection of anonymous identities is used by Abbasi and Chen (2008). Writeprints is a commonly used technique because of its high accuracy on a large datasets. Abbasi and Chen (2008) found that the individual-author-level feature set is rarely researched and could improve authorship categorization performance and scalability. They also found that their use of an extended set of features could improve the scalability of stylometric analysis by allowing greater discriminatory potential across larger sets of authors. Their extended feature set contained *lexical, syntactic and content-specific sets*. *Lexical features* include ***n-gram character, digit-level***, ***function words, word-length distributions and vocabulary richness measures***. *Content feature* set of ***n-gram word***, ***Parts of Speech Tag*** from the *Syntactic feature set* and *misspelling*. The data processing removed redundant characters and identifiers.

The classifier construction has two parts, creation and pattern disruption, the Writeprint creation produces a lower-dimensional usage variation pattern created based on the frequency of the writing

style features of an author's text. The writing style features are represented by a vector using the Karhunen-Loeve transform (K-L transform) which is used in pattern recognition experiments.

In pattern disruption, comparing identities requires a construction of a pattern of two sample text documents A and B. The pattern disruption stands as a comparison against identity A's Writeprint and vice versa. The overall similarity between identities A and B is the sum of the average n-dimensional Euclidean distance between Writeprint A and pattern B and Writeprint B and pattern A. It is preferable that A's zero-frequency writing style features act as pattern disruptors (Abbasi and Chen, 2008). This is where the presence of these features in identity B's text decreases the similarity for the particular A-B comparison. The distance between two identities' patterns can be used to determine the degree of stylistic similarity.

The Sequential Minimal Optimization (SMO) in SVM classifier with a 10-fold cross validation was used on the dataset that comprised of emails, website comments, code and chats of 100 authors collectively. The Writeprints technique had over an average of 85% on the data. Abbasi and Chen (2008) consider further improving the scalability of their proposed approach by increasing the number of documents and analyze proposed writing style features.

3.4 Summary

The related work reviewed how writing style features had been a marker for a writers' personal style since the 1800's. The initial data used in Authorship Identification tasks comprised of novels and articles written by a few writers making it easy to differentiate a writers' style. The increase of volume and variety of texts eventually led to cross-genre and cross-topic documents. The writing style features extracted from documents have increased in number due to the increase of genre/topic vocabulary and contexts. The review of related work demonstrated the use of the writing style features individually and as a combination of feature sets. The writing style features that were commonly used in most of the successful previous studies 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 classifiers also most commonly used in the previous works include

SVM (SMO), Multilayer Perception, Naïve Bayes, Decision trees, K Nearest Neighbour, Logistic Regression, Neural Network and Random Forest.

Most of the studies applied the writing style features because they were used in other studies. However, the writing style features they used were not compared and evaluated to determine whether they were ideal to use in their tasks. Due to the lack of evaluation comparisons of writing style features in cross-genre and cross-topic, there is no evidence that suggest that they are the best to be chosen. This dissertation sets out to empirically evaluate the writing style features used in the related works to measure their effectiveness on cross-genre and cross-topic documents which the previous studies have not attempted to do. The following section is the process taken to carry out the dissertation's plan.

4 METHODOLOGY

4.1 Introduction

The methodology outlined in this section will be used in the experiment to validate on the dissertation statement. Section 4.2 explains the Cross Industry Standard Process for Data Mining (CRISP-DM) process and section 4.2.1 describes how the dissertation is arranged following the CRISP-DM processes.

4.2 CRISP-DM

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a generic process model used as a management tool because of how volatile data mining projects are and is the go-to-process because of its efficiency in knowledge management experimentation. Bocko (2015) points out that the CRISP-DM methodology is widely considered as a standard for data analysis. This has been widely used by several authors such as Silipo and Zimmer (2015), Carnerud (2014) and De Waal et al., (2008) who used it in their probability data science projects. Badder (2005) describes the CRISP-DM methodology as helping to organize project annotations, streams and output according to the phases of a typical data mining project.

The CRISP-DM model has six phases. However, the phases do not necessarily follow one another sequentially. The cross functioning between the different phases is always required and at times necessary. The outcome of a phase determines which phase or particular task of a phase has to be performed next. At times, the previous phase is reiterated for a thorough outcome to be understood. Figure 9 depicts the sequence of the CRISP-DM model where the outer circle symbolizes the cyclical nature of a data mining process. The arrows between the phases indicate the frequent dependencies between phases. Figure 10 shows CRISP-DM model consisting of six phases. The rest of the section focuses on each of the phases' activities and how the dissertation and experiment is arrangement under each of the phases.

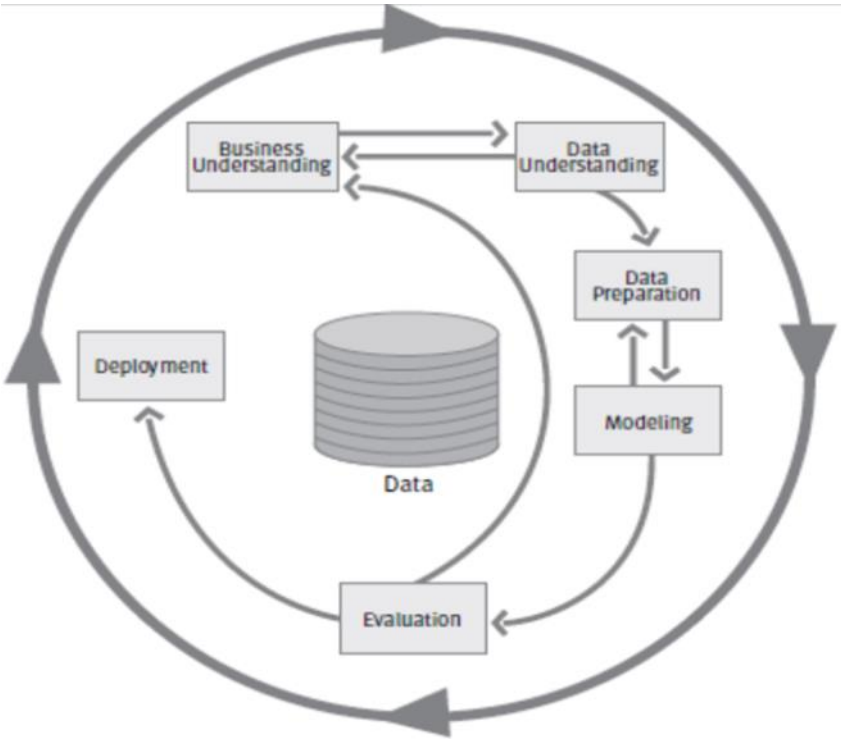


Figure 9: The phases of the CRISP-DM model, (Chapman, 2000).

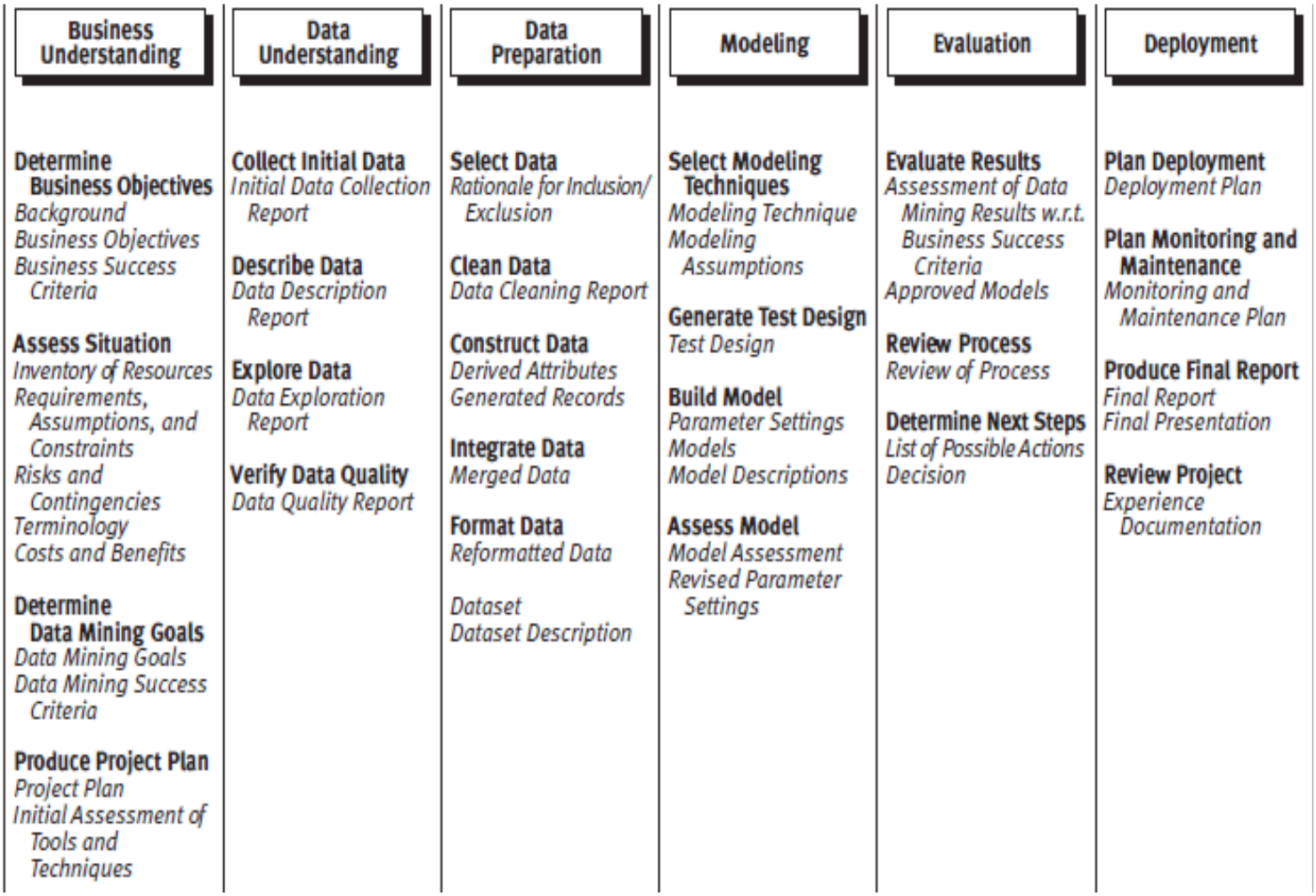


Figure 10: The Phases, tasks and the output in the CRISP-DM model, (Wirth and Hipp, 2000).

