ALGORITHM FOR ROBUST AUTHORSHIP ATTRIBUTION WITH OTPIMUM FEATURE SELECTION

A THESIS SUBMITTED TO

SAVITRIBAI PHULE PUNE UNIVERSITY

FOR THE AWARD OF DEGREE OF

DOCTOR OF PHILOSOPHY (PH.D.)

(COMPUTER ENGINEERING)

IN THE FACULTY OF

SCIENCE AND TECHNOLOGY

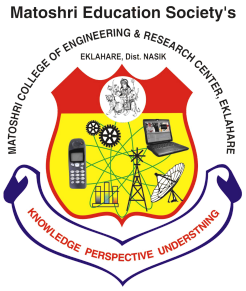
SUBMITTED BY

MUBIN SHOUKAT TAMBOLI

UNDER THE GUIDANCE OF

DR. RAJESH S. PRASAD

RESEARCH CENTRE



DEPARTMENT OF COMPUTER ENGINEERING

MATOSHRI COLLEGE OF ENGINEERING AND RESEARCH CENTER NASHIK, INDIA

JANUARY 2020

MATOSHRI COLLEGE OF ENGINEERING AND RESEARCH CENTER, Nashik

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that, the work incorporated in the thesis, **"Algorithm for Robust Authorship Attribution with Optimum Feature Selection"** is submitted by **Mubin Shaukat Tamboli** for the **Doctor of Philosophy (Ph.D)** in **Computer Engineering, Savitribai Phule Pune University**, has been carried out by the candidate at **Department of Computer Engineering, Matoshri College of Engineering and Research Centre Eklahare, Nashik** during the period from **August 2014 to July 2019** under the guidance of **Dr. Rajesh S. Prasad**.

Prof.(Dr.) Varsha Hemant Patil
Head of Department,
Computer Engineering,
MCOERC, Eklahare, Nashik

Prof.(Dr.) Gajanan K. Kharate
Principal,
MCOERC, Eklahare, Nashik

CERTIFICATE OF GUIDE

This is to certify that, the work incorporated in the thesis **"Algorithm for Robust Authorship Attribution with Optimum Feature Selection"** submitted by **Mubin Shoukat Tamboli** was carried out by the candidate for the Doctor of Philosophy (Ph. D) degree at **Department of Computer Engineering, Matoshri College of Engineering and Research Centre, Eklahare, Nashik** during the period from **August 2014 to January 2020** under my direct supervision and guidance.

Date:
Place:

Prof.(Dr.) Rajesh S. Prasad
Guide

DECLARATION BY THE CANDIDATE

I hereby declare that the thesis entitled **"Algorithm for Robust Authorship Attribution with Optimum Feature Selection"** submitted by me to the **Savitribai Phule Pune University, Pune** for the degree of **Doctor of Philosophy (Ph.D)** in **Computer Engineering,** is the record of work carried out by me during the period from **August 2014 to July 2019** under the guidance of **Dr. Rajesh S. Prasad** and has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other Institution of Higher learning. I further declare that the material obtained from other sources has been duly acknowledged in the thesis.

Date:
Place: Nashik

Mubin Shoukat Tamboli
(Eligibility No.: 12015232880)

ABSTRACT

Over the last few decades, there has been a tremendous growth in online communication which generate countless textual snippets. In these communication neither identities nor role of user is known. Under such circumstances it becomes a critical issue of identity tracing in forensic investigation. The identity of the user of such snippet is known by applying a text mining technique, named as 'authorship analysis'. Authorship analysis is possible due to statistical and computational methods. In the written document each author represent himself by distinctive writing style and it is extracted in terms of features.

Authorship analysis is of three types i) Author identification ii) Author verification and iii) Author characterization. Author identification is task to find ghostwriter of unknown text snippet. From the perspective of machine learning it is multiclass, single label text categorization task. Existing techniques has outlined the different methodologies and their improvements for the identification of anonymous authors using stylometry in the form of linguistic features. However a novel algorithm must consider the change in writing style over time period to improve author identification. This thesis provides a novel solution for authorship attribution which considers change in writing style of the author. This research work is carried out in two phases: i) To identify change in writing style of writer with respect to time and ii) To mitigate this change by a novel feature normalization technique.

A novel Transform Feature to Current Time function is presented for the normalization of features, where features are shifted to current time and given to the classification model building phase. Three types of features are used, namely, character n-gram, word n-gram, and PoS n-gram for the empirical evaluation. A SVM algorithm is used to attribute the real author. This novel system is successfully implemented and evaluated on a set of text samples written by several authors which were collected over a different time period and it results the identification of correct author. The research work shows that there is a remarkable improvement in the performance as compared to existing author identification techniques.

Another approach is presented to use word n-gram in author identification. Existing methods for word n-gram, keep the value of n constant, and then it is used for feature construction. Further, it is used in the authorship identification task. A novel approach presented in this thesis does not depend on the constant value of n; the value changes according to the occurrence of the word length of the current word token. The methodology is used on the collection of the text snippets, which are varied in the time domain. Dynamic

value of n is chosen to generate word sequences. The result shows that there is a significant improvement in accuracy when it is compared with the fixed value of n in word n-gram.

Each of these approaches are demonstrated in the research work and it shows notable improvement in the performance of the author identification task.

Keywords: *text mining, author identification, feature selection, feature extraction, feature transformation function, machine learning, n-gram*

ACKNOWLEDGEMENT

It gives me immense pleasure to acknowledge the help extended by many people to have directly and indirectly contributed for successful completion of this research work.

Gratitude is one of the hardest emotion to express and I sincerely dedicate this heartfelt feeling to my guide Dr. R. S. Prasad. He has influenced my view of the research process, and instilled in me the importance of aiming to produce quality research with the potential for impact. Most of all, I value his honest opinions, critics, calmness and clarity of advice during difficult times, and his patience and understanding over the past few years. I am indebted to have had guide who gave me all of the resources, guidance and support I could ever need during the period that lead up to this thesis completion.

My sincere thanks to Principal of our research center Dr. G. K. Kharate, whose valuable comments proved as a great help in my research work. His sound advice, support and few words of encouragement were used to fill me with enough energy to overcome all the obstacles. I owe special thanks to Dr. V. H. Patil our head of department and research coordinator for constant guidance and helpful nature. Thank you for making our days brighter with stimulating discussions, comment and insights on the topic.

I would also like to thank many of my colleagues at Amrutvahini College of Engineering, Sangamner for kind support and cooperation during this journey.

Finally, I am deeply indebted to my dear parents and family members for their love and encouragement throughout the years. I'm infinitely grateful for the values they have passed down to me, and for their continuous support throughout all my studies.

Mubin Shoukat Tamboli

Table of Contents

List of Figures

List of Tables

Abbreviations

SCAP	Source Code Author Profile
NLP	Natural Language Processing
PoS	Part of Speech
OCR	Optical Character Recognition
NP	Noun Phrase
PP	Preposition Phrase
VP	Verb Phrase
HTML	HyperText Markup Language
XML	Extensible Markup Language
KNN	K-Nearest Neighbors
RAR	Roshal Archive
LZW	Lempel–Ziv–Welch
GZIP	GNU zip
BZIP	Burrows-Wheeler transform ZIP
PCA	Principal Component Analysis
SVM	Support Vector Machine
GA	Genetic Algorithm
RDM	Recursive Data Mining
RCV1	Reuters Corpus Volume I
NN	Nearest Neighbour
PMSVM	Power Mean Support Vector Machine
SVC	Support Vector Classifier
TP	True Positive
TN	True Negative
FP	False Positive

FN	False Negative
TPR	True Postive Rate
FPR	False Positive Rate
NLTK	Natural Language Tool Kit
ASCII	American Standard Code for Information Interchange
URL	Uniform Resource Locator
TFCT	Tranform Feature to Current Time
SMO	Support Vector Machine Optimization

Chapter 1

Introduction

The information generated by many variants of digital devices such as computer system, tablets, mobiles which get connected to the world through the Internet. Countless information generated in second in the form of text over the web. This gives a pathway to rogue users to initiate malicious activities. These critical activities is then turn into cybercrime. Under such circumstances, there is a need to know who is communicating? As suspect always hide himself through false or nameless identity in communication. It can be restricted by identifying the participant in the dialogue. The on-line textual content repudiated by knowing who has written that. Most commonly online content is available in the form of email, chats, articles, forum post, blogs, reviews, social media sharings etc. Depending on source of the text, size of text gets varied. In some forensic cases disputed authorship need to be identified. This scenario brings a new concept to identify such users through authorship analysis.

Authorship attribution is a technique where the writing style of an author is presented in the form of linguistic features of textual content, and then it is decided that the author has written the content or not? Over the decades, various computational methodologies and linguistic stylometry methods have been studied by several researchers. A typical system of author identification shown in figure 1.1. In this system there are a set of known text sample for different authors, and the problem is to identify the author of unknown text. Who has written that text sample? The process is to find the authorship of unknown text from a set of existing writing samples of authors is called author identification.

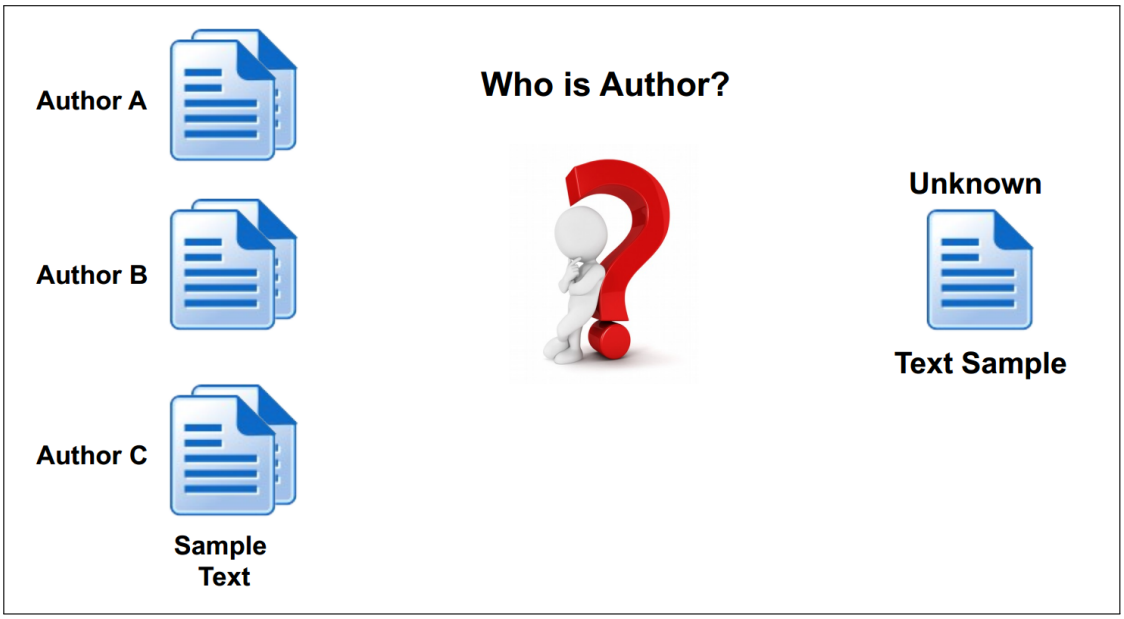


FIGURE 1.1: Author identification task

Nowadays, author identification has been repudiated by various agencies such as forensic science, research-funding, investigation, court to find guilt. Authorship of an unknown text is identified by analyzing the stylometry features of an author.

The content of the text and the writing style is responsible to retrieve the stylometry of an author. There are five different types of features are used to identify writing style of an author. Texts are split into characters or words to form tokens and that act as lexical type of features. Syntax-based features are in terms of the syntax used in writing such as tense, part of speech. Semantic type of features are based on semantics used in the text.

Application-specific type of features are based on the application and it extracted with the help of tools. These type of feature represented in terms of structural measures to get the author style. There are two different techniques used to find the authorship, namely, Profile-based and instance-based. In the profile-based approach is characterized by the similarity index to find the different between two texts. A machine learning and similarity based discrimination algorithms can be applied on instance-based approach.

Every author uses a unique style in his writing. It can be used to discriminate writings of different authors. But this writing style gets evolved over time hence it will get changed over time. Several factors are responsible for this change and those are age,

education, mother tongue, nationality, behavior and literacy. The change due to these factors are discussed by researchers but very few works go towards the solution.

This thesis provides two novel methodology for author identification system. In the first contribution, the solution is provided to identification task where writing style of author gets change with respect to the time. So this time parameter is taken into consideration for solving the problem. In the second contribution word gram approach is followed for the author identification task. To work with, an approach of variable length word gram is revealed.

1.1 Motivation

The growth of the Internet is full-on, and now every type of communication occurs through digital media. Textual communication over the Internet is anonymous, causing critical problems in identity tracing. This can be restricted by knowing the real writer of communication. The writer of unknown text can be known by authorship analysis. There is a challenge to find the author of such content through the analysis of communicated content. In authorship analysis, the characteristic of the writer is identified with the linguistic writing style.

Though a significant amount of research exists in the area of authorship attribution, still there is scope in this area where the performance of the system depends on number of authors, text sample size, algorithms, and methodology used. In the literature researcher have identified that the writing style of the author is affected due to age, nationality, literacy, and effect of their behavior. Existing studies quoted that the writing style of the author changes over time, but detail analysis is not done. Still, there is an open research problem to identify the owner where the writing style is changing over time.

1.2 Research Statement and Objectives

A systematic review of existing research work in the era of authorship analysis is done. In the survey, we came across various challenges regarding selecting a suitable dataset, to verify the impact of various type of features and the impact of time in the identification task. With consideration of all these facts, the problem statement is formulated as is, To

develop Algorithm for a Robust Authorship Attribution with Optimum Feature Selection. Research objectives of the proposed research work are as follows:

1. To study and review the work of feature selection construction and classification for author identification.
2. To identify and select features which dynamically change over time as style of author evolves and propose model for author identification on such selected features.
3. To implement and analyze the impact of selected feature and proposed model on authorship attribution.
4. To Validate the results.

1.3 Research Contributions

The main idea behind this research work is to build a system capable of handling the identification of true writers for unknown text, which beats the impact of change in writing style over time. This impact is removed in a novel authorship attribution system presented in this research work. Following are the summarized contributions of the work presented in this thesis:

1. This thesis presents extensive survey on author identification system where different facts of the system are studied. The effect of time reflected in the writing style of the author is identified and illustrated in this work.
2. In author identification, the effect of writing style over time is identified by regression function, and it is analyzed with different performance parameters. Changes in a style is calculated by a comparative study of the history of writing over a big-time period, which can be identified with a coefficient factor. A novel authorship attribution system is proposed which handles this change in writing style. The system is evaluated in terms of accuracy, precision, recall, f-measure, and variation in the accuracy of feature size.
3. Several author identification approaches reviewed in this thesis, which does not consider the variation that occurred in writing style due to various factors. This work

proves that there is an impact of time over writing style. In this research, a decay function is used to transform features used to overcome the effect of time on writing style.

4. The writing style of the author is extracted in terms of features, different types of features are reviewed in this thesis. A set of features are identified and used in the novel author attribution system.

5. This thesis describes the methodology to find a function that is responsible for translating the extracted features to the most recent time with the decay function. Algorithms are proposed for author identification through this feature transformation method. In our methodology, the features are transformed to the latest time period. This technique is never claimed so far in the available literature.

6. As per our study is concerned the dataset for the testing performance could not be found.

7. The experimentation system is evaluated with a machine learning discrimination algorithm and results obtained in terms of various performance parameters. The system is evaluated with three types of features named as characters, word, and PoS n-gram types.

8. In the literature, a machine learning algorithm is not used in the time-aware author attribution system. With proposed methodology, a novel method is developed and it uses a machine learning algorithm for the discrimination in time-based authorship attribution.

9. The research method used in this work gives higher accuracy than the existing methods with respect to time. Each author is evaluated with performance parameters.

10. A novel mutli-word gram method is implemented in this thesis. A performance is measured with respect to the use of stop words.

11. Finally, The effectiveness of the novel technique of author identification using TFCT function is validated in this work. The research also describes the implementation and evaluation of variable length word n-gram.

1.4 Organization of Thesis

The thesis is presented with five chapters and further divided in sections and subsections. Following is the overview of each chapter in this thesis.

Chapter 1 of the thesis provides an introduction about the need to know the authorship of digital communication and major work carried in author identification in digital forensics. It presents the motivation of the work and research objectives. Brief research contributions are described in this chapter.

Chapter 2 includes a detailed study of author identification. In this thesis, a survey is carried out in different types of features and discrimination of writing samples, according to the author. A novel state-of-the-art methodology, along with the performance parameters, are discussed in this chapter. It also focused on the existing approaches of author identification where time is a parameter. The basic framework of time-based author identification and ways to identify the effect of writing style are discussed.

Thereafter, chapter is focused on the challenges and limitations of the existing system through literature.

Chapter 3 gives detail about the new time-based author identification system framework. The drift in writing style of author is captured from the history of authors writing. A decay value that identifies the relationship between time and writing style is calculated and used to find the factor for transforming features in the form of TFCT function. The algorithm of a novel approach of variable length word n-gram is presented in this chapter.

Chapter 4 presents the experimentation setup, experiments and result analysis. The system is evaluated with performance parameters from the obtained output of the proposed method. Different facts are used to verify performance, such as feature count and the effect of stop-words. The result is compared with existing author identification systems. In comparison, the accuracy and the other parameters such as FP rate, precision, recall, F-measure are used.

Chapter 5 discusses the conclusion of the entire research work carried out. Guidelines to enhance the work towards the betterment of the system are also elaborated in this chapter.

Chapter 2

Literature Survey

2.1 Introduction

Authorship identification is a task of prediction where an unknown sample of writing attributed to a writer. It is in the field of text mining, where characteristics of the authors are mined from the writing samples of the known author and then these characteristics are used to predict authorship of unknown writing samples. Sometimes it is termed as fingerprinting. Author identification is one of the oldest problem which has numerous solutions explored by many researchers and still it is an open research problem. In this task, the process begins with extracting the characteristics of a writer from the written content.

These characteristics are in terms of stylometry, which are captured from the content of writings. Author identification plays a vital role in anonymous online communications, where to find the real author from written content is a challenge. The main challenge is to find a suspect from evidences. Each sample of writing acts as evidence, and a set of authors act as suspects, one suspect is to be find out from unknown written sample.

The figure 2.4 shows the general framework of author identification. There are different ways to consider evidences, which we will see in the upcoming sections. The first attempt was made in the 19th century to extract the writing style in the study of Mendelhall (1887) on the plays of Shakespeare. The facts were detailed in 1964 by Mosteller and Wallace. It was the initial footstep towards the characterizing the authorship.

2.2 Authorship Analysis

In the survey of Authorship attribution [1–3] has worked out in three domains named as author identification, authorship characterization, and similarity detection, as shown in figure 2.1.

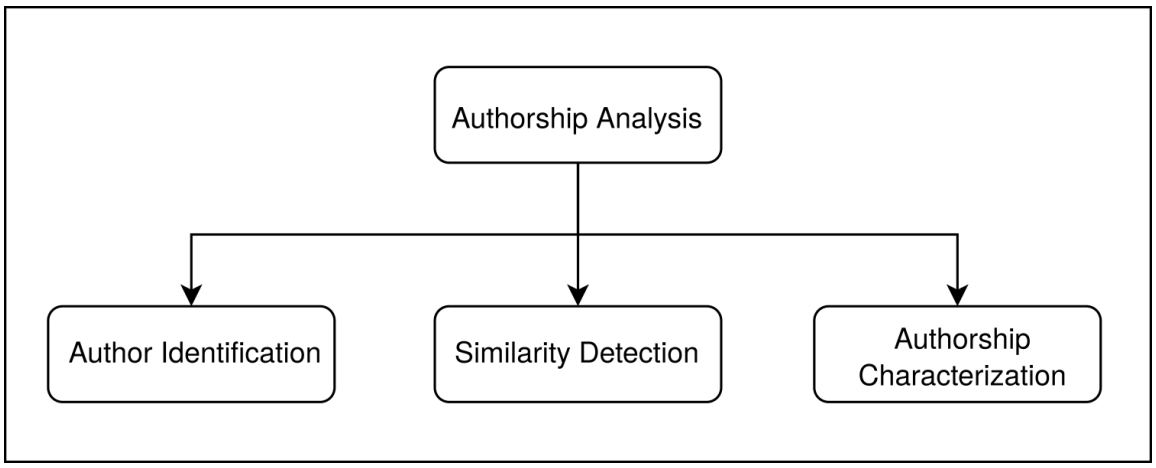


FIGURE 2.1: Authorship Analysis Domain

2.2.1 Author Identification

In this task, the main objective is to find the most plausible author of unknown text document from given set of authors. A typical representation of the system shown in figure 1.1 Anonymous text is compared with the text sample of all known authors to find the most probable writer. A traditional classifier model is built by extracting stylometry features from the sample documents. The style of the author is irrespective of the document content. This type of problem lies in one vs. the rest classes where each class is corresponding to a writer (author). In this type of problem most plausible author of an unknown sample is to find out. Identification problems were solved by supervised and unsupervised classification methods [4, 5].

2.2.2 Author Characterization

In this type, the profile of each author is built with the evidence of writing samples. The profile is represented in terms of the language model, which is created by accumulating characteristics of each writer. These characteristics are sometimes called as stylometry,

which is a direct and indirect form of contents of the sample text. The extracted characteristics of writer is a profile in terms of gender, age group, education, nationality, etc. [6, 7]. A system of author characterization is shown in figure 2.2

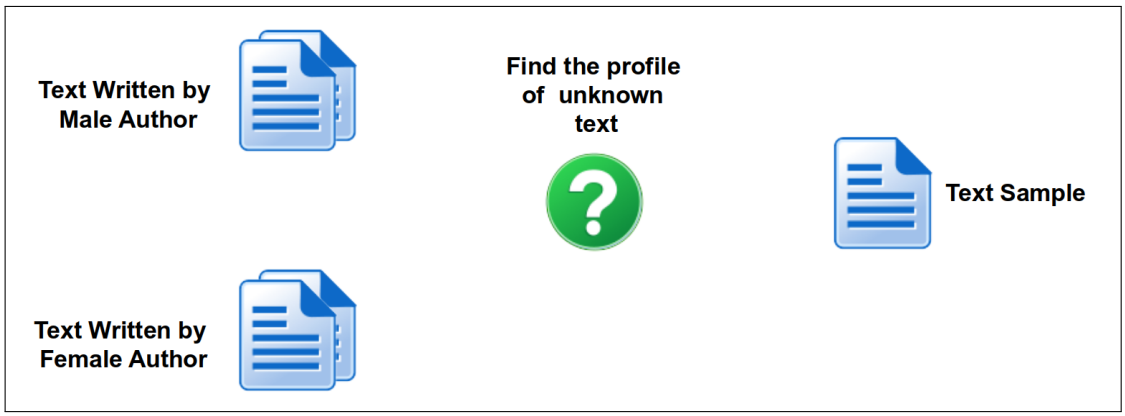


FIGURE 2.2: Author characterization/profiling task

2.2.3 Similarity Detection

Similarity detection is not directly related to the problem but can be a part. Here, how two writing samples are similar to find out in the extent of similarity.

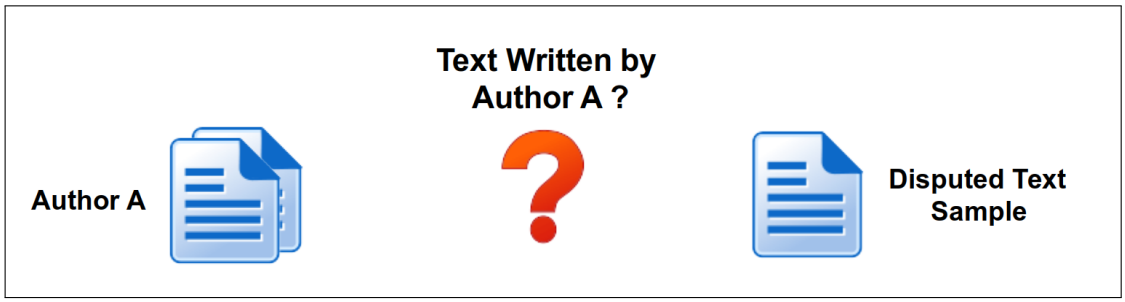


FIGURE 2.3: Author similarity detection task

A typical author similarity detection system is shown in figure 2.3. In this task, a set of author sample are given and task is to recognize that unknown text is written by same author or not? In the author identification problem, the work of different writers is compared with a single author to get its similarity score. This task uses the unsupervised method to find the suspect as no prior information is available. Plagiarism detection is an example of a similarity detection task where one's work is compared with others to evaluate the copied content. Different similarity measures are used to get the degree of similarity [1].

2.3 Author Identification Framework

A typical framework for authorship identification of textual messages is described in figure 2.4. There are mainly four steps involved in the framework, namely data collection, feature extraction, model generation, and author identification [8–12].

