University of Nairobi

SCHOOL OF COMPUTING AND INFORMATICS

PROJECT PROPOSAL

AUTHOR IDENTIFICATION SYSTEM

MBOGORI SALMA WANJA

P15/45122/2017

PROJECT SUPERVISOR

DR. MIRITI

This project proposal submitted in partial fulfillment of the requirements of the Bachelor of Science in Computer Science of the University of Nairobi.

Chapter One: Introduction

1.1. Background

Author identification strives to identify the authors of anonymously published texts. It is an emerging area of research associated with applications in literary research, cyber-security, forensics, and social media analysis. Author identification is a subfield in Natural Language Processing (NLP) that uses machine learning techniques to determine the author of a text based on identifying characteristics such as word frequency, vocabulary, etc. In the current era that we live in today, electronic communication has become one of the most popular modes of communication. These modes include email, text messaging, social networking sites such as Twitter and Facebook, forums and message boards, blogs, etc where most of them use the internet as the backbone that aids with the communication.

The majority of research in this field has focused on long formal texts such as excerpts from novels. However, as the current trends in information technology encourage an abundance of short, informal writing, it becomes increasingly important to determine to what extent author identification techniques also apply to such text. There are many potential practical applications for author identification in these areas, such as identifying the source of anonymous messages for security purposes or using identification as the basis for developing targeted advertising.

Author identification is a critical point to be ensured, because many people plagiarise the content belonging to other authors. Keeping in mind that every individual has a unique and distinctive way of speaking and writing, Stylometry which is the statistical analysis of variations in literary style between one writer and another, can be used for the author identification for text documents since non–repudiation and integrity of the message are becoming major concerns.

1.2. Problem Definition

Today, anyone with an internet access is able to anonymously post whatever they want whenever they want. Some individuals use this as an opportunity to spread hate and threats which may escalate to violence in the real world. The users also have access to multiple platforms where they can spread these vicious messages. Since these posts are done anonymously and over multiple platforms, it is difficult to identify who wrote what, when and where.

Also, in regards to the history of literature, there have been multiple publications on various topics that have been written anonymously often due to their political and controversial nature or the authors wanting to keep their privacy. Often these mysterious authors tend to release more than one publication and it becomes difficult to identify their work to someone who may be interested in it.

1.3 Goals

- To build a working system that would provide a means to identify the owner of a text excerpt based on the variation of literary style.
- To build a working system that can assist forensic experts and/or linguists to create profiles of writers.

1.4. Objectives

1.4.1 System Objectives

1. To use a machine learning algorithm to develop a system that is able to identify the owner of a specific written text.

1.4.1 Research Objectives

1. To understand the workings of existing algorithms in natural language processing used for author identification.
2. To establish a set of factors considered in the determination of the author of a text exert.
3. To evaluate the benefits and problems of existing author identification software.

1.5. Project Justification

With the advent of social media, in particular, the way our society communicates and exchanges information has changed. Social media opens up new opportunities to express opinion. The process of determining authors of online messages, especially those with offensive as well as threatening expressions, is given higher priority due to security concerns. Identifying and attributing authorship of different text passages can be beneficial for various tasks and areas including bibliometrics -- using statistical methods for analysis of text in books, articles and other publications -- information retrieval and plagiarism detection.

Chapter Two: Literature Review

2.1. Review

1. **Navoneel Chakrabarty**

In his article titled A Machine Learning Approach to Author Identification of Horror Novels from Text Snippets, Navoneel Chakrabarty uses the following algorithm to determine the authors of literary texts.

Steps:

1. Text processing:
    a. Removal of text punctuation
    b. Lemmatization - Lemma are inflected forms of a root word, e.g. going is a lemma of go. This step is done so that only distinct words are left.
    c. Removal of stopwords - stopwords include articles (a, an, the...) and other frequently occurring words that do not, by themselves, provide meaningful information.
2. Class encoding - Being a classification problem, the classes in his algorithm are the authors. These classes are non numeric therefore, for the dataset, he encodes them to numeric values.
3. Word Cloud Visualisation - This step involves creating a visualization of the most used to least used words within the text snippet.
4. Feature Engineering - Because machine learning algorithms operate on numeric data, this step involves the transformation of text data to numeric data using the Bag-of-words technique of future engineering. Bag of words is a simple method of extracting features from documents where the frequency of appearance of the feature in the document is put into consideration to help in training the model.
5. Training the model - The model is trained using the multinomial naive bayes algorithm.
6. Performance analysis - This step involves the evaluation of the performance of the trained model using testing data.