4.2.1 Business Understanding

This is the inceptive phase of the CRISP-DM process which defines the project objectives, requirements and assumptions of a project area. After producing a preliminary project plan, an initial assessment of tools and techniques is reviewed as a guide for direction (Wirth and Hipp, 2000). Under the business understanding, the dissertation objective is to identify the writing style features that are ideal in Authorship Identification for cross-topic and cross-genre documents. The data mining goals for the dissertation include identifying the writing style features that were used in previous successful Authorship Identification works and which writing style feature combinations can be used for cross-topic and cross-genre documents. The plan is to use the writing style features that have been identified on a cross-topic and cross-genre dataset to evaluate which writing style features produce outstanding results in Authorship Identification.

4.2.2 Data Understanding

The data understanding phase starts with an initial collection of data, describing the data either by categorising the data, exploring the data as well as verifying data quality or steps to be taken in terms of data cleaning (Wirth and Hipp, 2000).

The dataset used in this dissertation is taken from the 2015 PAN CLEF English collection which consists of 100 sets each containing a known author document and an unknown (questioned) document. The dataset consist of a variety of cross-topics and cross-genres including dialog lines from plays, excluding speaker names, stage directions and lists of characters from different authors. The documents within a set comprises of short texts having 350 words on average per document. Figure 11 shows the document sample of a known author and Figure 12 shows the unknown document from the same set. The text as it is needs pre-processing so that the data can be represented in a way they can be processed to be categorised into writing style features. The experiment follows text pre-processing techniques such as tokenising, normalising and stemming.

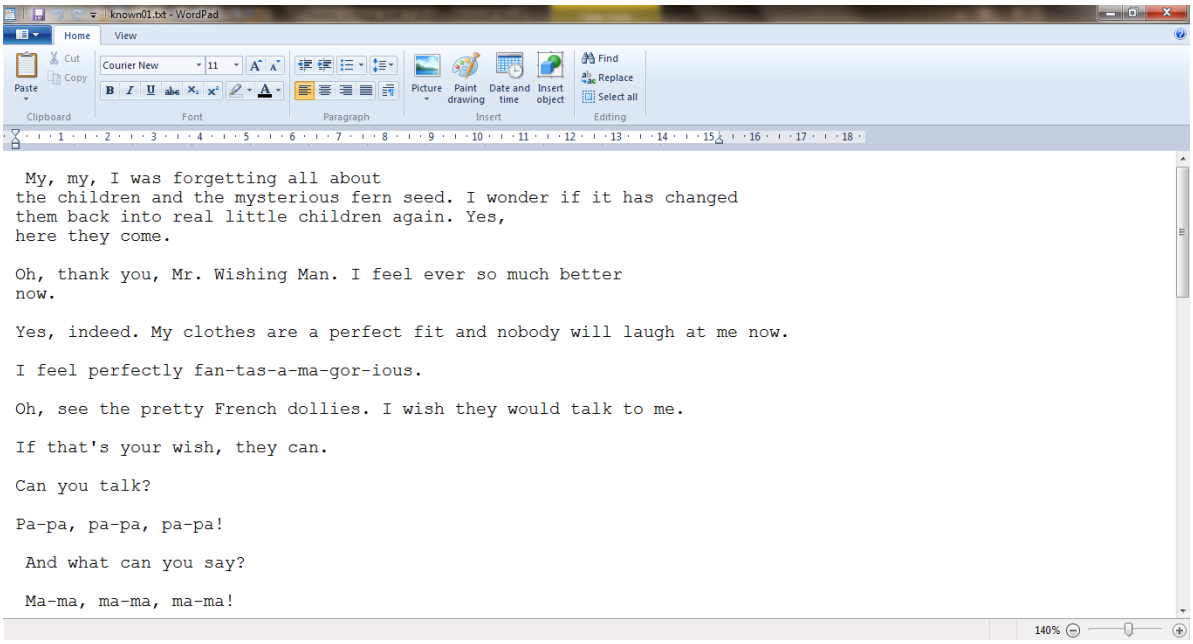


Figure *11*: A sample document of a known document, (PAN CLEF, 2015).

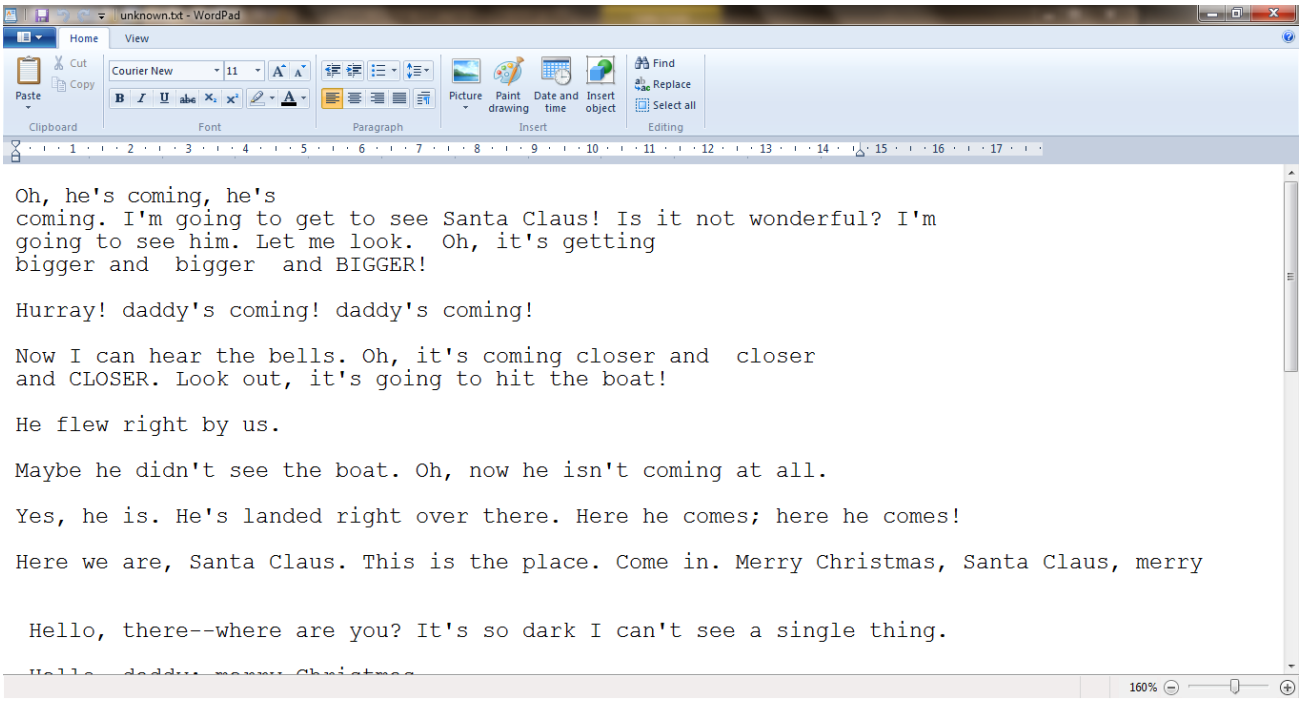


Figure *12*: A sample document of the questioned (unknown) document, (PAN CLEF, 2015).