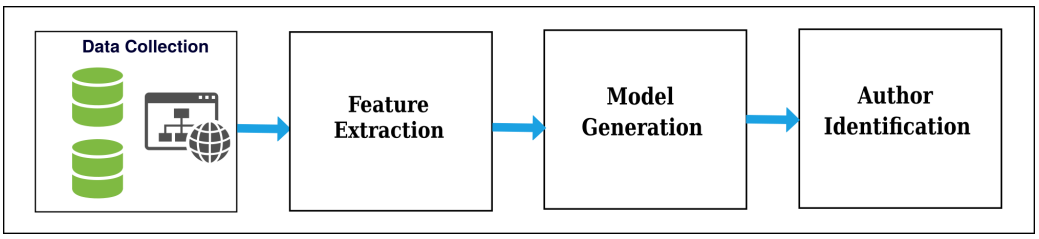


FIGURE 2.4: Author identification framework

2.3.1 Data Collection

The process begins with the collection of messages written by a group of authors from various sources. Primary sources are on-line messages such as email, chatting, blog, on-line newspapers, reviews, comments on on-line activities, novels, stories, etc. Sometimes the text contents are also captured through off-line messages such as diary writings, letters, etc. Offline messages are transformed to digital textual form by using various tools (e.g. OCR). This collected data is then converted to the required form so it can be used for the consecutive phases. This collected data can be used differently based on the profile-based approach or instance-based approach or hybrid approach [1].

In a profile-based approach, training samples of each author are concatenated to form one big text file. This big text file is then used to extract features [13]. In the instance-based approach, each training text sample participates in the process individually to build the attribution model. In the hybrid approach — the data samples are used in combinations for both profile and instance-based method. At first, for each author, one big file is created by concatenating all training samples, then several instances are built from this big file and then it is used to construct the attribution model [1, 14].

2.3.2 Feature Extraction

It is the second phase of the author identification framework. All collected corpus are in unstructured format i.e. raw text. Before extracting features, the writing style is defined. Then different types of features are extracted and represented in the feature vector form. Typically, all these attributes are in terms of the feature vector. There are several feature types, which are obtained and build from text samples. In this way, unstructured information is converted to a structured format.

2.3.3 Model Generation

In the model generation phase, the dataset is divided into two parts, one part is for training, and another is for testing. The training part is required for learning and model building, while the testing part is used to evaluate the generated model. The attribution model is evaluated by iterative partitioning of data into training and test set. It is called as cross-validation.

2.3.4 Author Identification

It is the last phase built model is validated. Now, it is time to check the writer of an unknown text sample. At this stage, the author of anonymous text is predicted from build model.

2.4 Feature Selection

Feature selection process is to identify the repeating pattern in the writing sample. The appropriate feature reduces the computation time for further processing. In the feature selection process, redundant and irrelevant features are removed. Based on the dataset, there are three types of the feature selection process. First is a supervised feature selection, a set of features are selected on the basis of the relation between features and label. It has high accuracy and high computational cost [15, 16]. The semi-supervised feature selection method is most relevant for labeled and unlabeled data [17]. In the unsupervised feature selection method, relevant features are selected without the prior knowledge and

then irrelevant features are removed by statistical measures [18]. Features are the attributes of repeating behavior of writer, and it is termed as writing style. Each writer has a unique writing style. The style of the author is gathered from the content which he has written and the method of representing the content is described in the next section.

2.4.1 Stylometry

The characteristic of the author can be captured from the text written by a writer from the content and style from the sample text. The writing style of an author is a collection of measurable patterns. A set of stylometry patterns are responsible for the discrimination of the text. In feature extraction, a measurable pattern is captured with several types of text processing tools and presented in the form of features. Previous work shows that writing style is characterized by style markers in the attribution task. Features are described in various types, such as words, a sequence of words, and characters, and its statistics. Some features give the information in terms of the count value, such as total token types, occurrences of unique words that represent the vocabulary richness, number of sentences, etc. Different categories of features are described in lexical, character, syntactic, semantic, application-specific, idiosyncratic.

2.4.2 Lexical Features

In the dataset, contents are distributed over sentences. A unit part of each sentence is a series of words, punctuation marks, special symbols, and digits. The contents from the text are split into linguistic tokens. The token is a unit part of the text, and it is part of the sentence in terms of word. The 'tokenizer' tool is used to split the text into tokens. It captures sentence information from text. Extracted tokens are used to build the feature set, and these features are directly included as a feature in terms of count (frequency) where tokens itself is involved for building features. Another type of feature are not directly related with the tokens of text but it holds the information about representation and it consists of token-based features such as word length [8], sentence length [19], count of hapexes occurred in the texts such as legomenon where words which occurred once in text sample, similarly word appeared twice is dislegomenon, tri-legomenon words occurred thrice are counted and act as a type of feature. These all represents the word usage in the text sample of an author [20]. There are several ways to measure the vocabulary richness

of an author. The token-type ratio is one way to measure the vocabulary richness from tokens [53]. The equation 2.1 shows calculation of token-type ratio.

$$R = \frac{V}{N} \tag{2.1}$$

Where R is token-type ratio,
V is the number of unique (type) tokens available in the sample,
N is total number of tokens in the text samples.

This information gives the vocabulary of the author, and it is a set of words known to the author. A lack of knowledge of vocabulary is identified with this measure. It can also be measured with some other types of terminology, such as Yule's measure, Simpson index, and entropy. Another kind of lexical feature is function words, which are not contributing towards the semantic information in the text sample. There are some merits to use function words in the author identification task. First is that it doesn't carry any information about topic and genre. The author is least conscious about function words in the writing. So function words are the measurable parameter to discriminate authorship [22].

Another type of lexical features is the occurrences of a token in the text sample. These occurrences are represented in terms of word frequency count, which are further utilized for discrimination among authors.

This word frequency is represented in bag-of-word format. To reduce the feature count, researchers have used the fixed number most frequent word occurred in the document samples [23]. One more type of lexical feature is sequences of n consecutive words (tokens).

In this scenario, the complete text is considered as a fixed number of consecutive words and it act as one token. In this way, a token vector of word n-gram is built. Along with the style, it captures the content-specific information from the writing sample. [1].

In this survey, word n-gram is applied for the author identification but it doesn't always give promising results like other types of features. In short text, the word pattern is not always effective because text is very small. It may not give correct information at all times because it is incapable when writing errors exist in the sample. It captures human behavior, but the behavior may change over time. For short text, there is less possibility of capturing such repeating behavior. In the research [14] word n-gram feature is used to

gather semantically meaningful information from the short text. The error made in the sample text is also considered as a type of feature. In error, mostly related to commonly misspelled words occurred in the text. An error occurred in the structure of the sentence, tokens, and its consecutiveness is termed as a feature, which can be utilized to discriminate among authors.

2.4.3 Character Features

A text samples are built with many sentences. Sentence are built with a set of words and words are made from several characters. A specific sequence of characters formulates a meaningful word. The smallest unit of a text sample is a character, and consecutive characters act as features that are used to discriminate among authors. There are many types of character features such as alphabets, lower and uppercase letter counts, or it measures digits, symbols, punctuation marks, etc. All these types of features are in terms of statistical measures. All these are directly related to statistics of writing rather than content. Character n-gram is one kind of feature which extracts the writing style of the author of both means, the statistical representation and content-specific. In linguistic arithmetic, character n-gram consist of a continuous sequence of n terms in a given text sample. It can be phoneme, which is a gesture of sound which differentiates one from another, syllable which distinguishes the sequence of speech sound in words e.g., mathematics composed of two syllables viz mathe and matics, letters, words, etc. In continual sequence, it is required to define whether it is in word or complete document. It means the consideration of the space, operator, and punctuation in words or not. Thousands of most frequent character n-gram represent the style of the author. It considered a language-independent stylometric methodology. Preprocessing is required for the character n-gram is negligible [24]. Instead of using character n-gram, variable length of character grams can also be used in the discrimination among authors [25]. In another way, the character n-grams are used in the compression method where discrimination among writings is compared bit by bit. In the compression model, initially, text documents are compressed with a compression algorithm, and then character-specific features are extracted [2].

2.4.4 Syntactic Features

Syntactic features are directly related to language grammar. It holds the syntactic constructs of text. In the writer identification technique, the fact assumed is, every author follows his syntactic patterns consciously or unconsciously. Hence, it is more reliable than other kind of features. Syntactic information about the author is captured in terms of part-of-speech, sentence constructs, phase structures, rewrite rule frequencies. Function word captures the syntactic information about the author. Different tools are capable of getting the syntactic information. The rules in writing are obtained from the texts are in the semantic form. NLP parser tools are used to extract the grammatical information from the content of the sample text. In python nltk library used to capture syntactic information from the documents. Parse tree is constructed for each sentence with the parser and which acts as one pattern. It acts as a feature. Grammatical information from the texts captured in the form of tags, these are called as part of speech. Tags are identified and attached to every token. Examples of some tag phrases are NP (noun phrase), PP (preposition phrase), VP (verb phrase). There is another perspective to use syntactic information as a feature, and it is the syntactic error. An example of a syntactic error is the use of tenses in the wrong way, structure representation error, grammatical error etc. [1, 26, 27].

2.4.5 Semantic Features

Semantic types of features are directly related to the content in the text. In [28], author uses the 'NLPWin' tool to generate features from the semantic graph. Semantic features extracted from these tools are noun, pronoun, tenses, various types of verbs, and its representations. All these features presented in the semantic graph with PoS. Another kind of feature is the use of synonyms and hypernyms of words. The WordNet tool is used to obtain the information. In [29], author uses 'ATMan' to extract semantic unit, called semantic lexicon of words, and used it for text classification.

2.4.6 Application Specific Features

There are three types of features assumed under this category named as structural, content-specific, and language-specific. The structural type of features are corresponding to the structure in which text is organized. This information is captured from the document

with the help of specific tools, such as HTML parser - to extract content from web pages, XML parser is used to extract content from these XML files. This information is extracted from a particular structural document with the help of a specific tool (parser). Structural information is the form of representation of any greetings, way to have a signature, indentation, justification of line, line spacing, paragraph starting. Along with this, how the text represented comes under the structure, such as font type, color, font size, etc. Another types of feature are content-specific; it belongs to specific content from the document. A particular tool and dictionaries are used to extract exact features. Language-specific features are more related to nationality and their presentations.

2.4.7 Idiosyncratic Features

These types of features consist of cause-effect such as wrong spelling, grammatical mistakes, social and cultural impact in chosen words. These features are very peculiar and appear in writing due to the odd habits of a person. This type of feature also represents the psychology of a person [30]. To interpret the writing style of feature chosen concerning to idiosyncratic features selected from simultaneous occurring, lexical sophistication, and lexical density [31].

2.4.8 Discrimination Procedure

This procedure is responsible for building the attribution model, which can be used to find the actual author of the unknown text. Discrimination is possible by finding similarity strength for each author, and another way is classification, where unknown text samples are assigned to the most probable class, each class label corresponds to the author. Alternative path to this is to find dissimilarity among the text which is used to prove whether contents are not written by a author or not? The consecutive occurances of PoS tags also participate in the discrimination of text which have been proved in researchers work. Many researchers in their work have been used consecutive occurrences PoS tags. It gives structural representation and its repetition over the sentences.

2.5 Attribution Methods

Depending on the representation of the text, the attribution approaches are distinguished. How all the text samples participate in the attribution process? In some example each sample is individually identified and involved in attribution, all the training samples of the same author are concatenated to form a big file, and then it contributes towards the attribution, this method is called profile-based approach. All attribution methods are discussed in consecutive sections.

2.5.1 Profile-Based Authorship Attribution Approach

In this type of approach, a training sample of each author is accumulated, and then it is concatenated to form a big file. Training sample consists of all writing samples of each author in a single text file. Equal number of writing samples are constructed as number of authors. From this single big file, the profile is built for each one by extracting the writing style of the author. Then the unknown text is compared with each author and most likely author will be determined.

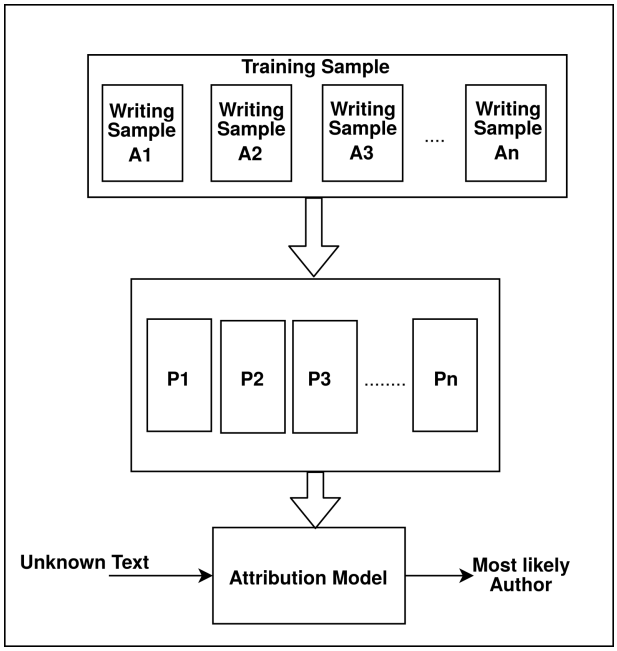


FIGURE 2.5: Profile based approach

In this way, unknown text is compared with distance metric to each author profile. A profile-based approach is shown in figure 2.5. Different types of features are used for

profiling. In the study [13, 32], an author identification method that uses character 4 gram as a feature to build author profile. Iqbal et al. [4] represents a method which uses a write-print - a pattern-based extraction. The profile-based approach gives a firm representation of author style for short text as well. There are different distance metrics used to compare the profile of each author with unknown text. Anwar et al. (2019) [33] come with a novel profile-based approach for author identification with n-grams as a feature and it uses cosine similarity distance metric to measure similarity and most similar author is assigned to unknown text.

2.5.2 Instance-Based Authorship Attribution Approach

The instance-based approach is typically used when machine learning algorithms are used for discrimination. Widely used machine learning algorithms are a neural network, Bayesian, decision tree, support vector machine, KNN to train and build the model. In this approach, each training sample is used separately to build the model where each sample is treated as instance hence called an instance-based approach. Instance-based method for author identification is elaborated in figure 2.6.

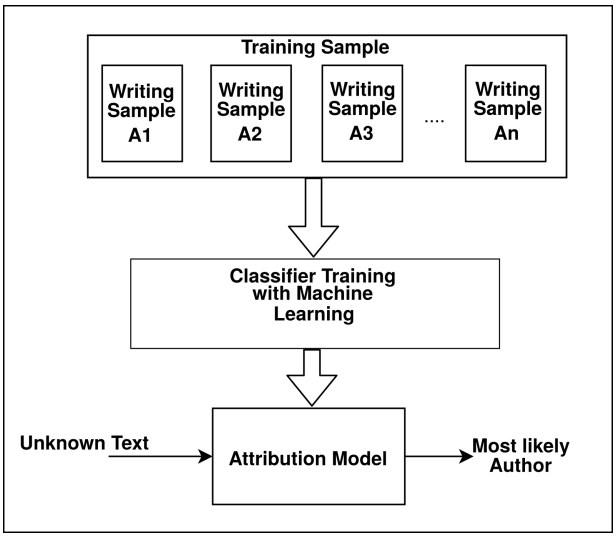


FIGURE 2.6: Instance based approach

It is possible that when the instance-based approach followed, the size of each sample may get varied, which can affect the performance of the model. Hence, to get a more accurate result, the size of each sample text kept the same. If it is variable length,

then it should get normalized in model building. Each text sample has its own identity and participate separately in the classification phase.

2.5.3 Probabilistic Model

It is an ancient model, which uses conditional probability to discriminate amongst authors. In this approach, the likelihood of text belonging to the author is maximized. The conditional probabilities are calculated for all texts belong to the author with the concatenation of remaining text for remaining all authors. Zamani et al. (2014) [34] uses the probabilistic distribution model for representing each document as a feature set and then distance measure used for discrimination and then probability is maximized to predict correct author.

2.5.4 Compression Model

As the name suggests text samples are compressed using a compression algorithm, then the similarity of unknown text calculated with each author on extracted features. This approach uses a profile-based approach, where all text written by the same author are concatenated, and then this big file is compressed using a compression algorithm. The compression algorithms used are RAR, LZW, GZIP, BZIP2, 7ZIP, etc. Tehan and Harper (2003) [35] has used the compression model for text categorization using the RAR compression algorithm.

2.6 Features in Author Identification Methods

There are three types of features used in different author identification methods named as content-based features, style-based features, topic-based features. Each person has his own style of writing. These differences are due to the subject interest of writer, behavior, vocabulary, and grammatical rules. The content-based feature consists explicitly of the content of writer, style based feature focused on the style which the writer follows in the writing; it is irrespective of the content and topic of writing. Topic-based features, identify the writer who writes differently based on his interest.

2.6.1 Content Based Features

Different people narrate the fact or story differently in their writing. This description is represented by the writer depends on the vocabulary richness that a person has, his interest, and the usage of words. So it becomes the distinguished feature to identify the author. There is also a difference in writing when a male and female author writes on the same fact. When discrimination amongst the author is done with a content-based feature, the contents are extracted from the writing samples of each author. These contents are extracted in the form of different n-grams from the document. Contents are in the form of a sequence of words or characters. When n consecutive characters or words taken, it is called n-gram. Several occurrences of n-gram are termed as 'frequency count' of each n-gram and identified as features. When such n-grams are very large, the dimension of features needs to be reduced. There are different techniques used to minimize such dimensions based on the importance of features. An example of a dimensional reduction algorithm is PCA (Principal Component Analysis). The content-based feature participates in distinguishing the age group of the writer from their writing. Rocha et al. [14], in the survey of author identification over short text (tweets) has used word n-gram and character n-gram as content-based features. In the work, n is varied from 1 to 5.

2.6.2 Style Based Features

Style based feature consists of n-grams, PoS tags, punctuation marks, symbols, statistical features on text such as total capital letters against all alphabet, token type ratio, vocabulary richness, the average number of characters per word, per sentence, etc. All these types of features are represented in terms of frequency or constant count value, and it needs to be normalized. In style-based features, the total number of features in the content-based feature type is less. Rong Zheng (2006) [9], utilized total 270 number of style-based features for author identification. Character n-gram can capture both content-based and style-based features.

2.6.3 Topic Based Features

The topic is related to the objective by which contents are written. Various people write differently on different topics. So topic on which content has been written can be

an identifiable feature to discriminate in their writings. On specific topics, many times, the same words are used in a different context. Such as writing on the same topic by a male and female candidate. Various topics are used in writing blogs, emails, chats. Topics can participate in discrimination in male and female writing. Writing can be distinguished based on the ages of the writer when topics-based features are used for discrimination [36].

2.7 Related Work

Author identification is a field where the analysis of authorship is done on existing writing of author. In this section we brief the review of research in authorship attribution. Authorship analysis process will proceed with the systematic study of the text written by author. Observation of text is made with extracting the attributes of text which termed as features.

A document is represented by a feature vector that contains one Boolean attribute for each word that occurs in the text. When this method is generalized by using word sequences and form consecutive occurrences termed as n-gram as a feature. To generate n-gram feature, we begin with the first token in the text and then n-gram is formulated with consecutive n tokens. Second n-gram created by repetition of same process from second token and so on. An efficient way to generate n-gram feature methodology is discussed in [37] which is based on value of n, the occurances of n-grams will be increased or decreased.

The algorithm utilizes three parameters document collection, MaxGramSize, Min-Frequency. The algorithm is based on the APRIORI-algorithm for discovering frequent items (features) subsets in databases. Outcomes from the project include a learning algorithm that removes stop words, word sequence of length 2, or 3. Longer sequences reduce performance. The results indicate that the addition of n-grams to the set-of-words representation frequently used by text categorization systems improves performance. However, sequences of length $n > 3$ are not useful and may decrease the performance.

The n-gram methodology is applied to categorize the text in the Bangla newspaper corpus. The work describes analyzing the efficiency of n-grams and shows that tri-grams have much better performance for text categorization in Bangla [38]. The use of n-gram is limited to two consecutive sequences called as bi-gram. The research work [39] focused on the use of bigram along with very less number of unigrams in solution of text categorization

problem. Based on information gain and frequency threshold features are selected in bi-grams and then two classifiers are used for text categorization. Bi-grams are selected in such a way that their occurrence and information gain is high.

Stylometric analysis techniques are categorized into supervised and unsupervised methods. In character-based features, there are several white characters, special characters, etc. The word-based features include words with vocabulary richness. It also indicates writing trends, emotional words, cognitive words, frequency of particular cues, the appearance of related words. Syntactic features include punctuation marks such as comma, semicolon, question mark, etc. These are very general and depend on facts like habits, mood, expression, etc. Structural based features rely on the layout, length of sentence, organization of writing skill. Function words show vocabulary richness, lexical meaning, personal styles, etc. [5]. A set of stylistic features, includes linguistic features, character-based features, word-based statistic features, syntactic features, structure-based features, and function words. The classification process includes normalized features using the max-min normalization method to put values between 0 and 1 as in [8].

The structure-specific feature includes the structure of documents in terms of representation. Normally it includes the length of sentences, separators used in sentences, length of a paragraph, for specific fact represented by the number of the sentence, repetition of part of sentence, and the layout of the whole document. Indirectly we can term it as a habit of author reflected by its writing style [9].

Along with the above basic features, there is the concern of originality, i.e., if the original document is of one author and is copied by some other author. At this time to find the real author of the document is difficult. There are two different approaches, one is writer dependent, and writer independent to build a robust method for identifying authorship. Again, it uses the same stylometric features described in the previous section based on conjunction and adverbs. Writer dependent model based on the individual author. Writer independent model is based on the forensic questioned document examination approach and classifies the writing in terms of authenticity, using the global model [40].

Over the internet, the huge amount of textual content are available in the form of blogs, emails, digital contracts, books, and many more. So the correct identity of this available data is difficult. That might incur cybercrime. The problem of anonymity in online communication addressed by applying authorship analysis techniques. In past, a lot

of research was there for the analysis and identification of the owner of data content. There are many approaches available for the authorship analysis.

In research [4], a unified data mining solution was proposed for authorship analysis. A stylometric pattern extracted from evidence text samples to generate frequent itemset. In this way author writing style represented in terms of writeprint then based on these writeprints a unified data mining algorithm used for author identification, similarity detection and author characterization.

Another specific category in which information grouped is topic modeling. In one description [41] author describes idea about topic modeling. The topic is identified by distribution of words and their frequency in the corresponding corpus. This topic distribution was found from the text document with the help of statistical techniques such as Dirichlet allocation or Gibbs sampling. Another way is Hierarchical topic modeling.

In research work [42], author uses hierarchical generative model. In this model, each word is associated with two variables one is an author, and the other is the topic. A set of N dimensional vector is used, indicating defined variables, topics, and author assigned for N words.

In the research study [43], a fully automated approach to identify the authorship on the unrestricted text that excludes lexical measures. The described method eliminates distributional lexical measures. Instead of using sentence length, punctuation marks and syntax-based, noun phrase count, verb phrase count were used. It provides the way of analyzing text and form of capturing information. This method lacks linguistic theory as it is based on statistical measures. This work also describes the approach to capture the diversity of an author's vocabulary, and one is the type-token ratio:

$$Token - type - raio = \frac{V}{N} \tag{2.2}$$

where V is the size of vocabulary of the sample text,

N is the number of tokens that form the sample text.

Another way of measuring the diversity of vocabulary is to count how many words occur once (i.e., hapaxlegomena), how many words occur twice (i.e., dislegomena), etc. These measures are strongly dependent on text-length.

The solution for disputed authorship is presented as in [44]. Due to causal basis likelihood judgment and conditional dependencies, the scholar makes critical errors. The study provides Bayesian inference in distributed authorship. Two hypotheses (H and $\sim H$) are used to examine the passages of each document and judge the extent to which each passage supports or refuses each hypothesis.

A method for source code authorship identification uses a SCAP (source code author profile) method [45]. Wide range of features are considered for java and common lisp, and depending on programs, comments, layout features, and packages selected naming influences, classification accuracy other like user-defined names, program-related features not appeared to influence accuracy. Features considered for the same are programming layout metrics style, metrics structure, and linguistic metrics. The SCAP approach makes use of contiguous n-gram sequences defined at the lower level attribute of a program. Program content categorizes into features like comment layout, identifier programming structure, etc.

Documents observed in a hierarchical fashion, stylistic characteristic of author and group of author-specific rules are used to build a classifier, and recursive data mining approach is performed as in [46]. The method used to perform identification was RDM (Recursive Data Mining). Using token and patterns as a feature, it performs well with Navie Bayes, SVM (Support Vector Machine), and RDM descrimination methods. The result of experiments shows the capturing stylistic pattern in the SEA and Enron dataset and also used for the organizational role of authors. The method divides the semantic knowledge for the semantically related pattern.

The research [47], describes a method as a Naive Bayes algorithm for a feature and word selection for text classification. This algorithm is for multidimensionality classification. For that, it uses feature clustering to reduce the dimensionality of the feature vector for classification and put a fuzzy similarity-based self-constructing algorithm for feature clustering. The described algorithm improves the performance of the algorithm with an elite strategy.

In the study [8], author discusses the gender identification, it is based on human psychology. Total of 545 psycholinguistic and gender-preferential cues, along with stylometric features are used to build the feature space for this identification problem. Three machine learning algorithms are designed for gender identification based on the proposed

feature. In the described technique, all features are collected and normalized using the max-min normalization method as described in equation 2.3 to ensure all feature values from 0 to 1.

$$Normalized(X_{ij}) = \frac{X_{ij} - min(X_j)}{max(X_j) - min(X_j)} \tag{2.3}$$

where

x_{ij} is j^{th} feature in i^{th} sample

$min(X_j)$ and $max(X_j)$ are the minimum and maximum feature values of the j^{th} feature For classification techniques used are Bayesian-based logistic regression, Ada-Boost decision tree and SVM classifiers separately on Reuters and Enron corpora.

