Detecting Intrinsic Plagiarism Using Text Analytics

B. B. D. S. Abeykoon

2020

Detecting Intrinsic Plagiarism Using Text Analytics

A Dissertation Submitted for the Degree of Master of Business Analytics

B. B. D. S. Abeykoon

University of Colombo School of Computing

2020

DECLARATION

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: B. B. D. S. Abeykoon

Registration Number: 2018/BA/001

Index Number: 18880012

20.09.2021

Signature: Date:

This is to certify that this thesis is based on the work of Mrs. B. B. D. S. Abeykoon under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

……………………………………..

Dr. Ruwan Weerasinghe
Supervisor
Date: 20.09.2021

External Supervisor:
T. Kartheeswaran
Department of Physical Science,
Faculty of Applied Science.
University of Vavuniya.

I would like to dedicate this thesis to my respectful parents and beloved husband without whose constant support this thesis thesis was not possible. They always inspire me and give words of encouragement and push for tenacity ring in my ears. They have never left my side and are very special. All of you have been my best cheerleaders.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr. Ruwan Weerasinghe and Mr. T. Kartheeswaran for all their help and advice with this research project. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. Finally, I would like to express my gratitude to my parents and my husband. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Abstract

With wide access to information and services, the detection of originality has become a serious problem that universities and other organizations have to increasingly pay attention to. Plagiarism is the term in used to categorize non-original content passed off as one's authentic contribution. Several services for the detection of plagiarism rely on massive archives of existing written work against which any original work is compared. While these services can be quite expensive, there is also a need to be able to detect plagiarism without access to such archives. This is known as intrinsic plagiarism detection, a document is analysed to distinguish any anomalies that exist in its own overall writing style. This study is focused on the identification of intrinsic plagiarism which aims to learn significant features that would help a machine learning algorithm to detect anomalous sections in a given document.

Documents were selected from the three broad domains of global warming, civics and health and rubber plantation, which were written by single authors. After initial preprocessing, paragraphs written by other authors on the same domain were added in order to simulate the intrinsic plagiarism scenario. The result was an imbalanced dataset and a model is built with the stylistic features. The One Class SVM algorithm was used for classification with the 'Author' class and the 'Non-author' class as labels. Lexical features and the POS tags were extracted from the text as features and the best ten features were selected among them. The model was implemented on all of the features and the best features were compared. The results were obtained with the performance measures of validation accuracy, f1 score, precision, and recall. In addition, the accuracies were compared with the Naive Bayes classifier, SVM classifier, and Logistic Regression classifier at character level, word level, tf-idf level, and n-gram level in the context of bag-of-words.

The final results were evaluated and the validation accuracy for the model built with the best features is 51.18% reagrding one-class svm classifier with stylistic features. Hence the ten best features we selected significantly impact the accuracy of the model. In the context of bag-of-words, the highest validation accuracy for Naive Bayes classifier was obtained for the count vectors and the value was 94.87%. The highest validation accuracy was retrieved as 93.59% for the count vectors in logistic regression classifier and regarding the svm classifier also, counter vectors showed 88.89% of highest validation accuracy. Though model accuracy is below expected, further improvements can be expected with more data and the application of newer deep learning models.

Table of Contents

List of Tables

List of Equations

List of Figures

Chapter 1

Introduction

1.1 Motivation

There are many computer-based automated plagiarism detection methods available to identify plagiarism offenses. Digitalized form of the documents is the basis for most of the plagiarism detection methods and it is a major issue identified. Hence intrinsic plagiarism detection can be named as a new method of identifying the plagiarism. And also it can be automated. Extrinsic plagiarism detection can be defined as the method which use to detect the similarity of a document against suspicious document collection and this method is also can be automated. The difference between the above two mentioned methods is, the extrinsic plagiarism detection method use collection of references while the intrinsic plagiarism detection does not use collection of resources in order to find the plagiarism.

Identification of intrinsic plagiarism was identified as the problem and this proposal presents a novel textual analytics approach to detect intrinsic plagiarism. The people can get the data or the information from various number of resources that is not available online and they can be an old book which is not digitalized. Hence, the extrinsic plagiarism tools are not capable of identifying the plagiarized content but, the writing styles can be analyzed and the intrinsic plagiarism tools will be capable of identifying. In another situation, the plagiarized content can be directly written by another author. For example, a student who asked someone else to write parts of a document for himself. This situation cannot be detected by reference to external sources. But can be resolved through analyzing writing styles. The writing style in the context of the number of words, the length of the sentences, and the symbol such as. 'Full stop', 'apostrophe' can be taken into consideration.

The problem identified here in the study is the check whether plagiarized paragraphs or sentences placed inside a document can be detected automatically when no collection of references given. For an example, how can we detect the plagiarized contents if the plagiarized paragraphs or sentences were taken from a book which is not available in digitalized form? This situation is known as intrinsic plagiarism detection and this study brings a text analytics approach to detect the intrinsic plagiarism by analyzing a single document concerning variations in writing style.

1.2 Statement of the Problem

Detection of text misuse and identifying suspicious texts which the authors are doubtful of having authored, has a long history in plagiarism detection. The related task of author identification also has improved the performance of plagiarism detection. But the solutions provided so far are still at an unsatisfactory level as the development of technology makes the problem worse. The challenge is to detect the reproduced new works in terms of ideas, findings and methodologies which have not given the proper credits to the original authors.

Most of the solutions provided for the plagiarism detection are based on the assumption of all the sources and related information being digitalized. Therefore, criticism of this assumption has been made as not all sources are digitalized and hence intrinsic plagiarism detection tools come into importance.

The aim of this study is similar to the aim of the intrinsic plagiarism detection method and the study mainly focus on the writing style of authors without aiming on collection of references.

1.2.1 Background

1.2.1.1 Defining Plagiarism

The border line between the research and the plagiarism is negligible and hence, a close relationship exists between them. There are several definitions for the word plagiarism and according to Plagiarism.org (What is plagiarism 2020), some of them can be interpreted as in the latter. Showing that a work done by another person as yours work, imitating the works and the ideas of another person without mentioning the ownership of them, make the quotations without the quotation marks, providing false information on the resource which the information was taken, etc. are the major type of plagiarism according to the resource. In addition, imitating the sentence structure of a resource also comes under this.

Plagiarism has become a common issue in the decade of digital era and most of the documents have been digitalized and online available. Therefore the researches on automated plagiarism detection also have been increased during the last decade which takes the benefit of the development of the many trending fields such as computational linguistics. There are exceptional scenarios such as the scholarly research papers whose primary objective is to verify or falsify their research statements by quoting a significant number of lines. The number of lines plagiarized may be hundreds of lines from other resources. But that should be ignored for such content with the context.

Plagiarism is of several types and it can be termed as finding the similarities between the original document and the suspected documents without referring to the citations. The use of computers has made it ease of grabbing the content from others as well as made possible the detection too. One example of a plagiarism detection tool is 'Turnitin'. However, it is impossible to make restrictions on accessing knowledge and information through the internet. Therefore, an efficient detection system is needed to maintain both academic integrity and research work quality.

The SkillScouter website reveals some of the statistics related to plagiarism in the world in the year 2020 as follows (Keegan, 2020).

- Research conducted for seven years by Donald McCabe among 70000 students has found that 58% of students confessed to plagiarized content.
- A Survey by a U.S. News and world report reveals that 90% of students didn't think that they would get caught for plagiarism

- Three-year research has been conducted in the United States based on 63,700 undergraduate students and 9250 postgraduate students have found that 36% among them admitted to copying sentences without referencing. 38% in the same survey admitted to copying from a written source without referencing it and 14% admitted to writing false and fabricated bibliography records. 7% of the group was reported on copied verbatim from written sources without any referencing and another 7% of the students have admitted that their work was completed by someone else.

The Same website reveals some other statistics mentioned below on plagiarized websites.

- According to Turnitin, the cases of academic dishonesty and cheating are dramatically increase with the adoption of online learning in more and more schools
- There is an increasing likelihood of submitting the essays written using artificial intelligence tools using the software that was developed with complex artificial intelligence technology
- More and more source code is stolen and copied from websites such as GitHub without permission and not giving the credits
- Third parties have increasingly persuaded the students through social media who aggressively try to get good grades

There are several plagiarism cases that can be used as inputs to the plagiarism detection system or software. Usage of synonyms to replace the wordings, shuffling of words, summarization and translation are used as intelligent manipulations. In addition, paraphrasing and idea adoption is also used. In the extrinsic plagiarism detection approach, the suspected document is featured, analyze and compare with similar documents or with the original source of documents.

The following Figure 1 represents the types of plagiarism mainly and according to the figure, style analysis can be done when there is no data source of references without considering the fact that whether it is an exact copy to the original document or modified copy. However, local similarity analysis and the local identity analysis lead to style analysis in these two instances.

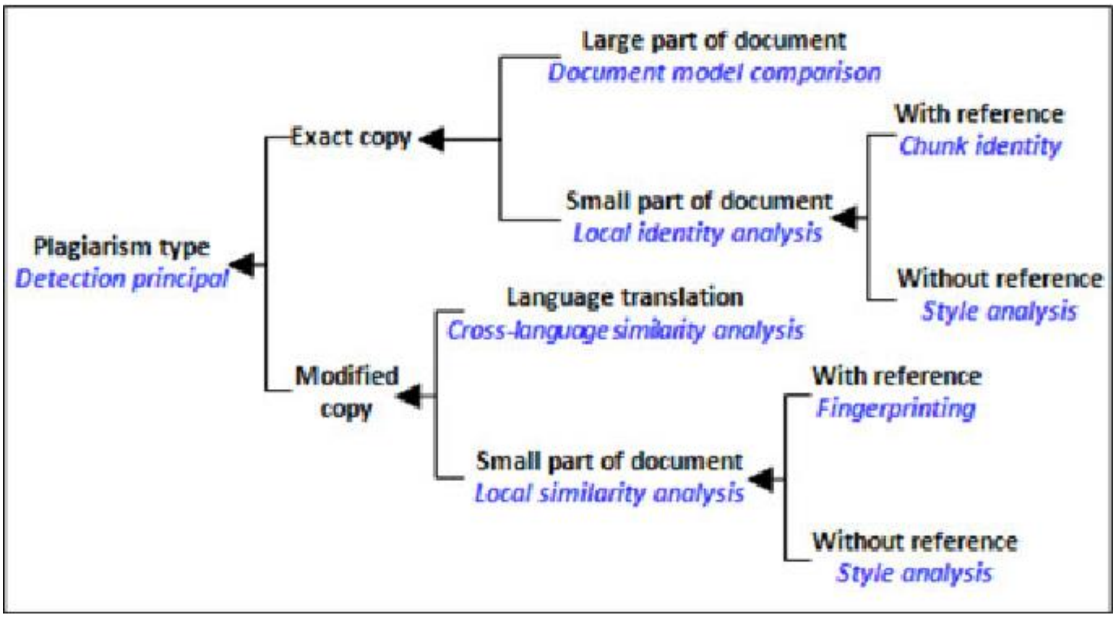


Figure 1: Plagiarism types with some related detection principles (Alzahrani et al., 2012:p.134)

A lot of researches have been conducted on plagiarism in academic activities and as well as on available software for detecting plagiarism. Hence, the researchers have found a new taxonomy of plagiarism as shown in below figure 3. According to the figure, plagiarism is of two types. The literal plagiarism can be divided into the categories as exact coy, near copy, and modified copy of restructuring. And intelligent plagiarism can be divided into text manipulation, translation, and idea adoption. This taxonomy is mainly based on the behaviors under the category of plagiarism.