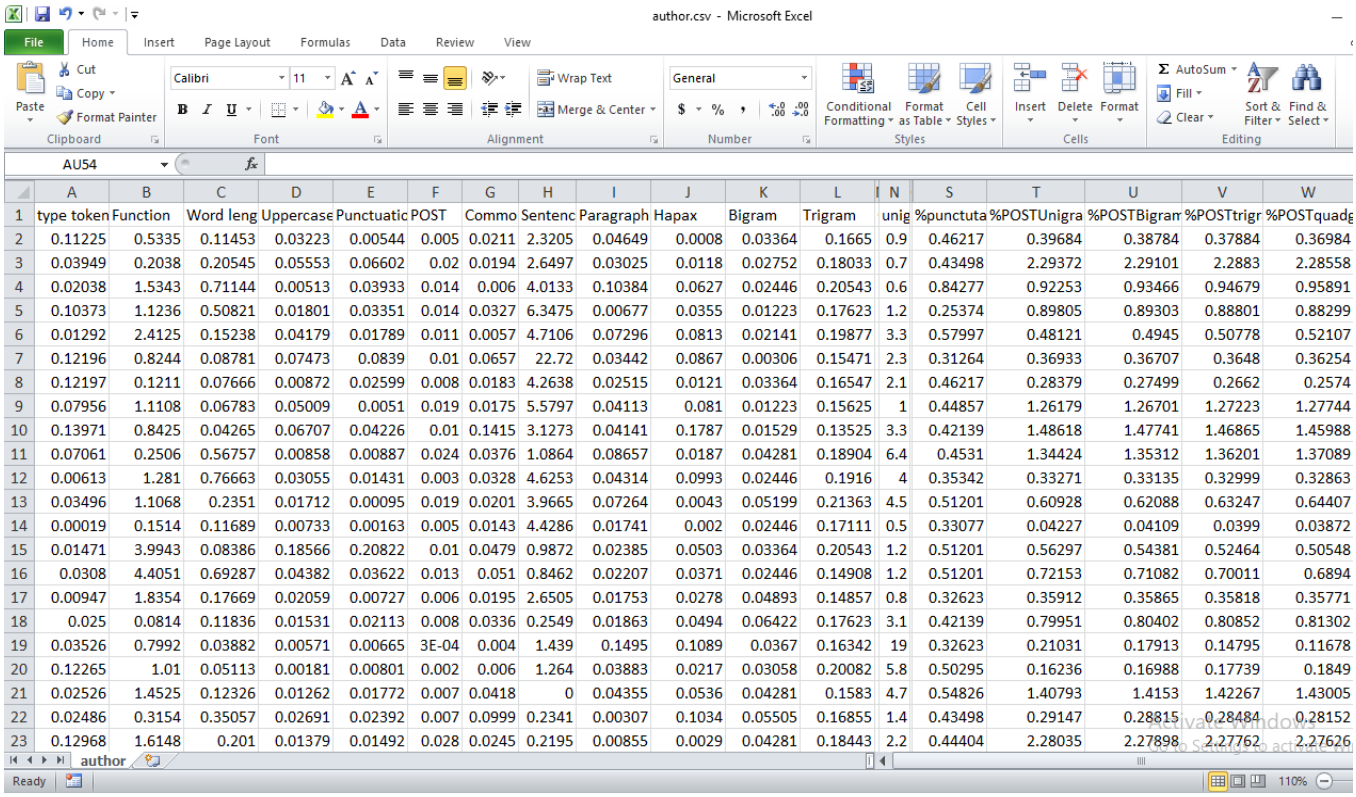
4.2.3 Data Preparation

According to Chapman et al., (2000) the data preparation phase covers all techniques needed to construct the final dataset that will be fed into the modelling tool(s) from the initial raw data. Data preparation functions are likely to be performed multiple times and are not in any preconditioned order. Tasks include feature selection, as well as transformation and cleaning of data for modelling tools. Other examples include aggregation and data manipulation to move from the raw data to a more structured and informative dataset. Another example can be the aggregation of values from a worksheet into two groups of values describing each attribute in question (Silipo and Zimmer, 2015).

The known document and unknown document within a set in the dataset used in the experiment was processed based on the writing style features. A formula was written for every writing style feature to get a numeric value from the text. The writing style feature difference between the known and unknown documents indicates whether the documents were written by the same author or not. If the difference is over 0.5, it is most likely written by the same author, otherwise, by different authors. For example, if the writing style feature *total length of the sentence* is 50 and the *total text length* is 300 in a known text file, the calculation would be 50/300=0.167. In an unknown

file, if the final calculation was 0.78, then the difference would be 0.78-0.167=0.613, indicating that the files were written by the same author. The scores were not rounded up and are calculated as the writing style feature frequency difference between the known and unknown document. All the data is stored in a excel spreadsheet.

The data is then converted from an excel spreadsheet into a Comma-Separated Value (CSV) file. A CSV is a file where each value is separated by a comma. Also known as a Comma Delimited file, it is a standard file type that a number of different data-manipulation programs can read and understand. Therefore, in the dissertation, the Comma Separated Value (CSV) format was chosen to be used. In a CSV file, the first line holds the writing style feature name into the header structure that makes up the beginning of the Comma Separated File. Each row represents a documents' writing style feature also separated from each other by commas. A sample of a Comma Separated Value document which is used in the dissertation shows some of the writing style features shown in figure 13.



	A	B	C	D	E	F	G	H	I	J	K	L	N	S	T	U	V	W
1	type token	Function	Word leng	Uppercase	Punctuatic	POST	Commo	Sentenc	Paragraph	Hapax	Bigram	Trigram	unig	%punctuta	%POSTUnigra	%POSTBigram	%POSTtrigr	%POSTquadg
2	0.11225	0.5335	0.11453	0.03223	0.00544	0.005	0.0211	2.3205	0.04649	0.0008	0.03364	0.1665	0.9	0.46217	0.39684	0.38784	0.37884	0.36984
3	0.03949	0.2038	0.20545	0.05553	0.06602	0.02	0.0194	2.6497	0.03025	0.0118	0.02752	0.18033	0.7	0.43498	2.29372	2.29101	2.2883	2.28558
4	0.02038	1.5343	0.71144	0.00513	0.03933	0.014	0.006	4.0133	0.10384	0.0627	0.02446	0.20543	0.6	0.84277	0.92253	0.93466	0.94679	0.95891
5	0.10373	1.1236	0.50821	0.01801	0.03351	0.014	0.0327	6.3475	0.00677	0.0355	0.01223	0.17623	1.2	0.25374	0.89805	0.89303	0.88801	0.88299
6	0.01292	2.4125	0.15238	0.04179	0.01789	0.011	0.0057	4.7106	0.07296	0.0813	0.02141	0.19877	3.3	0.57997	0.48121	0.4945	0.50778	0.52107
7	0.12196	0.8244	0.08781	0.07473	0.0839	0.01	0.0657	22.72	0.03442	0.0867	0.00306	0.15471	2.3	0.31264	0.36933	0.36707	0.3648	0.36254
8	0.12197	0.1211	0.07666	0.00872	0.02599	0.008	0.0183	4.2638	0.02515	0.0121	0.03364	0.16547	2.1	0.46217	0.28379	0.27499	0.2662	0.2574
9	0.07956	1.1108	0.06783	0.05009	0.0051	0.019	0.0175	5.5797	0.04113	0.081	0.01223	0.15625	1	0.44857	1.26179	1.26701	1.27223	1.27744
10	0.13971	0.8425	0.04265	0.06707	0.04226	0.01	0.1415	3.1273	0.04141	0.1787	0.01529	0.13525	3.3	0.42139	1.48618	1.47741	1.46865	1.45988
11	0.07061	0.2506	0.56757	0.00858	0.00887	0.024	0.0376	1.0864	0.08657	0.0187	0.04281	0.18904	6.4	0.4531	1.34424	1.35312	1.36201	1.37089
12	0.00613	1.281	0.76663	0.03055	0.01431	0.003	0.0328	4.6253	0.04314	0.0993	0.02446	0.1916	4	0.35342	0.33271	0.33135	0.32999	0.32863
13	0.03496	1.1068	0.2351	0.01712	0.00095	0.019	0.0201	3.9665	0.07264	0.0043	0.05199	0.21363	4.5	0.51201	0.60928	0.62088	0.63247	0.64407
14	0.00019	0.1514	0.11689	0.00733	0.00163	0.005	0.0143	4.4286	0.01741	0.002	0.02446	0.17111	0.5	0.33077	0.04227	0.04109	0.0399	0.03872
15	0.01471	3.9943	0.08386	0.18566	0.20822	0.01	0.0479	0.9872	0.02385	0.0503	0.03364	0.20543	1.2	0.51201	0.56297	0.54381	0.52464	0.50548
16	0.0308	4.4051	0.69287	0.04382	0.03622	0.013	0.051	0.8462	0.02207	0.0371	0.02446	0.14908	1.2	0.51201	0.72153	0.71082	0.70011	0.6894
17	0.00947	1.8354	0.17669	0.02059	0.00727	0.006	0.0195	2.6505	0.01753	0.0278	0.04893	0.14857	0.8	0.32623	0.35912	0.35865	0.35818	0.35771
18	0.025	0.0814	0.11836	0.01531	0.02113	0.008	0.0336	0.2549	0.01863	0.0494	0.06422	0.17623	3.1	0.42139	0.79951	0.80402	0.80852	0.81302
19	0.03526	0.7992	0.03882	0.00571	0.00665	3E-04	0.004	1.439	0.1495	0.1089	0.0367	0.16342	19	0.32623	0.21031	0.17913	0.14795	0.11678
20	0.12265	1.01	0.05113	0.00181	0.00801	0.002	0.006	1.264	0.03883	0.0217	0.03058	0.20082	5.8	0.50295	0.16236	0.16988	0.17739	0.1849
21	0.02526	1.4525	0.12326	0.01262	0.01772	0.007	0.0418	0	0.04355	0.0536	0.04281	0.1583	4.7	0.54826	1.40793	1.4153	1.42267	1.43005
22	0.02486	0.3154	0.35057	0.02691	0.02392	0.007	0.0999	0.2341	0.00307	0.1034	0.05505	0.16855	1.4	0.43498	0.29147	0.28815	0.28484	0.28152
23	0.12968	1.6148	0.201	0.01379	0.01492	0.028	0.0245	0.2195	0.00855	0.0029	0.04281	0.18443	2.2	0.44404	2.28035	2.27898	2.27762	2.27626

Figure *13*: Comma separated Value (csv) format sample of the features and feature scores saved in generated used in the study.