Author in [40] defines two approaches one is writer dependent, and writer independent. Because of this strategy, it becomes robust. This method used features as forensic stylistic, which is a subfield of forensic linguistics, which aims at applying stylistics to the context of author identification, where it is based on two writers who do not write in the same way, and writer himself does not write in the same way all the time. Proposed work uses conjunctions and adverbs of the Portuguese language to find the author. In this work, the authors extract features using a compression algorithm and achieve a success rate of 78%.

The work [48, 49] describes the feature extraction methods, which includes word similarity among sentence and their frequency occurred in a statement. Similar sentence repeated over document, paragraphs, and provide a solution for classification using evolutionary programming with the help of fuzzy logic and artificial neural network. The description in the study gives a view of the hybrid classification method, which provides a direction towards smart feature extraction.

In [5, 50], the research study uses stylistic features, including lexical, syntactic, structural, content-specific, and idiosyncratic attributes. Writeprint method also described in this work. This study describes that existing methods have focused on the author identification task, but there is a limitation for similarity detection and provide a summary of some features. Stylistic features represent lexical, syntactic, structural, content-specific, and idiosyncratic style markers. Lexical features include words, characters, their variance, and length distributions. The syntactic feature includes function words, punctuation, n-grams. Structural features are as file extension, font, colors, etc. Content-specific features which are keywords, phrases, and a topic name like word n-grams. Idiosyncratic features

have misspellings, grammatical mistakes, and other usage anomalies. The author introduces an extended feature set along with a baseline feature. Extended features include static and dynamic features. Writeprints methodology is used to construct the classifier. This technique has a creation and pattern disruption. For finding writing style variation, Karhunen-Loeve transforms applied with a sliding window to capture stylistic variation with a finer level of granularity.

A research study of [45, 51] introduces the method for classification of the author based on high-level programming features. Author describes author identification using high-level features that contribute to source code authorship identification using a tool the SCAP method. Source code author profile (SCAP) is based on the byte level feature, which is used to assess high-level programming features. The author describes the previous method of classification based on features, programming layout, style, structure, and linguistic metrics. In a set of experiments, the author uses the feature in an identifier, symbol name identifier, package name identifier, comments, layout metrics.

Genetic algorithm (GA) feature selection model is represented by [52], which is used to identify the writeprint features. In his model, features are represented in bits. A number of bits based on candidate features which define accuracy. These features are generated successively, and finally, this GA model generates different combinations and utilized for classification and termed as key writeprint features to discriminate the writing style of several authors.

2.7.1 Profile-Based Authorship Model

Koppel and Seidman (2017) [53] describes similarity measures as outlier identification among different documents. In the process, each document is represented in the form of the feature vector, then these vectors of various documents get compared with similarity measures and then based on a remarkable point as a threshold to decide the outliers. It also describes min-max similarity measures, which more effective than cosine similarity measures. The study also describes the aggregation methodology to measure similarity among feature vectors.

The work [27], uses syntactic information to find the identity. Author focused on baseline linguistic features, such as total words, sentences in documents, token-type ration,

standard deviation among length of words, different types of declarative, interrogative, imperative sentences, a count of active and passive voice, number of s-genitive, of-genitive, lack-genitive, count of negative words and uncertainty markers such as 'could', 'possibly' etc. All these are represented in the form of tf-idf weighted vector. Then cosine similarity measure is used to discriminate the identity. Data set used for experiments consist of chapters from different novels and used in terms of chapter parts. They have achieved accuracy from 12% to 95% on selected features.

In [54], the author reveals ideas of stylometric features rather than the content-based features. Stylometric features represent the characteristics of the author and independent of content. To get in the workplace, the author initially experimented by considering both types of features i.e., content-based and style-based. The cosine similarity function applied to calculate similarity among the candidates. The annotation process followed with the random evaluation to produces remarkable results.

In the research [55], the author deals with the programming source code written by the programmer. In this technique, whether the programming source code is written by a programmer or not is identified. Here programmer is an author. To deal with the situation, the author worked on byte-level n-gram to extract programmers source code style. Effective performance quoted by the author over 6 to 30 candidate authors. The dataset consists of a different programming language such as Java, CPP, C, etc. The proposed methodology in the study is independent of the language of the document. In the approach, each character (symbol, digit, alphabet) act as a byte value. For n sequence of bytes, frequencies are calculated and represented as the profile of author. As it is a profile based approach, so all the code for each author concatenated to form one big file. To find a similarity score, a SCAP approach is followed. The most frequent n-grams accumulated given in equation 2.4.

$$X = x_1, x_2, x_3, ...x_L \tag{2.4}$$

then similarity between author profile (SP_P) and code sample of unknown programmer (SP_A) is calculated with the equation 2.5.

$$similarity - score = \frac{SP_A \bigcap SP_P}{|X|} \tag{2.5}$$

where

$|X|$ is size of X

Accuracy obtained from this methodology was upto 95% for author profile size 1500.

Kocher and Savoy (2017) [56], describes a SPATIUM-L1 authorship verification model. Extracted features from the dataset were the most common 200 terms, which consist of an isolated word, punctuation symbols. The author evaluated methods by participating in the campaign PAN CLEF 2014. In the methodology, the profile of each author built, which denoted as A and the unknown text to which authorship found, indicated as Q. For the calculation of the similarity following function was used shown in equation 2.6.

$$score(Q, A) = \sum_{i=1}^{k} |P_Q[t_i] - P_A[t_i]| \tag{2.6}$$

where
k is number of term types (word, symbol)
$P_Q[t_i], P_Q[t_i]$ are the occurred probability of term ti in query Q and author profile A.
In this way similarity score is calculated and further decision made by applying rule. A notable results are achieved from the algorithm.

In [57], a profile-based authorship attribution followed on Chinese online messages. In Chinese online messages along with Chinese content, English is mixed; hence, the profile-based approach is suitable under these circumstances. The writing style of the author is captured in terms of character n-gram — the similarity measures are employed to determine the most probable author. The accuracy of the result was up to 88%.

Houvardas [58], supports through the revealed idea of the variable length of the n-gram in author identification, which helps in cybercrime. Where the study clears the idea about the use of character n-gram has the stylistic parameter of lexical, syntactical, and structural content. The illustrated work initially based on varying length of word sequences. A fixed set of the rules found on the length of the sequence of n characters where its glue, antecedent, successor value calculated, and compared with it. And only 3,4, and 5-grams used. Experimentation followed on English language RCV1, over 50 authors dataset where feature counts varied in size from 2000 to 10000 and selected most frequent 3, 4, and 5-grams features of the equal count. A support vector machine with a linear kernel classification method used on reduced features to discriminate among authors.

In [59], profile-based approach described for identification of authors. To capture the stylometry of writer, character n-gram used. A sequence of characters represented in a

bag of word form. All text for each author concatenated to one big file, and then the profile of each writer is constructed by extracting the features form such a big text file. Most frequent character n-gram organized in decreasing order and also normalized with respect to text length. Then unknown sample of the text compared with the profile of each author — such dissimilarity measure is calculated using equation 2.7.

$$d_0(P(x), P(T_a)) = \sum_{g \in P(x) \cup P(T_a)} \left(\frac{2(f_x(g) - F_{Ta}(g))}{f_x(g) - F_{Ta}(g)} \right) \tag{2.7}$$

Where

$f_x(g)$ frequencies of ngram g for writer training sample,

$f_{Ta}(g)$ for test sample,

$f(g) = 0$ when $g \notin P$.

Then KNN applied to find most probable author and value of $k = 1$.

$$author(x) = \underset{a \in A}{argmin}\, d_0(P(x), P(T_a)) \tag{2.8}$$

The corpus used in the experiment is collected from RCV1. In experiments value of n in n-gram was 3.

2.7.2 Instance-Based Author Identification

Rong Zheng [9], comes with one of the ideas to identify the author for online messages. English and Chinese languages are chosen for experimentation. Features considered in groups are lexical features, word-based features, syntactic features, content-specific features, structural features. A total of 270 features are adopted for building a language model . Experiments made in English and the Chinese language. The evaluation made on one to four types of features. C4.5, NN, SVM classification models were used for identification. Results based on feature types and techniques like SVM and NN produced challenging results. With all four types of features, SVM gives 88.33% accuracy on the Chinese dataset and 97.69% accuracy for the English language dataset. C4.5 technique has given the least accuracy.

Steven et al. [26] proposed three models of topical bias, local contextual bias, and lexical bias for learning. The final vector is the outcome from all these three modalities.

Topic relevancy capture by topic bias modality; irrelevancy in the topic of the sample captured in local contextual bias; the alternative word used in the context of the document represents lexical bias. Stylometry served by all of these types of modalities using a neural network model. The dataset consists of movie reviews of 62 different users and has 62000 movie reviews. The experiment result shows accuracy from 64% to 85%.

In [14] survey, represent a traditional authorship attribution model, and perform experimentation on tweet dataset. Detail review of character, word, and PoS n-gram features made, and along with these features, another set of the diversified feature was used. Five strategies of the SVM algorithm are described in the study. Power mean SVM (PMSVM) used for large scale datasets. The kernel used in this SVM is directly applicable to image and text classification. In the experimentation, a tweet set of 50 users is used as the dataset, and comparative performance is evaluated using various classifier along with all n-gram features, the value of n varied from 1 to 5.

In [60], a concept of signature was introduced in author identification. The signature is corresponding to the pattern of a unique writing style that appears in training samples. In the training sample set, at least one signature is present. Word n-gram where n varies from 2-5 and character n-gram, uniquely k signatures capture from the sample set. SVM machine learning algorithm was used to build such a k signature from all the samples. It also used for the characterization of the author.

Author in [61] describes SVM to solve author identification problems, which is capable of handling a vast number of features. The dataset consists of texts collected from the German newspaper. One-year text material collected with more than 2600 document with a length higher than 200 words. Two tests are performed in research work, in the first experiment word texts are extracted from the documents in the form of frequency count and SVM is applied with a different combination of kernels to achieve the best result. In the second experiment, nouns, verbs, and adjectives were replaced by tags and bi-grams, which causes reducing performance, but overall, SVM performs best. Experimentation uses $L1$ and $L2$ norms for feature normalization.

The authorship identification problem viewed from a different perspective in the research [62] by proposing one class author verification problem. The dataset made with the text samples, which was collected from twenty-one books written by ten different authors in the 19th century. Selected features from the document were 250 most frequent words for

each author. An unmasking applied to reduce feature set. This reduction made by applying SVM with linear kernel for each author is over text from each book hence termed as one class verification problem. Ten fold cross-validation used to check the accuracy for each author and in each fold most strongly positive as well as most strongly negative weighed features removed out to reduce feature dimensionality. The verification method described in this study has a very high accuracy.

In the research work [63], discriminating attributes are extracted by detecting patterns and then transformed the text into time series. These attributes gathered through the feature selection algorithm. To solve the authorship attribution problem, two modules were defined in the algorithm, and it was corresponding to feature selection, anomaly detection, classification, and visualization of algorithms. And the method performs best in the author identification task.

Na Cheng [8] recognizes the gender through proof from text in the form of an interaction between psycho-linguistics, nonspecific writing styles of men and women. In their studies, they have used three algorithms viz SVM, Bayesian logistic regression, and Adaboost decision tree. Accuracy captured is around 85%. Contributed features in discrimination were function words, word-based features, structural features. The model applied to Corpus from Reuters and Enron email dataset. The text was represented in the form of vectors. Many types of window algorithms are used to discriminate among several authors [64] to produce compromising accuracy. Writeprint, a new technique introduced [5] in which sliding window features were considered for the application of language model. Many types of features were accumulated and applied to this new model, which produces accuracy around greater than 90%. All these reviewed researches focused on the methods for author attribution and identification, the time when the document generated was not considered.

Azarbonyad [32] studied attribution where the writing nature of author changes. These temporal changes are observed concerning word distribution in writing samples. In the experiment dataset of tweets and Enron emails over five years of time span were used. Character 4-gram was the observed features. Temporal changes were captured with the algorithm defined in work [65, 66] time-based language model and calculated from linear regression techniques. Research work [67] elaborated the change in vocabulary usage by a writer and proved that the size of vocabulary goes on decreasing over time. The time

frame over the work was big about up to 35 years. Author [30, 68] also investigated and concluded that the writing style of author changes over time, and authorship verification accuracy increases nearer to the time at which text was written. A review of different methods of classification and their results were elaborated in the review [69]. And in [70], author uses a machine learning approach for content types of features where writing samples are shorts and irrespective of time. [71] shows the variation in writing over time in terms of features. Variation was not stationary in the work but indicates there is a change in the writing style of the author.

Houvardas [72] supports through the revealed idea of the variable length of the n-gram in author identification, which helps in cybercrime. Where the study clears the idea about the use of character n-gram has a stylistic parameter of lexical, syntactical, and structural content. The illustrated work initially based on varying length of word sequences. A fixed set of the rules found on the length of the sequence of n characters where its glue, antecedent, successor value calculated, and compared with it. And only 3,4, and 5-grams used. Experimentations are performed on English language RCV1, over 50 authors dataset where feature count experiment tested on the varied size from 2000 to 10000 and selected most frequent 3, 4, and 5-grams features of the equal count. A support vector machine with a linear kernel classification method used on reduced features to discriminate among authors. The system outperforms in terms of accuracy.

In [73], describes the author identification method over three types of features, character sequences, word gram, and PoS-tags features. Documents of several languages are shared in PAN 2018 contest. Three different models described to the author identification task, which used a machine learning algorithm and compared using F1 scores. In the process, initially, all document contents were preprocessed to remove non-required content from samples. Then on extracting features, five different machine learning algorithms were applied named multilayer perceptron, SVC, Linear SVC, Logistic regression, and Random forest. A comparison of the F1 score was made on all these types of the machine learning algorithms. The first model gives an average score of 0.582, over the features of character 6-gram, orthographic features, lexical richness, PoS tags, quantitative features. Another model produces an average 0.598 score over features character 3-8 grams, word unigrams, PoS-tags. The last model gives a 0.611 score over content-based features.

With a different perspective, the author identification problem solved by Helena

et al. [74] in PAN CLEF 2015 contest. Two types of author verification method named as intrinsic and extrinsic are described in work. The intrinsic method problem solved by considering the known sample of one author versus unknown text, where is to find whether the text written by the author or not. The extrinsic method uses a document written by all authors versus unknown documents to solve the problem. The methodology described in the work falls in the intrinsic category and the content of the document represented in terms of graphs as integrated syntactic graphs that were utilized for building linguistic features. Which on further compared with unknown text for each author employing similarity measures. The described method in the research outperforms.

Author identification by a neural network model was defined in [75] outperforms over traditional methods. N-gram features utilized with smoothing techniques. The work uses a minimal dataset and uses a neural network language model for attributing the author. The work [76] defines the threshold frequency if word frequency less then participates in the process. A convolution neural network model is used for the solving author identification problem. The dataset consists of the papers published by the author in the proceedings of the conferences. The accuracy achieved with the proposed model reaches up to 78% over text.

2.8 Research Approaches of Author Identification over Time

Author identification is a task where the owner of the unknown text is found by the analysis of published content of the same author. Several variables affects the writing style of the writer. These writing styles gets influenced due to elapsed time and it termed as style 'change over time'. The vocabulary of a writer grows with time [77]. The main factor affecting the shift in the writing style of the author is due to age, nationality, mother tongue, education, etc. While in author identification, these changes are taken into consideration to achieve better performance. In this chapter, several publications related to the modifications grasped in the writing of the author and their literature are described based on the review of a set of performance parameters selected to evaluate the system. The chapter is divided into two parts in the first literature regarding the style of the author changing over time and metrics to assess the system.

2.8.1 Style Change Over Time

Queries are classified on the basis of time, one is written recently generated and another written in the past. It indicates that a relationship exists in between time and relevancy in the queries. There is an effect of time over a document written at different time periods [78].

As age changes, the stylistic choice, usage of words, expression in writing changes. Then with the help of these characteristics, it is possible to predict the age of the writer. So to get such information, several types of features quoted in the work [79]. In this study, a set of feature shows the changes over time observed in the experimentation and found that emoticons used in writing and it decreases as age increases. Apart from this word usage, sentence length, word properties, slang, and punctuations have a cause of time. Such changes tracked with visualization of corresponding features over time. Some features show positive or negative growth over time, and some may fluctuates[71].

Changes caused in the features are due to the situation, surroundings, and psychology. These changes can be trapped by observing consecutively old documents and extracting the pattern form them in different forms. The change in writing style are found in the following means.

1. Hapax vs Token Ratio

When vocabulary in the document presented in terms of frequency count, a set of words that occurs in a specific amount of time in the document termed as hapaxes. Hapax legomena are word-groups that occur once in a document. The hapax and token ratio is dependent on sample size hence used with care. It represents lexical diversity, and it gets decreased as age increase. It also gives information about the vocabulary richness about writer [77, 80].

2. Token-Type Ratio

It is measured to find lexical diversity written by the writer in the context of the text. It can also be used for vocabulary richness. As there is an effect of vocabulary

richness over ages, as age grows, vocabulary increases [81]. Token-type ratio can be found with equation 2.9.

$$Token - type - ratio = fraclog(V)log(N) \tag{2.9}$$

3. Lexical Stylometry

Sentences written on the blog show a change in stylometry over time. The number of emoticons decreases over age, the number of capital words decreases over age, average sentence length increases over age. These were the general observation made in [79].

4. Content, Style and Topics-based

The writing style of each individual is different. These differences are in terms of topics of interest, grammar rules, selection of words of likeness. There is a mixed-use of adjectives and adverbs in writing. These types of words are used in discrimination of the writers, which shows variation at different age levels and genres [82].

5. N-gram

Consecutive occurrences of words and characters capture the content-based stylometry of the author. Similarly, the n-sequence of PoS captures the syntactic stylometry of the author. Due to several factors, the stylometric of writer changes over the period. This can be captured through the observation of differences over old writing versus new writing [83].

2.8.2 Publication and Review on Stylometry Change Over Time

In the survey of David Holmes [77], different authors suggested that the writing style of the writer changes over time. The review of the statistic analysis of such changes captured are discussed. The effect and stylometric techniques are described with respect to various feature types such as token-type ratio, hapaxes-token ration, lexical diversity, vocabulary richness is discussed and elaborated. But the solution to the problem was not addressed only it confirms that changes happened. Swan, in his work [84] identified the text written in which time span, hence named their work as “TimeMines.” A statistic

model was built over the usage of words to predict the time line of the document. This was based on the fact that in each document, the evidence of the time of the document creation is hidden in terms of the description of discrete events, scenarios, etc. In the process of feature extraction, the named entities were extracted, which are the phrases that describe the person name, organization, locations. Such phrases are labeled with PoS tag and represented in a bag of word form. Other types of features extracted, some of them are stationary and independent. Using the statistic analysis, the topics belonging to a specific time domain get tagged to specific time-tagged groups. In this work, time is directly extracted from time events and date tags.

First time, Xiaoyan [78], in the research focused on the language model with respect to time. In work, two queries categorized with respect to time, where relevancy decided based on time. It is more relevant to recent documents as compared to old documents. Research explains the relationship between time and ad-hoc title queries. The relevancy of the query is based on the time at which the document was created. It is calculated with the technique described in the research work. To find the similarity between document, a distribution factor is defined and assigned as prior probability so it can nullify the effect of time. Most probable documents assigned to more similar query as per normal distribution. In the research, it was assumed that there is an exponential distribution relationship exists between documents with respect to time. The time-based model gives good outcomes as compared to the baseline model to find likelihood.

In [85], author describes the writing style changes over time. Statistical methods are used to examine the repeating pattern of text and used to solve authorship attribution problems. The writing work of two Turkish authors were tested over a period of ten years. Changes observed over the old and new work in terms of repeating patterns through statistical observation. Three types of styles were used to distinguish old and new work. The first feature, "word length," average word length used in writing was larger than that of old work. This can be calculated using regression analysis. Logistic regression used to discriminate old and new work over token and type length frequencies. The third last type of feature to distinguish between the latest and old work was the use of most frequent words in writing through comparison made by a graphical illustration of PCA. Nattiya and Kjetil (2008) [86], on the web, there is no surety about the indication of timestamps. The author proposed a system that was capable to find the timestamps using a temporal language model. Online documents are not trustworthy for mentioning the timestamp and

to identify it is a big challenge. The solution to the problem is to identify the timestamp from written content, document creation time and the time mentioned in the document. This timestamp is recognized with two methods one is learning-based, and another is non-learning based. In the learning-based model, statistical hypotheses have been made and in a non-learning model timestamp determined by the time-tag in the document in the form of the date, event, etc. Statistical language model uses a word interpolation, recurring and non-recurring words.

The work [87] describes the relevance model based on times. Blogs content retrieved using a query expansion method based on time. The relevant feed of blogs retrieved based on the most recent days. Results analyzed on TREC09 datasets and which shows improved performance of the retrieval system. A query likelihood model based on the time proposed by the Bingjie et al. [88]. In the hypotheses the result of document retrieval, the document published more recent are more important. The ranking of documents made by assigning a higher score, which was done by applying hypotheses. In the work, a mixed time language model described with the following equation 2.10.

$$P(d) = \omega.P(d, q, t) + (1 - \omega).(d, t) \tag{2.10}$$

where,
q is query, d is document and ω smoothing parameter control.

The first authorship attribution model based on time was described by Hosein et al. [32]. In work, tweets and emails were collected from the Enron dataset. The method is based on the hypothesis that nearer document text to query document most similar than older. Initially, the drift of author similarity is calculated for every author using linear regression method. And to perform experiments character four-gram with similarity base language model was used. Character four-gram responsible for capturing content and stylometry of writer. Regression information results that there is a change in writing style with respect to time, but it is limited. From the linear regression, a decay factor is calculated, which is applied to the similarity score of the documents in evidences and query samples. The results shows an improvement using time aware approach when compared with time unaware approach. The dataset used in the experimentation was limited to four years of time span.

2.9 Gaps and Challenges in Existing Methods of Author Identification

The extensive literature survey identifies the following challenges which needs to bridge:

1. Dataset in Author Identification System

The state of the art methods of author identification depends on the available data set when the assumption is that the writing style of the author changes over time, the data set used with minimal time domain or having with two or three authors. So a big challenge is to have a dataset with a significant amount of time to track change over time.

What would be the minimum length of the text sample is a crucial issue regarding the captured the stylistic information. In the literature, most of the research work uses text sample less than 1000 words. There is no specific rule to set length of the text sample. In various study work published on a dataset that uses short text, that is collected from tweets, reviews, chats, etc. There is still no definition of the size of the training samples. If the text size is uneven in the training sample, then it affects the outcomes. A small amount of text samples does not provide the whole writing style of the author.

2. Feature Selection and Extraction

It is another challenge that directly affects the performance of the system. There are many types of features, and each of them gives a different performance on the different types of datasets. The most challenging part of the feature selection is that when selected features are of n-gram type, then feature dimensionality very high. There is no fixed threshold to set a number of features under such situations. The performance of the system varies on the feature set size as well as the types of training samples. In some work, it gives excellent results for character n-gram, but what kind of information produced by the n-gram is not justifiable.

3. Factors in Author Identification System

There are several factors that directly or indirectly affect the accuracy of the author identification problem. Age, education, nationality, behavior, literacy are the factors responsible for causing change. It is observed in the terms of content and style. Different types of features are figured out by several researchers. Character n-gram, token-type ratio, most frequent word, average word length are few of them. And all these factors are bound to time. But they are not used to build an author identification system, a very limited work carried out in this domain.

4. Feature Representation

There are different ways to represent features and used for the participation of author discrimination. Frequency count or the feature values are calculated and represented in tabular form. When working with content-based features, it is represented in the bag of words format. How the data presented in the bag-of-words form such as frequency count. To incorporate the factor which affects the writing style of the author due to time, very limited work found in the area.

5. Robustness in Author Identification System

Feature engineering is one big challenge for the problem of author identification. This directly makes an impact on accuracy. The challenges are how to beat the impact of change in writing style over time, which cause-effect on authorship attribution. The concern with word n-gram is to identify what value of n suitable to extract information about the writing style of the author. Existing methods doesn't note this fact. The behavior of the multi-sequence word gram required to verify in the author attribution task.

In this chapter, a review has been taken over the authorship attribution method. Feature selection is an essential criterion for author identification techniques. The writing style of an author gathered in terms of different kinds of features, which in turn used to recognize the uniqueness among each individual author. A detailed review has been conducted on different author identification techniques using different types of features and

commented on them. There are mainly three types of author attribution methods profile-based, instance-based, and mixed of both termed as a hybrid. The writing style of the author is extracted from the history of each individual author; hence, the problem called classification. Different feature selection and classification methods are reviewed in this chapter. The chapter is concluded by quoting various challenges in existing studies.

Chapter 3

A Novel Author Identification Methodology

Authorship attribution is one of the old problem domain to identify the author of text and also act as a tool to recognize cybercrime. The evaluation of writing in terms of style and the content in messages over a long period of time are not yet taken into consideration. In the previous research work, the issue of text size and accuracies were addressed. These challenges are handled with different feature engineering and discrimination approaches. The effect of time in writing style is quoted in the existing work but a significant solution is not available. The solution to the problem of change in writing style in author identification is presented in this thesis. In this research, the temporal changes occurred in terms of author style are analyzed over a long period and identifies the correct author based on analysis of text. The complete work is elaborated in this chapter. The experiment is conducted on the corpus, consisting of texts of authors over a big-time period.

Another challenge addressed in existing research is the significant use of consecutive word sequences in author identification. A novel way is introduced with successive word sequences and it is focused in this chapter. In experimentation, a novel algorithm is proposed to build consecutive word sequences.