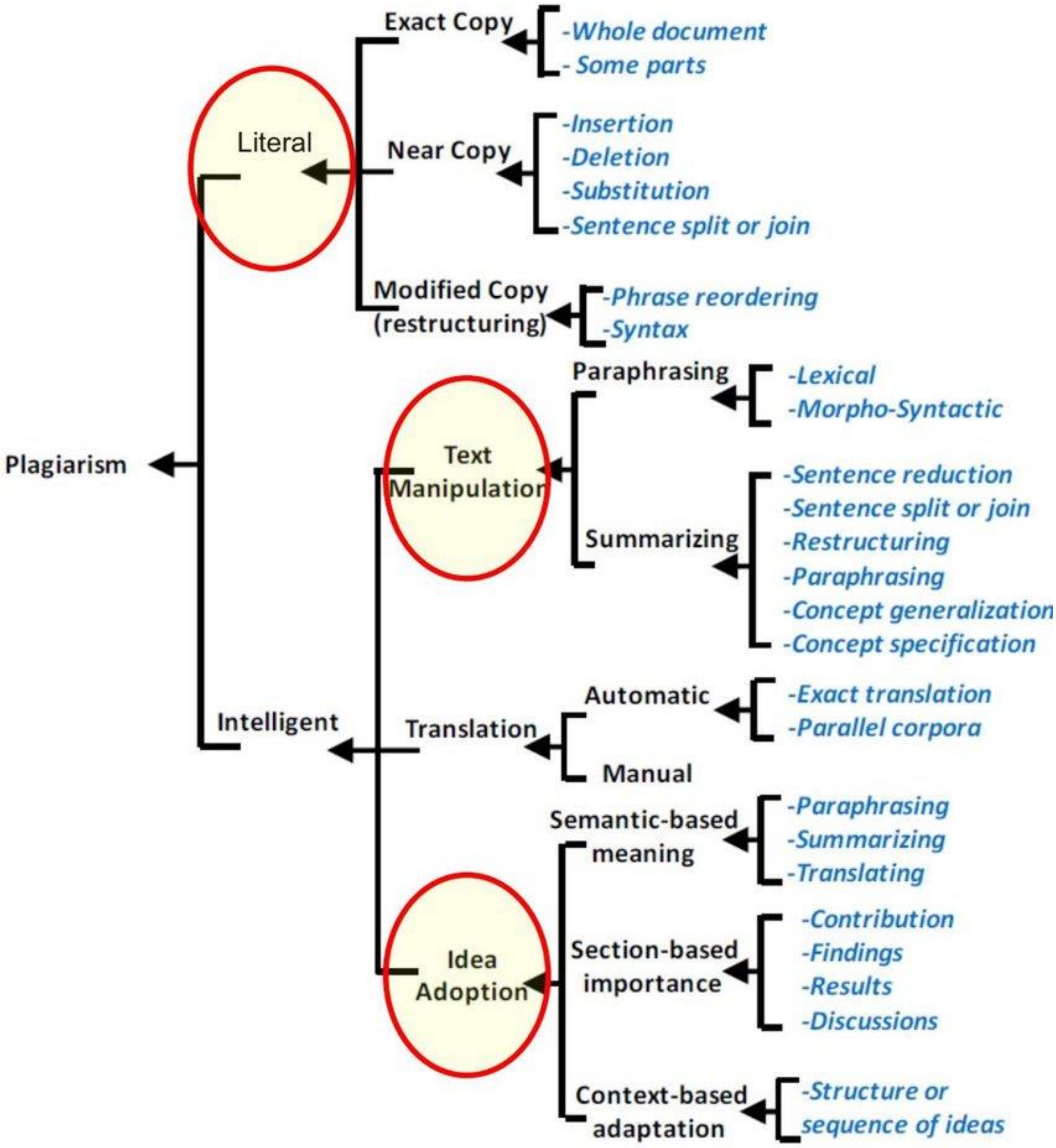


Figure 2: Types of Plagiarism and examples (Wakil et al., 2017:p.66)

Literal plagiarism has become a common practice among plagiarists which the plagiarized document is slightly different from the original text such as copying and pasting. In such a case, direct quotation is required around the content borrowed according to the academic law (Alzahrani *et al.*, 2012).

Intelligent plagiarism is where the plagiarist is trying to alter the original work with different methods such as translation, paraphrasing, summarizing, a combination of sentences and restructuring, etc. The mentioned ways are a form of plagiarism until they are properly cited (Alzahrani *et al.*, 2012).

1.2.1.2 Extrinsic Plagiarism Detection

The sources can be online or offline and the preprocessing of both documents, suspicious documents and sources should be done initially. When the resources are available offline, preprocessing a limited number of documents is not a complex procedure. But when the reference corpus is online, the initial preprocessing of a huge volume of sources sounds tedious.

A query processing technique is used currently to detect the plagiarized content extrinsically. This technique works as a search engine and provides the results for the requested query by comparing the sources and suspicious documents with similarity measures (Kanjirangat, 2016).

In extrinsic plagiarism detection, usually, the document is atomized into passages and then the passages are interpreted as a set of integers to process them finally before comparing the suspected documents with the reference corpus. Hence, high similarity values represent a high confidence value of the availability of plagiarism.

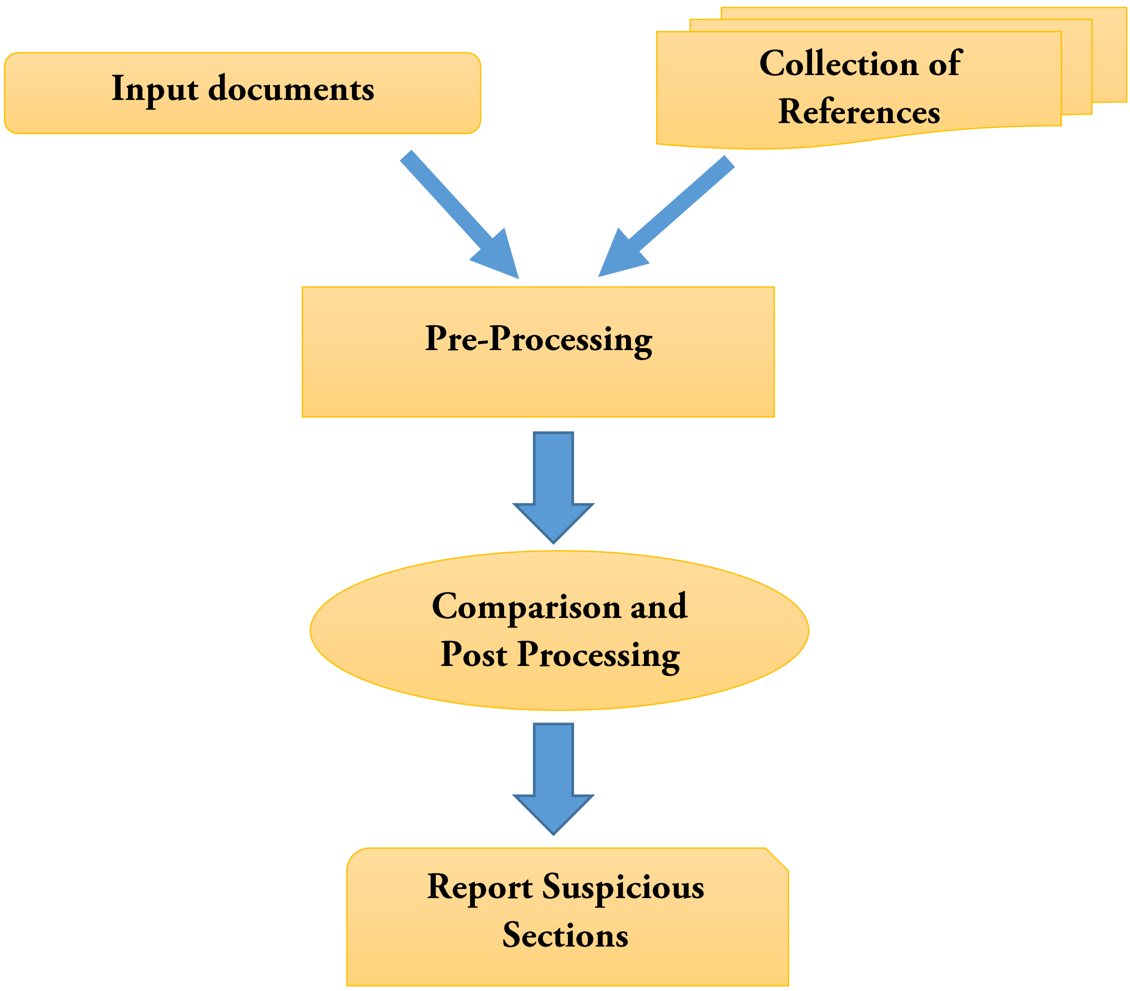


Figure 3: Extrinsic Plagiarism Detection

1.2.1.3 Intrinsic Plagiarism Detection

Intrinsic plagiarism detection refers to the idea that identifying the plagiarized content when no reference corpus is available and identifying techniques should be analyzed by the document itself to detect the plagiarism. This concept is closely related to authorship attribution which identifies written segments in a text by various authors. According to figure 4, the conventional method is to segment the document into passages, and then the features are extracted to make them classified as intrinsic plagiarism.

Intrinsic plagiarism detection has become a special interest in educational institutions as the traditional methods of plagiarism detection are using the document to document analysis. But the source of the documents is not always possible for these instances. Therefore text analysis can be done within the document to identify the deviation in writing styles. Hence the intrinsic plagiarism approach does not need any comparison with reference corpus and it is only depends on the words and the punctuations.

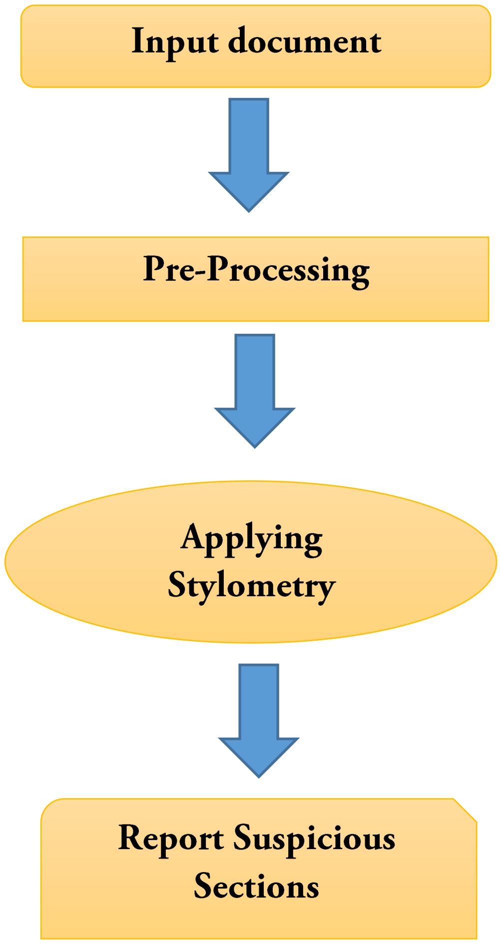


Figure 4: Intrinsic Plagiarism Detection

1.3 Research Aims and Objectives

- To find paragraphs or sentences within a document that appear to remain considerably dissimilar from the rest of the document

    - To collect an appropriate dataset for learning a model for detecting plagiarism
    - To annotate the dataset appropriately to enable supervised model building
    - To explore feature representation techniques and appropriate machine learning algorithms for detecting plagiarism
    - To evaluate the performance of the best algorithm and perform an error analysis on the results

1.4 Scope of the Study

The purpose of the researcher in the study is intrinsic plagiarism detection of a document that does not need a collection of reference documents to compare with the doubtful document. The scope of the study is limited to the English language and the number of authors are limited, owing to time constraints. English language sources are used by the researcher to analyze the writing style and hence writing styles three authors are analyzed in the study.

1.5 Structure of the Dissertation

This dissertation consists of five main chapters. The first chapter is the 'Introduction' and it describes the problem that is going to be addressed. It contains research background, motivation, aim & objectives, achievements, and the structure of the dissertation. The second chapter is the 'Literature Review' which describes the previous studies which are similar to existing works and the methodologies they have followed to carry out the research study. The third chapter is the 'Methodology' which clearly describes the methodologies which were adopted to solve the problem identified. Chapter four is the 'Evaluation'. The evaluation of the methodologies mentioned in the third chapter is described here. The final chapter is 'Conclusion' which describes the overall achievement of the research study, the problems and the limitations encountered, and the future work.

Chapter 2

Literature Review

2.1 Chapter Introduction

An introduction to the research is described in the previous chapter including the description of the types of plagiarism detection which are called extrinsic plagiarism detection and intrinsic plagiarism detection. Moreover, it discussed about the motivation to do the research, goals, and the objectives of the research and about the ultimate achievements. This chapter is related to the previous findings on intrinsic plagiarism detection and the approaches that have been used in various researches.

2.2 Intrinsic Plagiarism Detection

This section describes the related research projects and works similar to the author's research work of proposing a novel method to analyze the writing styles through text analytics.

The researchers (Polydouri et al., 2020) have done a study on intrinsic plagiarism detection and the method used was a machine learning approach under the category of supervised learning. An imbalanced dataset has been used for the purpose while engaging with the stylistic features. The sliding window method was used to document segmentation and it is different from the standard method of fixed window length and step size. Because the researchers have considered about three-level scale values. At last, the paper states that the experiment is not better than the ones of the standard method. Finally has achieved the best F-score of 0.42 for the PAN 2009 corpus and an F-score of 0.37 for the PAN 2011 corpus. Apart from these results, the researchers have used the data balancing technique called SMOTE (Synthetic Minority Oversampling Technique) technique to convert the imbalanced dataset to a balanced dataset. Their aim of using a data balancing technique was for good classification results.

According to the paper by researchers AlSallal and others, they have combined several techniques in their study to detect intrinsic plagiarism. A model is built with the help of statistical features of the common words used in the documents mostly after the extraction of them using the latent semantic analysis. Support Vector Machine (SVM), Random Forest (RF), Bayesian Network (BN) and Multi-layer Perceptron neural network (MLP) have been used and they have been trained as the classification algorithms. The study has achieved a 97% of prediction accuracy in terms of predicting author classes (AlSallal et al., 2019).

According to the researchers (Kuznetsov *et al.*, 2016), they have investigated a method for intrinsic plagiarism detection. The study also focuses on author diarization. They have developed it based on features of text sentences that construct an author style function in addition to outlier detection. The method consists of sentence splitting, vectorizing them, classification model training, finding outliers, etc. The model developed has achieved a 0.2 value of f1 measure for intrinsic plagiarism detection.

The job of intrinsic plagiarism detection was identified as recognizing the segments with in a document written by multiple authors by the researchers and the main goal has become to discover deviations in the writing style. Means, identifying the sections of the document written by another person. The study has followed a hybrid approach which combined with a style function generated and outlier detection. The method has achieved 0.686 of f1 value for PAN 16 corpus and 0.646 of f1 value for PAN 17 corpus (Elamine, Mechti, and Belguith, 2017).