2. Feature Selection for Enhanced Author Identification of Turkish Text

The author identification work in this study starts with collecting the documents of known authors from websites of Milliyet and Kıbrıs Gazetesi newspapers randomly. All the HTML tags are cleaned as a pre-processing step with the system developed in Java and plain text data are saved in MySQL. The database consists of 850 columns written by 17 columnists as a total, 50 columns from each columnist. The obtained datasets are; Dataset I: Milliyet, has 10 authors from www.milliyet.com.tr with 500 documents and have an average of 18,800 sentences and 305,000 words, Dataset II: Kibris, has 7 authors from www.kibrisgazetes.com with 350 documents and have an average of 265 sentences and 5,960 words, 50 columns per writer.

Feature extraction follows which is one of the most important stages of author identification since it finds distinctive features that exhibit the writing style of each author. From the list of stylometric features such as lexical, character, syntactic and semantic features, this study focuses on the lexical features. A significant advantage of such features is that they can be applied to any language and any corpus with no additional requirements [16]. After obtaining the corpus, 20 lexical features (style markers) were used to extract the feature sets. They are shown below;

UK

Table 1 Style Markers

No	Style Marker	No	Style Marker
1	# of sentences	11	# of question mark
2	# of words	12	# of punctuation marks
3	Average # of words	13	Average # of dots
4	# of words after stopword removal	14	Average # of commas
5	Average # of words after stopword removal	15	Average # of semicolon
6	# of dots	16	Average # of colon
7	# of commas	17	Average # of exclamation
8	# of semicolon	18	Average # of question mark
9	# of colon	19	# of Non-Turkish words
10	# of exclamation	20	Average # of non-Turkish words

Figure 1.0

Feature selection methods were used to utilize for more promising features and achieve better results.

For each feature set, all classifiers are trained and tested by applying 10-fold cross validation where it estimates the performance of a classifier by breaking the dataset into 10

partitions and then the model is trained on 9 datasets and tested on 1. The mean accuracy is obtained by repeating the process 10 times.

For classification, the study used four different classification algorithms and compared the performance of each to determine which classifier obtained a better accuracy. The algorithms studied include the following: ***Naïve Bayes, KNN, SVM , Decision Tree***

KNN was found to have the highest accuracy of all tested classification algorithms.

2.2. Other Existing Systems

1. AuthorClaim

This is a nonprofit open source library where people can claim to be the authors of a document. AuthorClaim provides a service where authors can register, leave personal data such as their names, affiliations and homepage URL. Then they can claim to be the author of document described in some contributing bibliographic databases. It covers all disciplines though most text excerpts are regarding economics. It Integrates with databases for institutions (ARIW) and publications (3lib.org).

2. ArXiv

This is an e-print service in the fields of physics, mathematics, non-linear science, computer science, quantitative biology, quantitative finance and statistics. Submissions to arXiv must conform to Cornell University academic standards. arXiv is owned and operated by Cornell University, a private not-for-profit educational institution.

Chapter Three: Methodology

3.1 System Analysis

This stage involves an undertaking to study the proposed system. The system was decomposed into its components which were then evaluated based on how well they interact to collectively accomplish their intended purpose.

3.1.1 Feasibility Analysis

The purpose of this stage is to determine the project's potential for success. Systems are a huge investment therefore highlighting the need to evaluate proposed solutions before implementation to minimize the risk of investment in a project without beneficial returns. Factors considered include:

I. Technical and Technological Feasibility

Technical feasibility-encompasses the study of the proposed system or a working model in terms of its inputs, processes, outputs, programs and procedures and is important in planning for the future and troubleshooting.
Technological feasibility-encompasses the evaluation of existence of sufficient and appropriate technological tools that support the creation and maintenance of the proposed system.

Data necessary for the implementation of this project is available and will be obtained from research. The data will be used to train the machine learning algorithm that will be used that will drive the functionality of the Author Identification System.

The project will be built using python for backend, react js for front end and data will be stored in a MySql database or read from file. These technologies are free and robust, with active support from the creators and the development community.

II. Economic Feasibility

Tools and technologies adopted for use within this project are open source, free to obtain and use. The value generated upon the completion of the proposed system will therefore surpass the cost incurred during development. The proposed system will be hosted on DigitalOcean cloud servers offering a 25GB SSD storage and 1GB RAM server at a monthly fee of 5 USD

III. Schedule Feasibility

The schedule for the project is indicated in section 3.3 below and it specifies the timeline for project implementation.

3.1.2 Functional and Non-Functional Requirements

A. FUNCTIONAL

- The system should be able to correctly match an unknown text excerpt to an author
- System should be able to group the text excerpts according to the authors.