Towards the solution of the authorship problem a framework is proposed and it is mainly divided into two parts, first part consist of feature selection and extraction, and in second part a machine learning SVM (Support Vector Machine) algorithm used for classification .

3.1 Feature Selection

The linguistic style of the author is recognized with stylometry in language writing. Features are used to represent styometry and are extracted from the text. This stylometry features are captured from writing by extracting features. In the literature, different types of a feature used to describe author writing style. Among all these types of features, three types of features are selected and those are a character, word, and part of speech. From all of these types, character and word types features can capture both content and stylometry of an author, and part of speech captures the syntactic writing behavior of the writer.

3.1.1 Character n-gram

Character n-gram is a contiguous sequence of n characters for a given sample of text. It is capable of capturing content and writing style from a given text snippet. Character sequences captures author's style as lexical and contextual information, which consist of capital letters and other symbols like special and punctuation characters. It can also mark noisy and erroneous content, a specific use of content sequence, and punctuations. Each writer follows his own traits during writing, which can be quickly pointed out from such types of characteristics. The value of n in character n-grams is to find out with a different number of consecutive characters. N should be big enough to capture contextual, lexical, and thematic information. The value of n is large to the language which has long words and for language which uses small word length n is small. For English like language, most of the research-work kept the value of n to 4. Feature size is increases for small n and reduces when n is large. N-gram is useful in the collection of phonemes, sounds of words. So, it is possible to use the character n-gram feature in language-independent authorship attribution. Table 4.1 shows the character n-gram for the sentence given below:
Text: I am a boy.

Table 3.1: Character n-gram

2-gram	I_,_a, am, m_, _a, a_, _b, bo, oy, y.
3-gram	I_a, am_, a_b, boy, etc.
4-gram	I_am, m_a_, _boy, etc.
5-gram	I_am_, _a_bo,

3.1.2 Word n-gram

Information gathered from the text in the form sequence of words named as word n-gram. Word n-gram is responsible for capturing the habit of the writer, where the writer commonly uses a sequence of words in the text. It also captures variations in vocabulary. Information gathered from the text is in the consecutive n sequences of the wordand termed as word n-gram. Table 4.2 shows construction of word n-gram.

Text: This is great thing.

TABLE 3.2: Word n-gram

Unigram	"This", "is", "great", "thing"
Bigram	"This is", "is great", "great thing"
Trigram	"This is great", "is great thing"

In this type of feature, tag counts are considered. In English language, the tags are limited; hence, feature size also reduces. PoS consist of a standard tagset used in NLTK [89]. Table 4.3 describes the PoS n-gram formation for the sentence given below. NLTK tagsets are used for following example.

Text: The grand jury commented on number

Tag: (AT, JJ, NN, VBD, IN, AT, NN)

TABLE 3.3: Part of Speech n-gram

PoS 1-gram	'AT', 'JJ', 'NN', 'VBD', 'IN', 'AT', 'NN', '.'
PoS 2-gram	'AT JJ', 'JJ NN', 'NN VBD', 'VBD IN', 'IN AT', AT NN', 'NN .'
PoS 3-gram	'AT JJ NN', 'NN VBD IN', 'IN AT NN'
POS 4-gram	' AT JJ NN VBD', 'VBD IN AT NN'

3.2 System Design

In this section, we present our feature transformation approach for the author identification problem. Figure 3.1 shows the proposed framework which consist of following

phases.

- Data collections
- Preprocessing
- Feature selection and Extraction
- Weight vector generations
- Feature Transformation
- Classification

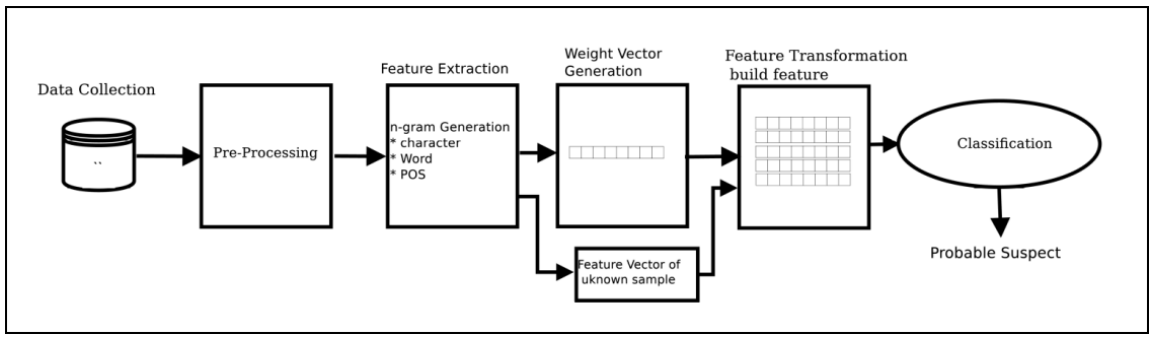


FIGURE 3.1: Author identification system

Very first, we look at the author identification problem, where the potential author of unknown snippet has to find.
Let a set of authors given below,

$A = \{a_1, a_2, a_3, ...a_N\}$ where,

A is a set of N authors.

$M = \{m_1, m_2, m_3, ...m_N\}$

where,

M is set of messages and m_i is set of known sample for author a_i.
The problem in author identification is to identify the authorship of unknown message $M_{unknown}$.

3.2.1 Preprocessing

All texts are collected from web media, newspaper columns, etc., so it contains non-ASCII characters, URLs, dates, etc. Which can not be used to get the writing style of the author; hence, it is not required and need to remove. The stylometry of the author in the text exists in the form of letter capitalization, word suffixes, grammatical mistakes, stop-words hence it should not remove it.

Stemming is also not useful in this scenario. If it is removed then these artifacts from the text are also get removed and it causes to disappear the unique stylometry of the writer. Hence, we should not remove it, and therefore, we can say that there is no need to have any deep preprocessing in this author identification task.

If all the facts in the texts are preprocessed, then it removes the stylistic information which is unique to each author. The information consists of repetitive usage of grammatical mistakes, the way by which contents are represented. There also no point in stemming the vocabulary in a text, as it shows the vital behavior concerning the unique writing style of the author. So we could remove following in preprocessing phase:

1. Non-ASCII characters
2. URL's
3. Date
4. Digits
5. Timestamp

All these facts are removed with the algorithm (1).

Algorithm 1: Preprocessing

Input : Dataset M, set of text sample for all authors, number of author N.

Output: Dataset M' text samples of N author after pre-processing.

begin

$P_e = \{non - asciicharacters, URL's, Date, Digits, timestamp\}$

foreach *m in M* **do**

    $M' = remove(P_e, m)$

end

return M'

end

In preprocessing, the style of the author is protected in the text itself, which on further used for identification. All the selected preprocessing entities have different structural variants and remain unchanged across all writers; hence, removing to this has significance. The process causes to reduce the text size, and it is capable of increasing the effectiveness of the system. In the system, all the removed entities represented in terms of the regular expressions.

3.2.2 Bag of Word Model

It is a kind of model where object categories are represented with a specific method. The main idea behind this is to represent extracted tokens in the form of a histogram. The histogram term describes the number of tokens that appeared and is also termed as frequency count. In the formation of bag-of-word, word frequency counted irrespective of occurrences of words in the document. This is very effective in the classification process. The procedure of creating a bag-of-words is as follows.

1. Represent text in terms of tokens.
2. Calculate the occurrences of words in terms of frequency after pre-processing.

The frequency count is calculated from the selected features in the text. A kind of content that is not contributing to the information retrieval task is removed, such as function words,

but it participates in the authorship identification task. Sometimes, these function words act as meaningful because it holds the usual behavior of the writer, which gets varied for each author. The following example shows the bag-of-words representation for sentences present in the document given below, and it is responsible for holding maximized information.

Sentences:

Document 1: Finally their holidays were over.

Document 2: After holidays they went back to their home.

TABLE 3.4: Dictionary for Bag-of-Words model

#Index	1	2	3	4	5	6	7	8	9	10	11
Token	After	back	Finally	holidays	home	over	their	they	to	went	were

The representation of tokens in the collection of the document is identified in the bag-of-words forms shown in table 3.4. This representation is like a dictionary form where index and tokens are defined.

Now, for every document we have feature vector that holds universal tokens from the corpus. Feature vector for the Document-1 shown below. For this purpose, feature extraction took place by tokenizing each sample from the corpus. Each unique tokens are identified as a feature. And then we represent it like a dictionary where each feature act as key and value is the number of occurrences of the word in the document. Let us called it frequency. The sequence by which tokens appeared in the text does not follow its relative position in the formation of the feature vector is followed. Table 3.5 and table 3.6 represents a bag-of-words form of feature vector for each document. Feature vector for Document-1 shown in Table 3.5.

TABLE 3.5: Bag-of-Words model for document-1

0	0	1	1	0	1	1	0	0	0	1

Similarly we will generate the feature vector for Document-2 shown in table 3.6

TABLE 3.6: Bag-of-Words model for document-2

1	1	0	1	1	0	1	0	1	1	0

The process described in the framework shown in figure 3.1, each document is represented as a feature vector of a bag-of-words form. The same representation used for all types of the feature described in the proposed work, for word n-gram, character n-gram, PoS n-gram, multi-word gram (variable length of word n-gram.

To represent the feature vector for the corpus, we have used the enlisted algorithm (2).

Algorithm 2: Generation of feature vector using bag of word format

Input : Dataset $M = m_1, m_2, m_3, m_4, ...m_n$ for respective author

$y = y_1, y_2, y_3, y_4, ...y_n$ where for each $y \in A$ and A is a set of author

Output: Feature vector V.

```
1  begin
2  Tokens = Tokenize(eachminM)
3  Bag = Unique(Tokens)
4  V = [.]
5  foreach m in M do
6      grams == Tokenize(m)
7      v = [0] * size(Bag)
8      foreach gram in grams do
9          index = getIndex(gram, Bag)
10         v[index]+ = 1
11     end
12     add(v, V)
13 end
14 return V
15 end
```

3.2.3 Hapax Legomena

In the area of linguistic, the hapax legomena is a term that appears for a single time for given entire corpus/documents. This information is sometimes not significantly used as it disregards the stylometry of the writer [90]. So, before beginning the work, such vocabulary in the bag has to remove. In proposed work, hapaxes are removed, which in turn reduces the size of the dataset. Using this algorithm, we removed all the features which appeared in the corpus at once as they appeared once hence not considered for the discrimination of writing parameter. Therefore, this is possible when all sample altogether forms a feature vector. Algorithm (3) shows the procedure to remove hapax legomena.

Algorithm 3: Remove hapax legomena from corpus

Input : Feature vector V.

Output: Feature vector V after removing all hapaxes.

```
1  begin
2  hapax_list = []
3  foreach feature in V do
4      if sum(frequency_count_feature) == 1 then
5          add(feature, hapax_list)
6      end
7  end
8  foreach featureinhapax_list do
9      foreach v in V do
10         if featureinv then
11             remove(feature, v)
12         end
13     end
14 end
15 return V
16 end
```

3.2.4 Feature Generation

It is essential to select the feature types to build the feature vector. As in feature selection, we used three types of features. And for each type, we generate consecutive occurrence of tokens where tokens are n-gram of characters, words, or part of speech tags. In this section, we build consecutive sequence of each feature type.

Character n-gram

Firstly, we select the value n for a consecutive number of characters. To build consecutive n characters as a token from the described algorithm (4).

Algorithm 4: Character n-gram generation

Input : Message m where $m \in M$

Output: T, character n-gram vector

1 **begin**

2 n = consecutive character size

3 charTokens = splitCharacterToken(m) //generate character tokens

4 $T = []$

5 add(T,charTokens$[i : i + n - 1]$) //add consecutive n character as a token to T

6 return T

7 **end**

In the algorithm, initially, the message is broken into characters and then it acts as a character token vector. From this character token vector, a new token is built by taking consecutive n character from the character token vector and add it to the final character n-gram vector. This procedure is repeated for every message for each author.

Word n-gram

In this section, we build a consecutive n sequence of words. To make consecutive word n-gram same steps followed as in character n-gram described in algorithm (5).

Algorithm 5: Word n-gram generation

Input : Message m where $m \in M$ and M is set of all messages

Output: T_w, word n-gram vector

begin

n = consecutive word size

wordTokens = split2wordToken(m) //generate word tokens

T_w = []

add(T_w,wordTokens[$i : i + n - 1$]) //add consecutive n word as a token to T

return T_w

end

The above algorithm shows that messages are split into words and act as word token vector. Then consecutive n-words are selected to build word n-gram. The same process is repeated, as described in the feature selection section.

Part-of-Speech

These types of features are not directly related to the content of the text. Rather than the content, it depends on the type of word or properties of the word used. In this type, tokens are generated from the text sample by the view of tag for each word. And a tagger is used to do so. In the proposed work, we used NLTK tools to tag each word from the text. Instead of textual content, we have a tag for each word. As tags are limited in number hence the number of unique tokens from the corpus are less than compared to the character and word n-gram. The universal tag part-of-speech tag-set are shown in figure 3.2.

Tag	Meaning
ADJ	adjective
ADP	adposition
ADV	adverb
CONJ	conjunction
DET	determiner, article
NOUN	noun
NUM	numeral
PRT	particle
PRON	pronoun
VERB	verb
.	punctuation marks

FIGURE 3.2: Universal PoS tags

The packages used in work is NLTK, it supports a set of part of speech tags defined shown in figure 3.3.

Following is an example of constructing a part-of-speech tag.

Text: They permit us to do something different

TABLE 3.7: PoS tag

Word =>	They	permit	us	to	do	something	different
Tags =>	PRP	VB	PRP	TO	VB	NN	JJ

Algorithm (6) is used to generate part-of-speech shown as below.

Algorithm 6: PoS n-gram generation

Input : Message m where $m \in M$ and M is set of all messages

$Tags$ = Standard Tags from nltk

Output: T_{PoS}, PoS tag n-gram vector

1 **begin**

2 n = consecutive PoS tag size

3 posTokens = getposToken(m,$Tags$) //generate PoS tag tokens

4 T_{PoS} = []

5 add(T_{PoS},posTokens[$i : i + n - 1$]) //add consecutive n PoS as a token to T

6 return T_{PoS}

7 **end**

CC coordinating conjunction
CD cardinal digit
DT determiner
EX existential there
FW foreign word
IN preposition/subordinating conjunction
JJ adjective
JJR adjective
JJS adjective
LS list marker
MD modal could,
NN noun, singular
NNS noun plural
NNP proper noun
NNPS proper noun
PDT predeterminer
POS possessive ending parent‘s
PRP personal pronoun
RBR adverb, comparative better
RBS adverb, superlative best
RP particle give up
TO to go ‘to‘ the store.
UH interjection
VB verb, base form take
VBD verb, past tense took
VBG verb, gerund/present participle
VBN verb, past participle taken
VBP verb, sing. present, non-3d take
VBZ verb, 3rd person sing. present
WDT wh-determiner which
WP wh-pronoun who, what
WP$ possessive wh-pronoun whose
WRB wh-abverb where, when
PRP$ possessive pronoun
RB adverb very, silently,

FIGURE 3.3: PoS tags in NLTK package

A distance metric is used to discriminate between the two texts. It shows how far a text is different from another. There are many types of such distance calculating metrics, which we discussed in the literature. In this work, for the calculation of the distance in terms of the score using cosine similarity.

3.2.5 Cosine Similarity

To find the similarity between two documents, each of them represents using vector. Hence, two vectors are there to show these two documents. Now the distance in terms of similarity score is identified with cosine angle between these two document vector. Each vector points in the same direction. When cosine angle between two vectors is 90 degrees then there is no similarity between these two document. When angle reaches to 0 degree then similarity score become one, as shown in figure 3.4 [91].

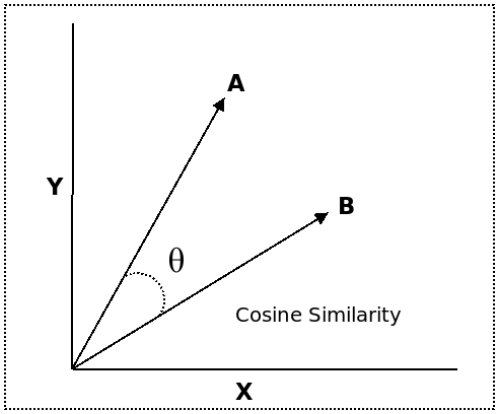


FIGURE 3.4: Cosine similarity

Let us have two document vectors named A and B for which a similarity has to find. We have an equation shown below to find similarity.

$$Similarity(x, y) = \frac{A.B}{\|A\| \, . \, \|B\|} \tag{3.1}$$

where,
$A = \sqrt{a_1^2 + a_2^2 + a_3^2 + a_4^2 + ...a_n^2}$ and
$B = \sqrt{b_1^2 + b_2^2 + b_3^2 + b_4^2 + ...b_n^2}$

3.3 Weight Vector Generation

To generate a weight vector, it is required to identify the simultaneous change that occur for each author over time. To do this, we have to investigate the style change parameter for every author. To deal with this situation, we have to compare each text sample for each author with the document written in the latest time. With the aggregated similarity score, we extract normalized factor as a weight, which in turn brings each writing sample to the most recent time. To generate a weight vector, a linear regression method is used to calculate the normalized factor. The normalized factor called 'Transform Feature to Current Time' (TFCT) function. The methodology used here is from the inspiration of the method described in [32]. Where a similarity-based approach was used. Initially, a scalar factor is calculated, then this scalar factor is applied on target similarity score to get final similarity index. And it act as a target score for discrimination and identification of the author. To calculate the similarity-based score, the algorithm uses in the existing

technique elaborated in the work [13]. The contribution described in this thesis does not uses a similarity-based approach it uses machine learning classification methodology. This could be done by introducing a novel technique where instead of scaling similarity score, features are transformed to the latest time period. This causes that all the features makes available in most recent time period hence scaling factor is not applied at the decision level.

The work begins with the calculation of the decay factor for every author as the decay rate and it is not same for all author. It represents the stylistic influence occurred due to time. The writing style change rates are more for modern authors and less for others. It depends on literacy, age, country, mother tongue. Various techniques were proposed to calculate the decay factor in the work [78, 92]. Exponential decay function proposed in [93]. The linear regression is used to calculate the scaling factor with equation 3.2.

$$decay(t) = \frac{1}{Z}(m.t + c) \tag{3.2}$$

where,
Z is normalized factor calculated from all decays so it is in-between 0 to 1.
M and c are parameters of linear regression function.

In the novel approach, linear regression is used to find the relationship between similarity score versus time at which the text content is written. Very first, all texts written by the same author are arranged according to the time at which they were written. Cosine similarity metric is used to calculate the similarity of all the document with the latest time period document. Cosine similarity measure is discussed in the previous section. Linear regression in equation 3.3 is used to get its parameters.

$$SimilarityScore = b_0 + b_1.t \tag{3.3}$$

where,
b_0 and b_1 chosen in such way that the error gets minimized and b_0 termed as intercept and b_1 considered as coefficient.

The calculated value of the intercept and coefficient are used to identify the decay factor. The hypothesis behind decay is that as age of document grows it decay value increases. In short, the document written at older time period are less similar to latest text as compare to the document written at nearer period.

The decay value calculated for each author from equation 3.2, which act as weight function for all author N, weight vector shown in equation 3.4.

$$W = \{w_1, w_2, w_3, ...w_N\} \tag{3.4}$$

where,
$w_i \in R^m$ calculated for each author a_i from equation 3.5.

$$w_i = decay_i(t_s - t) \tag{3.5}$$

where,
t_s is the time of an anonymous text sample and t is time for known text sample.

3.4 Transform Feature to Current Time

With the influence of decay function with respect to time and machine-learning concepts, we employ a novel feature transformation approach to build a feature vector and apply machine learning to construct a classifier for prediction of the most probable author of anonymous texts. After the weight vector generated, it employs feature transformation function to transform every feature to the latest time-space from all the documents. Our approach is based on the fact that the writing style of the author changes over time, and newer text is most similar as compared with older text. So assigned weight to older text is greater and reduces when applied to newer documents. We transform all these feature statistics to the current time, and then we use a classification algorithm to build the model which results in probable authorship. To design this approach, a function is formulated and it is called as a feature transformation function TFCT shown in Equation 3.6.

$$F_t = F_t + w_i.F_t \tag{3.6}$$

where,
F_{t_s} is the tranform feature to current time t_s,
F_t is extracted feature at time t and $w_i \in W$.

With the approach described in this thesis, features are transformed hence it is not bounded with the similarity-based approach. So, we can apply a machine-learning

algorithm to discriminate author style on transformed features. To find the right author candidate, the SVM classification algorithm is used, where a training set consists of a feature vector of known text predicted with a constructed SVM model and a feature vector of anonymous samples predicted with a build model.

3.5 Classification

This is the last step of the system, where a focus is made on classification strategies. As discussed previously, we represent our all text samples in the combination of profile-based as well as instance-based approach. We called it as profile-based because, at the beginning a group of text samples are brought altogether according to the year span. Hence, for each author, a profile is built based on each year. Then, these profiles for each author in each year act as instances. And now, it becomes an instance-base approach. Hence, we can utilize the machine learning classification algorithm, which works well. In the classification, the text is assigned to a known category based on its extracted attributes. There are several classification algorithms and each of them has its own significance and used in different types of problems. The classification algorithm used in this section is the support vector machine (SVM), which has its significance. A support vector machine is a type of supervised learning technique used in classification, outliers detection, and regression. It has the following advantages:

1. Effective in high dimensional spaces.
2. Very effective in the cases where the dimension of feature vector is greater than the number of samples.
3. Uses a subset of training points in the decision function (called support vectors), hence memory efficient.
4. Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

In support vector machine makes single or multiple hyperplane at a higher dimension for classification, regression, or another task. Separation among the data points of classes are achieved through hyperplane.

Larger the distance from the hyperplane indicates a good separation of data points. As larger margin have less generalization error in the classifier. SVM solves the binary classification problem as well as multi-class problems. In multiclass problem, it act as binary classifier with one class versus remaining classes. Kernel function in SVM is responsible for transforming non-linear data points into linear. The support vector machine supports different types of the kernel from polynomial to sigmoid function type. Before the use of SVM, the classifier user has to select appropriate kernel functions based on the problem domain.

After that, the classification tool will take care of the rest process [94]. The kernel of SVM proposed in this work given in equation 3.7.

$$f(x) = w_T.x + b \tag{3.7}$$

This equation solves binary classification problem the value of $f(x) > 0$ then the sample belongs to desired classifier i.e $y = 1$ otherwise it is not, and having set $y = -1$. The above equation can also be written as in equation 3.8.

$$f(x) = \sum_{1}^{m} \alpha_i.y^{(i)} K\left\langle x^{(i)}, x\right\rangle + b \tag{3.8}$$

where,
$K\left\langle x^{(i)}, x\right\rangle$ is kernel function.
Support Vector Machine Optimization (SMO) is an algorithm used to train support vector machines (SVM). This is the case of binary classification in multi-label classification problem; the problem evaluated for the class $C+$ and $C-$, $C+$ is class for which we get $f(x)$ higher than the rest of classes. The same process is repeated for each class and belonging find on versus others. At the end maximum score belong to a class is considered as a winner.

3.6 Algorithm of Proposed System

In this section, authorship attribution over a long period using the feature transformation method is described. The complete system is divided into three steps.

1. Feature Selection and Extraction.

2. Generate weight vector using decay function.

3. Predict the author of unknown text using classification method.

Firstly, there are few requirements for the dataset to get effective outcomes and it will be discussed in chapter 5.

Algorithm Steps

Step 1: Organize all training samples according to author.

Step 2: For each author group all samples according to year.

Step 3: Do the preprocessing as mentioned in section 4.3.1.

Step 4: Extract features of type $T = \{$Character n-gram — word n-gram — PoS n-gram$\}$

Step 5: Find decay function for each author shown in equation 3.5.

Step 6: Build feature vector V over n-gram features ng, where ng belong to T.

Step 7: Transform feature vector V to latest time using TFCT function to generate weight vector shown in equation 3.6.

Step 8: Use SVM classification algorithm to classify anonymous snippet from language model constructed in step 7.

3.7 Mathematical Model

According to set theory, the system is described as follows:
Let S be the proposed author identification system, such that

$$S = \{A, M, U, N, F_T, p_a^t, F, D_a, W, C, A_{unknownMSG}\}$$

where,
A is author set, and given as:

$$A = \{a_1, a_2, a_3, a_4, ..., a_n\}$$

M is a set of messages given below:

$$M = \{m_1, m_2, m_3, ...m_n\}$$

where,
$\exists m_i = \prod_{j=0}^{k} msg_j^i$ and,
msg_j^i set of all messages written by author a_i.

U is set of unknown messages.

N represents consecutive sequenses of F_T for $M \cup U$.
where,
Features F are of type T and

$$T = \{\text{character n-gram, word n-gram, PoS n-gram}\}$$

p_a^t set of profile of author a at time t for all time periods.
This profile is in terms of feature F_T.
In this way, for every instance that is message m_i, a feature vector F_T build and accumulated in F define as below:

$$F = \{f_1, f_2, f_3, ...\}$$

D is decay factor for each author a is calculated by equation 3.2 and weight vector is generated for every author a by equation 3.5 and which is generated as below:

$$W = \{w_1, w_2, w_3, ...\}$$

All features are now transformed into the latest time period with a weight vector, as shown in equation 3.6.

$$F_{M \cup U} = F + F * W$$

C is classification algorithm,

$$C = \{c_1\}$$

where c_1 is support vector machine classification algorithm.