A study on intrinsic plagiarism detection conducted by a researcher has trained a binary classifier with different feature sets. Then the performance has been observed for a set of 36 features in suspicious and non-suspicious documents. The mentioned feature set has achieved 0.85 or 85.10% value of f1 score. In addition, the researcher has found that features such as relative entropy and correlation coefficient are the most effective features (Rahman, 2015).

The researchers have conducted a study on two-step cluster-based mechanism for outlier detection in intrinsic plagiarism detection. The Naive Bayes algorithm has been used and the discretization is the procedure that has been followed to improve the performance of the algorithm. The study has used the tf-idf and query language model for the creation of features. The results are outperformed with values FP/FN (False Positive/ False Negative) threshold = 0.05 which have reduced the FP and FN rates. Hence the usage of the Naive Bayes algorithm is a success with the feature discretization based on the two--step clusters (Wijaya, A and Wahono, R. S. 2015).

The researchers (Bensalem, Rosso, and Chikhi, 2019) have done a study on intrinsic plagiarism detection only considering n-grams as an evidence, as the character n-grams has used so far in authorship attribution problems. The study has utilized five large document collections which have been written in English language and Arabic language. The results show that the least frequent n-grams are considerably impacting on the best n-grams frequency class features.

The researcher (Zurini, M, 2015) has researched on stylometric analysis which has led to the identification of authors to check the originality of the works. The writing styles of the authors provide the basis for the study and eight metrics for writing styles are considered. The result has become the best combination of values in terms of metrics. The average length of the words, the average length of the sentences in terms of words, the number of connection words, frequency of symbols, and the cultural affiliation are the lexical characteristics used in the study. The contextual meanings indicator, the weighted indicator of con-textual meanings, the richness of the Type-Token vocabulary, and the semantic richness of the vocabulary are the semantic characteristics used.

The researchers have done a study on the relationship between authorship attribution and different types of features under a variety of conditions. They have found that mostly the features based on the content are appropriate with high diversity datasets such as news, and datasets with less diversity such as movie reviews are more benefited from stylistic features. The proposed model shows highly effective and over-performed results (Sari, Stevenson, and Vlachos, 2018).

The researchers (Bensalem, Rosso, and Chikhi, 2014) propose a supervised classification-based method using a small number of features for the model built to discriminate the plagiarized and the original text fragments. The proposed method will segment each document into fragments and without considering the numerals, the n-gram class document model has been built while representing each segment with vectors.

Further, several classification algorithms have been used for training and testing in Weka software and different combinations of n-gram lengths has been provided. The Naïve Bayes algorithm has come up with the best results. The experiment has been conducted on three corpora which have had the documents in English Language and Arabic language. The method represents the best configuration of n-grams length as six (6) and the number of classes or features as four (4).

The researchers (Bensalem, Rosso, and Chikhi, 2014) introduce a language-independent intrinsic plagiarism detection technique which uses a text representation method called n-gram classes. According to the researchers, even though most of the intrinsic plagiarism detection approaches are analyzing the documents as a whole, it is crucial to analyze the writing styles of a document at the fragments level. Furthermore, the paper suggests the difficulties that occur in intrinsic plagiarism detection techniques such as multi-author related problems when a number of authors are there for the suspected document. The difficulty level increases when examined text and the potential author text are merged in the document with unknown boundaries. Moreover, fragmentation of a text is inevitable in reliable intrinsic plagiarism detection scenarios as coarse segmentation may lead to the prevention of identifying the short plagiarized text, and same time granular segmentation may cause undependable style analysis. Due to the mentioned difficulties, detecting intrinsic plagiarized content has become challenging.

The researchers (Oberreuter and Velasquez, 2013) have conducted a study and the main goal was to identify the deviations in writing style. The outliers are identified when the writing style get changes. A classification approach with self-based information is used and the ultimate results are low in precision (0.3). The model seems still unreliable and cannot be used for the corpora with less content.

Extrinsic plagiarism detection and intrinsic plagiarism detection are the two forms of plagiarism uncovering methodologies. The current literature has known classified documents which are the basis of the extrinsic plagiarism detection that use to compare the doubtful document (Alzahrani *et al.*, 2012). This plagiarism detection method performs at a good level when the copy and paste have been done as the detection is on the assumption of all related information is digitalized. Therefore, the assumption is always criticized and the topic of intrinsic plagiarism brings a new class for the theme (Meyer Zu Eissen, Stein and Kulig, 2007). Another study has mentioned that identifying the text author is a significant challenge, and also their recommendation is to develop approaches which can be used to analyze the stylistic variations and increase the performance of the current plagiarism detection techniques (Kakkonen and Mozgovoy, 2010).

The researchers spent their time to find about the authorship of some important documents and they wanted to find the most reliable ways. Moreover, the authorship wasn't agreed with the actual authors. For example, some of the corpora that pertained to Shakespeare was doubted to whether owned by Marlow (Zhao and Zobel, 2007). The features of the authorship can be revealed through the model guaranteed by intrinsic plagiarism detection methods. Usually, a small dataset is used for this purpose. The elementary intrinsic plagiarism recognition techniques are grounded on the assumption that the author is having a unique writing style and that writing style is invariant or not changing over time (Luyckx *et al.*, 2008). Mendenhell (1887) studied the works of Shakespeare and Marlow based on the defined assumptions. Mendenhell found that the plots drawing between the frequencies vs. word length for a particular author can discover the writing style uniqueness and invariant characteristic. Then these two features were the base for almost all the statistical approaches to identify the writing styles. Later the researchers were researching to find these unique features of the authors which are invariant over time (Liu, 2013).

Each author is having an individual writing style according to the fact-based on the stylometric features. An author consciously or subconsciously constructs patterns in the sentences as well as they have an individual vocabulary (Meyer Zu Eissen, Stein and Kulig, 2007).

The style changes in the writing style can be identified by studying stylometric features. The text segmentation algorithms are helpful in this and it helps to identify the author variations in the documents. The researchers have done an experiment in a small dataset of articles written by varying numbers of authors. The ultimate results of their study show that when there are more authors for an article, there is more potential is existing to identify the author changes (Rexha *et al.*, no date).

The stylometric features can be divided into the categories as lexical features, semantic features, and syntactic features. Basically, the frequency of words, words n-grams frequency, lexical errors, etc. are under the category of lexical features, and computational tools such as tokenizers, special dictionaries are the required tools. Part of speech, chunks and syntactic errors some of the stylometric features under the syntactic features and mainly the POS tagger and tokenizer can be used as the computational tools required. The semantic features are having the categories such as synonyms, hypernyms, and semantic dependencies which require computational tools such as the partial parser, semantic parser, tokenizer, etc (Alzahrani *et al.*, 2012).

Table 1: Lexical features (Alzahrani et al., 2012:p.140)

Lexical Features

Examples	**Required tools and resources**
Token-based: - Average word length - Average sentence length - Average syllables per word	Tokenizer, [Sentence splitter]
Vocabulary richness - Type-token ratio (i.e. total unique vocabulary/total tokens) - Hapax legomena/dislegomena	Tokenization
Frequency of words	Tokenizer, [Stemmer, Lemmatizer]
Frequency of function words	Tokenizer, Special dictionaries
Word n-grams frequency	Tokenizer
Averaged word frequency class	Tokenizer, [Stemmer, Lemmatizer]
Lexical Errors - Spelling errors (e.g. letter omissions and insertions) - Formatting errors (e.g. all caps letters)	Tokenizer, Orthographic spell checker

Table 2: Syntactic features (Alzahrani et al., 2012:p.140)

Syntactic Features	
Examples	**Required tools and resources**
Part-of-speech	Tokenizer, Sentence splitter, POS tagger
Part-of-speech n-gram frequency	
Chunks	Tokenizer, Sentence splitter, [POS tagger]
Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Partial parser
Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Full parser
Syntactic Errors - Sentence fragments - Run-on sentences - Mismatched tense	Tokenizer, Sentence splitter, Syntactic spell checker

Table 3: Semantic features (Alzahrani et al., 2012:p.140)

Semantic Features	
Examples	**Required tools and resources**
Synonyms, hypernyms, etc.	Tokenizer, [POS tagger], Thesaurus
Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Partial parser, Semantic parser
Functional	Tokenizer, Sentence splitter, POS tagger, Thesaurus, Specialized dictionaries

The two researchers Oberreuter and Velasquez (2013), have explored the difficulty of revealing text plagiarism and the solution of detecting the plagiarized content with the help of computer algorithms. According to them, the rise in the number of digitalized documents is increased day by day in huge amounts, hence significant progress in automatic plagiarism detection can be observed.

According to researchers (Bensalem, Rosso, and Chikhi, 2014) intrinsic plagiarism detection is an alternative solution to the situations when there is no digitalized version of the document is available. For example, when an author copied text from another non-digitalized old book or when there is no copying directly, but another author has written the content. E.g. a student asking another student to write on behalf of him. Therefore, the detection of intrinsic plagiarism is possible by analyzing the writing styles within the fragments of the document. The study mentions the following difficulties that come under the detection of intrinsic plagiarism.

- The document may have two or more unknown authors if the document contains plagiarism, which does not have any boundary to the number of authors.
- The plagiarized fragment of a document can be from multiple authors without any boundary.
- Segmentation of the document is a difficult task as granular segmentation brings undependable style analysis and coarse segmentation brings the prevention of short plagiarized texts.

The study is mainly composed of training a classification model.It consists of a less number of features through the supervised method with the n-gram classes. It has the phases of:

- Segmenting the document into fragments
- Building the N-gram class document model
- Representing each segment
- Combining the fragment vectors and label them
- Building the classifier

Further, this study fragment the suspicious documents based on the proportion of character N-gram classes which is a method to discover intrinsic plagiarism.

According to researchers, (Meyer Zu Eissen, Stein, and Kulig, 2007) plagiarism detection is categorized into two segments based on similarity assessment in global and local contexts. Among them, intrinsic plagiarism detection can be achieved through the Stylometry approach by analyzing the writing styles with in the document as in figure 5.

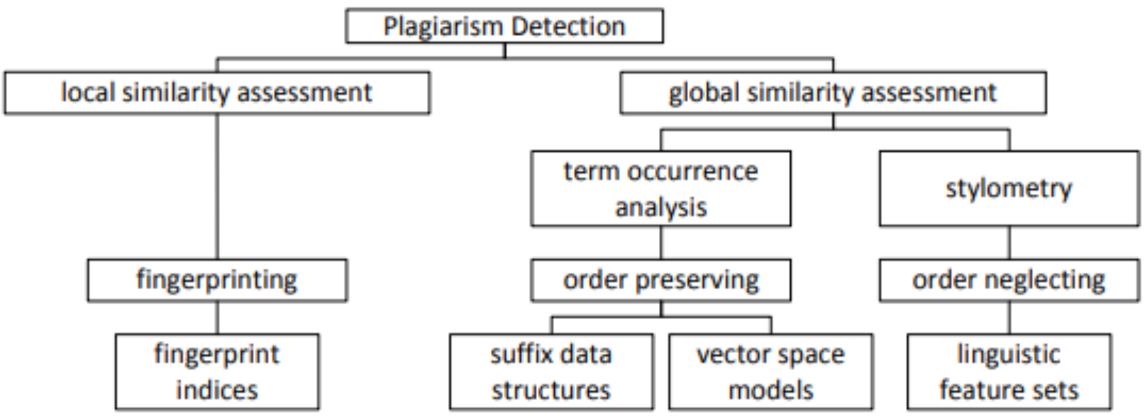


Figure 5: Overview of plagiarism detection (Meyer Zu Eissen, Stein, and Kulig, 2007)

According to Kestemont, Luyckx and Daelemans, 2011, the suspicious document is divided into equal size windows which are consecutive series and may be overlapping. The windows are represented as vectors with relative frequencies of character trigrams. Each of the documents' windows distance matrix is compared with each other window. However, the study was disappointed in terms of precision even though it returned a high recall value and the study approach does not perform well with the short and medium-length plagiarized sections.

The researchers have used several methods for intrinsic plagiarism detection. And they have been done using machine learning and deep learning approaches mainly. Latent semantic analysis and using N-gram classes have gained priority among them. In addition, Stylometry techniques and statistical approaches have been followed in the related study. However, eventhough a few methods show high performance measures, most of the measures show less precision and recall values which express the immature and unreliable nature of the approaches. Moreover, some of the approaches do not perform well with the short and medium-length plagiarized sections even though the precision and recall values are high. Specially, the approach with stylistic features using one-class algorithm is not used as an approach so far by the researchers and hence, the new approach is used in this research study.

Chapter 3

Methodology

3.1 Problem Domain

Intrinsic plagiarism detection is a way to analyze the suspicious document without any collection of references and identifying the plagiarized content by comparing the writing style variations in a single document. A number of methods have been introduced to detect intrinsic plagiarism by many researchers and the researcher in this study proposes a textual approach. The aim is to analyze the writing style of the document written by a writer and to detect intrinsic plagiarism.

Authors are having their own writing styles which makes their literature unique. Based on that fact, there is a possibility to recognize the writers from their writings, and various kinds of techniques can be used to verify the authors. Hence the researcher gets the usage of natural language processing in the initial steps to process the texts and later on the support of the machine learning to verify the author.

3.2 Methodology

Detection of intrinsic plagiarism is a task with several stages. The task can be performed initially by treating the research problem as an anomaly detection problem. Anomalies are also known as outliers and these outliers can be defined as the examples which do not fit with the rest of the data. This outlier detection or anomaly detection is a sub component of machine learning which is focused on one-class classification (OCC). The unsupervised learning algorithms can be used to model the examples given as either normal or abnormal.