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. The data is normalised between 0 and 1 to avoid numerical difficulties during experimentation as seen in figure 14. An ablation process was conducted in the experiment, which is the removal of some feature(s) of a model, or dataset and see how it affects performance before and after its removal. 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. Therefore, the experiment involves removing each writing style feature to monitor how it would increase or decrease performance. The removal of a writing style feature that increased performance was removed from the feature set, otherwise if it decreased performance it was kept in the feature set as seen in figure 14.

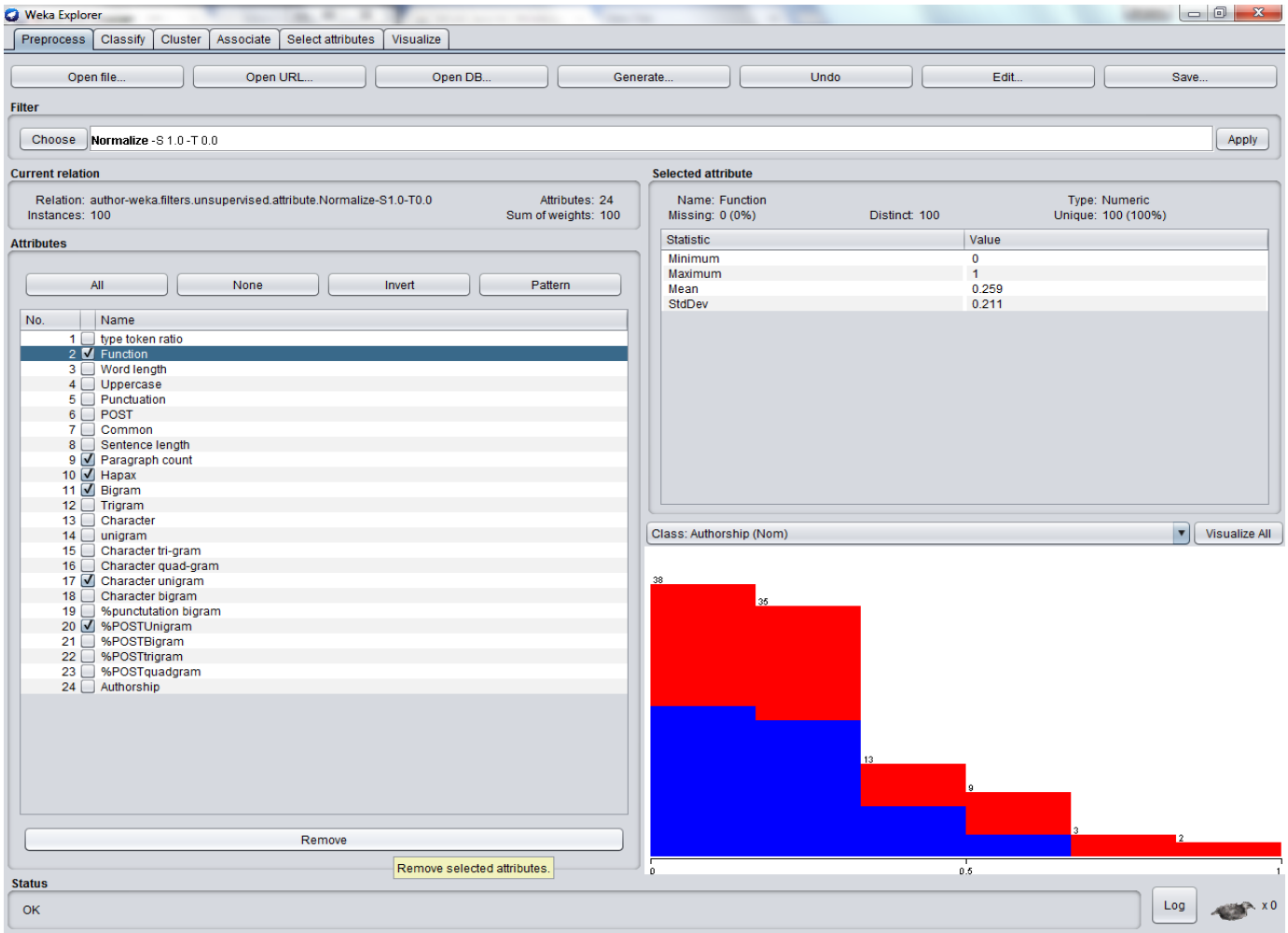


Figure *14*: Normalising writing style features from 0 to 1 and an ablation process on writing style features.

4.2.4 Modelling

The Modelling process is usually conducted with several models using their default parameters, then fine tune the parameters or revert to the data preparation phase for manipulations required by their model of choice. Different models follow a practical implementation guideline or optimal parameters in order to generate a high accuracy result. The use of different classifiers creates a generalised result of how a writing style feature and a learning method are performed.

The classifiers selected for the dissertation experimentation are *Support Vector Machine (SVM)* which uses the Kernel type functions; *Random Forest*, a decision tree based classifier and the *Naïve Bayes* classifier which builds a probabilistic model. Studies such as Bartoli et al., (2015) explored three different regressor algorithms; Decision trees (Tree), Random Forests (RF), and Support Vector Machines (SVM) for their experiment. SVM and Random Forest classifiers were chosen because they are well-known and popular supervised learning algorithms. Support Vector Machines (SVM) in particular, is a good approach to classification because they are designed to handle high-dimensional data, and it has been applied successfully to Author Identification in previous works. Classifier parameters have to be changed to obtain optimal classification accuracy performance. The experiment followed a Support Vector Machines (SVM) procedure formulated by Hsu et al., (2015) which ordinarily produces reasonable results.

In the Support Vector Machines classifier (SVM), the different kernel types used were *Linear kernel* and *Radical Basis Function* (RBF). The Linear kernel was used on writing style features that were many, that is, the number of writing style features that were more than the number of instances in the dataset did not need its data to be mapped to a high dimensional space. The RBF was used on writing style features that were few and needed a higher dimensional space (nonlinear kernel). The RBF parameters namely, Cost and Loss pair needed to be identified in order to accurately predict unknown data (test data). 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 was made up of 66% of the data while the remaining 34% of the dataset was used for testing because of the effective results produced.

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 Hsu et al., (2015) was used on Cost and Loss parameters and using the 10 fold cross validation method. To find the optimal parameters, an algorithm Grid is made following the process described in Figure 15.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. A flowchart to find the optimal parameters using a grid search with cross validation process is depicted in figure 15.

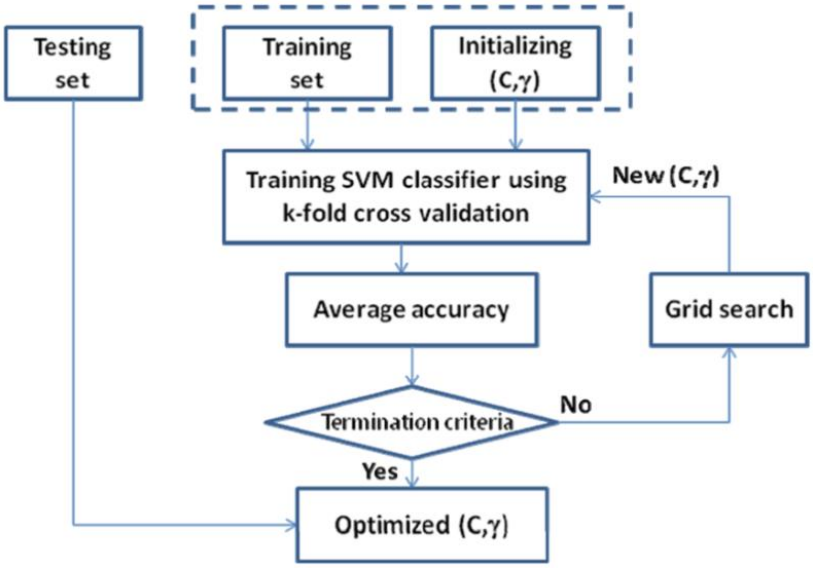


Figure 15: Flowchart Selection of optimal parameters of Kernel function using Grid search, (Thai et al., 2012).

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}.

4.2.5 Evaluation

In Chapman et al., (2000) literature on CRISP-DM, the evaluation step assesses the degree to which the model meets the business objectives and seeks to determine if there is some reason why a model used is deficient. Another application is to test the model on simulated real application. Moreover, the evaluation phase also assesses other data mining results generated. Data mining results involve models that are necessarily related to the original business objectives and all other findings that are not necessarily related to the original business objectives. However, they might also unveil additional challenges, information, or hints for future directions (Chapman et al., 2000). In the Results and Analysis section 5, the section describes the results obtained from the experiment on the dataset and a discussion of the research questions the dissertation set out to investigate.