B. NON-FUNCTIONAL

- Accuracy - The system correctly matches unknown text excerpts to authors.
- Consistency & Reliability - The results of the matching process should be reproducible to ensure reliability.
- Usability - The user interface should be appealing, with relevant content and easy to use.
- Speed - The algorithm for author identification should be fast enough to enable processing of multiple text excerpts.

3.1.3 Use Case Diagrams

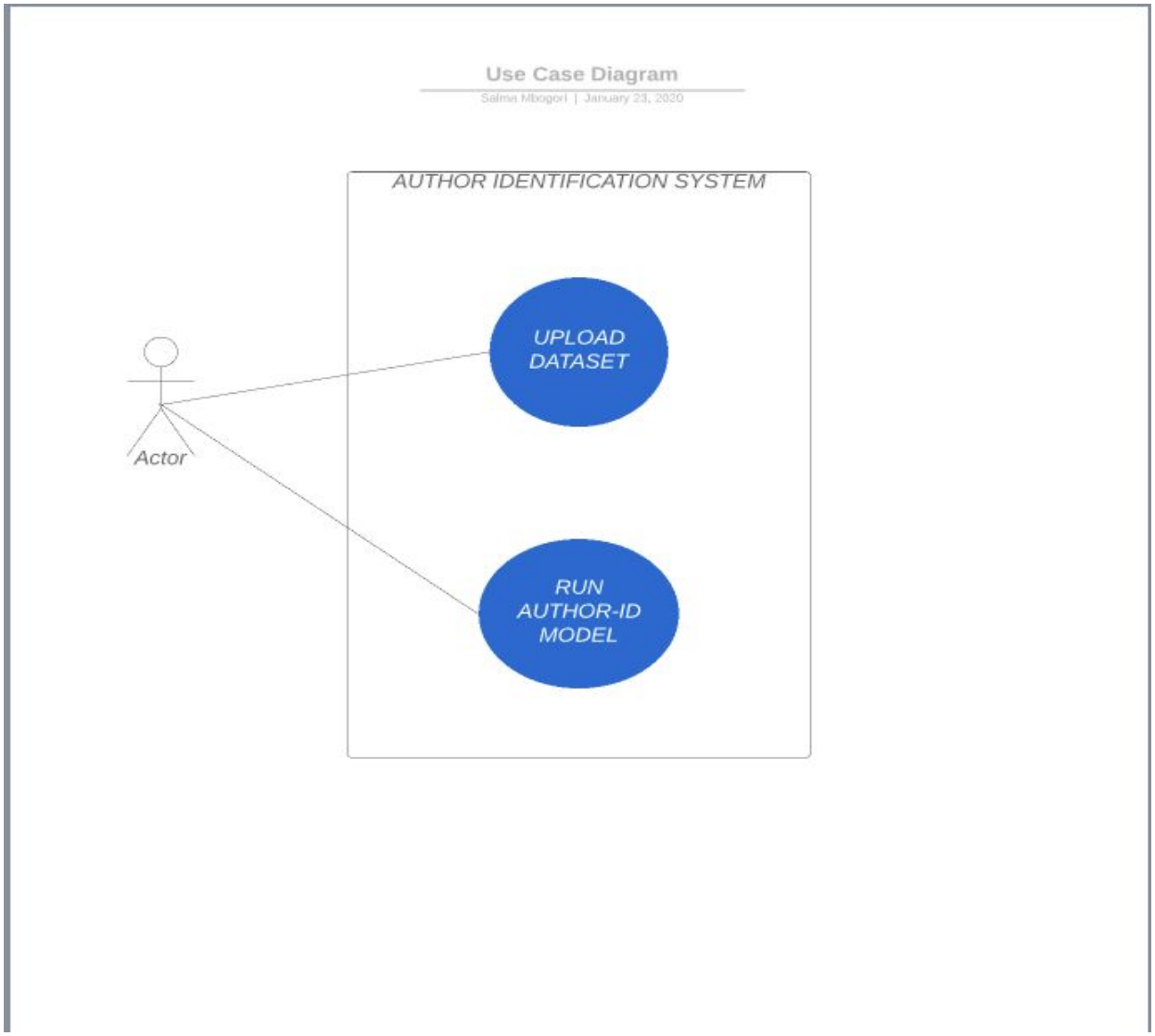


Figure 1.1

3.1.4 Data Flow Diagrams

1. LEVEL 0 DFD

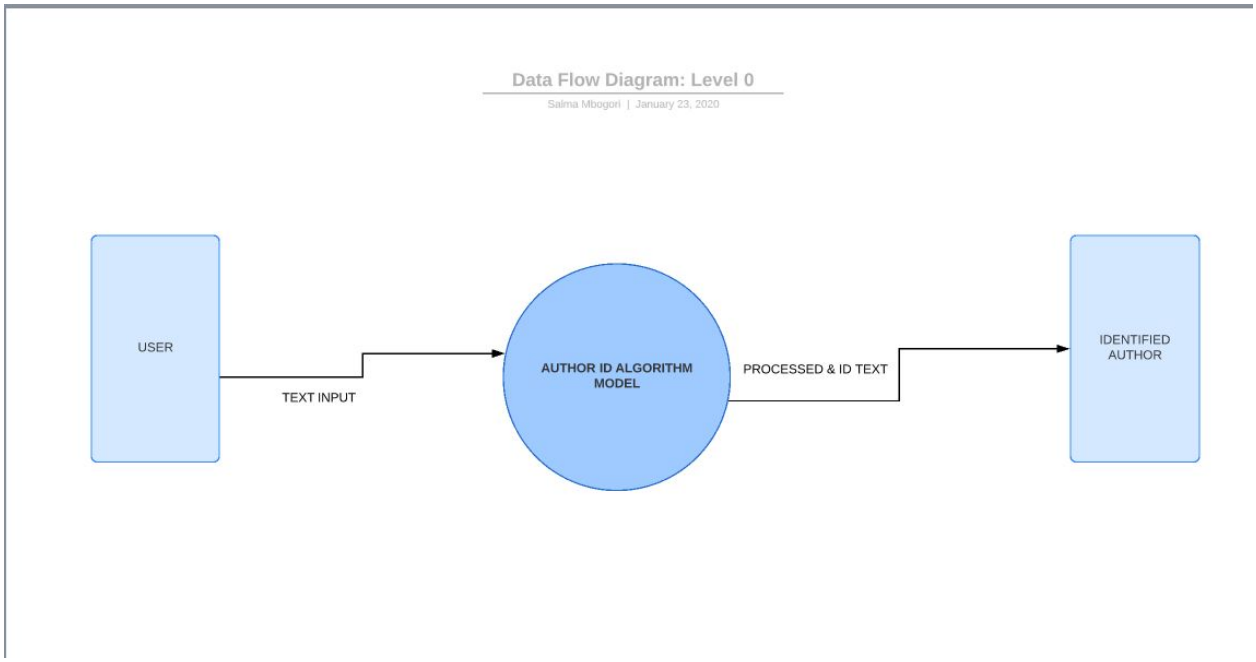


Figure 1.2

2. LEVEL 1 DFD

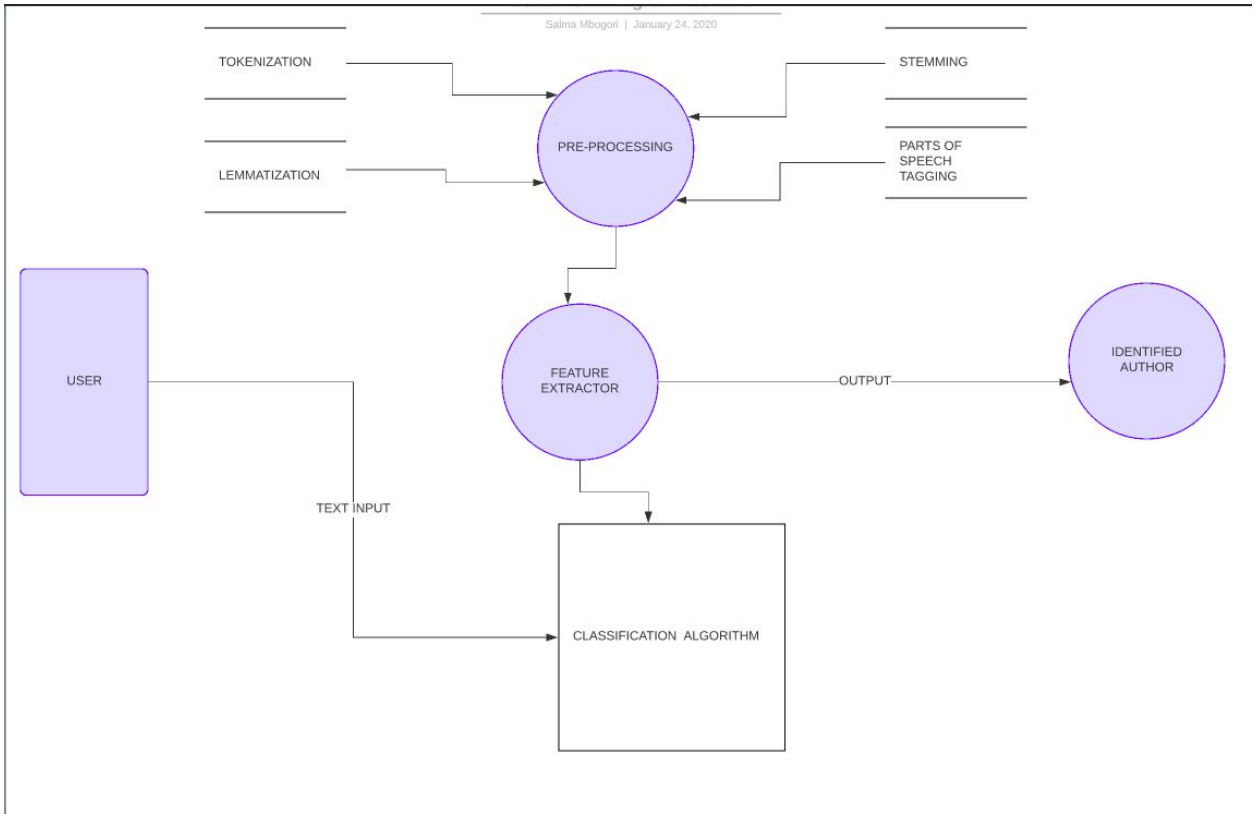


Figure 1.3

3. LOGICAL DFD

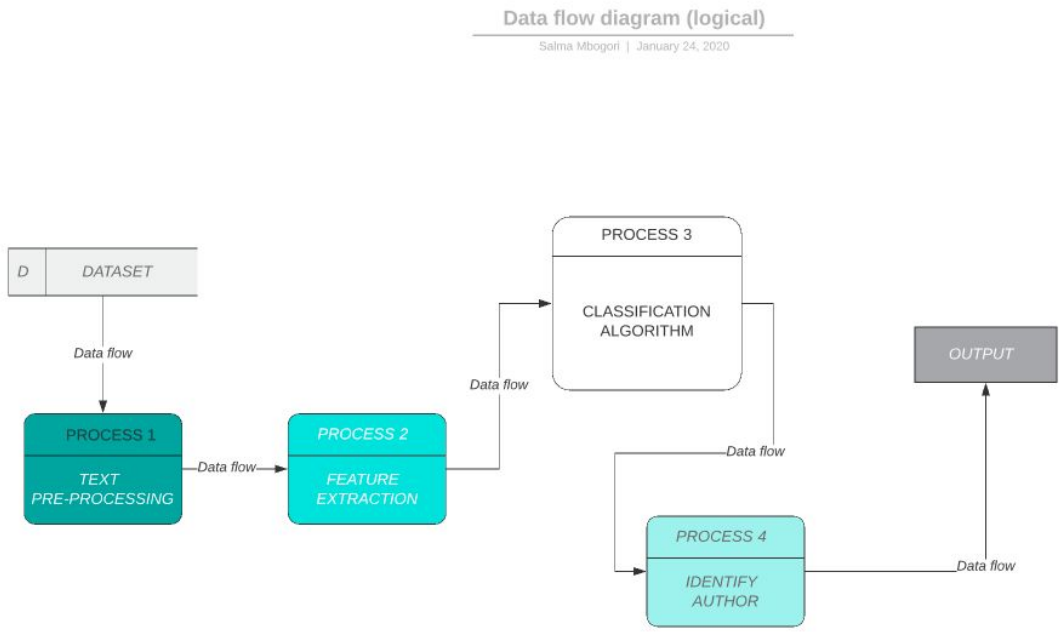


Figure 1.4

3.2 System Design

3.2.1 Text Analysis/Pre-processing

This involves cleaning the text to rid it of unnecessary words and punctuation before running it through the algorithm. There are a number of text attributes chosen depending on what is required or not required depending on the type of text to be analyzed. The following are just but a few;

A. **Tokenization-** There exists sentence tokenization which involves breakdown of a paragraph to sentences and word tokenization which is breakdown of the sentences to words.
B. **Stop-words Removal-** Includes getting rid of common language articles, pronouns and prepositions such as “and”, “the” or “to” in English. In this process some very common words that appear to provide little or no value to the NLP objective are filtered and excluded from the text to be processed, hence removing widespread and frequent terms that are not informative about the corresponding text.

C. **Stemming-** Refers to the process of slicing the end or the beginning of words with the intention of removing affixes (lexical additions to the root of the word).
D. **Lemmatization-** This involves reducing a word to its base form and grouping together different forms of the same word

3.2.2 Feature Extraction

In this step, the annotated text is converted into features which are simple objects that contain a name and a value which gives a machine learning algorithm a sample text that is easier to process. In this phase, certain attributes of the text are taken into consideration to create features.