In anomaly detection related to this study, needs to train the machine to identify a single author initially, and thereafter it can be extended for multiple authors. One-Class Support Vector Machine (One-class SVM) algorithm is used in the study with stylistic features and the model is built. In addition, Naive Bayes classifier, Logistic Regression classifier and SVM classifier are used for the bag-of-words in the data source. The methodology is shown in the figure 6 as below with the steps involving with the process. The methodology starts with document selection and text preprocessing needs to be done prior to feature extraction step. The best features are selected among the all features and the model is built with one-class svm classifier. The validation accuracies and other performance measures such as precision, recall and f1 score are evaluated and they are compared with the results of the other algorithms.

3.2.1 Selection of Documents

3.2.1.1 Document one

The study deals with text documents which has considerable number of text paragraphs within it. The model that is to be built is, first trained with a single author and the idea is to perform an anomaly detection as an initiative. A lengthy document is used and once the model is prepared, it can be used to classify new examples as either normal or anomaly.

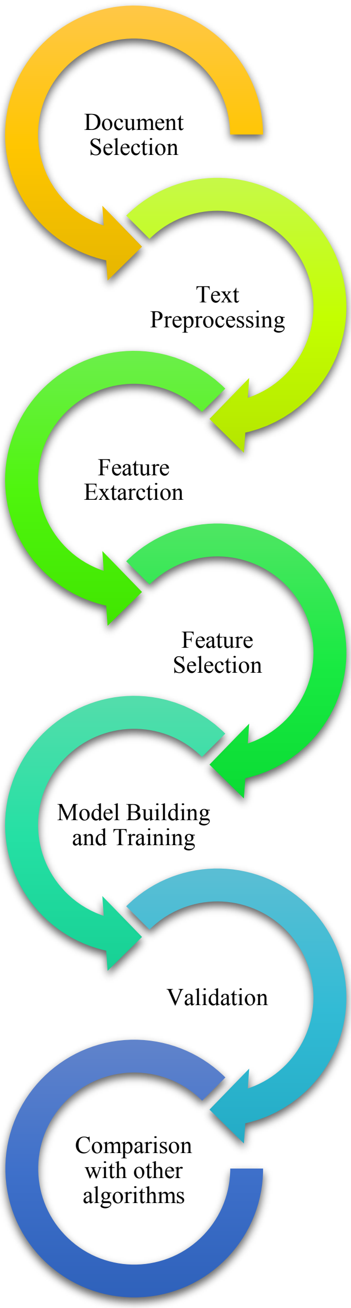
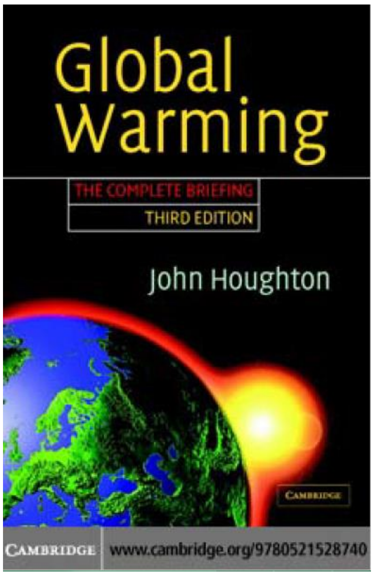
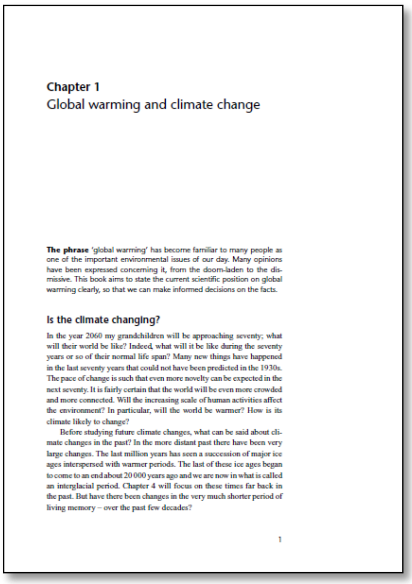


Figure 6: Methodology of the study

The sources taken from the authors were in the form of portable drive format (pdf) and they had to be converted into text format. While converting it is assured that the content is not changed and it is exact to the original document. The researcher selected the book "Global Warming" by John Houghton which was the third edition published in the year 2004 as the document one. The book has been published in the United States of America and the press was the Cambridge University Press, New York.





Chapter 1
Global warming and climate change

The phrase 'global warming' has become familiar to many people as one of the important environmental issues of our day. Many opinions have been expressed concerning it, from the doom-laden to the dismissive. This book aims to state the current scientific position on global warming clearly, so that we can make informed decisions on the facts.

Is the climate changing?

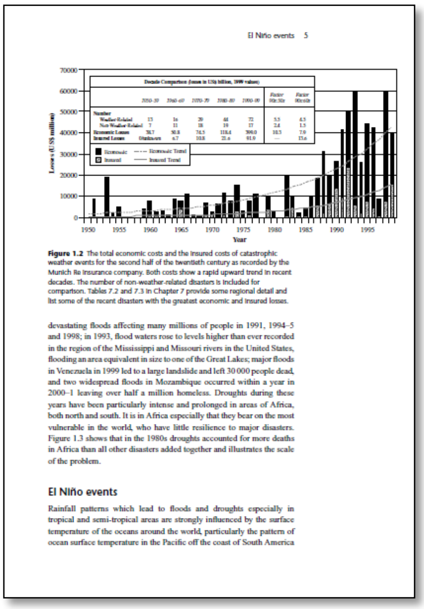
In the year 2060 my grandchildren will be approaching seventy; what will their world be like? Indeed, what will it be like during the seventy years or so of their normal life span? Many new things have happened in the last seventy years that could not have been predicted in the 1930s. The pace of change is such that even more novelty can be expected in the next seventy. It is fairly certain that the world will be even more crowded and more connected. Will the increasing scale of human activities affect the environment? In particular, will the world be warmer? How is its climate likely to change?

Before studying future climate changes, what can be said about climate changes in the past? In the more distant past there have been very large changes. The last million years has seen a succession of major ice ages interspersed with warmer periods. The last of these ice ages began to come to an end about 20 000 years ago and we are now in what is called an interglacial period. Chapter 4 will focus on these times far back in the past. But have there been changes in the very much shorter period of living memory – over the past few decades?

1

Figure 7: Global Warming Book by John Houghton I

The book one selected for the one-class classification is consisted with texts, titles and headings, figures and figure captions, tables and table captions, punctuations, special characters, etc.



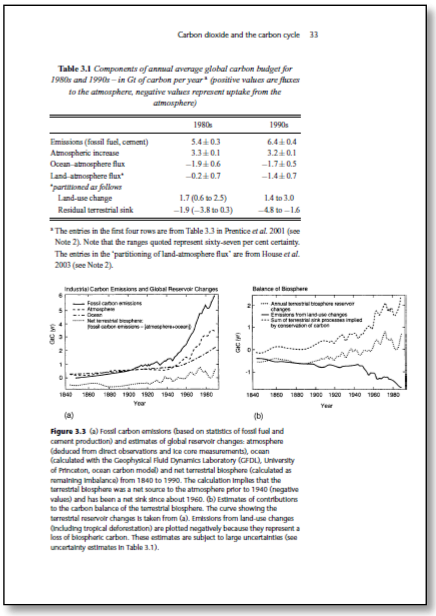
El Niño events 5

Figure 1.2 The total economic costs and the insured costs of catastrophic weather events for the second half of the twentieth century as recorded by the Munich Re insurance company. Both costs show a rapid upward trend in recent decades. The number of non-weather-related disasters is included for comparison. Tables 7.2 and 7.3 in Chapter 7 provide some regional detail and list some of the recent disasters with the greatest economic and insured losses.

devastating floods affecting many millions of people in 1991, 1994–5 and 1998; in 1993, flood waters rose to levels higher than ever recorded in the region of the Mississippi and Missouri rivers in the United States, flooding an area equivalent in size to one of the Great Lakes; major floods in Venezuela in 1999 led to a large landslide and left 30 000 people dead, and two widespread floods in Mozambique occurred within a year in 2000–1 leaving over half a million homeless. Droughts during these years have been particularly intense and prolonged in areas of Africa, both north and south. It is in Africa especially that they bear on the most vulnerable in the world, who have little resilience to major disasters. Figure 1.3 shows that in the 1980s droughts accounted for more deaths in Africa than all other disasters added together and illustrates the scale of the problem.

El Niño events

Rainfall patterns which lead to floods and droughts especially in tropical and semi-tropical areas are strongly influenced by the surface temperature of the oceans around the world, particularly the pattern of ocean surface temperature in the Pacific off the coast of South America



Carbon dioxide and the carbon cycle 33

Table 3.1 *Components of annual average global carbon budget for 1980s and 1990s – in Gt of carbon per year [a] (positive values are fluxes to the atmosphere, negative values represent uptake from the atmosphere)*

	1980s	1990s
Emissions (fossil fuel, cement)	5.4 ± 0.3	6.4 ± 0.4
Atmospheric increase	3.3 ± 0.1	3.2 ± 0.1
Ocean–atmosphere flux	–1.9 ± 0.6	–1.7 ± 0.5
Land–atmosphere flux*	–0.2 ± 0.7	–1.4 ± 0.7
**partitioned as follows*		
Land-use change	1.7 (0.6 to 2.5)	1.4 to 3.0
Residual terrestrial sink	–1.9 (–3.8 to 0.3)	–4.8 to –1.6

[a] The entries in the first four rows are from Table 3.3 in Prentice *et al.* 2001 (see Note 2). Note that the ranges quoted represent sixty-seven per cent certainty. The entries in the 'partitioning of land-atmosphere flux' are from House *et al.* 2003 (see Note 2).

(a) (b)

Figure 3.3 (a) Fossil carbon emissions (based on statistics of fossil fuel and cement production) and estimates of global reservoir changes: atmosphere (deduced from direct observations and ice core measurements), ocean (calculated with the Geophysical Fluid Dynamics Laboratory (GFDL), University of Princeton, ocean carbon model) and net terrestrial biosphere (calculated as remaining imbalance) from 1840 to 1990. The calculation implies that the terrestrial biosphere was a net source to the atmosphere prior to 1940 (negative values) and has been a net sink since about 1960. (b) Estimates of contributions to the carbon balance of the terrestrial biosphere. The curve showing the terrestrial reservoir changes is taken from (a). Emissions from land-use changes (including tropical deforestation) are plotted negatively because they represent a loss of biospheric carbon. These estimates are subject to large uncertainties (see uncertainty estimates in Table 3.1).

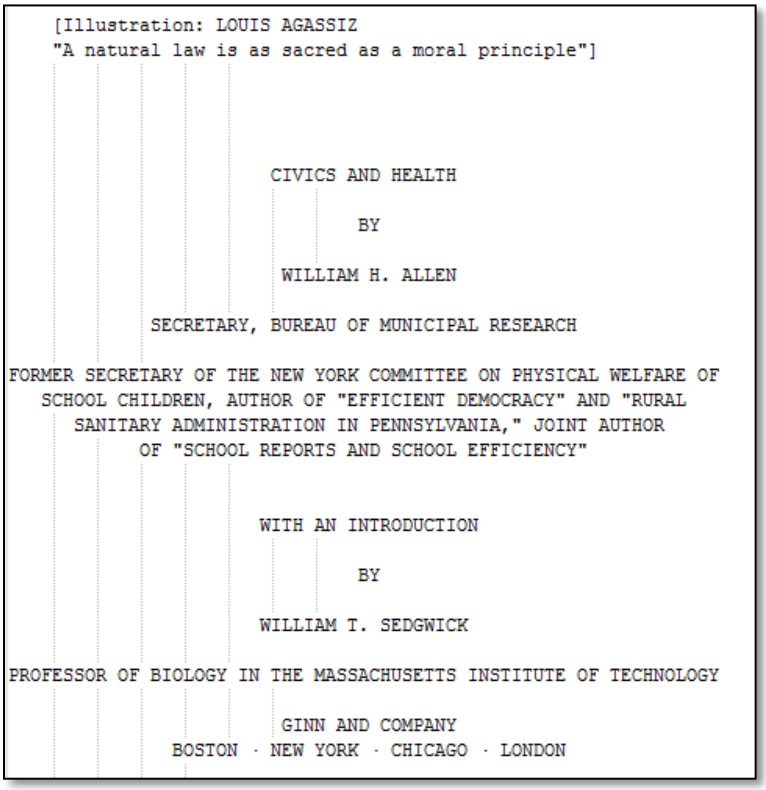
Figure 8: Global Warming Book by John Houghton II

After the conversion of the documents to the text format, the researcher is able to do the preprocessing.

The document one is prepared with text book “Global Warming” by John Houghton as mentioned in the previous section and after preprocessing, the content was 845 total paragraphs. An assumption made in the study is that the majority of the documents is written by one author. Thereafter, paragraphs from another book written by a separate author are added randomly with the help of a systematic random number generation method to implement the intrinsic plagiarism detection concept. Two random numbers are generated and one number among them is used for after how many paragraphs, the foreign paragraph/s should be inserted. Other generated random number is used to decide the number of paragraphs that should be inserted from the foreign document at once.