4.2.6 Deployment

The knowledge gained from a project will need to be organized and presented in a way that the users can use it and readers can understand it. The findings and experiment approach can bring about lessons learnt, advice to novices and recommendations to other work done which was also reviewed for related work (Deshpande and Thakare, 2010). A broader perspective by the Badder (2005) say that deployment can mean that you use the insights gained from data mining projects to elicit change in a field of study. A data mining process does not end once a solution is deployed. The lessons learned during the process and from the deployed solution can bring about new, often more focused business enquiries. Future data mining processes can benefit from the experiences of previous ones (Chapman, 2000).

4.3 Summary

The methodology looks at the CRISP-DM process the dissertation will be following because of the nature of data mining projects and reviews examples of how previous works used CRISP-DM in their own respectable capacities particularly in Authorship Identification. The CRISP-DM phases were reviewed and the dissertation was arranged according to the CRISP-DM phases. The business phase covered the dissertation objectives and plan, data understanding phase investigated the data content and whether it needed any pre-processing. The data preparation covers the data

pre-processing using stemming, tokenisation and normalising implementation and feature extraction from the documents from the dataset. The modelling phase describes the experimental setup using the classifiers and finding the optimum parameter values. The evaluation and deployment were also explained with the evaluation explanation seen in Chapter 5 Results and Analysis.

5 RESULTS AND ANALYSIS

In this study, an empirical evaluation is conducted to determine which writing style features can be used for cross-topic and cross-genre Authorship Identification. This is achieved by answering the following research questions;

RQ1 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?

RQ2 Which type of writing style features work best for cross-genre and cross-topic documents and which cannot work best used?

RQ3 Which writing style features can be combined for cross-genre and cross-topic document in Authorship Identification?

RQ4 Do the results from this study generalise across the three different family of classifiers?

In the empirical evaluation, the following evaluation measures were used namely Sensitivity (TP), Specificity (TN), Accuracy, ROC (AUC) and Kappa coefficient. Recall that Sensitivity (TP) measures the proportion of actual positives that are correctly identified and that 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. The c@1 performance measure was not used in the empirical evaluation because the data was only labelled into positive (more than 0.5) and negative (less than 0.5) values. The C@1 measures data that are positive, negative and unlabelled (data that are not categorised into positive and negative).

5.1 Discussion of Research Question 1

Table 3 shows all the writing style features identified from previous works in their respectable individual feature sets used for this study's empirical evaluation on the PAN CLEF 2015 English dataset. An initial evaluation was performed with the full feature sets to generate initial results that are in table 4. 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 *3*: 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}, Type token ratio, Word length, Hapax Legomena,
Syntactical	Parts of Speech Tag{ Unigram, Bigram, Trigram and Quad-gram}, Parts of Speech Tag, Punctuation, Punctuation Bigram, Function word
Structural	Paragraph frequency, Sentence Length
Content	Common words, Word {Unigram, Bigram, Trigram}

The full individual feature sets with all the writing style features shown in table 3 were used to generate the initial evaluation results in table 4. 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 4 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 in table 4 show that

Syntactical writing style features had the best results of all the feature sets with an average AUC of 0.669. The *Lexical* and *Content* feature sets had moderate results with AUCs of 0.599 and 0.547 respectively, while the *Structural writing style* features generated the lowest results with an AUC of 0.528. The results of 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 4: 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
	SVM	0.980	0.120	55	0.550	0.1
	Random Forest	0.389	0.688	53	0.583	0.07
Syntactical	Naïve Bayes	0.780	0.580	68	0.738	0.36
	SVM	0.940	0.260	60	0.600	0.2
	Random Forest	0.660	0.580	62	0.669	0.24
Structural	Naïve Bayes	0.333	0.625	47	0.483	-0.04
	SVM	0.143	0.875	53	0.509	0.02
	Random Forest	0.278	0.750	50	0.528	0.03
Content	Naïve Bayes	0.500	0.688	58	0.583	0.18
	SVM	0.580	0.460	52	0.520	0.04
	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 4.2.1 that the Data Preparation phase explains the experiment implemented an ablation process that removes a writing style feature from a feature set to identify whether it improves or decreases the experiment 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 checked in the same way. Therefore, the experiment involves removing each 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 3.

Table *5*: The writing style features in each feature set after the ablation process that increased performance.

Feature Set	**Writing Style Features Identified as Increasing Performance**
Lexical	Word Length, Uppercase frequency, Character level {Unigram, Bigram, Trigram and Quad-gram}
Syntactical	Parts of Speech Tag{Unigram, Bigram, Trigram and Quad-gram}, Punctuation Bigram
Structural	Paragraph frequency, Sentence Length
Content	Common words, Word {Bigram; Trigram}

The writing style features shown in table 5 were kept after it was shown that their removal from a feature set decreases performance. This means that their presence contributes to the generation of high performance. The other writing style features that are not in table 5 were removed from the feature sets because they showed to increase performance by their removal meaning that their

presence in a feature set brings down the performance result. The following writing style features were removed include;

- (Lexical set) ***Type token ratio, Word length, Hapax Legomena*** and ***Character Unigram*** in the Lexical features set.
- (Syntactical set) ***Parts of Speech Tag, Punctuation*** and ***Function word***
- (Structural set) ***Word Unigram***

Table 6 shows the results of the features sets with the writing style features that were kept which generated high results after the ablation process. The *Syntactical* set still 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 verifies 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 qudgram)*** 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.

The *Lexical* and *Content* feature sets had general moderate results, with the Lexical set having a higher AUC than Content with 0.714 and the Content set with an AUC of 0.647. The feature set with the lowest results was the Structural set with the lowest AUC of 0.554. This means that the *Structural* writing style features are not suitable for cross-genre and cross-topic document Authorship Identification. To answer ***research question 4*** *(Do the results from this study generalise across the three different family of classifiers?)* in table 6, the results generally showed that most of the highest results were generated from Syntactical set, then secondly Lexical, then Content followed by Structural set. This generalisation is the same as the initial evaluation and after ablation process results in table 4 and table 6 respectively.

Table 6: 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
	SVM	0.920	0.280	60	0.600	0.2

	Random Forest	0.556	0.750	64.7	0.635	0.3
Syntactical	Naïve Bayes	0.800	0.580	69	0.745	0.38
	SVM	0.660	0.800	73	0.730	0.46
	Random Forest	0.660	0.740	70	**0.750**	0.4
Structural	Naïve Bayes	0.857	0.438	63	0.554	0.29
	SVM	0.857	0.438	63	0.647	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
	SVM	0.380	0.840	61	0.610	0.22
	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?)* a combination feature set analysis is applied to see how it affects the experiment performance, the process is as follows;

i. The writing style features in the feature set that were found to work the best in the experiment in table 5 were merged with another to make a feature set combination pair,

ii. then with another one to make another feature set combination pair. For example, a *Lexical* set combined with a *Syntactical set,* and then a *Lexical* set combined with a *Structural* set.

iii. 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.

Table 7 shows all possible combination feature sets with the writing style features.

Table 7: The combination feature sets with their writing style features.

Combination Feature set category	**Writing Style Features used for performance**
Lexical and Syntactical	Word Length, Uppercase frequency, Punctuation Bigram, Character {Unigram, Bigram, Trigram, Quad-gram}, Parts of Speech Tag {Unigram, Bigram, Trigram, Quad-gram}
Lexical and Structural	Word Length, Uppercase frequency, Sentence Length, Paragraph frequency, Character {Unigram, Bigram, Trigram, Quad-gram}
Lexical and Content	Word Length, Uppercase frequency, Common words, Character {Unigram, Bigram, Trigram, Quad-gram}, Word {Bigram, Trigram}
Lexical, Structural and Syntactical	Parts of Speech Tag {Unigram, Bigram, Trigram and Quad-gram}, Character {Bigram, Trigram, Quad-gram}, Common words, Uppercase frequency, Word Length, Punctuation Bigram, Sentence Length, Paragraph frequency
Lexical, Structural and Content	Word Length; Common words; Character {Unigram, Bigram, Trigram, Quad-gram} Word{Bigram, Trigram}, Part of Speech Tag{Unigram, Bigram, Trigram, Quad-gram}
Lexical, Syntactical and Content	Word Length, Punctuation bigram, Common words, Word {Bigram, Trigram}, Character {Unigram, Bigram, Trigram, Quad-gram} POST {Unigram, Bigram, Trigram, Quad-gram},
Lexical, Structural, Content and Syntactical	Word Length, Uppercase frequency, Common words, Word{Bigram, Trigram}, Part of Speech{Unigram, Bigram, Trigram, Quad-gram}, Punctuation Bigram, Character {Unigram, Bigram, Trigram, Quad-gram}, Sentence Length,

	Paragraph frequency
Syntactical and Structural	Punctuation Bigram, POST{Unigram, Bigram, Trigram, Quad-gram}, Paragraph frequency, Sentence length
Syntactical and Content	POST{Unigram, Bigram, Trigram, Quad-gram}, Punctuation Bigram, Common words; Word{Bigram, Trigram};
Syntactical, Structural and Content	Common words; Word {Bigram, Trigram}, Sentence Length, Paragraph frequency, Punctuation Bigram, POST {Unigram, Bigram, Trigram, Quad-gram}
Structural and Content	Paragraph frequency, Sentence length, Common words, Word {Bigram, Trigram}

Table 8 shows the initial evaluation results of the better performing writing style features in combined feature sets from table 7. The initial evaluation results in table 8 generated higher results than the individual feature sets results in table 6 because the better performing writing style features from table 5 were combined. A combination of writing style features that increase performance can create high results. 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* set had the highest results with an AUC of 0.751. The other sets that had higher results include *Lexical, Syntactical* and *Content* set with 0.762, *Syntactical* and *Content* had 0.740 as well as the *Lexical, Syntactical and Structural* set with 0.760. 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 *8*: 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
	SVM	0.664	0.688	67	0.665	0.33
	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
	SVM	0.786	0.500	63	0.643	0.27
	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
	SVM	0.786	0.563	67	0.674	0.34
	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
	SVM	0.780	0.740	76	**0.760**	0.52
	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
	SVM	0.722	0.688	70	0.705	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
	SVM	0.74	0.540	64	0.640	0.28
	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
	SVM	0.778	0.688	73	0.733	0.47
	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
	SVM	0.960	0.200	58	0.580	0.16

	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
	SVM	0.800	0.680	74	**0.740**	0.48
	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
	SVM	0.722	0.750	74	0.736	0.47
	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
	SVM	0.620	0.480	55	0.550	0.1
	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 features sets. In this case, the writing style features were removed and put back one by one from their combination feature sets to see whether they generate high or low results. 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 to 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. These common writing style features as well as other writing style features that were shown as ideal for performance are shown in table 9.