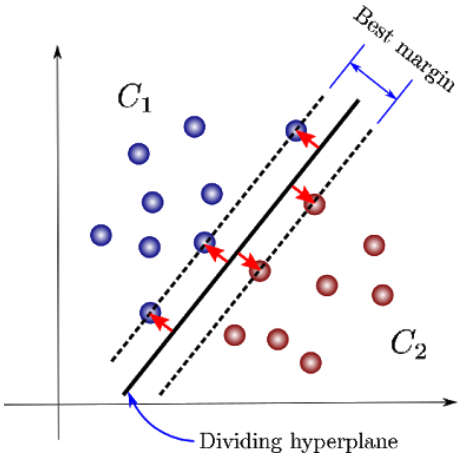
3.2.3 Classification

Support Vector Machines

This technique is based on finding the maximal margin hyper-plane which separates the data in two sets. Finding this hyper-plane is based on structural risk minimization, a principle that tries to minimize the generalization error while minimizing the training error and avoiding a model that is too complex. Other machine learning techniques only minimized the training error, but this does not necessarily mean that the generalization error is minimized. SVM can better generalize over unseen data. And in contrast with decision trees and neural networks, SVM do not use a greedy approach, therefore it can find the globally optimal solution.

A SVM tries to find the hyper-plane with the largest margin because this improves the generalization error and a small margin is prone to overfitting. The hyper-plane is positioned so that the margin between the classes is as large as possible. Only the data points that are necessary to determine the largest margin are considered, these are called the support vectors. Support vectors are the points that touch the line were the margin stops (dotted).

The SVM as described only separates two classes. But in many real life situations, like author identification, one wants to distinguish between more classes. This can be performed using pairwise classification. This classification method constructs classifiers for each pair of classes, while ignoring the data that does not belong to one of these two classes. So, for C classes, C(C-1)/2 binary classifiers need to be constructed. The unseen data sample gets the class label that is predicted most by the classifiers. As with the nearest neighbor algorithm, there can be a tie between class labels