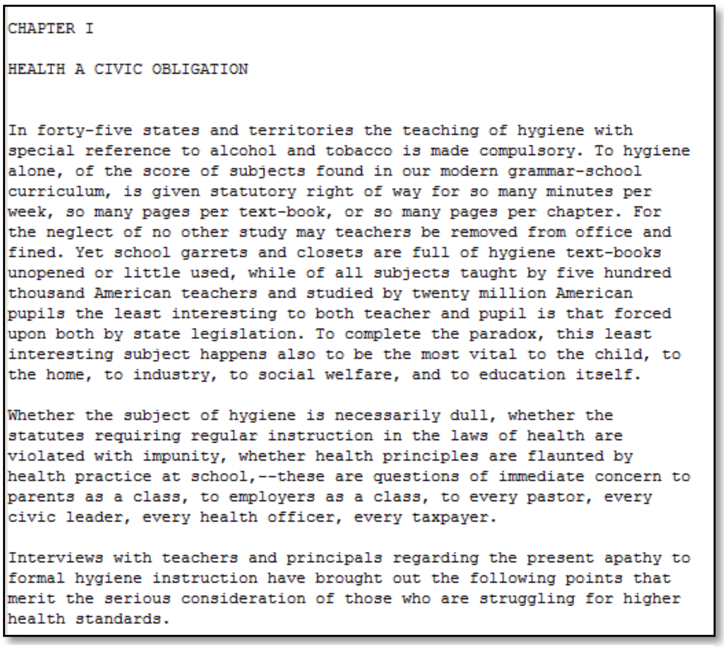
3.2.1.2 Document two

The topic of the document two taken for the research study is “Civics and Health” by the author William H. Allen. It was taken under the license of the Project Gutenberg. The initial number of lines of the downloaded text document is 14,267 and the initial length was 774,936.



```
    [Illustration: LOUIS AGASSIZ
    "A natural law is as sacred as a moral principle"]

                       CIVICS AND HEALTH

                               BY

                        WILLIAM H. ALLEN

              SECRETARY, BUREAU OF MUNICIPAL RESEARCH

FORMER SECRETARY OF THE NEW YORK COMMITTEE ON PHYSICAL WELFARE OF
   SCHOOL CHILDREN, AUTHOR OF "EFFICIENT DEMOCRACY" AND "RURAL
      SANITARY ADMINISTRATION IN PENNSYLVANIA," JOINT AUTHOR
            OF "SCHOOL REPORTS AND SCHOOL EFFICIENCY"

                      WITH AN INTRODUCTION

                               BY

                      WILLIAM T. SEDGWICK

PROFESSOR OF BIOLOGY IN THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

                        GINN AND COMPANY
                BOSTON · NEW YORK · CHICAGO · LONDON
```

Figure 9: Civics and Health - Document Two I



CHAPTER I

HEALTH A CIVIC OBLIGATION

In forty-five states and territories the teaching of hygiene with special reference to alcohol and tobacco is made compulsory. To hygiene alone, of the score of subjects found in our modern grammar-school curriculum, is given statutory right of way for so many minutes per week, so many pages per text-book, or so many pages per chapter. For the neglect of no other study may teachers be removed from office and fined. Yet school garrets and closets are full of hygiene text-books unopened or little used, while of all subjects taught by five hundred thousand American teachers and studied by twenty million American pupils the least interesting to both teacher and pupil is that forced upon both by state legislation. To complete the paradox, this least interesting subject happens also to be the most vital to the child, to the home, to industry, to social welfare, and to education itself.

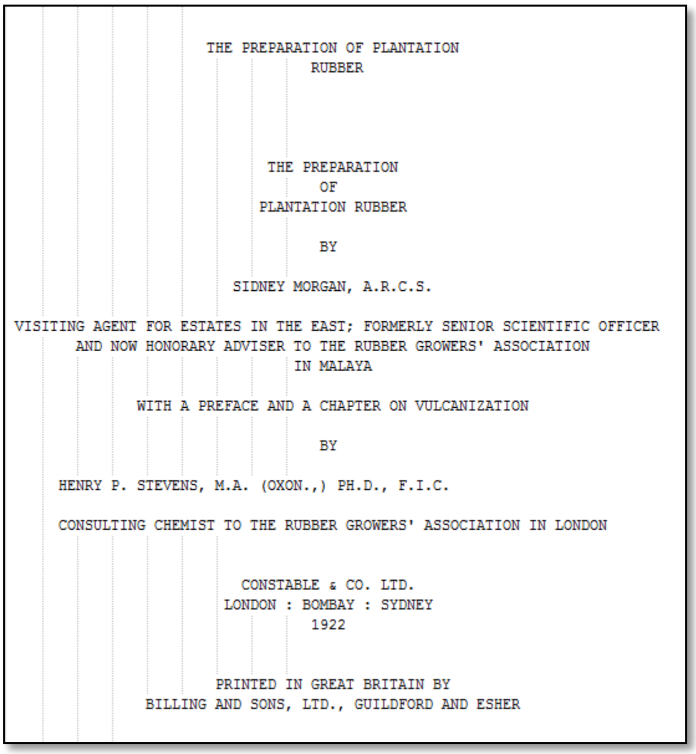
Whether the subject of hygiene is necessarily dull, whether the statutes requiring regular instruction in the laws of health are violated with impunity, whether health principles are flaunted by health practice at school,--these are questions of immediate concern to parents as a class, to employers as a class, to every pastor, every civic leader, every health officer, every taxpayer.

Interviews with teachers and principals regarding the present apathy to formal hygiene instruction have brought out the following points that merit the serious consideration of those who are struggling for higher health standards.

Figure 10: Civics and Health - Document Two II

3.2.1.3 Document three

The topic of the document three taken for the research study is "The Preparation of Plantation Rubber" by the author William H. Allen. It was taken under the license of the Project Gutenberg. The initial number of lines of the downloaded text document is 12,345 and the initial length was 673,744.



THE PREPARATION OF PLANTATION
RUBBER

THE PREPARATION
OF
PLANTATION RUBBER

BY

SIDNEY MORGAN, A.R.C.S.

VISITING AGENT FOR ESTATES IN THE EAST; FORMERLY SENIOR SCIENTIFIC OFFICER
AND NOW HONORARY ADVISER TO THE RUBBER GROWERS' ASSOCIATION
IN MALAYA

WITH A PREFACE AND A CHAPTER ON VULCANIZATION

BY

HENRY P. STEVENS, M.A. (OXON.,) PH.D., F.I.C.

CONSULTING CHEMIST TO THE RUBBER GROWERS' ASSOCIATION IN LONDON

CONSTABLE & CO. LTD.
LONDON : BOMBAY : SYDNEY
1922

PRINTED IN GREAT BRITAIN BY
BILLING AND SONS, LTD., GUILDFORD AND ESHER

Figure 11: The Preparation of Plantation Rubber - Document Three I

CHAPTER I

PLANTING

To criticise the methods of the pioneer planters of _Hevea Brasiliensis_ presents no difficulty in the light of present comparative knowledge, and to be "wise after the event" is a failing which is not confined to those interested in modern planting methods. Looking at the matter broadly, however, it must be acknowledged that the pioneers, wrong though they may have been on some points, did remarkably well, considering that there existed no real knowledge on the subject and that the methods employed were perforce of an empirical nature. Although we know a little more concerning the scientific aspects of rubber planting, the sum total of that knowledge does not justify any drastic criticism of the methods employed by our predecessors. In fact, although we may be of opinion that on general lines there is little now to be learned regarding the planting of _Hevea Brasiliensis_, our present knowledge does not preclude the possibility that future investigations may bring against us charges similar to those sometimes levelled at the earlier planters.

The main theme of the present volume is that of the preparation of rubber for the market. Hence it is not proposed to deal in detail with the work attaching to the opening and development of rubber estates. For this the reader is referred to the literature dealing specifically with rubber planting. Certain points in connection with planting may advantageously be treated in a general way according to modern knowledge, and of these it is proposed to discuss a few in the following pages.

[Illustration: SEEDS, SHOWING VARIABLE SIZE, SHAPE, AND MARKING.]

Figure 12: The Preparation of Plantation Rubber - Document Three II

3.2.2 Programming Environment

The study uses Python language for the programming purose and the Spyder 3.0 version is used in the Anaconda environment. Hence the libraries are needed for the operations in the process. NLTK or the natural language tool kit is a suite of libraries which can be specially used for text processing such as stemming, tokenization, tagging, etc. Numpy, Pandas and Scikit learn also are libraries needed for the study. Numpy library is used for array processing and it provides high performance in scientific computing while the Pandas library uses for the data analysis in python. Numpy arrays can be easily converted to data frames using the Pandas library functions. Module for the regular expressions is imported for functions related with strings such as search function comes with a regular expression.

3.2.3 Text Preprocessing

The selected documents needs to be preprocess in order to obtain a document with text paragraphs only. Hence following are the preprocessing steps followed:

- **Remove figures and figure captions**

  The figures and figure captions in the document are manually removed by the researcher.

- **Remove tables and table captions**

The tables and table captions in the document are manually removed by the researcher.

- **Remove page formatting**

  The page formatting such as page numbers, headers and the footers exist in the document are removed from the document manually.

- **Remove the line breaks generated from pdf to text conversion**

  The line breaks are generated in the process of converting a .pdf document into .txt document and hence they are removed and make them as paragraphs manually.

- **Convert into lower case**

  All the texts in the document are converted into lowercase.

- **Remove stop words**

  There are common words can be seen abundantly in any language. They provide low level information to the predictions. The examples for this kind of stop words are like conjunctions, prepositions, articles, etc. in the English language such as ‘a’, ‘an’, ‘the’, ‘as’, etc.

- **Remove numbers**

  The numbers in the document are removing

- **Stemming**

  Stemming is known as reduce the words into their stems and the texts in the document need to be stemmed.

- **Lemmatization**

  Lemmatization makes the words with the use of vocabulary and according to morphological analysis.

- **Tokenization**

  In the process of machine learning, the texts have to be converted into numbers as the machine learning algorithms take numbers as inputs. Breaking the texts into elements such as words, phrases, sentences, etc. is known as tokenization and it should be perform in order to identify the features of the text for further steps.

There are a number of preprocessing steps involved in the text analytics, but this study will not focus on most of them without few preprocessing steps as preprocessing more and more will reduce the chance of training the machine learning model with the writing styles of authors.

For example, the punctuations are not removed from the text due to the ability to identify the features with the number of separate punctuations used within a document such as number of commas, number of stops, number of apostrophes, etc.

The study has three documents called document one, document two and document three as mentioned previously. The initial documents were prepared by having two columns which one column has the category of the text and the other column has the text. As of that, figure 13, figure 15 and figure 17 show the initial representation of the documents before preprocessing was done.

3.2.3.1 Document One – before preprocessing

The document one was prepared with the book "Global warming" by John Houghton. The book which was available online was converted to the text format and put into the .CSV format for the coding purposes. The document 'Combined1.csv' has all the text paragraphs taken from the particular book and in addition, the text paragraphs taken from web and research papers related to the topic global warming. The text paragraphs taken from the "Global Warming" book were named as "Author", the text paragraphs taken from web were named as "Web" and the text paragraphs taken from research papers were named as "RP" in the first column of the .csv file. The figure 13 shows a portion form the document as follows.

Author	Many recognise this lack of will to act as a 'spiritual' problem (using the word spiritual in a general sense), meaning that we are too obsessed with the 'material' and the immediate and fail to act according
Author	Those with religious belief tend to emphasise the importance of coupling together the relationship of humans to the environment to the relationship of humans to God.*? It is here, religious believers wo
Author	One of the main messages of this chapter is that action addressing environmental problems depends not only on knowledge about them but on the values we place on the environment and our attitudes t
Web	Some of the most immediate impacts of global warming are beneath the waves. Oceans act as carbon sinks, which means they absorb dissolved carbon dioxide. That's not a bad thing for the atmosphere, b
Web	Corals, in particular, are the canary in a coal mine for climate change in the oceans. Marine scientists have observed alarming levels of coral bleaching, events in which coral expel the symbiotic algae that p
Web	Despite overwhelming scientific consensus about the causes and reality of global warming, the issue is contentious politically. For instance, deniers of climate change have argued that warming slowed be
Author	The perspectives of balance, interdependence and unity in the natural world generated by the underlying science.
Author	A recognition — some would argue suggested by the science — that humans have a special place in the universe, which in turn implies that humans have special responsibilities with respect to the natura
Author	A recognition of the importance of the cultural and religious basis for the principles of stewardship — humans as 'gardeners' of the Earth is a possible 'model' of such stewardship. A recognition that, just a
Web	Unfortunately for the planet, the hiatus never happened. Two studies, one published in the journal Science in 2015 and one published in 2017 in the journal Science Advances, reanalyzed the ocean tempe
Author	I shall return to the practical outworking of some of these issues in later chapters especially Chapter 12. Finally, let me recall some words of Thomas Huxley, an eminent biologist from last century, who en
Author	In the next chapter we shall reflect on the uncertainties associated with the science of global warming and consider how they can be taken into account in addressing the imperative for action. For instanc
Author	This book is intended to present clearly the current scientific position on global warming. A key part of this presentation concerns the uncertainty associated with all parts of the scientific description, espe
RP	The last decade has witnessed increasing interest in possible connections between historical global warming and individual extreme climate events (1–9). This interest is grounded in both scientific and pr
RP	Effective management of climate-related risks therefore requires robust quantification of the probability of extremes in the current and future climate (10). For example, quantification of risk and liability
Author	Before considering the 'weighing' process and the cost of action, we begin by explaining the nature of the scientific uncertainty and how it has been addressed by the scientific community.
Author	In earlier chapters I explained in some detail the science underlying the problem of global warming and the scientific methods that are employed for the prediction of climate change due to the increases
Author	However, the situation is complicated by feedbacks and regional variations. Numerical models run on computers are the best tools available for addressing these problems. Although they are highly comp
RP	Although the tails of climate distributions have been analyzed for many years (e.g., ref. 18), quantifying the contribution of historical warming to unprecedented events presents an imposing scientific cha
RP	Some methods have matured to the point that "rapid" analyses are now being undertaken (e.g., ref. 39), creating a pathway to operationalize single-event attribution (5, 40). Approaches to evaluate opera
RP	We find that 79% of the observed area exhibits a statistically significant trend in peak summer monthly temperature (Table 1 and Fig. S1). The trend has increased the severity and probability of the maxim
Author	However, model limitations remain, which give rise to uncertainty (see box below). The predictions presented in Chapter 6 reflected these uncertainties, the largest of which are due to the models' failur
Author	With uncertainty in the basic science of climate change and in the predictions of future climate, especially on the regional scale, there are bound also to be uncertainties in our assessment of the impacts o
Author	The Intergovernmental Panel on Climate Change! has described the scientific uncertainty as follows.
Author	There are many uncertainties in our predictions particularly with regard to the timing, magnitude and regional patterns of climate change, due to our incomplete understanding of:
Author	sources and sinks of greenhouse gases, which affect predictions of future concentrations, clouds, which strongly influence the magnitude of climate change, oceans, which influence the timing and patter

Figure 13: Document One - before preprocessing

3.2.3.2 Document One – after preprocessing

The document one is preprocessed by following several steps such as making them all lower case, removing numbers, removing punctuations, etc. The figure 14 shows the label of the paragraph in the first column, the original text paragraph in the second column and the preprocessed text paragraph in the third column.