Now lets build the classification model from all known message M, which are considered in terms of feature vector F as defined above.

$$\theta = modelSVM(F_M, A)$$

and then predict the author for every unknown messages

$$A_{unknownMSG} = Predict(\theta, U)$$

In this way, authors of all unknown messages are predicted.

3.8 Variable Length Word Gram for Author Identification

Overview

Character and word n-gram are the most followed method for feature construction and participates in the authorship identification task. In this section, we point out on word n-gram. The approach described in the chapter does not depend on the constant value; the value changes according to the occurrence of word sequences. The methodology applied to the collection of text which is from the varied time domain. Dynamic value of n chosen to generate word sequence. The number of features generated by this method are less as compared to the feature created on a constant value of n, in word n-gram.

Introduction

Character n-gram and word n-gram are ways to capture the style of the author. In this type, most repetitive n sequences of the text gives stylistic information on the lexical, syntactical, and structural level [1]. Hassan in [95] shows the effectiveness of character n-gram, where bi-gram and tri-gram were used to verify author of texts . Character n-gram captures emoticons - use of punctuation from text documents. It is responsible for accepting important features without describing subject in the text documents. And it captures both styles and content and it performs well and robust than other types of features. It is also

capable to handle syntactic information from texts [1, 14]. But when used, it consists of consecutive n characters, which has break-in words and join two words from spaces. Hence, it is difficult to understand at how far to capture the stylometry of author. Rather than looking into this, we found word n-gram more related to content. Word n-gram captures semantically meaningful information from the text in the form of short phrases. It can point out short repetitive phrases and similarly it captures syntactic and semantic features. The punctuation sequence is also one of the captured properties from texts. In this section, we worked out on word n-gram features. Initially, we found the sequence count of n affects the performance of authorship analysis. Our main motto is to identify the behavior of the system when the constant value of n taken. Is really performance get affected on such variable constant values? Is feature space affected when we cumulatively use a variable length of n in word n-gram? Whether feature count affect the performance on the dynamic value of n in multiword gram features? Our task is to find the answer of all these questions through experimentation. By taking a previously used variable length of n-gram, we redefine the way of finding variable length in word n-gram. On proposed variation followed by a set of experimentation the performance affected by selecting the different number of features is described in the later part of the thesis. Contribution to variable length word gram is as below:

1. Defining an approach for word n-gram feature selection based on the dynamic selection of n consecutive word.
2. Generate set of rules depends on number of consecutive word sequence value n.

Proposed methodology expected to outperform state-of-the-art of system with respect to accuracy and effect on feature count.

3.9 Variable Length Word n-gram Approach

In the author identification problem, our main contribution is towards feature construction in word n-gram. Towards the solution of the problem, words extracted from the corpus, and then for finalizing the features, the proposed approach is applied to build final feature set which is then used for classification. This methodology is applied to

consecutive words and encouraged from the technique used in [75]. The framework of the system described in figure 3.5.

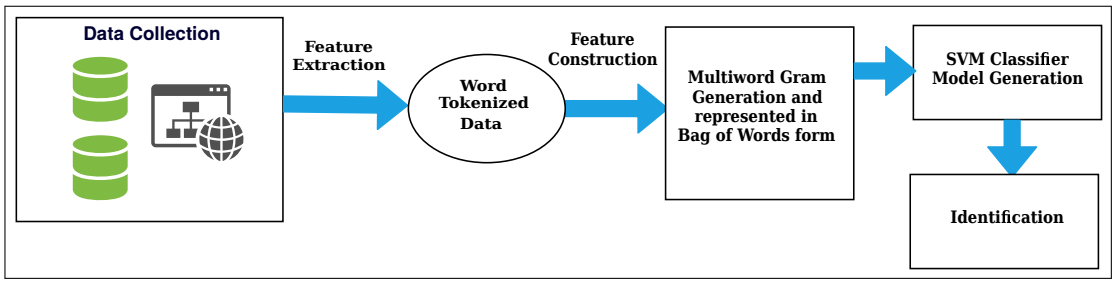


FIGURE 3.5: Variable length word n-gram system

The system shown in figure 3.5 consist of five steps:

1. Data collection
2. Feature extraction
3. Feature construction
4. Model building
5. Identification (Author prediction)

Data Collection

In this section, we described data, and the data is in terms of textual writing of authors at different time periods. The sources of data are from newspaper columns, articles, and letters. Experimentation section describes the dataset briefly. The dataset used in this approach is same as used in author identification with feature transformation methodology and it is described in section 5.1. A TFCT function used to bring all the features to the current time so the effect of the time gets nullified.

Feature Extraction

The identified feature of our problem domain is the word. So, we split the text sample into words, act as a token. In this work, again, we used consecutive word sequences termed as word n-gram. After the extraction of words as tokens, it is represented in vector.

For every author, we have a text vector in terms of the token in sequence as it appears in the text sample. These extracted tokens are ready for the next step, where features constructed as consecutive tokens.

Feature Construction

In the review, the author identification used a variable length of character n-grams, where instead of a fixed length of character sequences, a variable length of n-gram features represented. In this approach instead of using constant length of n, all possible length n, from 1 to n is used in character n-gram. We applied a similar type of technique to formulate variable sequences of consecutive words.

Algorithm 7: Variable length word n-gram generation

Input : Message m where $m \in M$ and M is set of all messages

Output: T_{vw}, variable length word n-gram vector

```
1  begin
2  wordTokens = split2wordToken(m)
3  i = 0
4  T_vw = []
5  foreach i in range(0 to len(wordTokens)) do
6      n = variableSize(len(wordTokens[i])
7      add(T_vw,wordTokens[i : i + n − 1]) //add consecutive n word as a token to T
8  end
9  return T_vw
10 end
```

In our work, we have selected word-based features, in which consecutive words were chosen as a feature set. Word sequences in the document are capable of capturing syntactic and semantic properties along with regular occurrences of the vocabulary. Then it is used to discriminate the writing style[96]. To construct the features from the word vector, we brief a set of rules to formulate a variable size of n in multiword gram, which depends on the size of the word. The rules-based on the specific hypothesis states that short length word sequence discriminates more than the consecutive words appear of big length. Hence, the value of n is large for short length word and for big length words the

value of n is et to small. In the described approach n varies from 1 to 3. The algorithm (7) shows the generation of variable-length word n-gram.

3.9.1 Variable Length n-gram

In proposed approach variable length word n-gram is used for author identification task. The definition of a variable value of n is based on the length of each word occurred in the document. A focus made on the fundamental use of word size. Word length has the strength to identify the vocabulary richness of the writer. In regular writing big length words appear less than the smaller length words. Based on this fundamental consideration the value of n decided to build feature set. There is a relationship between smaller length words versus big length words in documents. Short length words appears more number of times than the big length words. With this approach the feature size also gets reduced. Higher the value of n for short length words and the value of n more towards 1 for bigger length words.

Methodology

The problem is to identify the author of writing samples. A is set of author and known document set D. Our approach takes writing sample of author set $A = \{a_1, a_2, a_3, ...\}$, for each author a known Document set $D_a = \{d_1, d_2, d_3, ...\}$. We have to find author of unknown document d_u from author set A. To identify the author using variable-length word n-gram, we go through the following steps.

1. For each writing sample from D, the data is preprocessed by removing non ASCII characters along with punctuation marks as we focused only words rather than text structure and syntax.

2. Tokenize each writing sample, where each token is a word.

3. Build Feature vector W for each sample by considering consecutive n words or tokens, with dynamically varied value n. This n value is calculated by our proposed approach. This is done for training and testing documents.

4. Represent extracted features W into the bag of words form.

5. Build feature vector from bag of word representation.

6. Apply various classification strategies for verification of suspectable author.

Generating Multi-word Gram

In multiword gram word sequencing is varied from one to three. As our assumption based on the length of word. As the length of word is fewer, occurrences of sequencing are more, and big length indicates the vocabulary richness and fewer occurrences of words. So by considering this assumption, we formulate a set of rules to determine the value of n as below. For each document D_a, of every author a we have word tokenizer set W_{D_a}.

$$D_a = (w_1, w_2, w_3, ...w_{n-1}, w_n) \tag{3.9}$$

From the above document representation, we have to build a feature set W_{D_a}. Each element in the sequence in feature set W_{D_a} created from following a set of rules.
Rules Described as below:

1. if $length(w_i) < 4$, then $W_{seq} = (w_i, w_{i+1}, w_{i+2})$ where $i <= n-2$

2. if $length(w_i) > 3$ and < 8 then $W_{seq} = (w_i, w_{i+1})$ where $i <= n-1$

3. if $length(w_i) > 7$ then $W_{seq} = (wi)$ where $i <= n$

Now we have each document represented as word sequences set W_{D_a} for each docuement D written by author a.

Bag of Words Representation

Bag-of-words is a model commonly used in categorizing documents by representing words occurring in terms of frequencies. It is used to represent the document feature vector. We can consider a feature set F as below.
$F = [this, is, one, and, only, one, which, is, one]$
Its equivalent bag of word representation is shown in table 3.8.

Table 3.8: Bag of word representation

Word	this	is	one	and	only	which
Frequency#	1	2	3	1	1	1

Once each document represented in a bag-of-words form, then it is used to build a feature vector, which on further used for finding authorship.

Model Building and Prediction

In particular classification process, all instances are partitioned into two subsets, one is training, and the other is testing. The training sample utilized to build and learn the classification model. The prediction power of the classifier validated on the test subset. Accordingly, the author identification classification model is generated. In the scenario, we used a support vector machine and naive Bayes classification model. Depending on the performance of the classification model which are verified on the testing samples, unknown messages are applied to the model to predict the correct authorship.

3.9.2 Mathematical Model

Let S be the author identification system using a variable-length word n-gram approach.

$$S = \{A, M, U, W, N, F, C, A_{unknownMSG}\}$$

where,
A is author set, and given as:

$$A = \{a_1, a_2, a_3, a_4, ..., a_n\}$$

M is a set of messages given below:

$$M = \{m_1, m_2, m_3, ...m_n\}$$

where,

$\exists m_i = \prod_{j=0}^{k} msg_j^i$ and,

msg_j^i set of all messages written by author a_i.

U is set of unknown messages.

For each text sample, we have tokenized word vector:

$W = \{w_1, w_2, w_3, ...\}$

where,

W is a word token vector set

Feature vector F is generated as below:

$$F = ConsructFeature(N_{dynamic}, W) \tag{3.10}$$

Where,

The value of N calculated for each token vector and it is based on the length of current token word from word vector hence termed as $N_{dynamic}$

$N = findN(Length(w))$

C is classification algorithm,

$$C = \{c_1\}$$

where c_1 is support vector machine classification algorithm.

Now let build the classification model build from all known message M, which are considered in terms of feature vector F as defined above.

$$\theta = modelSVM(F_M, A)$$

and then predict the author for every unknown messages

$$A_{unknownMSG} = Predict(\theta, U)$$

In this way, authors of all unknown messages are predicted.

This chapter focused on the detail implementation of the methodology in this research. The features used in the system are explained in detail. The research component

of the system is a 'Transform Feature to Current Time' function. TFCT function build with a decay factor, which calculated through a change in writing style with respect to time. The classification algorithm used in the system discussed. Each stage of the algorithm briefed in this chapter which starts from preprocessing phase.

Word grams applied in a novel way in another set of experiments. It termed as variable-length word gram, wherein word n-gram, n is dynamically changed based on the length of the current word. A novel algorithm is proposed to create variable-length word grams.

Finally, mathematical model of each methodology is described in this chapter.

Chapter 4

Experiments and Results

The writing style of the author changes over the time, and this fact has to be considered in the authorship attribution system. This research gives a direction to handle such change over time. In this section, a corpus over the significant period described and then style change over time is observed and evaluated, which later on is used in feature transformation. When all features are transformed to the latest time, the original writer of the unknown text sample is identified. In this way, experiment is performed in two stages. First is to identify style change factor, and in the second stage, it is applied in author identification task. In the experimentation, different types of features are evaluated with several performance parameters.

With another aspect, instead of using character n-gram, the word gram feature chosen for further experimentation. Variable-length word n-gram approach selected for feature building. A series of experimentation is made with a different configuration for analysis and observations.

For the comparative analysis, SCAP and naive-based similarity method algorithm is used. All experiments are followed by corpus selection, experiment setup, experimentation, and performance evaluations.

4.1 Corpus

There are hundreds of documents and articles used in training and testing which are collected from various online web sources. The goal of the experiment is to identify the

Table 4.1: Dataset Description

Sr. No.	Author	Years
1	Abidgil Smith	38
2	Benjamin Franklin	35
3	Bill Keller	10
4	Coomi Kapoor	20
5	David Brook	29
6	John Adams	40
7	Mahathma Gandhi	48
8	Michael Cooper	28
9	Pratap Bhanu Mehta	10
10	Ron Rolheiser	25
11	Swami Vivekanand	10
12	Thomas Friedman	11
13	Tanveel Singh	10

change in writing style with respect to time. To prove this fact, we chose the documents whose writing time known. Following are specific requirements of the dataset.

- Dataset in the form of text document written in English language.
- Enough Time span for significant result, we assume long time period, so good variation in writing style can be captured.
- To get better result, handwritten documents are preferred, but should be in textual soft-copy form (not image). As hand written documents have pure writing style of writer.
- Enough document length is required to gather features. Text size should not too short in terms of word count.

To satisfy all the above requirements, we didn't find any dataset available to work on. In time aware author identification [32] uses the Enron email and tweeter dataset and which also belong to a small-time period up-to four years only. It is not sufficient to capture the change in writing style. To satisfy all our requirements, we have chosen both types of writing, handwritten, and typed written textual content. Typewritten documents are accumulated from a columnist of the newspaper. For data set collection, we used the NY Times, The Indian Express newspaper, and handwritten documents of well-known authors. The list of authors and their writing collection are described in the table 4.1.

TABLE 4.2: Dataset Statistics

No. of Authors	Total Messages #	Average Message per author #	Average words per message #	Time span in years #
13	774	60	680	25

We have a data corpus of 13 authors over the period from 6 years to 48 years, on average, 22 years time span for authors. Each document name identifies its date of creation.

In the dataset, we assumed that the handwritten document could have many realistic features. Because while writing author can put his effort into writing rather than the helping tools available at on-line editor such as Microsoft Word,text editors, etc. We have collected numerous data from real-world entities. We have collected digital as well as handwritten docs whose digital copy available on-line. We have collected data of 13 authors of various time spans. Mix dataset was considered.

To perform the experiment, we build a corpus in English language text collected from a newspaper columnist. NY Times, Indian Express, etc. Corpus consists of columns written by the author over a long period, so variation is there due to the factor of time, which may cause changes in the writing style of the author. For every author, we have more than 50 documents files, and each file consists of more than 500 words. To work with more real writing, handwritten letter converted into texts are used for experimentation. Total of 13 author's corpus used in the experiment. Statistics of the dataset are shown in table 4.2. The average period of the collected document is 25 years, and on average, 60 messages available for each author.

4.2 Experiment Setup

In this section, we discuss the experimental setup, training and testing data, results, and observations. All experiments are performed on the system with configuration core i5 2.50 GHz, 8 GB RAM. And to perform experimentation, we used python 2.5 and WEKA toolkit. In the first contribution experiments are performed in two steps. The first step used to evaluate the change occurs over time for each author, and in second step, the author identification system over selected features. In both the steps the feature selection and extraction performed in the same manner.

4.2.1 Author Identification with TFCT Function

We divide the experiment into two steps, the first step is to find decay function parameters, and the second is the author identification. For experimentation, we used three types of features:

1. Character n-gram
2. Word n-gram
3. Part of speech n-gram

For all the experimentation above types of features are used. And feature extraction process is carried in python till we get end feature vector data. It is carried out in the following steps.

1. For each author, corpus grouped into years.
2. Feature extraction
   (a) Feature vector is built for each sample.
   (b) It is instance based approach hence each instance represents each sample.
3. For each author find drift in terms of decay factor.
4. Calculation of TFCT function for each author.
5. Transform feature of each sample to current time using TFCT function.
6. Python used for building features.
7. Features are stored in csv file.
8. WEKA tool is used for classification model building and performance evaluation.

4.2.2 Time-aware Author Identification Performance Parameters

The research is carried out in two phases, in first the impact of time is evaluated over the writing style of an author with a set of parameters, and then a function proposed and implemented and evaluated with another set of parameters. One of the important

task is feature selection to measure the effect of time correctly. There is a need to extract effective style change; hence, feature selection is an important criterion to proceed in the right direction. Numerous types of features and their extraction methodologies described in study [1, 71].

Similarity Measures

To differentiate among text, it must be required to determine the similarity measure. The measure used to find the level of closeness and distinguishable to set of the desired object through extracted properties from text data. In this context, similarity measures used to find the change in two writings of an author. Different types of similarity measures are used to calculate the difference among various writings [97]. As to find the drift in the writing style of an author over time, cosine is the most used, having the capability to identify the distance between two texts.

In the system, to calculate the difference between two text documents, each document represented as document vectors, and the similarity between these is a correlation among these vectors. In cosine similarity, the cosine angle between the vectors is calculated. For two given documents d1 and d2 be the vectors, their cosine similarity expressed as in equation 4.1.

$$Sim_{cosine}(d_1, d_2) = \frac{d_1.d_2}{|d_1| \times |d_2|} \tag{4.1}$$

where,

d_1, d_2 be the m-dimensional term vectors, where term are the feature extracted from each documents.

The significance of cosine similarity is that, it doesn't depend on document length.

Linear Regression

When the collection of text samples are present for each author then it is require to find the relationship among writing samples over time. This rate of change is calculated with the regression method, where the variable is a similarity score.

Linear regression is a statistical method for the analysis of observed data. It describes how the value of response changes as the predictor value changes. With the

linear regression, we find the parameters as intercept and slope. The equation of linear regression is shown in equation 4.2 [98].

$$E(Y|X = x_1) = B_0 + B_1.x_1 \tag{4.2}$$

where,

B_0 is intercept, it is value of function above when x_1 is zero.

B_1 is slope, it is rate of change of above function as value of x_1 changes.

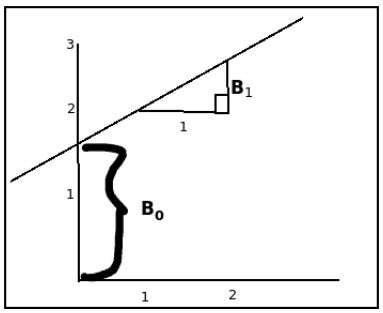


FIGURE 4.1: Linear regression

Slope Parameters

This metric is used to measure the continuous change occurred in writing of authors. This slope is calculated with the help of linear regression. Slope calculated for all authors and it represents how changes are captured. To evaluate change the value of slope analysed for all authors in terms of following.

Minimum slope: It is minimum slope accountability among all author.

Maximum slope: It is maximum slope accountability among all author.

Average slope: It is mean slope value for all authors.

Standard deviation: It is deviation [99] from the mean value of slope and calculated from equation 4.3.

$$S = \frac{1}{N}.\sqrt{\sum_{i=1}^{N} x_i - \overline{X}} \tag{4.3}$$

4.2.3 Evaluation Parameters for Author Identification System

The research is carried out in two phases, the first one is to finalize the performance evaluation parameters of the system and second is to implement author identification technique using feature transformation method. To evaluate the performance of the proposed system different parameters related to accuracy are described in this section.

Confusion Matrix

In author identification, an unknown writing sample will be assigned to a probable author from a set of known authors using a classification algorithm. Such prediction accuracy evaluated through correctly classified instances against incorrectly assigned instances. A confusion matrix is a useful tool for analyzing the performance of these assignments. All the parameters for the evaluations are extracted from the confusion matrix.

Confusion matrix is used to assess the performance of the classification model. It gives the information about the prediction versus actual result, and it represents in a square matrix form where columns are corresponding to the predicted class, whereas row represents actual class [100]. A typical representation for binary class is shown in figure 4.2.

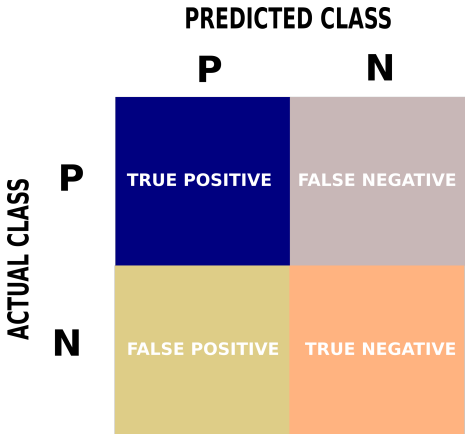


FIGURE 4.2: Confusion Matrix

Positive: When observation is right (yes/correct).

Negative: When observation is wrong (no/not correct).

True Positive (TP): When predicted and actual results are positive.

True Negative (TN): When both, predicted and actual result is negative.

False Positive (FP): When prediction is positive but in actual it is negative.

False Negative (FN): When actual result is negative but predicted as positive.

Accuracy

In the author identification system, a vital metric used to validate the performance of the system is accuracy. Accuracy is one which is calculated from the average accuracy of the classification method for all n folds, and it is estimated as in equation 4.4.

$$Accuracy(\%) = \frac{number.of.correctly.classified.entities}{all.entities} * 100 \tag{4.4}$$

It can also be easily calculated from confusion matrix as from equation 4.5.

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{4.5}$$

True Positive Rate (TPR) / Recall

In the era of information retrieval, it is also termed as recall or sensitivity. It identifies correctly recognized classes (labels). In author identification, for the significant performance of the system, the value of TPR should be higher and reaching towards 1. To evaluate the aggregate performance of the system in terms of TPR; the average value of TPR is considered for all the classes. It is calculated with equation 4.6 .

$$TPR = \frac{TP}{TP + FN} \tag{4.6}$$

Precision

It is one of the important parameters in the author identification system, and it defines the proportion of correctly identified to an author, from overall identification to an author. Along with the accuracy of the system, the individual performance for each class

can be evaluated with precision values. It is calculated in equation 4.7.

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

False Positive Rate (FPR)

In author identification, this term belongs to wrongly classified items to an author against all the items which do not belong to an author. It is represented with equation 4.8.

$$FPR = \frac{FP}{FP + TN} \tag{4.8}$$

It can also be represented as below.

$$FPR = 1 - TNR \tag{4.9}$$

Where,
TNR (True negative rate) is given in 4.11.

$$TNR = \frac{TN}{FP + TN} \tag{4.10}$$

F-measure

It is also called as F1 measure or F-score. It is used to measure the predictive power of the classification procedure. More the value of the F-score, better the classification procedure performance. For perfect classification, the value reaches 1. It is a combination of precision and recall and assumes the harmonic mean of both. The harmonic mean (h) is shown in figure 4.3.

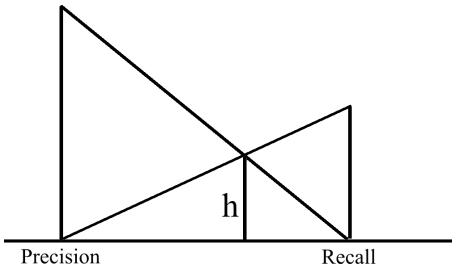


FIGURE 4.3: Harmonic mean

F-measure is calculated from the equation given below.

$$F - measure = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4.11}$$

Where,

Precision is calculated in equation 4.7, and

Recall is calculated in the equation 4.6.

Feature Size

Feature size directly affects the performance of the system. Feature size indicates number of attributes exist in each sample. Features participate in model building, and are used for predication. Feature size is always considered as performance parameter along with accuracy. Accuracy is calculated from equation 4.5. Feature in the system can be described in the equation given below.

$$X = \{\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, ...\} \tag{4.12}$$

where,

$\overline{x_i} \in \mathbb{R}^m$

The impact of feature size can be ovserved with accuracy obtained from system.

4.3 Results and Evaluations of Author Identification with TFCT

The goal of the first step is to find the change that occurred in writing style due to various aspects. After preprocessing all texts samples, features are extracted to build feature vector. The evaluation made on three types of features character n-grams, word n-grams, and PoS n-grams. To see the effect of time over text written at different time periods, we measure these changes with the linear regression method. The similarity score is calculated between the latest known sample and the rest of the samples. Linear regression is applied to the similarity score to get the drift. This drift is in terms of parameters of regression for all authors.

TABLE 4.3: Slope of different features

Feature Type	**Max. slope**	**Min. slope**	Average slope	Standard Deviation
PoS 3 Gram	0.02	0.001	0.01	0.009
PoS 4 Gram	0.06	0.001	0.02	0.019
Char 4 Gram	0.02	0.002	0.01	0.009
Char 5 Gram	0.04	0.003	0.01	0.015
Word 2 Gram	0.07	0.003	0.03	0.025
Word 3 Gram	0.1	0.004	0.03	0.04
Word 4 Gram	0.1	0.003	0.04	0.04

Slope showed in the Table 4.3 confirms that the changes occurred over time is limited over time. However, there is a significant difference between the minimum and maximum value of the slope. For each type of feature, we found different values as shown table. Over time, there is a change in writing style and vocabulary usage in the author's writing. All the values are shown in table 4.3 are cumulatively presented for all authors.

The size of the training corpus is 774 texts of 13 authors. In traditional author identification, all features are considered without concerning about time. Our approach includes a parameter time, which improves the final result of the identification process. Previous methods are based on similarity, hence at the end, similarity strength was weighed by the time parameter. In our methodology, all extracted features from the texts are normalized to the latest period with feature transformation method, and then machine learning classification SVM is applied to the text written by author. We have collected differential results described in next section.