3.3 Conceptual Design of the system

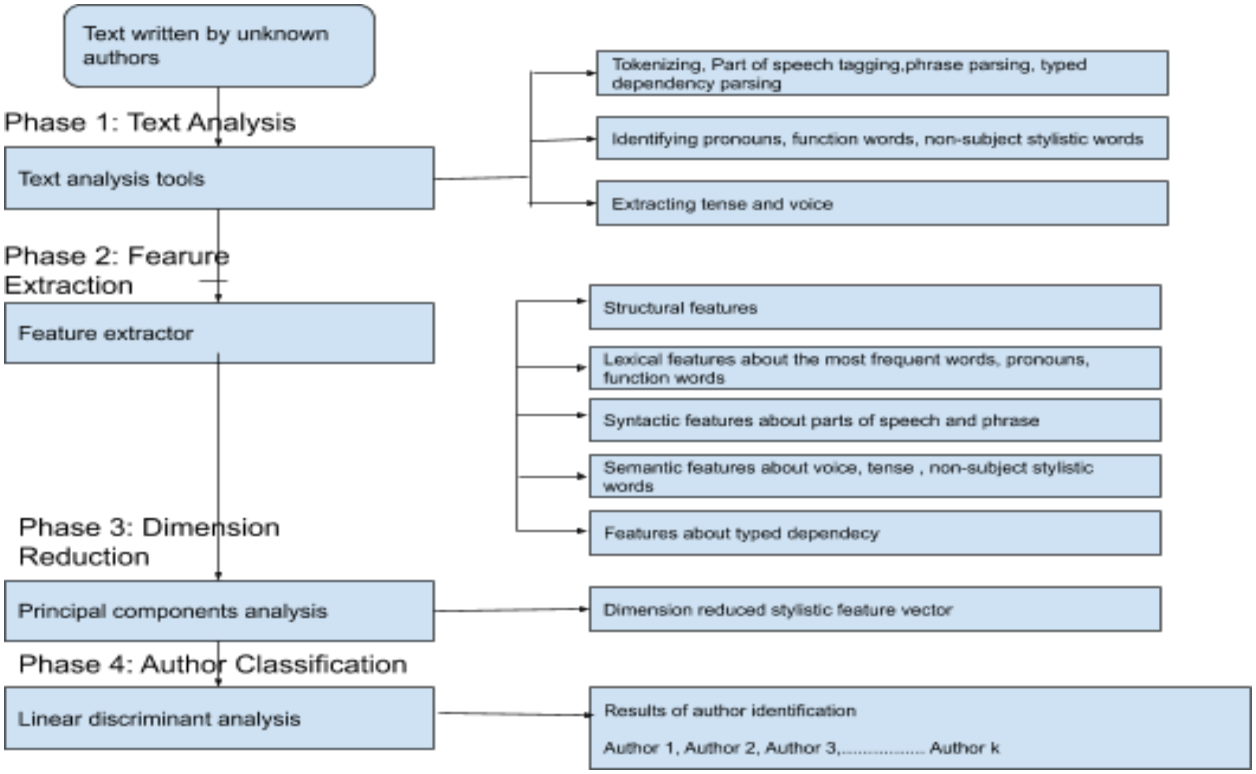


Figure 1.5

3.4 System Development Methodology to be used

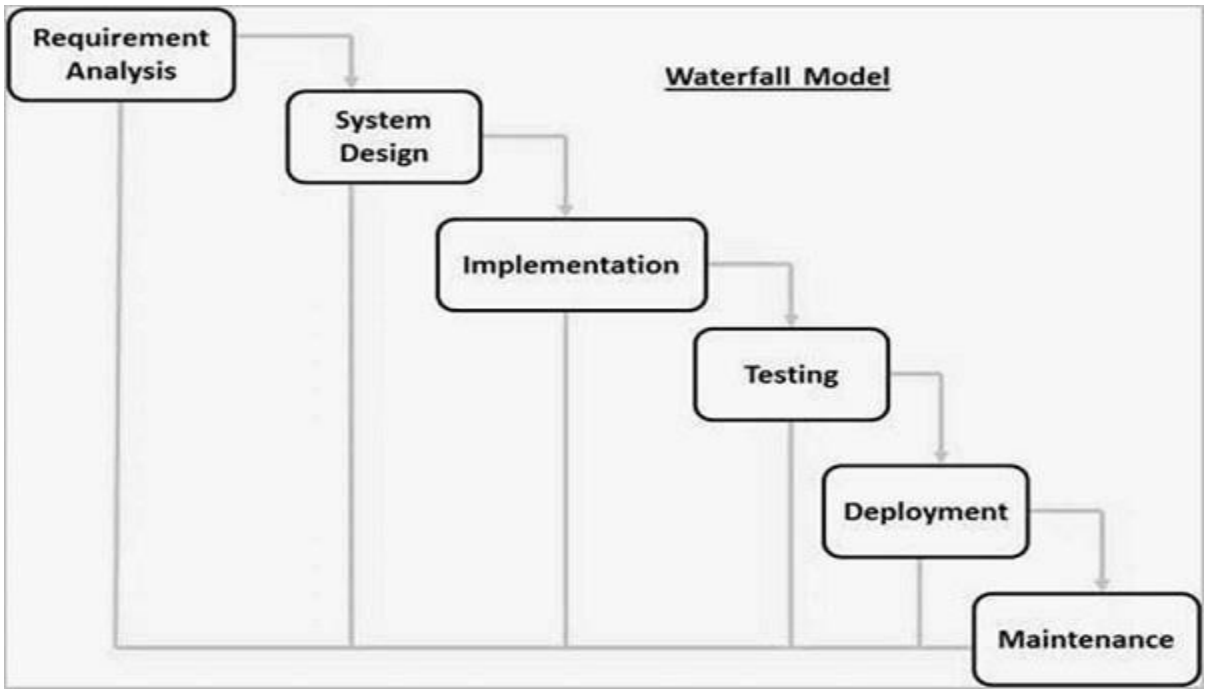


Figure 1.6

For this project, the waterfall approach and found best to be used. The orderly sequence of development steps and strict controls for ensuring the adequacy of documentation and design reviews helps ensure the quality, reliability, and maintainability of the developed software. Progress of system development is measurable and it conserves resources. The project was divided into sequential phases, with some overlap and splashback acceptable between phases. The phases included:

I. Requirement Specification

Requirement gathering was done through;

• Study of the already existing systems that more or less function in a similar manner

Requirement specification also involved elicitation which is the definition of the system in terms understood by the user and the developer.

II. Analysis

The requirements that were specifies were grouped into functional and non-functional requirements and were analyzed using use cases and scenarios.

III. Design

Requirements were created into high level detailed design diagrams that would be used for the

logical understanding of how the system is expected to be created and work.

IV. Coding

The designs from the Design phase were used to generate the actual code in the respective

languages used in order to actually develop the systematic

V. Implementation/ Testing

The different forms of code was combined to develop the final working system and testing was

done to ensure the system was working correctly.