Web	We often call the result global warming, but it is causing a set of changes to the Earth's climate, or long-term weather patterns, that varies from place to place. While many people think of global warming and climate change as synonyms, scientists use Ã¬climate changeÃ® when describing the complex shifts now affecting our planetÃ-s weather and climate systemsÃ³in part because some areas actually get cooler in the short term.	often call result global warming causing set change earth climate long term weather pattern varies place place climate change synonym scientist use Ã¬climate changeÃ® describing complex shift affecting planetÃ-s weathe actually get cooler short term
Author	Figure 3.5 shows that these fractions may change substantially in the future.	figure show fraction may change substantially future
Web	Climate change encompasses not only rising average temperatures but also extreme weather events, shifting wildlife populations and habitats, rising seas, and a range of other impacts. All of those changes are emerging as humans continue to add heat-trapping greenhouse gases to the atmosphere, changing the rhythms of climate that all living things have come to rely on.	climate change encompasses rising average temperature also extreme weather event shifting wildlife populat change emerging human continue add heat trapping greenhouse gas atmosphere changing rhythm climate livir
Author	About ninety-five per cent of fossil fuel burning occurs in the northern hemisphere, so there is more carbon dioxide there than in the southern hemisphere. The difference is currently about two parts per million and, over the years, has grown in parallel with fossil fuel emissions, thus adding further compelling evidence that the atmospheric increase in carbon dioxide levels results from these emissions.	ninety five per cent fossil fuel burning occurs northern hemisphere carbon dioxide southern hemisphere differ year grown parallel fossil fuel emission thus adding compelling evidence atmospheric increase carbon dioxide
Author	We turn now to what happens in the oceans. We know that carbon dioxide dissolves in water; carbonated drinks make use of that fact. Carbon dioxide is continually being exchanged with the air above the ocean across the whole ocean surface (about 90 Gt per year is so exchanged Ã³ Figure 3.1), particularly as waves break. An equilibrium is established between the concentration of carbon dioxide dissolved in the surface waters and the concentration in the air above the surface. The chemical laws governing this equilibrium are such that if the atmospheric concentration changes by ten per cent the concentration in solution in the water changes by only one-tenth of this: one per cent.	turn happens ocean know carbon dioxide dissolve water carbonated drink make use fact carbon dioxide contin whole ocean surface per year exchanged figure particularly wave break equilibrium established concentration water concentration air surface chemical law governing equilibrium atmospheric concentration change ten per change one tenth one per cent

Figure 14: Document One - after preprocessing

3.2.3.3 Document Two – before preprocessing

The document two was prepared with the book “Civis and Health” by William H. Allen. The book which was available online was converted to the text format and put into the .CSV format for the coding purposes. The document ‘Combined2.csv’ has all the text paragraphs taken from the particular book and in addition, the text paragraphs taken from web and research papers related to the topic global warming. The text paragraphs taken from the “Civis and Health” book were named as “Author”, the text paragraphs taken from web were named as “Web” and the text paragraphs taken from research papers were named as “RP” in the first column of the .csv file. The figure 15 shows a portion form the document as follows.

Author	Natural law points to a Nature Fore as well as a Nature Back, to a Nature Up and Beyond as well as a Nature Down and Behind. The Nature that was yesterday will not do for to-morrow, any more than a m
Author	But every experiment in turning back exalts the present and the future. Gifts as well as problems are seen to come with complexity, and civilization flatly refuses to relinquish these gifts. Sound maturity
Author	Problems of health and of civics can never be solved by appealing to Nature Back, when only the few could be healthy, when one baby in three died in infancy, when old age was toothless and childish, w
Author	By using numerous tests which have been suggested in preceding chapters we can learn how far we and our communities obey natural law when working and playing. Health for health's sake has nowher
Author	Fashions, tastes, mannerisms, personal indulgences, have been left for Agassiz to deal with. Generally speaking, we all know of numerous acts committed and numerous acts omitted in our daily routine
Author	Last night I went to a dinner party at eight. I ate and ate a great variety of palatable foods that Nature Back never knew. After two hours of eating I imbibed for two hours the tobacco smoke of the gentlen
Author	Nature Back says I should not have gone to this dinner. But I was compelled to go. I know I am going to others. I cannot do my work unless I overdraw my current health account. Nature Fore tells me that
Author	Nature Back demands "dress reform." Nature Fore tells me that I can march in step with my contemporaries without either attracting attention or discrediting and affronting natural law. Passion for the na
Author	Nature Back throws little light upon conditions necessary for modern labor. It can do nothing but demand the abolition of the factory, the big store, the tenement, the school. Nature Fore says we cannot
Web	Civic Health Index (CHI) is at the center of our work. We think of “civic health” as the way that communities are organized to define and address public problems. Communities with strong indicators of civ
Web	CHI partnerships have changed the way governments go about their work, reintroduced civics to our classrooms, redirected investments, influenced national and local conversations resulting in enhancin
Web	NCoC currently works with cross-sector partners in over 30 states and communities to strengthen civic life in America. The Civic Health Initiative uses engaging reports, infographics, fact sheets, and forun
Web	While our civic health research has been conducted annually ever since 2006 on a national level, we quickly realized that we are not the experts on the ground. In order for the data to have the most impa
Web	These partnerships have grown exponentially over the past few years, and we now work in over 30 communities nationwide.
Web	We don’t purport to know all the answers, nor do we assert that we are the best tellers of these local stories. That’s why we partner with organizations throughout the country who can tell the local story
Web	Strategy: Supporting partners through the project development process by supporting fundraising, identifying local stakeholders, developing strategy, helping determine goals, and creating timelines an
Web	Research: Managing the national research partnerships with CNCS, US Census, and our Civic Indicators Working Group to establish survey questions, advocate for the data collection and manage prelimina
Web	Data: Providing our local partners with preliminary findings and ongoing consulting on data analysis, research questions, and narrative.
Web	Design: Leading the report production process from copy editing through layout, design, printing and shipping.
Web	Communications: Supporting our partners through their communications and dissemination efforts by drafting press releases, outreach to the media, advising on and attending launch events, and consul
RP	Education affects mortality. One US study shows that an additional year of study reduces the probability of dying in the next 10 years by 3.6 years; another Swedish study shows that an additional year red
RP	Although precise calculations have to be very tentative, some of these benefits can be costed. A UK study estimates that taking women without qualifications to a Level 2 qualification would lead to a red
RP	The health productivity of learning requires considerably more attention from policy makers. Measurement of education depends too heavily on quantity and qualifications. More emphasis should be pla
RP	Not all learning is good for health! At a collective level education can increase inequalities, with negative health consequences; and can raise stress levels.
RP	While policy makers widely recognise the fact that education serves as an engine for economic growth through the accumulation of human capital, education is also strongly associated with boosting leve
RP	Anyone with even a cursory familiarity with the literature on civic and social engagement may assume that linking education and CSE is an easy task, and can be summarised tidily: education has a univers

Figure 15: Document Two - before preprocessing

3.2.3.4 Document Two – after preprocessing

The document two is preprocessed by following several steps such as making them all lower case, removing numbers, removing punctuations, etc. The figure 16 shows the label of the paragraph in the first column, the original text paragraph in the second column and the preprocessed text paragraph in the third column.

Author	Nature Back says I should not have gone to this dinner. But I was compelled to go. I know I am going to others. I cannot do my work unless I overdraw my current health account. Nature Fore tells me that effective coË†peration with others will frequently require me to eat at the dinner hour of others, to retire at others' sleeping time, to wear what others will approve, to violate natural law. But Nature Fore also tells me how to build up a health reserve so that I can meet these emergencies without endangering my health credit.	nature back say gone dinner compelled know going others cannot work unless overdraw current health ac others retire others sleeping time wear others approve violate natural law nature fore also tell build healt
Author	Nature Back demands "dress reform." Nature Fore tells me that I can march in step with my contemporaries without either attracting attention or discrediting and affronting natural law. Passion for the natural has effected numerous reforms in dress, diet, and social habits, until commerce provides a natural adaptation of practically every fashion. With regard to few things is it necessary to-day for any one who reads magazines to do violence to bodily health for fashion's sake. We may wear what we will, eat what we prefer, decline what is unnatural for us, without inviting censure. The debauches of those unfortunate people who live an unnatural, purposeless existence, affect such a small number that their laws need not be considered here. Natural law makes obedience to itself attractive; hence commerce is rapidly learning to cater to distaste for the unnatural. With few exceptions, only temporary concessions to unnatural living are required in order to dress and act conventionally.	nature back demand dress reform nature fore tell march step contemporary without either attracting atte dress diet social habit commerce provides natural adaptation practically every fashion regard thing necess decline unnatural without inviting censure debauch unfortunate people live unnatural purposeless existe hence commerce rapidly learning cater distaste unnatural exception temporary concession unnatural livir
	Nature Back throws little light upon conditions necessary for modern labor. It can do nothing but demand the abolition of the factory, the big store, the tenement, the school. Nature Fore says we cannot abolish the means of working out the highest forms of coË†peration. But we can make them compatible with natural living. We can modify conditions so that earning a livelihood will not compel workers to violate natural law at any or all times. The greatest need of factory and tenement reform is for parents and teachers to make a religion of Nature Fore and to instill its principles in the minds of children. Parents and teachers must live the natural before they can make children love the natural. Parents and teachers cannot possibly be natural in this day,	nature back throw little light upon condition necessary modern labor nothing demand abolition factory bi coË†peration make compatible natural living modify condition earning livelihood compel worker violate n

Figure 16: Document Two - after preprocessing

3.2.3.5 Document Three – before preprocessing

The document three was prepared with the book “The Preparation of Plantation Rubber” by the author William H. Allen. The book which was available online was converted to the text format and put into the .CSV format for the coding purposes. The document ‘Combined3.csv’ has all the text paragraphs taken from the particular book and in addition, the text paragraphs taken from web and research papers related to the topic global warming. The text paragraphs taken from the “Global Warming” book were named as “Author”, the text paragraphs taken from web were named as “Web” and the text paragraphs taken from research papers were named as “RP” in the first column of the .csv file. The figure 17 shows a portion from the document as below.