When the PoS feature type used in author identification, the result enumerated in table 4.4 in terms of accuracy. We compare the accuracies, when feature transformation function is applied and when not applied. As a result, we found that when feature transformation TFCT function used, the result is improved for both type, PoS 3-gram, and PoS 4-gram. For PoS 3-gram and PoS 4-gram we get accuracy 74.93 %, and 81.65% respectively. For PoS n-gram feature types, the using proposed feature transformation function, the accuracy increased.

Table 4.4: Accuracy for feature type: PoS tag

Features	Without TFCT	With TFCT
PoS 3-gram	71.32	74.93
PoS 4-gram	75.19	81.65

Character n-gram shows higher accuracy than PoS n-gram. When character 4-gram and 5-gram used as a feature then we get accuracy shown in table 4.5. Again comparing accuracies with and without the use of feature transformation method we found that the accuracy with the proposed feature transformed method shows improvement. Character 5-gram has greater accuracy as compare to character 4-gram, which is about 94.83 %. As a result, Character n-gram having more accuracy than PoS n-gram. So here we can say that character n-gram identifies greater stylometry than PoS n-gram where PoS n-gram is not directly related to the content. It is more concerned with the grammatical stylometry.

Table 4.5: Accuracy for character 4, 5 gram

Features	Without TFCT	With TFCT
Character 4-gram	89.86	92.11
Character 5-gram	91.60	94.83

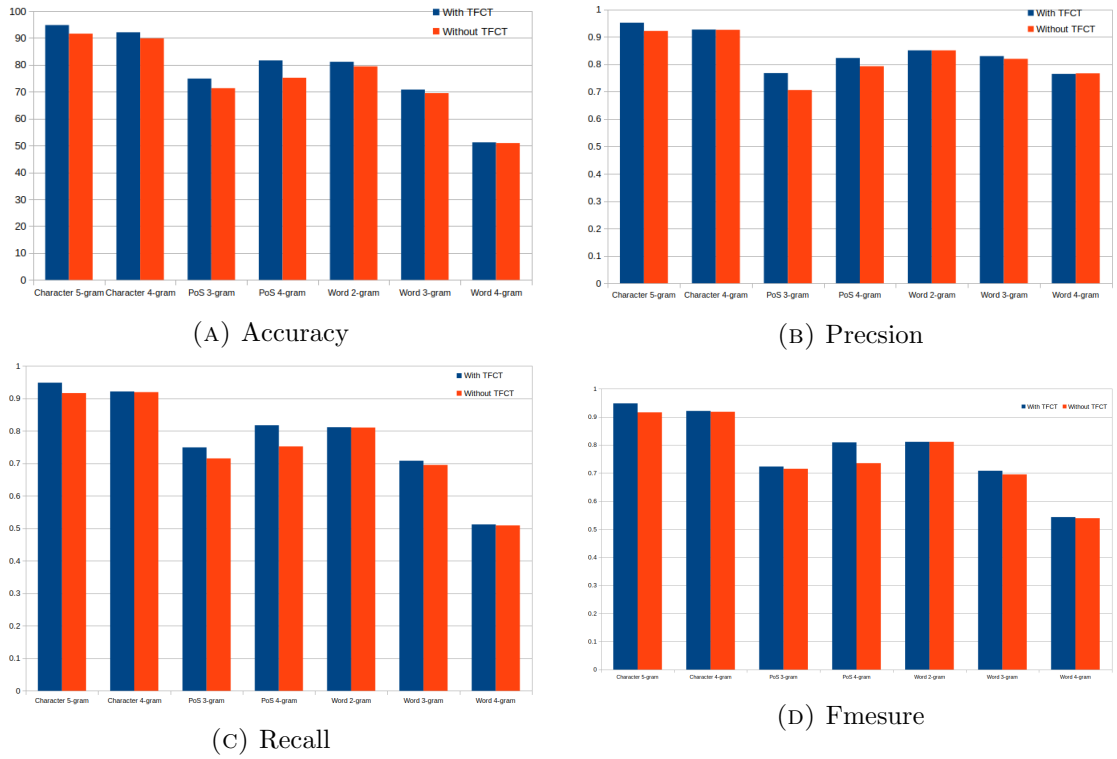
The table 4.6 shows the performance of word n-gram in terms of accuracy. Where consecutive sequences of word appearance are considered as features in the author identification system.

Table 4.6: Accuracy for word 2,3 and 4 gram

Features	Without TFCT	With TFCT
Word 2-gram	79.45	81.14
Word 3-gram	69.50	70.80
Word 4-gram	50.90	51.16

Again the accuracy with and without feature transformation method compared in this table. It shows improvement in accuracy for all word 2-gram, 3-gram, and 4-gram, but there is higher difference value observed in word 2-gram. For words 2-gram, 3-gram

and 4-gram, the accuracies are 81.14%, 70.80%, and 51.16% respectively. From the result, we can say that consecutive two words are contributing more to the stylometry of the author as compared to words 3 and 4 gram. The comparative result of different perfromance parameters are shown in figure 4.4.



(A) Accuracy

(B) Precsion

(C) Recall

(D) Fmesure

FIGURE 4.4: Comparision of Accuracy, precsion, recall and fmeasure for author identification system with and without using TFCT function

Among all these three types of features, two types of feature character and word grams are concerned with the content. It is capable to handle structural and contextual stylometry, whereas PoS capable of touching with grammatical behavior in terms of tagging. Table 4.6 shows accuracies of word 2-gram, 3-gram and 4-grams. Table 4.7 indicates the performance parameters for the proposed system.

TABLE 4.7: Accuracy, Precision, recall, F-measure of author identification with TFCT function

Feature Type	Accuracy	Precision	Recall	F-measure
Character 5-gram	**94.83**	**0.952**	**0.948**	**0.948**
Character 4-gram	92.11	0.927	0.921	0.921
PoS 3-gram	74.94	0.768	0.749	0.723
PoS 4-gram	81.65	0.823	0.817	0.809
Word 2-gram	81.14	0.851	0.811	0.811
Word 3-gram	70.8	0.83	0.708	0.708
Word 4-gram	51.16	0.765	0.512	0.543

As in character n-gram, the accuracy shown is more because characters in the English language are limited; hence, their occurrences are more obvious and so contributing more. The comparative performance of various features with and without the use of the TFCT function is presented in figure 4.4a. A comparative measure of all types of features with feature transformation function shown in figure 4.5 where character 4-gram, character 5-gram and PoS 4-gram gives higher values which indicates the correctness of authorship attribution.

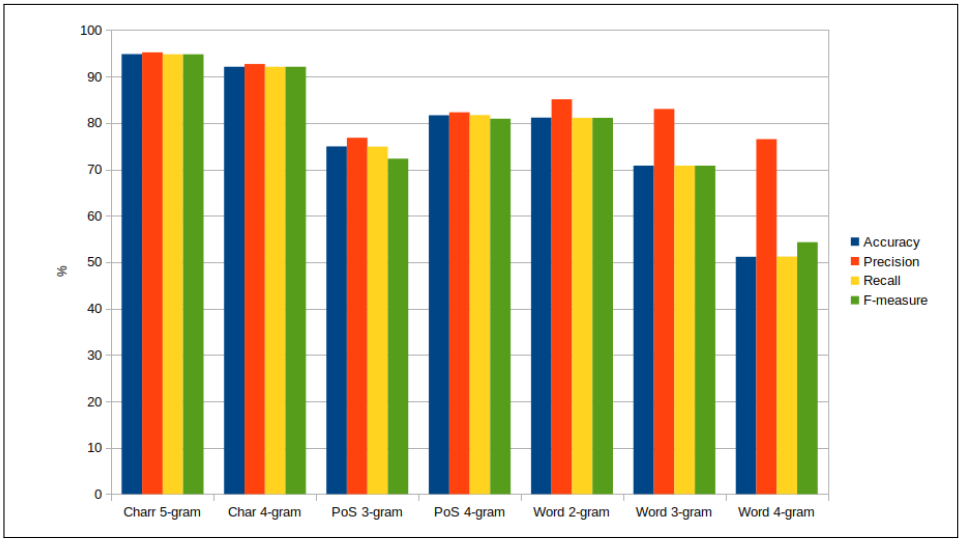


FIGURE 4.5: Accuracy, precision, recall, f-measure for character, word and PoS

When we cumulatively take a look at all types of features, character four and five gram shows the highest accuracy in both cases. With the proposed feature transformation method PoS 3, 4-grams and word 2, 3 grams shows accuracy from 65% to 80% range. And

in all these cases, we have improved accuracy with the proposed methodology. When we compare the differences in improvement, the PoS n-grams shows more improvement in the accuracy as compared to the other two feature types. Hence, we can say that instead of content and structural stylometry, the grammatical rule contributes more in style change over time. As age grows, maturity increases, the vocabulary increases and gramatically rules become more correct and hence shows improvement. As consecutive word sequences grow, the accuracy decreases a large. This happens because, in the content more lengthy consecutive reappearance of the word contributes less to recognize and distinguish the writing style of the author.

TABLE 4.8: Author identification with TFCT function for different classification methods

Features	**SVM**	**Naive Bayes**	**Random Forest**
Character 4-gram	**92.11**	86.43	79.32
Character 5-gram	**94.83**	86.43	79.45
PoS 3-gram	74.94	61.36	**79.45**
PoS 4-gram	**81.65**	63.3	63.82
Word 2-gram	81.14	**84.62**	73.12
Word 3-gram	70.8	**82.17**	63.95
Word 4-gram	51.16	**59.94**	48.44

Table 4.8 shows the accuracy when different types of features are used with various classification techniques. Support vector machine classifier supports higher feature dimensional; naive Bayes is a type of probability-based classifier, and the last random forest is decision tree based. According to accuracy support vector machine classifier does best for character 4, and 5 gram as it has higher dimensional than others, which accuracies are above 90% and rest two classifiers having less accuracy, which is less than 81%. In PoS, 3,4-gram SVM does best for PoS 4 gram, which accuracy 81.65% and random forest did good for PoS 3 gram. Word 3,4 and 5 grams types of features give their best accuracy in naive Bayes classifier and which are 84.62%, 82.17%, and 59.94%, respectively. For the same, when the support vector machine classifier used word 2 gram has the highest accuracy than rest two, which is 81.14%.

4.3.1 Author-wise Results

In this section, we discuss the performance of the proposed methodology where TFCT function used to transform feature to latest time period. We evaluated the performance of the system on thirteen different author datasets. Every writer has his own writing traits, and different style changes over time due to their knowledge, education level, personality, etc. We have various performance parameters discussed in chapter 4 derived from the confusion matrix.

TABLE 4.9: Author-wise performance for character 4-gram type of feature

Author	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
Coomi Kapoor	0.957	0.003	0.978	0.957	0.968
Benjamin Franklin	0.984	0.01	0.897	0.984	0.938
Pratap B. Mehta	0.988	0.001	0.988	0.988	0.988
Michael Cooper	0.838	0	1	0.838	0.912
Abigil Smith	0.791	0.003	0.944	0.791	0.861
Bill Killer	0.765	0.001	0.929	0.765	0.839
Swami Vivekanand	0.682	0	1	0.682	0.811
Mahatma Gandhi	0.84	0	1	0.84	0.913
Tanveel Singh	0.989	0.001	0.989	0.989	0.989
David Brook	0.911	0.029	0.783	0.911	0.842
Thomas Friedman	0.896	0.013	0.87	0.896	0.882
John Adams	0.952	0.018	0.819	0.952	0.881
Ron Rolheiser	0.926	0.007	0.946	0.926	0.935

Table 4.9 indicates the performance measures of different authors for character 4-grams type of feature. From the table, Michael Cooper, Swami Vivekanand, and Mahatma Gandhi have the highest precision, which is 1. This indicates the content written by these three authors are classified false. They have a unique style and very different from the rest. We look correctness at (TPR), which indicates that Benjamin Franklin, Pratap Bhanu Mehta, and Tanvil Singh having the highest accuracy 98%. Pratap Bhanu Mehata and Tanveel Singh have the highest F1-score.

TABLE 4.10: Author-wise performance for character 5-gram type of feature

Author	TP Rate	FP Rate	Precision	Recall	F-Measure
Coomi Kapoor	0.989	0.003	0.979	0.989	0.984
Benjamin Franklin	0.887	0.004	0.948	0.887	0.917
Pratap B. Mehta	1	0	1	1	1
Michael Cooper	1	0	1	1	1
Abigil Smith	0.86	0.003	0.949	0.86	0.902
Bill Killer	0.706	0.003	0.857	0.706	0.774
Swami Vivekanand	0.773	0	1	0.773	0.872
Mahatma Gandhi	0.88	0	1	0.88	0.936
Tanveel Singh	1	0.001	0.989	1	0.994
David Brook	0.962	0.016	0.874	0.962	0.916
Thomas Friedman	0.97	0.006	0.942	0.97	0.956
John Adams	0.984	0.02	0.813	0.984	0.891
Ron Rolheiser	0.926	0.001	0.989	0.926	0.956

Accuracy of all the author is improved in the character 5-grams. From table 4.10, we observed that Pratap Bhanu Mehta and Michael Cooper's all writing content are correctly classified, and no others are grouped with these authors. Similarly, Tanveel Singh's content is also not misclassified but other classified with this author; hence, we do not have precision "1" for the author. Bill Killer has the least accuracy 70.6% but higher precision. Though David Brook has the highest FP rate, it has a high accuracy of about 96.2%. Bill Keller and Swami Vivekanand have the least accuracy, and it shows that their writing style follows fewer traits among all other listed writers.

PoS is not directly used the content as feature type, and it is tag based, which is predefined for each word type. Table 4.11 shows the evaluation performance of the proposed system when PoS 3-gram as features applied. PoS 3-gram give 74.93% accuracy. For individual author accuracy described in table 4.11. In tabulated result, Coomi Kapoor, Pratap B. Mehta, Tanveel Singh, and John Adams gives more than 90% correct prediction. And Bill Keller, Swami Vivekanand, Mahatma Gandhi, Thomas Friedman, Abigail Smith shows less accuracy and Swami Vivekanand least. This indicates that writing style in terms of three consecutive tag sequencing is not a firm attribute for getting a unique writing style.

All test samples belong to John Adams are correctly classified using PoS 3-gram. Bill Keller and Swami Vivekanand indicate the highest precision means other authors samples are classified with these writers.

TABLE 4.11: Author-wise performance for PoS 3-gram type of feature

Author	TP Rate	FP Rate	Precision	Recall	F-Measure
Coomi Kapoor	0.957	0.007	0.947	0.957	0.952
Benjamin Franklin	0.629	0.025	0.684	0.629	0.655
Pratap B. Mehta	0.988	0.01	0.921	0.988	0.953
Michael Cooper	0.568	0.005	0.84	0.568	0.677
Abigil Smith	0.488	0.005	0.84	0.488	0.618
Bill Killer	0.059	0	1	0.059	0.111
Swami Vivekanand	0.045	0	1	0.045	0.087
Mahatma Gandhi	0.2	0.005	0.556	0.2	0.294
Tanveel Singh	0.978	0.023	0.845	0.978	0.906
David Brook	0.785	0.085	0.512	0.785	0.62
Thomas Friedman	0.418	0.027	0.596	0.418	0.491
John Adams	1	0.037	0.705	1	0.827
Ron Rolheiser	0.862	0.047	0.717	0.862	0.783

As compared to the PoS 3-gram feature type, the correctness of the author identification system is increased in PoS 4-gram. Specifically for Bill Keller and Swami Vivekanand, which was very less in PoS 3-gram feature type. The highest accuracy obtained for Coomi Kapoor and John Adams. Similarly, Bill Killer shows the least value of F-measure, and Tanveel Singh shows the highest F-measure. The precision value of Bill Killer was the largest. The detail evaluation report is presented in table 4.12.

Word 2-gram gives more than 81% correct prediction, and the sincere way result statistics are described in table 4.13. As a result, seven authors among thirteen give maximum precision values for which samples of other writers are not predicted as them. Pratap B. Mehta shows 100% accuracy along with precision and recalls value 1. This indicates that any of the samples of the author not classified as other authors, and in the same way, other authors' samples also not classified as Pratap B. Mehta. The Smallest accuracy is obtained for Bill Killer, about 58%.

Table 4.12: Author-wise performance for PoS 4-gram type of feature

Author	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
Coomi Kapoor	0.968	0.01	0.929	0.968	0.948
Benjamin Franklin	0.79	0.038	645	0.79	0.71
Pratap B. Mehta	0.952	0.01	0.919	0.952	0.935
Michael Cooper	0.946	0.003	0.946	0.946	0.946
Abigil Smith	0.814	0.008	0.854	0.814	0.833
Bill Killer	0.235	0	1	0.235	0.381
Swami Vivekanand	0.318	0.001	0.875	0.318	0.467
Mahatma Gandhi	0.64	0.003	0.889	0.64	0.744
Tanveel Singh	1	0.009	0.937	1	0.967
David Brook	0.658	0.045	0.627	0.658	0.642
Thomas Friedman	0.746	0.028	0.714	0.746	0.73
John Adams	0.968	0.008	0.909	0.968	0.937
Ron Rolheiser	0.691	0.04	0.707	0.691	0.699

Table 4.13: Author-wise performance for word 2-gram type of feature

Author	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
Coomi Kapoor	0.968	0.003	0.978	0.968	0.973
Benjamin Franklin	0.839	0.007	0.912	0.839	0.874
Pratap B. Mehta	1	0	1	1	1
Michael Cooper	0.919	0	1	0.919	0.958
Abigil Smith	0.837	0	1	0.837	0.911
Bill Killer	0.588	0	1	0.588	0.741
Swami Vivekanand	0.591	0	1	0.591	0.743
Mahatma Gandhi	0.76	0	1	0.76	0.864
Tanveel Singh	0.989	0	1	0.989	0.994
David Brook	0.962	0.063	0.633	0.962	0.764
Thomas Friedman	0.836	0.011	0.875	0.836	0.855
John Adams	0.806	0.025	0.735	0.806	0.769
Ron Rolheiser	0.904	0.006	0.955	0.904	0.929

When word 3-gram used as a feature, then we get author wise statistics shown in table 4.14. The overall accuracy for the proposed system using these types of features gives 70.8% correct prediction. But according to the author, prediction rate varies, for Pratap B. Mehta shows 95.2% highest correct prediction, but none of the samples correctly classified for Bill Killer. Among all six writers have the highest precision value shows the uniqueness of their writings. Minimum f-measure values indicated by Swami Vivekanand for word 3-gram. Cumulatively three authors show more than 88% accuracy.

TABLE 4.14: Author-wise performance for word 3-gram type of feature

Author	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
Coomi Kapoor	0.84	0.006	0.952	0.84	0.893
Benjamin Franklin	710	0.111	0.846	0.71	0.772
Pratap B. Mehta	0.952	0	1	0.952	0.975
Michael Cooper	0.649	0	1	0.649	0.787
Abigil Smith	0.419	0	1	0.419	0.59
Swami Vivekanand	0.227	0	1	0.227	0.37
Mahatma Gandhi	0.32	0	1	0.32	0.485
Tanveel Singh	0.888	0	1	0.888	0.94
David Brook	0.886	0.19	0.347	0.886	0.498
Thomas Friedman	0.552	0.031	0.627	0.552	0.587
John Adams	0.548	0.031	0.607	0.548	0.576
Ron Rolheiser	0.755	0.056	0.651	0.755	0.7

Table 4.15 represents individual author performance when consecutive four-word uses as a feature type. Amongst all the feature type word 4-gram has the least accuracy, which indicates the more consecutive word can not be used to discriminate the writing style of the author as this consecutiveness increases the incorrectness in prediction rises. The tabulated result indicates the least accuracy for Swami Vivekanand, which is 9.1%, and Ron Rolheiser shows the highest 98.9% accuracy. Six author shows the highest precision.

TABLE 4.15: Author-wise performance for word 4-gram type of feature

Author	TP Rate	FP Rate	Precision	Recall	F-Measure
Coomi Kapoor	0.596	0.021	0.8	0.596	0.683
Benjamin Franklin	0.452	0.008	0.824	0.452	0.583
Pratap B. Mehta	0.602	0	1	0.602	0.752
Michael Cooper	0.649	0	1	0.649	0.787
Abigil Smith	0.302	0	1	0.302	0.464
Bill Killer	0.176	0	1	0.176	0.3
Swami Vivekanand	0.091	0	1	0.091	0.167
Mahatma Gandhi	0.12	0.003	0.6	0.12	0.2
Tanveel Singh	0.618	0	1	0.618	0.764
David Brook	0.278	0.027	0.537	0.278	0.367
Thomas Friedman	0.358	0.007	0.828	0.358	0.5
John Adams	0.371	0.017	0.657	0.371	0.474
Ron Rolheiser	0.989	0.471	0.225	0.989	0.367

4.3.2 Author-wise Performance for all Features

When author-wise performances are evaluated with precision,recall and f-measure parameters, Character 4-gram and character 5-gram are capable to distinguish the stylometry in all the cases hence an impressive outcome shown by them. The performance of word-gram is worst at all the cases. For best performing system all the value of precision recall and f-measure should reach to score 1.0. Bill Killer, Swami Vivekanand and Mahatma Gandhi shows least result for all the types of feature compare to other feature type this indicates that the writing style of them are unique and non repetitive. There are authors showing the good recognition of stylometry to identify their writing in all the feature types. The distinguished result for all authors to all seven types of feature are shown in figure 4.6 and figure 4.7. None of samples are correctly classified for Bill Killer of word 3-gram type of features. PoS 4-gram shows relatively good peformance as compared to PoS 3-gram. Among all types of feature the performance are improved from word n-gram to PoS n-gram to character n-gram.