3.3. Schedule

This project is anticipated to span 7 months, starting October 2019 and ending in April

2010. The schedule and Gantt chart is shown in the tables below.

3.3.1 Project Schedule

TASKS DESCRIPTION	TIME		
	START DATE	END DATE	DURATION (Days)
Problem definition and research	14/10/2019	13/11/2019	30
Requirements Elicitation	14/11/2019	24/11/2019	10
Requirements Specification	25/11/2019	2/12/2019	7
Algorithm Design	3/12/2019	23/12/2019	20
User Interface Design	3/01/2020	13/01/2020	10
Implementation	14/01/2020	18/03/2020	50
Testing	19/03/2020	29/03/2020	10
Review and Refinement	30/03/2020	6/04/2020	7
Deployment	07/04/2020	12/04/2020	5

Table 1.1

3.3.3 Gantt Chart

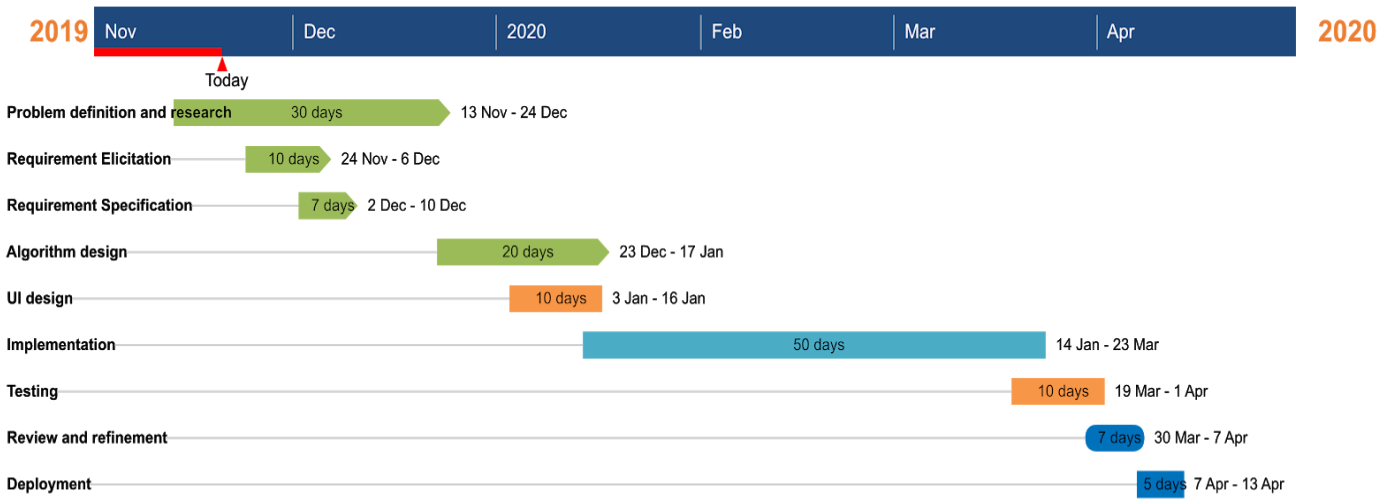


Figure 1.7

3.4. Budget

ITEM DESCRIPTION	SPECIFICATION	QUANTITY	UNIT PRICE (KSH)	AMOUNT
Computer	Intel® CoreTM i5-6200U CPU @ 2.30GHz × 4 Intel® HD Graphics 520 (Skylake GT2)	1	0	0
Printing	N/A	500 Pages	10	5000
Binding	N/A	5	100	500
Notebook	A4 200 Pages	1	100	100
Home /School internet	20 Mbps	7 Months	4000	28,000
Deployment Server	4GB RAM 4 Cores 300 GB HDD	6 Months	2,500	15,00
Transport	Data Collection, Testing, Review	10 trips	300	3000
TOTAL				**51,600**

Table 1.2

References

- Rachel M. Green, John W. Sheppard. *Comparing Frequency- and Style-Based Features for Twitter Author Identification.* Available at: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS13/paper/viewFile/5917/6043

- Lakshmi M, Pushpendra Kumar Pateriya. *A Study on Author Identification through Stylometry.* Available at: https://pdfs.semanticscholar.org/30a2/6b6c5ce991e62cc0c12f3308451161c9b6a7.pdf?_ga=2.13307703.162710293.1574080265-56545054.1574080265
- George K. Mikros. (2012) *Authorship Attribution and Gender Identification in Greek Blogs.* Available at: https://www.researchgate.net/publication/236583622_Authorship_Attribution_and_Gender_Identification_in_Greek_Blogs
- Efstathios Stamatatos. *A Survey of Modern Authorship Attribution Methods* . Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1634&rep=rep1&type=pdf
- Juliane Witte. (2011-2012) *Author Identification Techniques*. Available at: https://www.academia.edu/1198018/Author_Identification_Techniques
- Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein and Martin Potthast5. (2018) *Overview of the Author Identification Task.* Available at: http://ceur-ws.org/Vol-2125/invited_paper_2.pdf