Author	Many recognise this lack of will to act as a ‘spiritual’ problem (using the word spiritual in a general sense), meaning that we are too obsessed with the ‘material’ and the immediate and fail to act according
Author	Those with religious belief tend to emphasise the importance of coupling together the relationship of humans to the environment to the relationship of humans to God.*? It is here, religious believers wo
Author	One of the main messages of this chapter is that action addressing environmental problems depends not only on knowledge about them but on the values we place on the environment and our attitudes t
Web	Some of the most immediate impacts of global warming are beneath the waves. Oceans act as carbon sinks, which means they absorb dissolved carbon dioxide. That's not a bad thing for the atmosphere, b
Web	Corals, in particular, are the canary in a coal mine for climate change in the oceans. Marine scientists have observed alarming levels of coral bleaching, events in which coral expel the symbiotic algae that p
Web	Despite overwhelming scientific consensus about the causes and reality of global warming, the issue is contentious politically. For instance, deniers of climate change have argued that warming slowed be
Author	The perspectives of balance, interdependence and unity in the natural world generated by the underlying science.
Author	A recognition — some would argue suggested by the science — that humans have a special place in the universe, which in turn implies that humans have special responsibilities with respect to the natural
Author	A recognition of the importance of the cultural and religious basis for the principles of stewardship — humans as ‘gardeners’ of the Earth is a possible ‘model’ of such stewardship. A recognition that, just a
Web	Unfortunately for the planet, the hiatus never happened. Two studies, one published in the journal Science in 2015 and one published in 2017 in the journal Science Advances, reanalyzed the ocean tempe
Author	I shall return to the practical outworking of some of these issues in later chapters especially Chapter 12. Finally, let me recall some words of Thomas Huxley, an eminent biologist from last century, who em
Author	In the next chapter we shall reflect on the uncertainties associated with the science of global warming and consider how they can be taken into account in addressing the imperative for action. For instance
Author	This book is intended to present clearly the current scientific position on global warming. A key part of this presentation concerns the uncertainty associated with all parts of the scientific description, espe
RP	The last decade has witnessed increasing interest in possible connections between historical global warming and individual extreme climate events (1–9). This interest is grounded in both scientific and pr
RP	Effective management of climate-related risks therefore requires robust quantification of the probability of extremes in the current and future climate (10). For example, quantification of risk and liability
Author	Before considering the ‘weighing’ process and the cost of action, we begin by explaining the nature of the scientific uncertainty and how it has been addressed by the scientific community.
Author	In earlier chapters I explained in some detail the science underlying the problem of global warming and the scientific methods that are employed for the prediction of climate change due to the increases
Author	However, the situation is complicated by feedbacks and regional variations. Numerical models run on computers are the best tools available for addressing these problems. Although they are highly comp
RP	Although the tails of climate distributions have been analyzed for many years (e.g., ref. 18), quantifying the contribution of historical warming to unprecedented events presents an imposing scientific cha
RP	Some methods have matured to the point that “rapid” analyses are now being undertaken (e.g., ref. 39), creating a pathway to operationalize single-event attribution (5, 40). Approaches to evaluate opera
RP	We find that 79% of the observed area exhibits a statistically significant trend in peak summer monthly temperature (Table 1 and Fig. S1). The trend has increased the severity and probability of the maxim
Author	However, model limitations remain, which give rise to uncertainty (see box below). The predictions presented in Chapter 6 reflected these uncertainties, the largest of which are due to the models’ failur
Author	With uncertainty in the basic science of climate change and in the predictions of future climate, especially on the regional scale, there are bound also to be uncertainties in our assessment of the impacts o
Author	The Intergovernmental Panel on Climate Change! has described the scientific uncertainty as follows.
Author	There are many uncertainties in our predictions particularly with regard to the timing, magnitude and regional patterns of climate change, due to our incomplete understanding of:
Author	sources and sinks of greenhouse gases, which affect predictions of future concentrations, clouds, which strongly influence the magnitude of climate change, oceans, which influence the timing and pattern

Figure 17: Document Three - before preprocessing

3.2.3.6 Document Three – after preprocessing

The document three is preprocessed by following several steps such as making them all lower case, removing numbers, removing punctuations, etc. The figure 18 shows the label of the paragraph in the first column, the original text paragraph in the second column and the preprocessed text paragraph in the third column.

Web	We often call the result global warming, but it is causing a set of changes to the Earth's climate, or long-term weather patterns, that varies from place to place. While many people think of global warming and climate change as synonyms, scientists use Ã¬climate changeÃ® when describing the complex shifts now affecting our planetÃ-s weather and climate systemsÃ³in part because some areas actually get cooler in the short term.	often call result global warming causing set change earth climate long term weather pattern varies place place climate change synonym scientist use Ã¬climate changeÃ® describing complex shift affecting planetÃ-s weath actually get cooler short term
Author	Figure 3.5 shows that these fractions may change substantially in the future.	figure show fraction may change substantially future
Web	Climate change encompasses not only rising average temperatures but also extreme weather events, shifting wildlife populations and habitats, rising seas, and a range of other impacts. All of those changes are emerging as humans continue to add heat-trapping greenhouse gases to the atmosphere, changing the rhythms of climate that all living things have come to rely on.	climate change encompasses rising average temperature also extreme weather event shifting wildlife populat change emerging human continue add heat trapping greenhouse gas atmosphere changing rhythm climate livir
Author	About ninety-five per cent of fossil fuel burning occurs in the northern hemisphere, so there is more carbon dioxide there than in the southern hemisphere. The difference is currently about two parts per million and, over the years, has grown in parallel with fossil fuel emissions, thus adding further compelling evidence that the atmospheric increase in carbon dioxide levels results from these emissions.	ninety five per cent fossil fuel burning occurs northern hemisphere carbon dioxide southern hemisphere differ year grown parallel fossil fuel emission thus adding compelling evidence atmospheric increase carbon dioxide
Author	We turn now to what happens in the oceans. We know that carbon dioxide dissolves in water; carbonated drinks make use of that fact. Carbon dioxide is continually being exchanged with the air above the ocean across the whole ocean surface (about 90 Gt per year is so exchanged Ã³ Figure 3.1), particularly as waves break. An equilibrium is established between the concentration of carbon dioxide dissolved in the surface waters and the concentration in the air above the surface. The chemical laws governing this equilibrium are such that if the atmospheric concentration changes by ten per cent the concentration in solution in the water changes by only one-tenth of this: one per cent.	turn happens ocean know carbon dioxide dissolve water carbonated drink make use fact carbon dioxide contin whole ocean surface per year exchanged figure particularly wave break equilibrium established concentration water concentration air surface chemical law governing equilibrium atmospheric concentration change ten per change one tenth one per cent

Figure 18: Document Three - after preprocessing

3.2.4 Feature Extraction

The pipeline extends with the next step of feature extraction and this is also a crucial phase in the process. The features are extracted in order to build the model and hence they should be extracted carefully. The preprocessed documents are used to extract the features and the features used for the study are shown below with the description and the figures are attached.

Table 4: Features extracted and the notations

Feature Extracted	***Notation in .csv document***
Number of sentences per paragraph	#sentences
Number of words per paragraph	Total#words
Average number of words per sentence	Avg#words
Lexical diversity	lexical_diversity
Dots per paragraph	Total#dots
Commas per paragraph	Total#comma
Semicolons per paragraph	Total#semicolon
Colons per paragraph	Total#colon
Exclamation per paragraph	Total#Exclamationmark
Question marks per paragraph	Total#Questionmark
Hyphens per paragraph	Total#Hyphens
% per paragraph	Total#percentage
> per paragraph	Total#lessthan
< per paragraph	Total#greaterthan
Average Dots per paragraph	Avg#dots
Average Commas per paragraph	Avg#comma
Average Semicolons per paragraph	Avg#semicolon
Average Colons per paragraph	Avg#colon
Average Exclamation per paragraph	Avg#Exclamationmark
Average Question marks per paragraph	Avg#Questionmark
Average Hyphens per paragraph	Avg#Hyphens
Average % per paragraph	Avg#percentage
Average > per paragraph	Avg#lessthan
Average < per paragraph	Avg#greaterthan

In addition to the above mentioned lexical and punctuation features, the POS taggers were added in order to increase the accuracy of the features. POS taggers are the annotations for the sentence structures available in the NLTK library. They are helpful to identify the structure of the sentences in a document. The POS taggers are used to recognize the writing styles of the documents in the study. The POS taggers that added for the feature extraction were as follows in the table 5.

Table 5: POS feature notation and description

Notation Used	*Description*
CD	cardinal Digit
JJ	adjective
NN	noun
NNP	proper Noun
NNS	noun Plural
RB	adverb
VBD	verb, past tense
VBG	verb, gerund
VBN	verb, past participle
VBP	verb, present
VBZ	verb, 3rd person
IN	preposition
MD	modal
VB	verb
JJR	adjective, comparative
FW	foreign word
WP$	possessive wh-pronoun
JJS	adjective, superlative
DT	determiner
RBR	adverb, comparative
WDT	wh-determiner
CC	coordinating conjunction
PRP$	possessive pronoun
EX	existential there
WRB	wh-abverb
RP	particle
WP	wh-pronoun
RBS	adverb, superlative
PRP	personal pronoun
POS	possessive ending
UH	interjection

The Python programming language was used for the study and the feature extraction was done as shown in figure 19 after the preprocessing of the documents. The figure shows the extracting of the features: number of sentences per paragraph, number of words per paragraph, average number of words per sentence, lexical diversity, etc.

```
df['lexical_diversity'] =  df['cleaned_text'].apply(lambda x:  lexical_diversity(x)  )

df['Total#dots'] =  df['text'].apply(lambda x: x.count('.'))
df['Total#comma'] =  df['text'].apply(lambda x: x.count(','))
df['Total#semicolon'] =  df['text'].apply(lambda x: x.count(';'))
df['Total#colon'] =  df['text'].apply(lambda x: x.count(':'))
df['Total#Exclamationmark'] =  df['text'].apply(lambda x: x.count('!'))
df['Total#Questionmark'] =  df['text'].apply(lambda x: x.count('?'))
df['Total#Hyphens'] =  df['text'].apply(lambda x: x.count('-'))
df['Total#percentage'] =  df['text'].apply(lambda x: x.count('%'))
df['Total#lessthan'] =  df['text'].apply(lambda x: x.count('<'))
df['Total#greaterthan'] =  df['text'].apply(lambda x: x.count('>'))

df['Avg#dots'] = round(df['Total#dots']/df['#sentences'])
df['Avg#comma'] = round( df['Total#comma']/df['#sentences'])
df['Avg#semicolon'] =  round(df['Total#semicolon']/df['#sentences'])
df['Avg#colon'] =  round(df['Total#colon']/df['#sentences'])
df['Avg#Exclamationmark'] = round( df['Total#Exclamationmark']/df['#sentences'])
df['Avg#Questionmark'] =  round(df['Total#Questionmark']/df['#sentences'])
df['Avg#Hyphens'] = round(df['Total#Hyphens']/df['#sentences'])
df['Avg#percentage'] =  round(df['Total#percentage']/df['#sentences'])
df['Avg#lessthan'] = round( df['Total#lessthan']/df['#sentences'])
df['Avg#greaterthan'] =  round(df['Total#greaterthan']/df['#sentences'])
```

Figure 19: Feature extraction

3.2.4.1 Extracted lexical features and punctuation features

The below figure 20 shows the part of extracted features from the dataset. According to the figure, first few lexical features are the total number of sentences per paragraph, total number of words per paragraph, average number of sentences per paragraph, lexical diversity. The initial punctuation features are total number of dots per paragraph, total number of commas per paragraph, total number of semicolons per paragraph and the total number of colons per paragraph.

1	class	text	cleaned_text	#sentences	Total#words	Avg#words	lexical_diversity	Total#dots	Total#comma	Total#semicolon	Total#colon
2	Author	The phrase Ã«global warmi	phrase Ã«global warmingÃ- b	3	32	11	0.96875	3	2	0	0
3	Author	In the year 2060 my grandcl	year grandchild approaching s	8	47	6	0.787234043	3	2	1	0
4	Author	Before studying future clim	studying future climate chang	6	48	8	0.729166667	4	1	0	0
5	Author	Variations in day-to-day we	variation day day weather occ	10	99	10	0.767676768	10	13	2	0
6	Author	The 1980s and 1990s were u	unusually warm globally spea	3	38	13	0.736842105	3	2	0	0
7	Author	The period has also been re	period also remarkable remar	5	66	13	0.772727273	5	3	1	0
8	Author	But those storms in Europe	storm europe mild compariso	6	115	19	0.782608696	6	5	1	0
9	Author	The increase in storm inter	increase storm intensity recei	9	112	12	0.732142857	9	7	1	0
10	Web	Global warming is the long-	global warming long term hea	3	52	17	0.826923077	3	3	0	0
11	Author	Windstorms or hurricanes a	windstorm hurricane mean w	10	135	14	0.777777778	10	8	3	0
12	Author	Rainfall patterns which lea	rainfall pattern lead flood drc	5	75	15	0.8	5	2	1	0
13	Author	A particularly intense El Nif	particularly intense nifio secc	8	100	12	0.76	8	9	2	0
14	Author	Studies with computer moc	study computer model kind d	4	59	15	0.627118644	4	0	1	0
15	Author	Volcanoes inject enormous	volcano inject enormous quar	4	51	13	0.901960784	4	2	0	0
16	Author	One of the largest volcanic	one largest volcanic eruption	5	72	14	0.875	5	2	0	0
17	Author	Over the centuries differer	century different human com	2	40	20	0.825	2	0	1	0
18	Web	Since the pre-industrial pe	since pre industrial period hu	5	42	8	0.785714286	5	2	0	0
19	Web	Climate change is a long-te	climate change long term cha	2	22	11	0.818181818	2	1	0	0
20	Author	But the question must be a	question must asked remarka	3	25	8	1	0	0	0	1
21	Author	Here a note of caution mus	note caution must sounded ra	6	37	6	0.837837838	5	1	0	0
22	Author	However, we know for sure	however know sure human ac	2	44	22	0.909090909	2	4	0	0
23	Author	exist. Although, therefore,	exist although therefore certa	3	39	13	0.923076923	3	7	1	0
24	Author	The generally cold period v	generally cold period worldw	3	35	12	0.885714286	3	0	0	0
25	Author	What is important is contin	important continually make c	4	54	14	0.87037037	4	5	0	0
26	Author	Human activities of all kind	human activity kind whether	6	74	12	0.810810811	6	7	0	0

Figure 20: Extracted features I

The punctuation features: total number of question marks (?) in a paragraph, total number of hyphens (-) in a paragraph, total number of percentage marks (%) in a paragraph, total number of less than (<) symbols in the paragraph, total number of greater than symbols in the paragraph (>), average number of dots (.) in a document, average number of commas (,) in a document, average number of semicolons (;) in a document are the next few features taken for the model as shown in the figure 21.