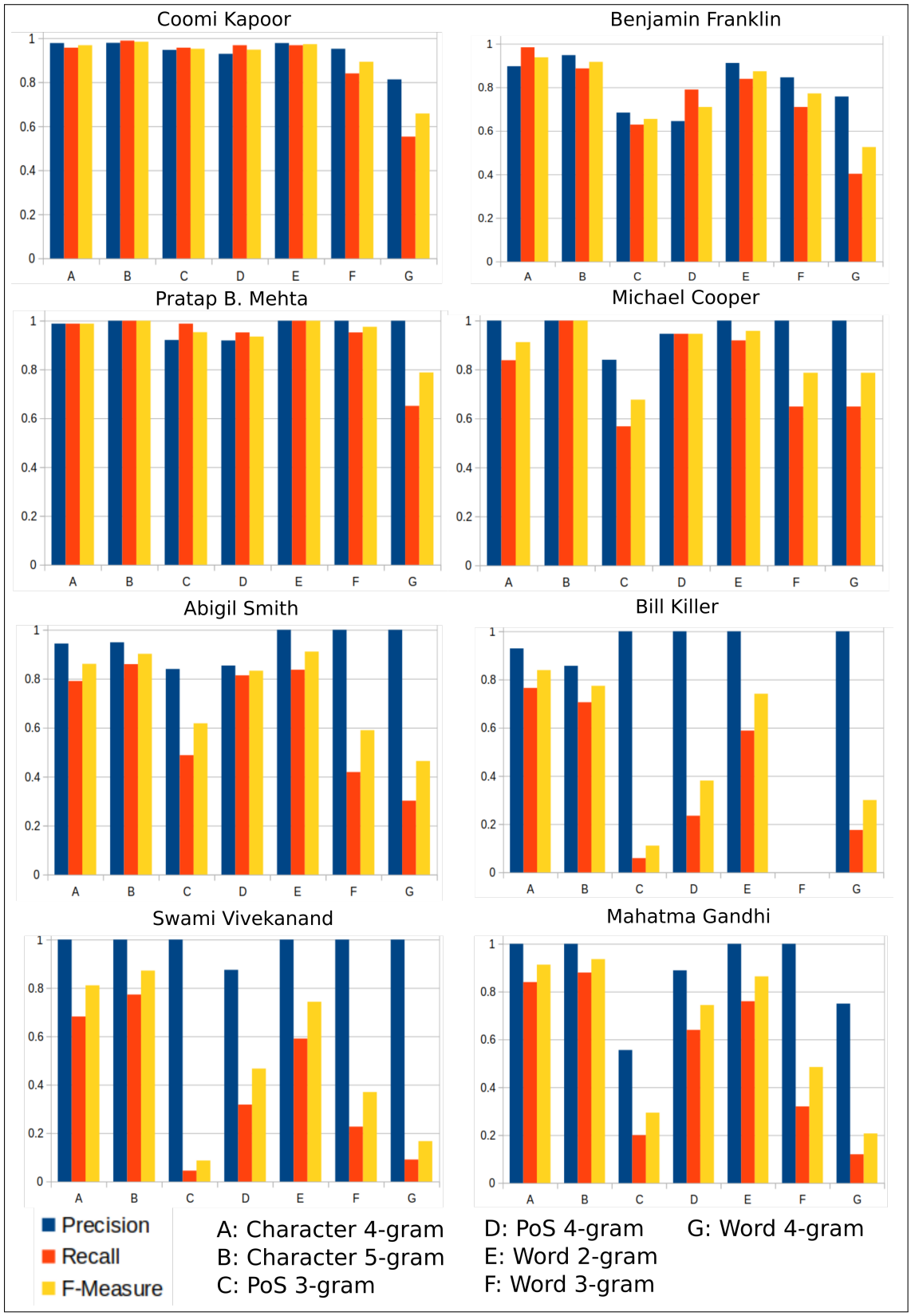


FIGURE 4.6: (a) Author wise precision recall and f-measure for all types of features

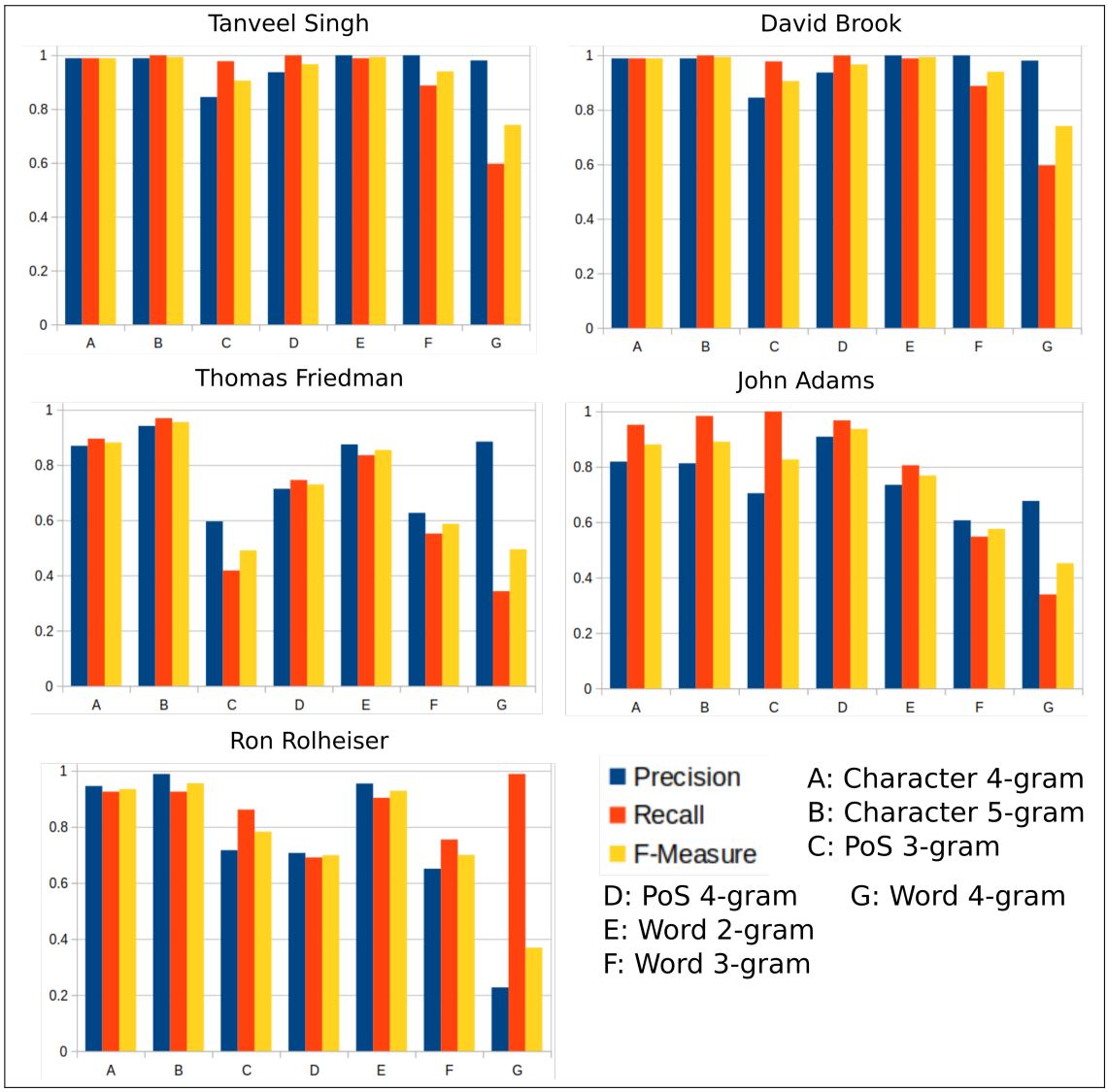


FIGURE 4.7: (b) Author wise precision recall and f-measure for all types of features

4.3.3 Comparative Results

In this section, we discuss the result obtained by the proposed methodology and the existing methods. Source code authorship profiling (SCAP) [45] and naive-based similarity method [13, 32] methods are chosen for comparison. When the proposed technique of TFCT used to build different types of features and compared the results of methodology with mentioned two methods, first is SCAP and second naive-based similarity method. Figure 4.8 shows the comparative result. When proposed methodology compared with SCAP, then all have better performance except word 4-gram type of method. The result of SCAP methodology shows 57.44%, and word 4-gram has 51.16% rest type of features shows greater

prediction accuracy than SCAP. The detailed statistics described in table 4.16. From the figure 4.8, it is clear that the character four and 5-gram gives better performance than naive-based similarity method. The prediction accuracy of the proposed method when character 4, 5-grams are 92.11% and 94.83%, respectively, but the naive based similarity method has 87.23% prediction accuracy. Hence, we proved that the proposed method gives more accuracy. But for all other methods (part-of-speech and word n-gram), the accuracies are less than the naive based similarity method, but the type of feature used in naive based similarity method is different. Among all the methods PoS 3-gram, word 3-gram word 4-gram, and SCAP shows prediction accuracy less than 80%.

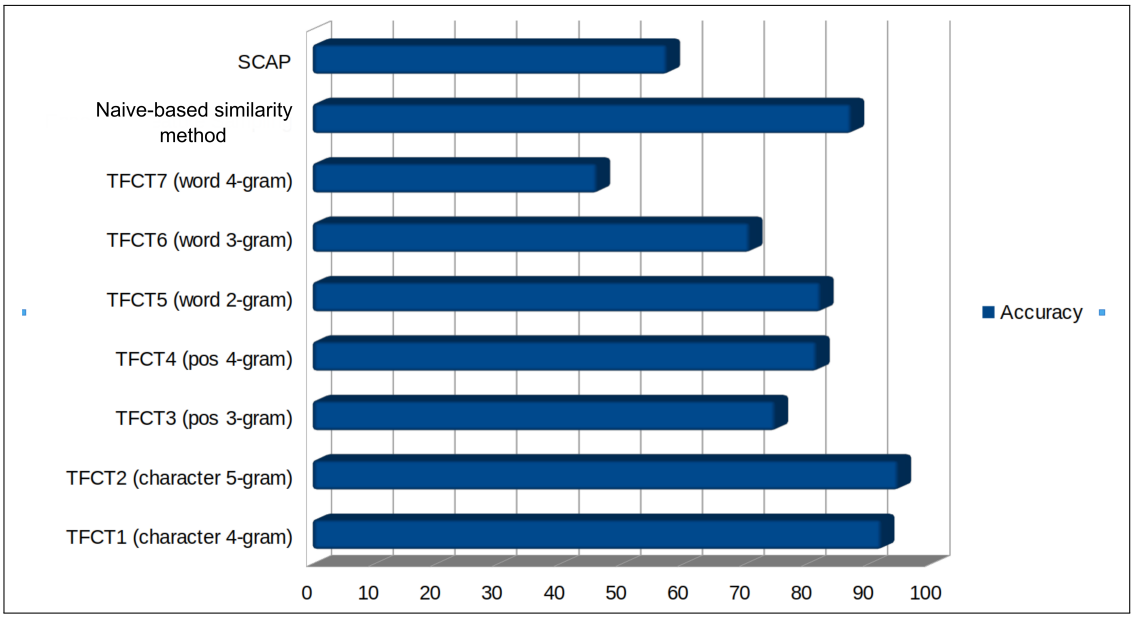


FIGURE 4.8: Comparative results

When the FP rate compared, for naive based similarity method, it is 1.5% and SCAP 3.78. And maximum FP rate is shown by word 4 gram, which also has the least prediction accuracy. In the case of precision, SCAP has the least 57.44%, and the character 5-gram feature has the highest 95.2%. TFCT with word 4-gram has the least recall 51.16%, while character four and five gram have 92.1% and 94.8 recall.

Table 4.16: Comparative results statistics

Method	Accuracy	FP Rate	Precision	Recall	F-Measure
TFCT1 (character 4-gram)	92.11	0.8	92.7	92.1	92.1
TFCT2 (character 5-gram)	**94.83**	**0.5**	**95.2**	**94.8**	**94.8**
TFCT3 (PoS 3-gram)	74.94	2.7	76.8	74.9	72.3
TFCT4 (PoS 4-gram)	81.65	2	82.3	81.7	80.9
TFCT5 (word 2-gram)	81.14	1.9	85.1	81.10	81.10
TFCT6 (word 3-gram)	70.8	3.33	77.15	70.8	62.86
TFCT7 (word 4-gram)	51.16	6.5	76.5	51.2	48.3
Naive-based similarity method [13]	87.23	1.3	87.23	87.23	87.23
SCAP [45]	57.44	3.78	57.44	57.44	57.44

4.3.4 Effect of Feature Size

The size of the feature has an impact on the accuracy of the identification of the correct author. In the proposed algorithm to select the feature count for processing, all features are ranked and then top fixed number of them selected for further processing. To check the performance of the system, the impact of feature count observed on word 2-gram, PoS 4-gram, and character 4, 5 gram, which have given higher accuracies. Figure 4.9 shows the accuracy plot for different feature count. In the experimentation in between 500 to 6000 feature count was observed.

Character 4-gram types of features give good accuracy in the proposed system until the count becomes up to 4000. And then goes on a slight decreasing and for the character, 5-gram have a simultaneous increase in accuracy up to count 6000. For all types of the feature when the count is 500 then it has less accuracy but on increasing count causes rapid improvement in prediction. Cumulatively, when the count reaches 5000, the result becomes constant and then shows a slight reduction in prediction accuracy.

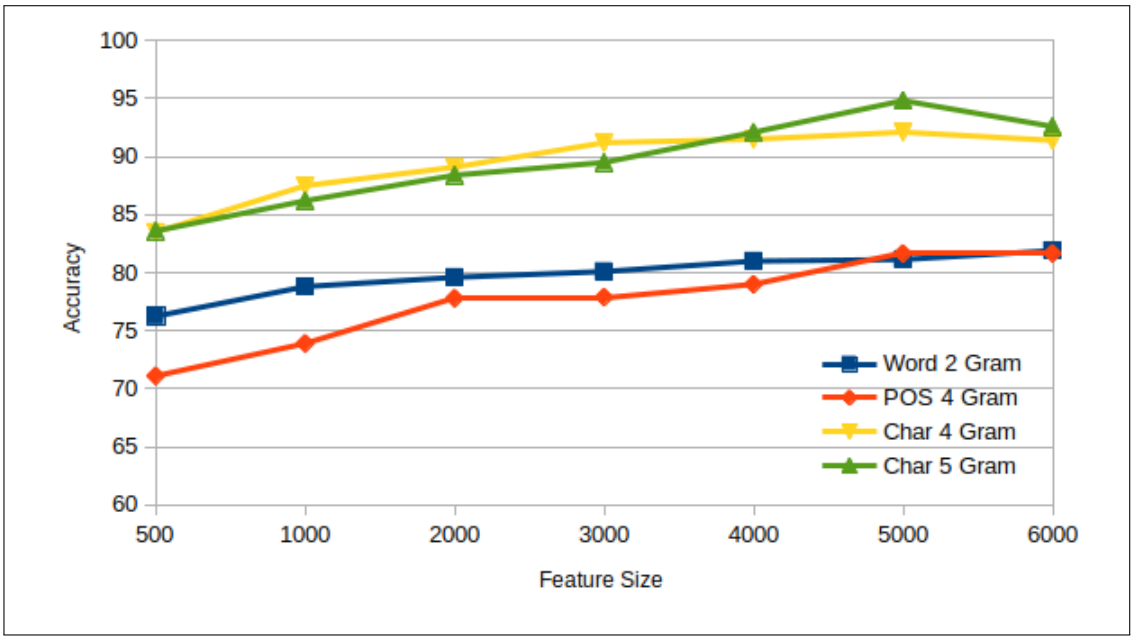


FIGURE 4.9: Impact of feature size on word 2-gram, PoS 4-gram, character 4 and 5 grams

4.4 Result and Evaluation of Author Identification with Variable Length Word Gram

In variable-length word n-gram, the value of n is not constant; it depends on the current word length. The value of n varies from 1 to 3 in the experimentation. The same corpus is used as in author identification with the feature transformation method. Thirteen different writers whose writing collected from on-line textual content from news columns, letters, and articles. Statistics of dataset described in table 4.1, 4.2.

Initially, the content by transforming the text into one case lowercase and need to remove non-ASCII characters. Feature extraction is done using python and analysis of attribution with the WEKA tool [101]. All extracted features are stored in a bag-of-words format and act as a feature vector for every sample. Initially, four-fold cross validation method is used to validate the performance of the classifier. In experiments, SVM is used for classification and performance evaluated on the basis of the dataset described in section 5.1. A SVM machine learning algorithm is used to identify probable author. The use of different descrimination algorithms to identify authorship are reviewed in [69]

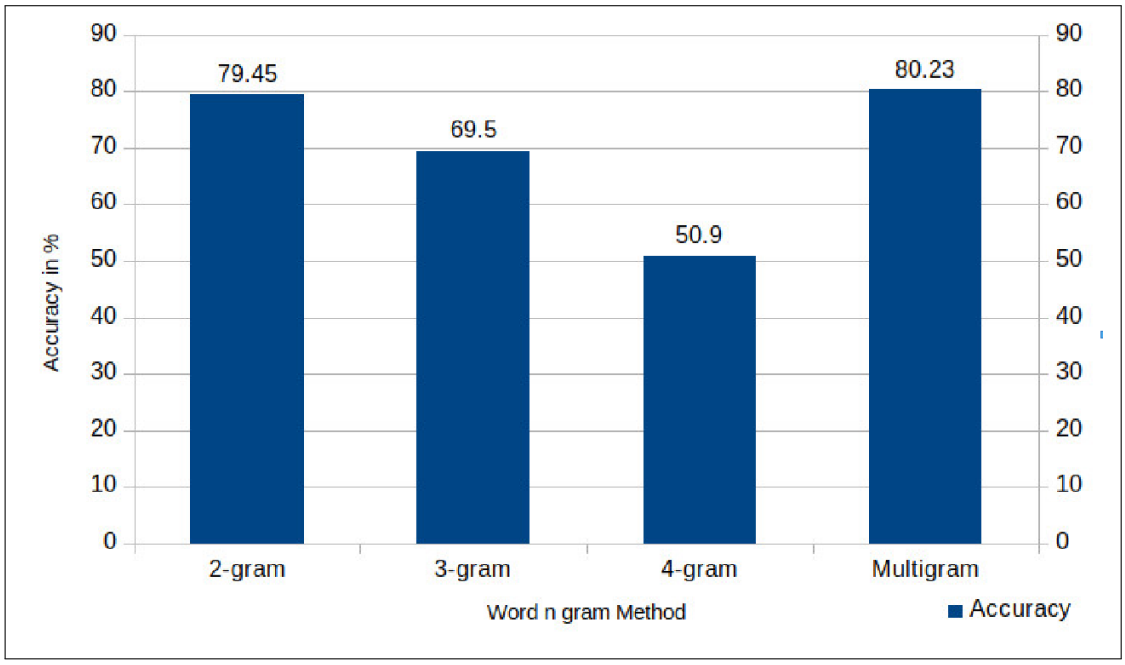


FIGURE 4.10: Accuracy of word 2, 3, 4 gram and multiword grams

Firstly we performed experiments on word 2-gram, word 3-gram, and word 4-gram. The comparative result for word n-gram shown in figure 4.10 We found that word 2 gram individually generates a good result in terms of accuracy than the remaining two. But there was a big difference in accuracy among them. Word 4 gram shows about 50.9%, and word 2-gram shows accuracy 79.45%. When we applied our proposed methodology for variable sequence word n-gram (multiword n-gram), a better achievement indicated by all existed ones. Multi-word gram indicates 80.23% accuracy. Comparative result for constant word n-gram and variable word n-gram is shown in figure 4.10.

Figure 4.11 shows the effect of feature size on accuracy for variable sequence word, let it call as a multi-word gram. To find the effect of the feature set, we consider variable training and testing sets of documents for cross-validation of the result. We choose the best feature from all accumulated features, and its count varies from 500 to 4000. We gradually verify the accuracy result by the SVM classifier and plot shown in figure 4.11. Initially, accuracy gradually increases as feature count till it reaches up to 2000 in the count and attains maximum accuracy 81.39% then on increasing feature size it gets slightly decreases, we can say, it is nearly constant after that.

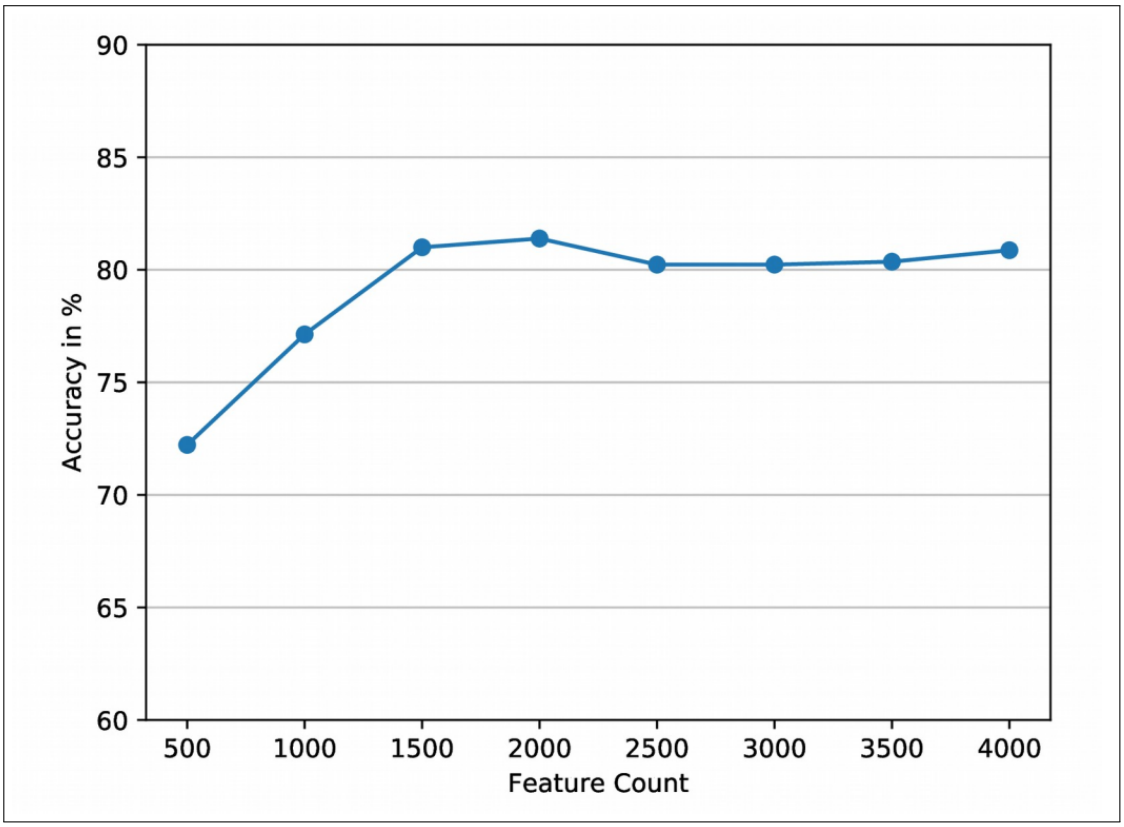


FIGURE 4.11: Effect of feature size on variable length word gram

4.4.1 Impact of Stop Words

There are common words that frequently appear in textual contents; those called as stop words. Looking at this type of content, it seems that they are useless and not participating in distinguishing the writing style of the author. In this section, we evaluate the impact of stop word in a variable length of the word gram approach. In the experiment 'variable length word gram' extracted in the same way as described in the previous section. The finite list of 154 stop words considered in this experimentation is enlisted in the frame 1.

Figure 4.12 shows the obtained comparative result of the author identification system with and without stop words used in the feature. In case of without stop word, in the preprocessing stage, all the stops get removed. Experiments are carried out on words 2, 3, 4, and multiword sequences. In all the cases, when stops get removed prediction accuracy of the system goes down. The differences in with and without stops is extreme in word 3-gram and least at multi-word gram model. The impact of stops in to capture writing

style with multi-word gram is very less.

'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', 'couldn', 'didn', 'doesn', 'hadn', 'hasn', 'haven', 'isn', 'ma', 'mightn', 'mustn', 'needn', 'shan', 'shouldn', 'wasn', 'weren', 'won', 'wouldn'

Frame 1: List of all stop words

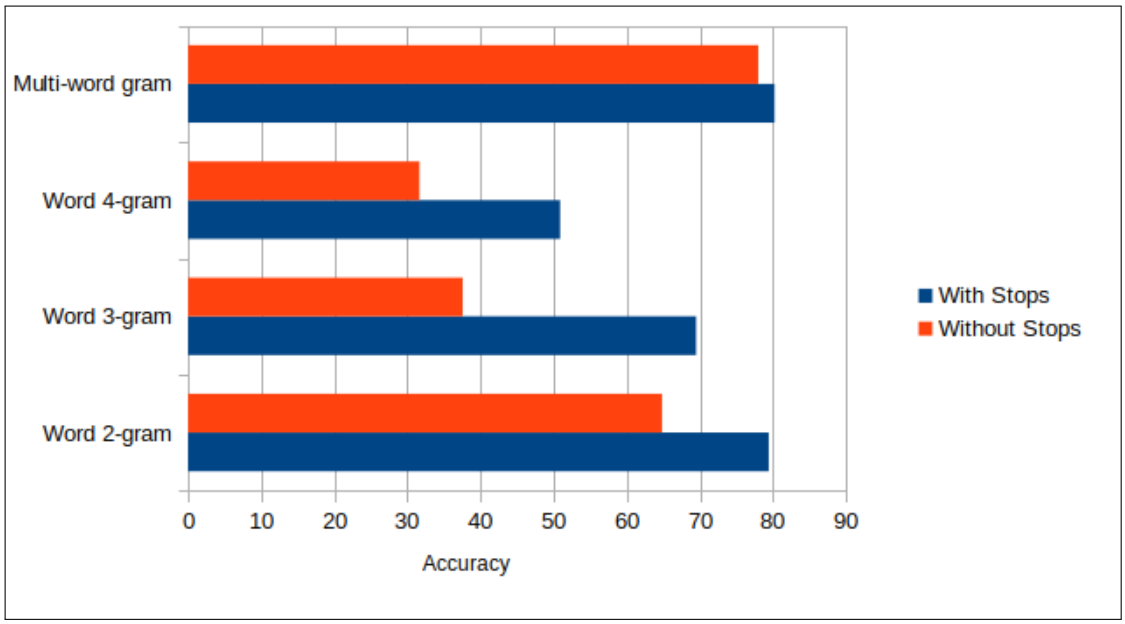


FIGURE 4.12: Effect of stop words on variable length word gram

Rather than taking all text in sequence, we build a window of fixed word size to catch up with multiword grams to create a feature vector. The extraction process acts separately for each of these windows to build vector. The performance evaluation of the system with the different option are shown in table 4.17. For any chunk size, prediction results become constant, but without chunk accuracy of correct identification increases. In the tabulated result, various performance measures are quoted. From results, it is clear that the impact of stops seen, the accuracy of without removing stops, and using the proposed approach (TFCT) gives better results than other combinations. Along with accuracy, the results of different parameters are challenging.

TABLE 4.17: Result statistics for multiword gram with chunk size 20

Method	Accuracy	Precision	Recall	F-Measure
With removing stops and without TFCT (A)	76.48	82.7	76.5	76.4
Without removing stops and with TFCT (B)	80.23	84.1	80.2	80.5
With removing stops and with TFCT (C)	76.61	82.7	76.6	76.6
Without removing stops and without TFCT (D)	78.94	82.9	78.9	79.3

Figure 4.13 shows the graphical representation of the performance of the system along with all option given table 4.17.

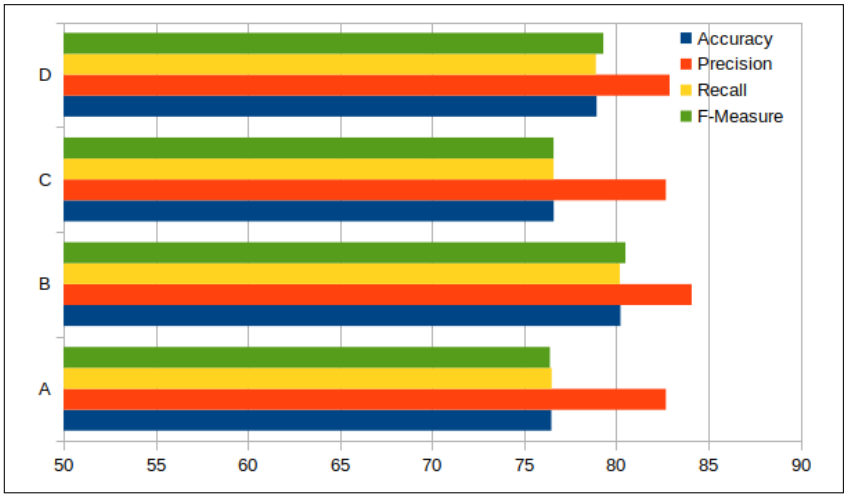


FIGURE 4.13: Comparative result of variable length word gram

The results of five experiments are presented in this chapter. In the first experiments, the writing style of the author extracted over time. Statistics obtained by comparing the writing style of the author with its latest content in terms of slope stats and standard deviation. Each feature type shows the impact of time in the writing style of the author. This on further used in the next experimentation where the feature transformation function is proposed.