Table 9: The writing style features in the combination feature set used to increase performance

Combination Feature set category	**Writing Style Features used for performance**
Lexical and Syntactical	Word Length, Character {Bigram, Trigram, Quad-gram}; Parts of Speech Tag{ Trigram, Quad-gram}
Lexical and Structural	Word Length, Uppercase frequency; Character {Bigram, Trigram}, Sentence Length
Lexical and Content	Word length, Uppercase frequency; Character{Bigram, Trigram}; Word Trigram ; Word Length
Lexical, Structural and Syntactical	Parts of Speech Tag{Bigram, Trigram}; Character {Trigram, Quad-gram}; Common words; Uppercase frequency; Word Length; Punctuation frequency
Lexical, Structural and Content	Word Length; Common words, Paragraph count, Character {Bigram, Trigram}, Word Trigram; Part of Speech Tag{Bigram, Trigram}

Lexical, Syntactical and Content	Word Length; Punctuation Bigram, Common words; Character trigram; POST{Bigram, Trigram}; Word Trigram
Lexical, Structural, Content and Syntactical	Word Length, Common word, Part of Speech{ Bigram, Trigram}; Character{Bigram, Trigram}; Word Trigram; Sentence Length;
Syntactical and Structural	Sentence length, Paragraph frequency; POST{Bigram, Trigram }
Syntactical and Content	Punctuation Bigram, Common words; POST; Word Trigram; POST{Unigram, Bigram, Trigram, Quad-gram}
Syntactical, Structural and Content	Common words; Word Trigram; Sentence Length; Paragraph count; Punctuation bigram, POST{Bigram, Trigram, Quadgram}
Structural and Content	Paragraph count; Common words, Word Trigram

Table 10 shows the results of the feature set combination after the ablation process with the writing style features from table 9. The combination feature sets that had the highest results was the *Lexical, Syntactical, Structural* and *Content* set with an AUC of **0.837** and had impressively general high results of all the combination feature sets. 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 answers ***research question 3*** *(Which writing style features can be combined to work best for cross-genre and cross-topic documents in Authorship Identification?).* 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 results 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 Luyckx and Daelemans (2005) 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. To answer ***research question 4*** *(Do the results from this study generalise across the three different family of classifiers?),* the results in table 10 generally show across the classifiers that the combination feature sets that had Syntactical and Lexical features generated the highest results. The combination feature sets that had mostly Content features performed moderately and combination sets that had mostly Structural sets had the lowest results across the classifiers.

Table 10: 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
	SVM	0.760	0.760	76	0.760	0.52
	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
	SVM	0.786	0.563	66	0.674	0.34
	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
	SVM	0.857	0.563	70	0.710	0.41
	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
	SVM	0.820	0.680	75	0.750	0.5
	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
	SVM	0.860	0.700	78	**0.780**	0.56
	Random Forest	0.556	0.658	61	**0.781**	0.24
	Naïve Bayes	0.556	0.875	70	**0.795**	0.42

Lexical, Structural and Content	SVM	0.857	0.563	70	0.710	0.41
	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
	SVM	0.860	0.680	77	0.770	0.54
	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
	SVM	0.820	74	0.740	0.48	0.660
	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
	SVM	0.833	0.750	79	**0.792**	0.58
	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
	SVM	0.833	0.750	79	**0.792**	0.59
	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
	SVM	0.857	0.500	66	0.679	0.35

	Random Forest	0.643	0.563	60	0.670	0.20

5.4 Results Comparison with Related Works

In this section the dissertation results are compared with the related works. In particular, the dissertation had similar writing style features with Zheng et al., (2006) who performed Authorship Identification task on a single-genre dataset of online messages. In their work, Zheng et al., (2006) generated an AUC of 0.977. The similar writing style features used were ***character (uni to quad-grams), word length, frequency of function words, common word count, sentence length*** and ***punctuation frequency***. The dissertation obtained an AUC of 0.837 with these similar writing style features which means that the features can be used most effectively in cross-topic and cross-genre documents. The difference in writing style features is that Zheng et al., (2006) did not use were ***uppercase frequency, Parts of Speech Tag (uni to quad-grams)*** and ***word (bi and trigrams)*** that the dissertation used. Their data collection comprised of 20 authors each having on average 48 messages and 169 words in each message. The dissertation on the other hand had 100 documents each containing a known author document and an unknown (questioned) document with each document having 350 words on average per document. Zheng et al., (2006) had a thorough accuracy with more documents per author than the dissertation. Zheng et al., (2006) used Neural Network classifier in their experiment, whereas the dissertation used SVM, Random Forest and Naïve Bayes.

Gómez-Adorno et al., (2018) had cross-topic dataset made up of a single newspaper which had 50 text samples of Politics, Society, World, UK, and Book reviews topics from each of the 13 authors they had collected. However, the dissertation had cross-genres of books, essays and web forums in addition to the cross-topic dataset and had less author documents trained on. Gómez-Adorno et al., (2018) used the same ***character (uni to quad-grams), word (bi and trigrams)*** and ***Parts of Speech Tag (uni to quad-grams)*** writing style features as the dissertation to achieve their AUC result of 0.90. The difference in writing style features used is that Gómez-Adorno et al., (2018) did not use ***punctuation frequency, word count, common word count, function words*** and ***uppercase frequency*** which may probably be the reason for their higher AUC than the dissertation

obtained. This also shows that these writing style features are good for their dataset used. Gómez-Adorno et al., (2018) used the td-idf to represent the writing style features as well as the use of the Logistic regression classifier chosen to be used to create their model. The dissertation used more classifiers for its experiment, namely SVM, Random Forest and Naïve Bayes.

Another related work that used similar writing style features as the dissertation was Abbasi and Chen (2008) who achieved an AUC OF 0.85. Abbasi and Chen (2008) used cross-topic four datasets that comprised of emails, website comments, code and chats each having 100 authors. The dissertation on the other hand used a cross-topic and cross-genre dataset with a fewer number of documents per author. Abbasi and Chen (2008) used ***character (uni to quad-grams), word (bi and trigrams)***, ***function words, word-length distributions*** and ***Parts of Speech Tag (uni to quad-grams)***. The similar writing style features used showed that they are ideally used for cross-topic and cross-genre documents because of the results they generate. The difference in Abbasi and Chen (2008) work is that they did not use *Structural* ***sentence length*** and ***paragraph length*** writing style features in their experiment as the dissertation did. The dissertation results found that *Structural* writing style features produces the lowest results which is probably why Abbasi and Chen (2008) did not use them and had a higher AUC of 0.85. Abbasi and Chen (2008) had more documents per author for their experiment having a higher accuracy than the dissertation.

In 2015, Bagnall (2015) had the highest AUC result of 0.811 at the PAN CLEF. Bagnall (2015) worked on a cross-topic collection of 100 documents whereas the dissertation experimented on 100 documents in a cross-topic and cross-genre dataset. Bagnall (2015) only use of cross-topic documents where there is no evaluation for overall different genres of documents. Bagnall (2015) also only used ***character (uni to quad-grams)*** in his Neural Network work to predict word formation. The dissertation experiment also used ***character (uni to quad-grams)*** to calculate text. The ***character (uni to quad-grams)*** writing style feature shows to be a strong writing style feature because it is the only writing style feature used in Bagnall (2015) which generated a result of 0.811. However, Bagnall (2015) doesn't use any other writing style features in his work whereas the dissertation's use of more writing style features such as ***Parts of Speech Tag (uni to quad-grams)*** and ***word (bi and trigrams)*** which may have contributed to the dissertations result. The dissertation also uses SVM, Random Forest and Naïve Bayes for its experiment while Bagnall (2015) use of Neural Network concentrated on the formation of every. This thorough analysis may have assisted

in the generating of 0.811 with the singular use of ***character (uni to quad-grams)*** writing style feature.

Castro et al., (2015) achieved an AUC of 0.75 that worked on a cross-topic collection of 100 documents. The dissertation experimented on 100 documents in a cross-topic and cross-genre dataset whereas Castro et al., (2015) only used cross-topic documents which shows that there is no evaluation for overall different genres of documents. Castro et al., (2015) used the same writing style features as the dissertation namely ***character (uni to quad-grams)***, ***word (bi and trigrams)*** and ***Parts of Speech Tag (uni to quad-grams)***. The dissertation had more writing style features such as ***Function words, word length, character (uni to quad-gram), uppercase frequency*** and ***punctuation count.*** The dissertations use of more writing style features may have contributed to obtaining the result.

Table 11 shows the AUC results of related work in single genre/topic, cross-genre/topic documents and PAN CLEF Works in Authorship Identification works as compared to the dissertation results.

Table 11: AUC results of previous works of Authorship Identification.

Previous Works	**Document Comprehension**	**AUC**
Zheng et al., (2006)	Single Genre	0.977
Raju et al., (2017)	Single Genre	0.972
Viswanathan and Mooney (2014)	Single topic and Single Genre	0.956
Eder (2011)	Single Genre	0.95
Suh (2016)	Single Genre	0.945
Gómez-Adorno et al., (2018)	Cross Topic	0.90
McDonald et al., (2012)	Single Genre	0.86
Ramyaa and Rasheed (2004)	Single Genre	0.853
Lou et al., (2017)	Single Genre	0.85
Abbasi and Chen (2008)	Cross Topic	0.85
Dissertation results	**Cross topic and Cross Genre**	**0.837**
Ghaeini (2013)	Single Genre	0.837
Pavelec et al., (2009)	Single Genre	0.832
Kešelj et al., (2003)	Single Genre	0.83
Coyotl-Morales et al., (2006)	Single Genre	0.83
Bagnall (2015)	Cross Topic	0.81
Witten et al., (1999)	Single Genre	0.80
Seidman (2013)	Single Topic	0.792
Solorio et al., (2011)	Single Topic and Single Genre	0.774
Ruseti and Rebedea (2012)	Single Genre	0.77
Moreau and Vogel (2013)	Single Topic	0.767
Pacheco et al., (2015)	Cross Topic	0.763
Halvani and Winter (2015)	Cross Topic	0.762
Feng and Hirst (2013)	Single Topic	0.75
Castro et al., (2015)	Cross Topic	0.75
Khonji and Iraqi (2014)	Cross Genre	0.75
Guiterrez et al., (2015)	Cross Topic	0.74
Kosher and Savoy (2012)	Cross Genre	0.738
Tanguy et al., (2011)	Single Genre	0.737
Jankowska et al., (2013)	Single Topic	0.733
Fréry et al., (2014)	Cross Genre	0.723
Moreau et al., (2014)	Cross Genre	0.703

6 CONCLUSION AND FUTURE WORK

6.1 Conclusion

The dissertation proposed to identify and evaluate the writing style features to be used in Authorship Identification for cross-topic and cross-genre documents. There have been a few related work that evaluate writing style features for Authorship Identification for cross-topic and cross-genre documents. The remainder of this section discussed are as follows.

The **Background** of the dissertation in chapter 2 discusses the classification techniques namely Tree based (Decision tree), Kernel (SVM) and Probabilistic based (Naïve Bayes) methods used in the dissertation. The document representations include data pre-processing techniques such as Normalising, Tokenising and Stemming and feature extraction that are performed on the documents to be processed by a learning method. The various evaluation methods that assess how well classification is performed for a task were also reviewed such as the Kappa coefficient, Accuracy, Sensitivity, Specificity and ROC (AUC).

In chapter 3, a review is done of the **writing style features** that have been used in Authorship Identification task since the nineteenth century. Certain writing style features were used in Authorship Identification for single-topic and single-genre documents because of the expressions and words attached to a domain. The increased variety of topic and genre also increased the variety of writing style features to be used for Authorship Identification in multiple topic/genre documents.

The study reviewed successful related works and extracted the writing style features they had used 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 in most of the successful previous studies 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 related works 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 Cross Industry Standard Process for Data Mining (CRISP-DM) was presented in chapter 4 as the methodology the dissertation followed. The nature of the Authorship Identification task is a data mining process therefore a model had to be followed to complete the process. The dissertation was mapped according to the CRISP-DM phases, for example, Business understanding phase clearly defined the dissertations' objectives and the Modelling phase planed how the classification techniques are to be used and parameter setup.

The data preparation phase of the CRISP-DM model describes the evaluation dataset used in the dissertation was taken from the PAN CLEF 2015 English training phase which comprised of 100 documents where each document consists of a known author sample and an unknown sample for authorship comparison. The texts in the documents were categorised into numeric writing style features and loaded into WEKA. The data preparation phase also explains the experiment implemented an ablation process to monitor the writing style features that could improve or decrease performance. This is done by removing a writing style feature and putting it back into the feature set to see how it affects performance.

The **results and analysis** in chapter 5 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 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. The overall performance of the results

shows that the highest generated results were from Naïve Bayes, followed by Random Forest and then lastly SVM. This generalisation does not mean that SVM produces the lowest results, rather in terms of performance of the writing style feature sets.

6.2 Discussions for Future Work

As part of future work, there are some aspects of the empirical evaluation that may require some further investigations and improvements made to the experimental setup to get higher results. The following could be explored in future studies;

- An improvement could be adding more writing style features from related work to cover all possible writing style features as being ideal for Authorship Identification in cross-genre and cross-topic documents. Writing style features such as the occurrence of special character (e.g., @#$%ˆ) and letter count (e.g., a, b, c). Additional feature set category such as Idiosyncratic, which has misspellings as a feature (e.g., “beleive”, “thougth”) can represent an authors’ common spelling mistake that they make.

- An investigation can be made on whether a larger dataset consisting of 1000 documents could possibly generate the same impressive results generated as the empirical evaluation. The documents used in the experiment comprised of only one known document and one unknown sample, which raises the idea whether it would be more effective if more known documents could be used produce better results.

- The ablation process used in the experiment was done manually, later work will use an automatic ablation process which will be very convenient for many writing style features.

- Another plan as part of future work is to apply a clustered approach and compare the results with the classification approach the dissertation was using to see which learning method produces a higher recall rate and is better used in Authorship Identification.

- Possible improvements in the experimental setup include a finer grid search in parameter pairing value selection in Cost and Loss for SVM classifier as well as gamma parameters for Random Forest which could acquire better classification accuracy.

7 REFERENCE

ABBASI, A. and CHEN, H. (2008). Writeprints*: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS),* 26**,** 7.

ARGAMON, S., and JUOLA, P. (2011), *Overview of the International Authorship Identification Competition at PAN-2011*. Paper presented at the CLEF (Notebook Papers/Labs/Workshop).

BADDAR, M. (2015). A Framework for Text Classification using IBM SPSS Modeler. *IBM, Learning Center, 11*.

BAGNALL, D. (2015). Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*.

BARTOLI, A., DAGRI, A., LORENZO, A., MEDVET, E., and TARLAO, F. (2015). An author verification approach based on differential features. *Working Notes Papers of the CLEF*.

BOCKO (2015). WEB BASED DATA-MINING ASSISTANT, P. J. Safarik University, Faculty of Science; Košice.

BOZKURT, I. N., BAGLIOGLU, O., and UYAR, E. (2007). *Authorship attribution. Performance of various features and classification methods*. Presented at the Computer and Information science 2007. ISCIS 2007. 22nd international symposium on.

BRADLEY, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145-1159.

BREIMAN, L.: Random forests. Machine learning.(2001). 45(1), 5–32.

CARNERUD, D. (2014). Exploration of text mining methodology through investigation of QMOD-ICQSS proceedings. In QMOD-ICQSS Prague, Czech Republic, 2014 (Sep. 3-5).

CASTRO, D., ADAME, Y., BRIOSO, M. P., and MUÑOZ, R. (2015). *Authorship Verification, combining Linguistic Features and Different Similarity Functions.* Paper presented at the CLEF (Working Notes).

CHAPMAN, P., CLINTION, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., and WIRTH, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

CLEARY, J. G., & TEAHAN, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, *40*(2_and_3), 67-75.