Second experiment is for the prediction system, which is carried out in different phases. A feature transformation function applied to the feature vector, which is calculated from the decay factor. This causes to transform all features to the latest period of sample. It can also be termed as feature normalization. Then the performance of the system evaluated with different audit parameters accuracy, precision, recall, F-measure. In the result statistics, accuracy for every feature type compared when the transformation function used and when not used. To validate the resulting outcome of three classifiers are compared. With the audit parameters, the prediction accuracy for every author compared for all feature types. The result of the proposed system is compared with two different author identification methods. The impact of feature size also verified for the proposed system.

In the third experiment, a variable-length word gram author identification system is implemented. It works on dynamic sequence of consecutive word hence termed as multi-word gram. Accuracy is the primary outcome of experimentation. The comparative results are discussed with words 2, 3, and 4 grams with the multiword-gram model. The performance of the system is validated by varying feature sizes in the author attribution system.

The third experiment extended to forth one where same multi-word gram system used for prediction but here, the impact of stops is evaluated. A list of 154 stop words identified and used in the experimentation. The main audit parameter in this experiment is accuracy and the impact of stops. Sometimes stops also termed as function words.

In the last experiment, the implementation of variable word length has been done with slight modification in which a chunk size is defined and evaluated with audit parameters such as accuracy precision-recall and F-measure. The impact of feature transformation function and stop words evaluated with these parameters.The method which uses TFCT function and includes stop word gives high accuracy than others.

Chapter 5

Conclusions and Future Scope

The thesis is concluded with remarkable outcome from the research work in author identification system which become robust by mitigating the impact of timely changed writing style as well as the novel approach of variable sequence word gram. Then, it provides path to future work.

5.1 Conclusions

Author identification is a technique to recognize the ownership of an unknown text document. The limitation existing methodologies addressed in this thesis is the writing content of author changes over time due to the various factors such age, education, behaviour, place, mother tongue, etc. Hence, over the time these changes affects the performance of the system. This work provides a research solution to the problem by removing this impact of style change over time. A set of performance parameters such as accuracy, precision, recall and impact of feature size are used to evaluate the performance of the system. The significance of the system and result of evaluation parameters are discussed.

1. The novel research method modifies the feature values according to time aimed at distinguishing the author with the use of individual features as part of speech n-gram, word n-gram, and character n-gram. The novel method is responsible for normalizing features to the current time so that it is applied to a SVM for classification. The main difference with the generic approach is that the features considered as it is irrespective of the time period, which can be responsible for changing or evolving the writing

style of the writer. The focus is made to track such changes by observing features in different time zone and weighted them to bring at the current time. The higher weight assigned to old text features and less weight for newer text. The experiment made on the proposed method shows that there is a positive change in accuracy when classification algorithm is applied. A support vector machine algorithm used for classification which works fine on high dimensional features. The research shows that the word n-gram can participate in identification prominently so the resultant accuracy is impressive.

2. The maximum accuracy obtained by this system is upto **94.83%** and it is for character 5-gram type of feature. The lowest accuracy obtained in the system is by word 4-gram type of features. This shows that the character 4-gram and character 5-gram type of feature are more capable to discriminate writing style among authors.

3. The impact of the TFCT function is analyzed with and without applying this function in the system. It is observed that for each type of the feature, the TFCT function shows the improvement in accuracy. The highest improvement found in PoS 4-gram type of feature. This feature type quotes the enhancement in accuracy upto **6.46%**.

4. Author identification system with TFCT is compared with two methodologies that were implemented and tested on the same dataset. The first methodology is source code authorship profiling (SCAP), and the other is naive-based similarity method. The results of the experiment are compared and it is observed that novel approach gives better accuracy than these two methods.

5. With all three types of features, each authors performance is evaluated by performance parameters like precision, recall and accuracy.

6. However, in the system when the dimension of the feature vector increased upon a certain limit, the negative impact seen on accuracy for all three types of features.

A novel approach with a variable sequence of consecutive words to identify the author of an unknown text sample is also presented in this thesis. The feature used in the approach is word n-gram, where the value of n is not constant, and it is dynamically changing. This approach is used on a dataset of the different authors whose text sample written in the variable period, and it shows an effective result. In this methodology, the value of n is selected in word n-gram by introducing a set of rules. The rules are defined

based on variable-length word n-gram, where the value of n varied from 1 to 3. The value of n is dynamically changing and depends on the length of the currently appearing word in the text. A notable improvement in the performance has been seen in this approach. In the experiments, the performance of a constant value of n in word n-gram evaluated with the implemented method and the algorithm in this approach beats the existing consistent word n-gram. In this research work, the effect of feature count on accuracy is evaluated, which shows that the accuracy remains constant on specific feature size. Function words are also crucial to style markers in the writing of the author. The impact is evaluated and found that it affects the strength of correctness in author identification. The algorithm of this approach evaluated with stopwords and with TFCT function to transform feature to the latest time and produces the highest accuracy **80.23%**.

The empirical performance of each these approaches evaluated in terms of accuracy, precision, recall and f-measure, with varying feature size the accuracies are measured for all three types of features. Experiments were conducted and evaluation of each individual author is confirmed and it is found performance for each author is different based on repetitiveness of writing style. The performance and evaluations are provided in this research.

5.2 Future Scope

In this section, we conclude with research by introducing the probable way to extend this work. The main contribution to this research is to remove impact of time in the writing style of the author. In this direction, very limited amount of work exists, including ours, hence it is possible to extend this work in the future. Work can be extended as follows:

1. Huge amount of work of author identification exists which uses machine learning algorithms so various combination of machine learning algorithm possible to use for higher accuracy.

2. Extension to work is possible by enhancing feature set with feature engineering techniques.

3. The problem is yet open to extend this work by combining different techniques at feature selection and discrimination level.

4. This work uses a machine learning approach for discrimination. Further, it can be possible to enhance the approach by using deep learning method with a comprehensive use of features to improve the result.

References

[1] Stamatatos, Efstathios, "A survey of modern authorship attribution methods", Journal of the American Society for information Science and Technology 60.3 (2009): 538-556.

[2] Pavelec D, Oliveira LS, Justino E, Nobre Neto FD, Batista L V, "Author identification using compression models", In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. ; 2009. doi:10.1109/ICDAR.2009.208

[3] Nirkhi S, .R.V.Dharaskar, "Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis", International Journal of Advanced Computer Science and Applications(IJACSA) 2013;4(5):32-35, doi:10.14569/ijacsa.2013.040505

[4] Iqbal, F., Binsalleeh, H., Fung, B. C., & Debbabi, M, "A unified data mining solution for authorship analysis in anonymous textual communications", Information Sciences 231 (2013): 98-112.

[5] Abbasi, Ahmed, and Hsinchun Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace", ACM Transactions on Information Systems (TOIS) 26.2 (2008): 7.

[6] Argamon, Shlomo, Marin Šarić, and Sterling S. Stein, "Style mining of electronic messages for multiple authorship discrimination: first results", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[7] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J., "Automatically profiling the author of an anonymous text", Commun. ACM 52.2 (2009): 119-123.

[8] Cheng, Na, Rajarathnam Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text", Digital Investigation 8.1 (2011): 78-88.

[9] Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang, "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, no. 3 (2006): 378-393.

[10] Fatima, Mehwish, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab, "Multilingual author profiling on Facebook", Information Processing & Management 53, no. 4 (2017): 886-904.

[11] Estival D, Gaustad T, Pham SB, Radford W, Hutchinson B., "Author profiling for English emails", In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics 2007 Sep 19 (pp. 263-272).

[12] Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, "Personality, gender, and age in the language of social media: The open-vocabulary approach", PloS one 8, no. 9 (2013): e73791.

[13] Koppel, Moshe, Jonathan Schler, and Shlomo Argamon, "Authorship attribution in the wild", Language Resources and Evaluation 45.1 (2011): 83-94.

[14] Rocha, Anderson, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos, "Authorship attribution for social media forensics", IEEE Transactions on Information Forensics and Security 12, no. 1 (2016): 5-33.

[15] Shang, Ronghua, Yang Meng, Chiyang Liu, Licheng Jiao, Amir M. Ghalamzan Esfahani, and Rustam Stolkin, "Unsupervised feature selection based on kernel fisher

discriminant analysis and regression learning", Machine Learning 108, no. 4 (2019): 659-686.

[16] Villar-Rodriguez, Esther, Javier Del Ser, Miren Nekane Bilbao, and Sancho Salcedo-Sanz, "A feature selection method for author identification in interactive communications based on supervised learning and language typicality", Engineering Applications of Artificial Intelligence 56 (2016): 175-184.

[17] Sheikhpour, Razieh, Mehdi Agha Sarram, Sajjad Gharaghani, and Mohammad Ali Zare Chahooki, "A survey on semi-supervised feature selection method", Pattern Recognition 64 (2017): 141-158.

[18] Singh, Danasingh Asir Antony Gnana, Subramanian Appavu Alias Balamurugan, and Epiphany Jebamalar Leavline, "An unsupervised feature selection algorithm with feature ranking for maximizing performance of the classifiers", International Journal of Automation and Computing 12.5 (2015): 511-517.

[19] Madigan, David, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye, "Author identification on the large scale", In Proc. of the Meeting of the Classification Society of North America, vol. 13. 2005.

[20] Brocardo, Marcelo Luiz, Issa Traore, and Isaac Woungang, "Authorship verification of e-mail and tweet messages applied for continuous authentication", Journal of Computer and System Sciences 81.8 (2015): 1429-1440.

[21] Koppel, Moshe, and Shachar Seidman, "Detecting pseudepigraphic texts using novel similarity measures", Digital Scholarship in the Humanities 33.1 (2017): 72-81.

[22] Kestemont, Mike, "Function words in authorship attribution. From black magic to theory?", Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL). 2014.

[23] Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis, "Automatic text categorization in terms of genre and author", Computational linguistics 26.4 (2000): 471-495.

[24] Stamatatos, Efstathios, "Ensemble-based author identification using character n-grams", Proceedings of the 3rd International Workshop on Text-based Information Retrieval. Vol. 36. 2006.

[25] Niesler, Thomas R., and Philip C. Woodland, "A variable-length category-based n-gram language model", 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Vol. 1. IEEE, 1996.

[26] Ding, Steven HH, Benjamin CM Fung, Farkhund Iqbal, and William K. Cheung, "Learning stylometric representations for authorship analysis", IEEE transactions on cybernetics 49, no. 1 (2017): 107-121.

[27] Uzuner, Özlem, Boris Katz, and Thade Nahnsen, "Using syntactic information to identify plagiarism", Proceedings of the second workshop on Building Educational Applications Using NLP. Association for Computational Linguistics, 2005.

[28] Gamon, Michael, "Linguistic correlates of style: authorship classification with deep linguistic analysis features", Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[29] Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan, "Stylistic text classification using functional lexical features", Journal of the American Society for Information Science and Technology 58, no. 6 (2007): 802-822.

[30] Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard, "urveying stylometry techniques and applications", ACM Computing Surveys (CSUR) 50, no. 6 (2018): 86.

[31] Kajzer-Wietrzny, Marta, "Idiosyncratic features of interpreting style", New Voices in Translation Studies 9.1 (2013): 38-52.

[32] Azarbonyad, Hosein, Mostafa Dehghani, Maarten Marx, and Jaap Kamps, "Time-aware authorship attribution for short text streams", In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 727-730. ACM, 2015.

[33] Anwar, Waheed, Imran Sarwar Bajwa, and Shabana Ramzan, "Design and Implementation of a Machine Learning-Based Authorship Identification Model", Scientific Programming 2019 (2019).

[34] Zamani, Hamed, Hossein Nasr Esfahani, Pariya Babaie, Samira Abnar, Mostafa Dehghani, and Azadeh Shakery, "Authorship identification using dynamic selection of features from probabilistic feature set", In International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 128-140. Springer, Cham, 2014.

[35] Teahan, William J., and David J. Harper, "Using compression-based language models for text categorization", Language modeling for information retrieval. Springer, Dordrecht, 2003. 141-165.

[36] Santosh, K., Romil Bansal, Mihir Shekhar, and Vasudeva Varma, "Author profiling: Predicting age and gender from blogs", Notebook for PAN at CLEF 2013 (2013).

[37] Fürnkranz, Johannes, "A study using n-gram features for text categorization", Austrian Research Institute for Artifical Intelligence 3.1998 (1998): 1-10.

[38] Mansur, Munirul, "Analysis of n-gram based text categorization for bangla in a newspaper corpus", Diss. BRAC University, 2006.

[39] Tan, Chade-Meng, Yuan-Fang Wang, and Chan-Do Lee, "The use of bigrams to enhance text categorization", Information processing & management 38.4 (2002): 529-546.

[40] Pavelec, Daniel, Edson Justino, Leonardo V. Batista, and Luiz S. Oliveira, "Author identification using writer-dependent and writer-independent strategies", In Proceedings of the 2008 ACM symposium on Applied computing, pp. 414-418. ACM, 2008.

[41] Inches, Giacomo, and Fabio Crestani, "Online conversation mining for author characterization and topic identification", Proceedings of the 4th workshop on Workshop for Ph. D. students in information & knowledge management. ACM, 2011.

[42] Rosen-Zvi, Michal, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers, "Learning author-topic models from text corpora", ACM Transactions on Information Systems (TOIS) 28, no. 1 (2010): 4.

[43] Stamatatos, Efstathios, Nikos Fakotakis, and Georgios Kokkinakis, "Computer-based authorship attribution without lexical measures", Computers and the Humanities 35.2 (2001): 193-214.

[44] Burns, K., "Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support", Information Sciences, (2006) 176(11), 1570-1589.

[45] Frantzeskou, G., MacDonell, S., Stamatatos, E., & Gritzalis, S. (2008), "Examining the significance of high-level programming features in source code author classification," Journal of Systems and Software, 81(3), 447-460.

[46] Chaoji, V., Hoonlor, A., & Szymanski, B. K. (2010), "Recursive data mining for role identification in electronic communications", International Journal of Hybrid Intelligent Systems, 7(2), 89-100.

[47] Sreemathy, J., & Balamurugan, P. S., "An efficient text classification using knn and naive bayesian", International Journal on Computer Science and Engineering, (2012) 4(3), 392.

[48] Shardan, Rajesh, and Uday Kulkarni, "Implementation and evaluation of evolutionary connectionist approaches to automated text summarization", (2010).

[49] Prasad, R. S., Kulkarni, U. V., & Prasad, J. R., "A novel evolutionary connectionist text summarizer", (ECTS). In 2009 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication (2009, August) (pp. 606-610). IEEE.

[50] Qian, Tieyun, Bing Liu, Li Chen, Zhiyong Peng, Ming Zhong, Guoliang He, Xuhui Li, and Gang Xu, "Tri-Training for authorship attribution with limited training data: a comprehensive study", Neurocomputing 171 (2016): 798-806.

[51] Frantzeskou, Georgia, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas, "Source code author identification based on n-gram author profiles", In IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 508-515. Springer, Boston, MA, 2006.

[52] Li, Jiexun, Rong Zheng, and Hsinchun Chen, "From fingerprint to writeprint", Communications of the ACM 49, no. 4 (2006): 76-82.

[53] Koppel, Moshe, and Shachar Seidman, "Detecting pseudepigraphic texts using novel similarity measures", Digital Scholarship in the Humanities 33, no. 1 (2017): 72-81.

[54] Rexha, Andi, Mark Kröll, Hermann Ziak, and Roman Kern, "Authorship identification of documents with high content similarity", Scientometrics 115, no. 1 (2018): 223-237.

[55] Frantzeskou, Georgia, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas, "Effective identification of source code authors using byte-level information", In Proceedings of the 28th international conference on Software engineering, pp. 893-896. ACM, 2006.

[56] Kocher, Mirco, and Jacques Savoy, "A simple and efficient algorithm for authorship verification", Journal of the Association for Information Science and Technology 68, no. 1 (2017): 259-269.

[57] Ma, Jianbin, Bing Xue, and Mengjie Zhang, "A Profile-Based Authorship Attribution Approach to Forensic Identification in Chinese Online Messages", In Pacific-Asia Workshop on Intelligence and Security Informatics, pp. 33-52. Springer, Cham, 2016.

[58] Houvardas, John, and Efstathios Stamatatos, "N-gram feature selection for authorship identification", In International conference on artificial intelligence: Methodology, systems, and applications, pp. 77-86. Springer, Berlin, Heidelberg, 2006.

[59] Stamatatos, Efstathios, "Author identification using imbalanced and limited training texts", In 18th International Workshop on Database and Expert Systems Applications (DEXA 2007), pp. 237-241. IEEE, 2007.

[60] Schwartz, Roy, Oren Tsur, Ari Rappoport, and Moshe Koppel, "Authorship attribution of micro-messages", In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1880-1891. 2013.

[61] Diederich, Joachim, Jörg Kindermann, Edda Leopold, and Gerhard Paass, "Authorship attribution with support vector machines", Applied intelligence 19, no. 1-2 (2003): 109-123.

[62] Koppel, Moshe, and Jonathan Schler, "Authorship verification as a one-class classification problem", In Proceedings of the twenty-first international conference on Machine learning, p. 62. ACM, 2004.

[63] Neme, Antonio, J. R. G. Pulido, Abril Muñoz, Sergio Hernández, and Teresa Dey, "Stylistics analysis and authorship attribution algorithms based on self-organizing maps", Neurocomputing 147 (2015): 147-159.

[64] Kestemont, Mike, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast, "Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection", In Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al., pp. 1-25. 2018.

[65] Zhai, Chengxiang, and John Lafferty, "A study of smoothing methods for language models applied to information retrieval", ACM Transactions on Information Systems (TOIS) 22, no. 2 (2004): 179-214.

[66] Nini, Andrea, "An authorship analysis of the Jack the Ripper letters", Digital Scholarship in the Humanities 33, no. 3 (2018): 621-636.

[67] Lancashire, Ian, and Graeme Hirst, "Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: a case study", In 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice, pp. 8-10. 2009.

[68] Van Dam, Michiel, and Claudia Hauff, "Large-scale author verification: temporal and topical influences", In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 1039-1042. ACM, 2014.

[69] Tamboli, Mubin Shaukat, and Rajesh S. Prasad, "Authorship analysis and identification techniques: A review", International Journal of Computer Applications 77, no. 16 (2013).

[70] Hughes, James M., Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore, "Quantitative patterns of stylistic influence in the evolution of literature", Proceedings of the National Academy of Sciences 109, no. 20 (2012): 7682-7686.

[71] Tamboli, Mubin Shoukat, and Rajesh S. Prasad, "Feature Selection in Time Aware Authorship Attribution", In 2018 International Conference On Advances in Communication and Computing Technology (ICACCT), pp. 534-537. IEEE, 2018.

[72] Kešelj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas, "N-gram-based author profiles for authorship attribution", In Proceedings of the conference pacific association for computational linguistics, PACLING, vol. 3, pp. 255-264. sn, 2003.

[73] HaCohen-Kerner, Yaakov, Daniel Miller, Yair Yigal, and Elyashiv Shayovitz, "Cross-domain Authorship Attribution: Author Identification using char sequences, word unigrams, and POS-tags features", Working Notes of CLEF (2018).

[74] Gómez-Adorno, Helena, Grigori Sidorov, David Pinto, and Ilia Markov, "A graph based authorship identification approach", Working notes papers of the CLEF (2015).

[75] Ge, Zhenhao, Yufang Sun, and Mark JT Smith, "Authorship attribution using a neural network language model", In Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[76] Hitschler, Julian, Esther van den Berg, and Ines Rehbein, "Authorship attribution with convolutional neural networks and POS-Eliding", In Proceedings of the Workshop on Stylistic Variation, pp. 53-58. 2017.

[77] Holmes, David I, "Authorship attribution", Computers and the Humanities 28, no. 2 (1994): 87-106.

[78] Li, Xiaoyan, and W. Bruce Croft, "Time-based language models", In Proceedings of the twelfth international conference on Information and knowledge management, pp. 469-475. ACM, 2003.

[79] Rosenthal, Sara, and Kathleen McKeown, "Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 763-772. Association for Computational Linguistics, 2011.

[80] Luyckx, Kim, and Walter Daelemans, "Shallow text analysis and machine learning for authorship attribtion", LOT Occasional Series 4 (2005): 149-160.

[81] Nieto, Victoria Guillén, Chelo Vargas Sierra, María Pardiño Juan, Patricio Martínez Barco, and Armando Suárez Cueto, "Exploring state-of-the-art software for forensic authorship identification", International Journal of English Studies 8, no. 1 (2008): 1-28.

[82] Brocardo, Marcelo Luiz, Issa Traore, and Isaac Woungang, "Toward a framework for continuous authentication using stylometry", In 2014 IEEE 28th International Conference on Advanced Information Networking and Applications, pp. 106-115. IEEE, 2014.

[83] Zlatkova, Dimitrina, Daniel Kopev, Kristiyan Mitov, Atanas Atanasov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov, "An ensemble-rich multi-aspect approach for robust style change detection", CLEF 2018 Working Nots of CLEF (2018).

[84] Swan, Russell, and David Jensen, "Timemines: Constructing timelines with statistical models of word usage", In KDD-2000 Workshop on Text Mining, pp. 73-80. 2000.

[85] Can, Fazli, and Jon M. Patton, "Change of writing style with time", Computers and the Humanities 38, no. 1 (2004): 61-82.

[86] Kanhabua, Nattiya, and Kjetil Nørvåg, "Improving temporal language models for determining time of non-timestamped documents", In International Conference on Theory and Practice of Digital Libraries, pp. 358-370. Springer, Berlin, Heidelberg, 2008.

[87] Keikha, Mostafa, Shima Gerani, and Fabio Crestani, "Time-based relevance models", In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1087-1088. ACM, 2011.

[88] Wei, Bingjie, Shuai Zhang, Rui Li, and Bin Wang, "A time-aware language model for microblog retrieval", CHINESE ACADEMY OF SCIENCES BEIJING INST OF COMPUTING TECHNOLOGY, 2012.

[89] Web: www.nltk.org

[90] Weeber, Marc, Rein Vos, and R. Harald Baayen, "Extracting the lowest-frequency words: Pitfalls and possibilities", Computational Linguistics 26, no. 3 (2000): 301-317.

[91] Han, Jiawei, Jian Pei, and Micheline Kamber, "Data mining: concepts and techniques" Elsevier, 2011.

[92] Kanhabua, Nattiya, and Kjetil Nørvåg, "A comparison of time-aware ranking methods", In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1257-1258. ACM, 2011.

[93] Kanhabua, Nattiya, and Kjetil Nørvåg, "Determining time of queries for re-ranking search results", In International Conference on Theory and Practice of Digital Libraries, pp. 261-272. Springer, Berlin, Heidelberg, 2010.

[94] Yang, Jie, Chen-zhou YE, Yong Quan, and Nian-yi CHEN, "Simplified SMO algorithm for support vector regression", Infrared and Laser Engineering 5 (2004).

[95] Hassan, FI Haj, and Mousmi A. Chaurasia, "N-Gram Based Text Author Verification", International Proceedings of Computer Science & Information Technology 36 (2012).

[96] Peng, Fuchun, Dale Schuurmans, and Shaojun Wang, "Augmenting naive bayes classifiers with statistical language models", Information Retrieval 7, no. 3-4 (2004): 317-345.

[97] Huang, Anna, "Similarity measures for text document clustering", In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, vol. 4, pp. 9-56. 2008.

[98] Weisberg, Sanford., "Applied linear regression", Vol. 528. John Wiley & Sons, 2005.

[99] Wan, Xiang, Wenqian Wang, Jiming Liu, and Tiejun Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range", BMC medical research methodology 14, no. 1 (2014): 135.

[100] Kantardzic, Mehmed, "Data mining: concepts, models, methods, and algorithms", John Wiley & Sons, 2011.

[101] Web: www.cs.waikato.ac.nz/ml/weka

Publications

Journal Papers

1. Tamboli, M. S., & Prasad, R. S. (2013). Authorship analysis and identification techniques: A review. International Journal of Computer Applications, 77(16).
   Indexing: ProQuest, Google Scholar, EBSCO, Open J-Gate etc.

2. Tamboli, M. S., & Prasad, R. (2019). A robust authorship attribution on big period. International Journal of Electrical & Computer Engineering (2088-8708), 9.
   Indexing: Scopus, Google Scholar Profile, ProQuest, EBSCO etc.

3. Tamboli, M. S., & Prasad, R. (2019). Author identification with feature transformation method. Digital Scholarship in the Humanities.
   Indexing: Scopus, Computer Science Index ,PROQUEST DATABASE, Social Sciences Citation Index

Conference Paper with Publications

- Tamboli, M. S., & Prasad, R. S. (2018, February). Feature Selection in Time Aware Authorship Attribution. In 2018 International Conference On Advances in Communication and Computing Technology (ICACCT) (pp. 534-537). IEEE.
  Published at IEEE Xplore digital library
  Indexing:Scopus

- Tamboli, M. S., & Prasad, R. S. (2018, December). Authorship Identification with Multi Sequence Word Selection Method. In International Conference on Intelligent Systems Design and Applications (pp. 653-661). Springer, Cham.
  Published as chapter in Intelligent Systems Design and Applications book published by Springer

Indexing: ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink

- Tamboli, M. S., & Prasad, R. S. (2016). Feature Selection in Author identification. In CAASR International Conference on Innovative Engineering and Technologies (CAASR-ICIET 16), Kuala Lumpur, MALAYSIA