COYOTL-MORALES, R. M., VILLASENOR-PINEDA, L., MONTES-y-GOMEZ, M., & ROSSO, P. (2006). Authorship attribution using word sequences. In Iberoamerican Congress on Pattern Recognition (pp. 844-853). Springer Berlin Heidelberg.

de WAAL, A., VENTER, J., & BARNARD, E. (2008). Applying topic modeling to forensic

data. In IFIP International Conference on Digital Forensics (pp. 115-126). Springer US.
DESHPANDE, S. and THAKARE, D. V. 2010. Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS),* 1**,** 32-44.
ELAYIDOM, M. S., JOSE, C., PUTHUSSERY, A., and SASI, N. K. (2013). *Text classification for authorship attribution analysis*. *arXiv preprint arXiv:1310.4909*.
ESTIVAL, D. (2008). Author attribution with email messages. *Journal of Science, Vietnam National University*, *1*, 1-9.
FENG, V. W., and HIRST, G. (2013). *Authorship verification with entity coherence and other rich linguistic features notebook for PAN at CLEF 2013*.
FISSETTE, M.V.M. (2010). *Author Identifcation in Short Texts.*
Fréry, J., Largeron, C., & Juganaru-Mathieu, M. (2014, September). *UJM at CLEF in Author Verification based on optimized classification trees*. In *CLEF 2014* (p. 7p).
Fu, Y. (2017). *Combination of Random Forests and Neural Networks in Social Lending*. *Journal of Financial Risk Management*, *6*(04), 418.
GAIGOLE, P. C., PATIL, L. H., and CHAUDHARI, P. M. (2013). *Preprocessing Techniques in Text catagorization. In National Conference on Innovative Paradigms in Engineering and Technology (NVIPET-2013),* Proceedings published by International Journal of Computer Applications (IJCA).
GHAEINI, M. (2013). *Intrinsic author identification using modified weighted knn*. Notebook for PAN at CLEF.
Gómez-Adorno, H., Posadas-Durán, J. P., Sidorov, G., & Pinto, D. (2018). *Document embeddings learned on various types of n-grams for cross-topic authorship attribution*. Computing, 1-16.
GREEN, R. M., and SHEPPARD, J. W. (2013). *Comparing Frequency-and Style-Based Features for Twitter Author Identification*. In FLAIRS Conference.
GUTIERREZ, J., CASILLAS, J., LEDESMA, P., FUENTES, G., and MEZA, I. (2015). *Homotopy Based Classification for Author Verification Task. Working Notes Papers of the CLEF*.
HALVANI, O. and WINTER, C. (2015). *A Generic Authorship Verification Scheme Based on Equal Error Rates*. Working Notes Papers of the CLEF.
HANLEIN. H. (1999).“Studies in Authorship Recognition: a Corpus-based Approach”. Peter Lang.
HILL, T. and LEWICKI, P. (2007). *STATISTICS: Methods and Applications*. StatSoft, Tulsa, OK.
HOWEDI, F., and MOHD, M. (2014). Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data. *Computer Engineering and Intelligent Systems*, *5*(4), 48-56.
HSU, C.-W., CHANG, C.-C., and LIN, C.-J. (2003). A practical guide to support vector classification.
INCHES, G. and CRESTANI, F.(2012). *Overview of the International Sexual Predator Identification Competition at PAN-2012*. CLEF (Online Working Notes/Labs/Workshop).
JANKOWSKA, M., KESELJI, V. L. A. D. O., & MILIOS, E. (2013). *Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task.* In CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers (pp. 23-26).
JUOLA, P. and STAMATATOS, E. (2013). *Overview of the author identification task at pan Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. 4th International Conference of the CLEF Initiative, CLEF, 2013. 23-26.

KESELJ, V., PENG, F., CERCONE, N., & THOMAS, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING* (Vol. 3, pp. 255-264).

KHONJI, M. and IRAQI, Y. (2014). A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF). *Notebook for PAN at CLEF*.

KOCHER, M. and SAVOY, J. 2015. UniNE at CLEF 2015: Author Identification. *Working Notes Papers of the CLEF*.

LAHIRI, S., & MIHALCEA, R. (2013). Authorship attribution using word network features. *arXiv preprint arXiv:1311.2978*.

LOU et al.(2017). *Which Author Authored Which: Predicting Authorship from Text Excerpts.* University of Stanford.Los Angeles.

LUYCKX, K. and DAELEMANS, W. 2005. *Shallow Text Analysis and Machine Learning for Authorship Attribtion*. LOT Occasional Series, 4, 149-160.

MCDONALD, A. W., AFROZ, S., CALISKAN, A., STOLERMAN, A. and GREENSTADT, R. (2012).Use fewer instances of the letter "i": Toward writing style anonymization. Privacy Enhancing Technologies, 2012. Springer, 299-318.

MENDAENHALL, T. C. (1887). The characteristic curves of composition. Science, IX, 237–49.MOHAMMAD, S. 2003. *Combining lexical and syntactic features for supervised word sense disambiguation.* UNIVERSITY OF MINNESOTA.

MOHTASEEB, H., & AHMED, A. (2009). More blogging features for author identification.

MOREAU, E., JAYAPAL, A., and VOGEL, C. (2014). Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm-Notebook for PAN at CLEF 2014. Paper presented at the Working Notes for CLEF 2014 Conference.

MOREAU, E., & VOGEL, C. (2013, September). 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).

MOSTELLER, F. and WALLACE, D.L. (1964). *Inference and disputed authorship: The Federalist. Addison-Wesley*.

NIRKHI, S. and DHARASKAR, R. V. (2013). Comparative study of authorship identification techniques for cyber forensics analysis. arXiv preprint arXiv:1401.6118.

OZHGüR, A. (2004). *Supervised and unsupervised machine learning techniques for text document categorization.* Citeseer.

PACHECO, M. L., FERNANDES, K. and PORCO, A. (2015). Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. *Working Notes Papers of the CLEF*.

Pavelec, D., Oliveira, L. S., Justino, E., Neto, F. N., & Batista, L. V. (2009, June). Compression and stylometry for author identification. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on* (pp. 2445-2450). IEEE.

PIMAS, O., KROLL, M., and KERN, R. (2015). Know-Center at PAN 2015 author identification. *Working Notes Papers of the CLEF*.

Polamuri.S.(2017).Machine Learning. Dataaspirint. forest-algorithm-machine-learing/

Raju, N. G., Tejaswini, P., & Mounica, Y. (2017). *Style based Authorship Attribution on English Editorial Documents*. International Journal of Computer Applications, 159(4).

RAMYAA, C. H. and RASHEED, K. *Using machine learning techniques for stylometry*. Proceedings of International Conference on Machine Learning, 2004.

RASCHKA, S. (2014). *Naive bayes and text classification i-introduction and theory*. arXiv preprint arXiv:1410.5329.

RASCHKA, S. (2015). Machine Learning FAQ. https://sebastianraschka.com/faq/docs/evaluate-a-model.html.

ROSSO, P., RANGEL, F., POTTHAST, M., STAMATATOS, E., TSCHUGGNALL, M., and STEIN, B. (2016). *Overview of PAN'16.* Paper presented at the International Conference of the Cross-Language Evaluation Forum for European Languages.

RUSETI, S., & REBEDEA, T. (2012). Authorship Identification Using a Reduced Set of Linguistic Features. In *CLEF (Online Working Notes/Labs/Workshop)*.

SARI, Y., and STEVENSON, M. (2015). A Machine Learning-based Intrinsic Method for Cross topic and Cross-genre Authorship Verification. *Working Notes Papers of the CLEF*.

SEIDMAN, S.(2013).*Authorship verification using the impostors method*. CLEF 2013 Evaluation Labs and Workshop-Online Working Notes, 2013.

SILIPO. R, ZIMMER.M.A (2015). *Data and Machine Architecture for the Data Science Lab Workflow Development, Testing, and Production for Model Training, Evaluation, and Deployment*. Copyright © 2015 by KNIME.com AG.

SOKOLOVA, M., JAPKOWICZ, N. and SZPAKOWICZ, S. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence.* Springer.

Solorio, T., Pillay, S., Raghavan, S., & Montes-Gomez, M. (2011). Modality specific meta features for authorship attribution in web forum posts. In *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 156-164).

STAMATATOS, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology,* 60**,** 538-556.

STAMATATOS, E., DAELEMANS, W., VERHOEVEN, B., POTTHAST, M., STEIN, B., JUOLA, P., SANCHEZ-PEREZ, M. A. and BARRÓN-CEDEÑO, A. (2014). Overview of the Author Identification Task at PAN 2014. *analysis,* 13**,** 31.

STAMATATOS, E., DAELEMANS, W., VERHOEVEN, B., POTTHAST, M., STEIN, B., JUOLA, P., Lopez-Lopez, A., Stein, B.(2015). Overview of the Author Identification Task at PAN 2015. analysis.

STAMATATOS, E., POTTHAST, M., RANGEL, F., ROSSO, P. and STEIN, B. (2015). Overview of the PAN/CLEF 2015 Evaluation Lab. *Experimental IR Meets Multilinguality, Multimodality, and Interaction.* Springer.

Suh, J. H. (2016). Comparing writing style feature-based classification methods for estimating user reputations in social media. *SpringerPlus*, *5*(1), 261.

TAN, P. N., STEINBACH, M., and KUMAR, V. (2006). Classification: basic concepts, decision trees, and model evaluation. *Introduction to data mining*, *1*, 145-205.

Tanguy, L., Urieli, A., Calderone, B., Hathout, N., & Sajous, F. (2011, September). A multitude of linguistically-rich features for authorship attribution. In *PAN Lab at CLEF*.

THAI, K. M., NHUYEN, T. Q., NGO, T. D., TRAN, T. D., and HUYNH, T. N. P. (2012). *A support vector machine classification model for benzo [c] phenathridine analogues with topoisomerase-I inhibitory activity. Molecules*, 17(4), 4560-4582.

Viswanathan, V., Mooney, R., & Ghosh, J. Detecting Useful Business Reviews using Stylometric Features.

WIRTH, R. and HIPP, J. CRISP-DM.(2000): Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. Citeseer, 29-39.

WITTEN, I. H., BRAY, Z., MAHOUI, M., & TEAHAN, B. (1999). Text mining: A new frontier for lossless compression. In *Data Compression Conference, 1999. Proceedings. DCC'99* (pp. 198-207). IEEE.

XIER, L. (2010). *Kappa—A Critical Review*. Uppsala University, Sweden.

YEDIDIA. A.(2016).*Against the F-score*-WordPress.com

YULE, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

ZIPF, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press, Cambridge, MA.

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 for information Science and Technology, 57*(3), 378-393.