1	class	text	cleaned_text	Total#Questionmark	Total#Hyphens	Total#percentage	Total#lessthan	Total#greaterthan	Avg#dots	Avg#comma	Avg#semicolon
2	Author	The phrase Ã«global	phrase Ã«global warr	0	1	0	0	0	1	1	0
3	Author	In the year 2060 my g	year grandchild appro	5	0	0	0	0	0	0	0
4	Author	Before studying futu	studying future clima	2	0	0	0	0	1	0	0
5	Author	Variations in day-to-(	variation day day wea	0	5	0	0	0	1	1	0
6	Author	The 1980s and 1990s \	unusually warm globa	0	2	0	0	0	1	1	0
7	Author	The period has also b	period also remarkab	0	2	0	0	0	1	1	0
8	Author	But those storms in E	storm europe mild cc	0	1	0	0	0	1	1	0
9	Author	The increase in storm	increase storm intens	0	3	0	0	0	1	1	0
10	Web	Global warming is th	global warming long	0	4	0	0	0	1	1	0
11	Author	Windstorms or hurric	windstorm hurricane	0	3	0	0	0	1	1	0
12	Author	Rainfall patterns whi	rainfall pattern lead f	0	1	0	0	0	1	0	0
13	Author	A particularly intense	particularly intense n	0	3	0	0	0	1	1	0
14	Author	Studies with comput	study computer mod	0	3	0	0	0	1	0	0
15	Author	Volcanoes inject eno	volcano inject enorm	0	0	0	0	0	1	0	0
16	Author	One of the largest vo	one largest volcanic e	0	0	0	0	0	1	0	0
17	Author	Over the centuries di	century different hur	0	0	0	0	0	1	0	0
18	Web	Since the pre-industr	since pre industrial p	0	1	0	0	0	1	0	0
19	Web	Climate change is a l	climate change long t	0	1	0	0	0	1	0	0
20	Author	But the question mu	question must asked	3	0	0	0	0	0	0	0
21	Author	Here a note of cautio	note caution must so	0	1	0	0	0	1	0	0
22	Author	However, we know f	however know sure h	0	0	0	0	0	1	2	0
23	Author	exist. Although, ther	exist although theref	0	0	0	0	0	1	2	0
24	Author	The generally cold pe	generally cold period	0	0	0	0	0	1	0	0
25	Author	What is important is	important continually	0	0	0	0	0	1	1	0
26	Author	Human activities of a	human activity kind v	0	0	0	0	0	1	1	0

Figure 21: Extracted Features II

The features: average number of colons (:) per document, average number of exclamation marks (!) per document, average number of question marks (?) per document, average number of hyphens (-) per document, average number of percentage marks (%) per document, average number of less than symbols (<) per document, average number of greater than symbols (>) per document are the rest of the punctuation features taken into consideration for the model creation as in figure 22 below.

1	class	text	cleaned_text	Avg#colon	Avg#Exclamationn	Avg#Questionmar	Avg#Hyphens	Avg#percentage	Avg#lessthan	Avg#greaterthan
2	Author	The phrase Ã«global	phrase Ã«global warr	0	0	0	0	0	0	0
3	Author	In the year 2060 my g	year grandchild appro	0	0	1	0	0	0	0
4	Author	Before studying futu	studying future clima	0	0	0	0	0	0	0
5	Author	Variations in day-to-	variation day day wea	0	0	0	0	0	0	0
6	Author	The 1980s and 1990s	unusually warm glob	0	0	0	1	0	0	0
7	Author	The period has also b	period also remarkab	0	0	0	0	0	0	0
8	Author	But those storms in E	storm europe mild cc	0	0	0	0	0	0	0
9	Author	The increase in storm	increase storm intens	0	0	0	0	0	0	0
10	Web	Global warming is th	global warming long	0	0	0	1	0	0	0
11	Author	Windstorms or hurric	windstorm hurricane	0	0	0	0	0	0	0
12	Author	Rainfall patterns whi	rainfall pattern lead f	0	0	0	0	0	0	0
13	Author	A particularly intense	particularly intense n	0	0	0	0	0	0	0
14	Author	Studies with comput	study computer mod	0	0	0	1	0	0	0
15	Author	Volcanoes inject eno	volcano inject enorm	0	0	0	0	0	0	0
16	Author	One of the largest vo	one largest volcanic e	0	0	0	0	0	0	0
17	Author	Over the centuries di	century different hur	0	0	0	0	0	0	0
18	Web	Since the pre-industr	since pre industrial p	0	0	0	0	0	0	0
19	Web	Climate change is a l	climate change long t	0	0	0	0	0	0	0
20	Author	But the question mu	question must asked	0	0	1	0	0	0	0
21	Author	Here a note of cautio	note caution must so	0	0	0	0	0	0	0
22	Author	However, we know f	however know sure h	0	0	0	0	0	0	0
23	Author	exist. Although, ther	exist although theref	0	0	0	0	0	0	0
24	Author	The generally cold pe	generally cold period	0	0	0	0	0	0	0
25	Author	What is important is	important continually	0	0	0	0	0	0	0
26	Author	Human activities of a	human activity kind v	0	0	0	0	0	0	0

Figure 22: Extracted Features III

In addition to the lexical features and the punctuation features, the Part of speech taggers were considered for more accuracy when building the model. The POS taggers used are shown in both the figures 23 and 24 as below. The studying the structure of the sentences and providing the particular sentence structure possess by a document to the model building was the main objective of introducing POS taggers. The nouns, verbs, adjectives, pronouns, conjunctions and other many types of building blocks in the sentence structures are identified here.

1	class	text	cleaned_text	CD	JJ	NN	NNP	NNS	RB	VBD	VBG	VBN	VBP	VBZ	IN	MD	VB	JJR
2	Author	The phrase Ã«global	phrase Ã«global warr	1	11	11	1	1	1	1	2	1	1	1	0	0	0	0
3	Author	In the year 2060 my g	year grandchild appro	0	13	17	0	0	5	3	2	2	1	0	2	1	1	0
4	Author	Before studying futu	studying future clima	1	10	22	0	0	3	4	2	1	0	0	3	0	0	2
5	Author	Variations in day-to-	variation day day wea	0	20	58	0	1	8	1	2	3	1	0	1	1	1	0
6	Author	The 1980s and 1990s	unusually warm glob	1	6	16	0	0	5	2	2	0	0	0	3	0	0	0
7	Author	The period has also b	period also remarkab	3	12	27	0	0	7	6	1	2	0	1	4	0	0	2
8	Author	But those storms in E	storm europe mild cc	1	28	54	0	4	7	8	3	3	2	1	1	0	1	1
9	Author	The increase in storn	increase storm intens	9	22	52	0	0	9	10	1	5	2	0	1	0	1	0
10	Web	Global warming is th	global warming long	0	12	23	0	1	5	2	4	2	0	1	2	0	0	0
11	Author	Windstorms or hurric	windstorm hurricane	5	32	59	0	3	10	12	5	1	3	1	2	0	0	1
12	Author	Rainfall patterns whi	rainfall pattern lead f	2	14	35	0	2	6	2	1	2	3	1	5	0	0	1
13	Author	A particularly intense	particularly intense n	1	33	35	0	3	10	10	2	3	1	0	1	0	0	0
14	Author	Studies with comput	study computer mod	0	15	32	0	0	5	5	0	0	2	0	0	0	0	0
15	Author	Volcanoes inject eno	volcano inject enorm	0	10	23	0	0	8	3	2	0	2	0	1	0	0	1
16	Author	One of the largest vo	one largest volcanic e	5	16	34	0	0	4	4	3	1	0	0	2	0	0	1
17	Author	Over the centuries di	century different hur	1	7	23	0	0	2	2	1	0	1	1	1	0	0	0
18	Web	Since the pre-industr	since pre industrial p	0	11	19	0	0	2	1	3	1	0	0	4	0	0	1
19	Web	Climate change is a l	climate change long t	0	7	12	0	0	0	1	0	1	0	1	0	0	0	0
20	Author	But the question mu	question must asked	0	4	13	0	0	1	1	2	1	1	0	0	1	1	0
21	Author	Here a note of cautio	note caution must so	0	7	21	0	0	2	1	0	1	1	0	1	2	1	0
22	Author	However, we know f	however know sure h	1	9	21	0	0	4	2	2	2	2	0	1	0	0	0
23	Author	exist. Although, ther	exist although theref	1	12	18	0	0	2	0	2	1	1	0	1	0	1	0
24	Author	The generally cold pe	generally cold period	0	6	14	1	0	5	3	1	2	0	0	1	1	1	0
25	Author	What is important is	important continually	0	18	24	0	0	5	1	1	1	4	0	0	0	0	0
26	Author	Human activities of a	human activity kind v	4	9	37	0	0	6	2	7	0	1	1	2	1	1	1

Figure 23: Extracted Features IV

The next part of speech taggers used for the study is shown in the below figure 24 and the taggers represent the sentence grammatical building blocks such as foreign words (FW), determiners (DT), coordinating conjunctions (CC), interjections (UH), possessive pronouns (PRP$), etc.

1	class	text	cleaned_text	FW	WP$	JJS	DT	RBR	WDT	CC	PRP$	EX	WRB	RP	WP	RBS	PRP	POS	UH
2	Author	The phrase Ã«glc	phrase Ã«global	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Author	In the year 2060 ı	year grandchild	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Author	Before studying t	studying future	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Author	Variations in day	variation day da	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Author	The 1980s and 19	unusually warm	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Author	The period has al	period also rem	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Author	But those storms	storm europe m	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	Author	The increase in s	increase storm i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Web	Global warming i	global warming	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Author	Windstorms or h	windstorm hurri	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Author	Rainfall patterns	rainfall pattern l	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
13	Author	A particularly int	particularly inte	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Author	Studies with com	study computer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Author	Volcanoes inject	volcano inject e	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
16	Author	One of the larges	one largest volc	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
17	Author	Over the centuri	century differer	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
18	Web	Since the pre-inc	since pre indust	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Web	Climate change i	climate change l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Author	But the question	question must a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Author	Here a note of ca	note caution mu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Author	However, we kn	however know s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Author	exist. Although, t	exist although tl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	Author	The generally col	generally cold p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Author	What is importar	important conti	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	Author	Human activities	human activity k	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0

Figure 24: : Extracted Features V

3.2.5 Feature Selection

The total numbers of the features are 58 and the best features need to be identified in the study to build the model. The best features are selected by the Chi-Squared test provided by the 'SelectKbest' class in the Scikit-learn library, calculating the score as follows. The scores are arranged in descending order to identify the best set of columns for features and their values are taken into use. The figure 26 and 27 show all the features with their scores obtained and the figure 28 shows the best 10 features selected.

The highest Chi-Squared test score among the features has obtained by the feature 'lexical_diversity' and it is about 8430.52. The feature 'Total#words' also giving a considerably higher value and the range of the rest of the features is between 100 and 700. Figure 29 shows the selected best features with their values in the context of the text document.

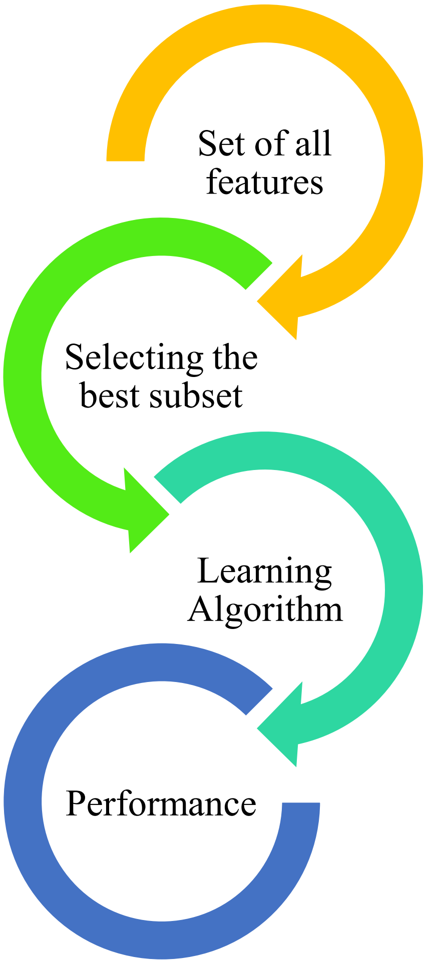


Figure 25: Feature Selection

Specs	Score
lexical_diversity	8430.519187
Total#words	1622.019521
JJ	645.530770
Total#comma	625.858786
NN	532.759558
Total#dots	307.294802
Total#percentage	263.758794
#sentences	258.508304
VBG	179.376242
Total#Exclamationmark	121.124031
CD	104.648649
NNS	81.386702
Total#Hyphens	54.785578
Avg#percentage	54.000000
Avg#words	18.751099
JJS	15.157895
Total#semicolon	15.155738
Avg#comma	13.714286
Avg#Hyphens	13.172043
Avg#colon	11.000000
VBN	10.658041
Total#lessthan	10.000000
Total#stop_words	6.250000
Total#colon	6.145455
RP	5.555556
VBD	5.487805
IN	4.655493
WP	4.571429
Avg#Questionmark	4.500000

Figure 26: Features and their scores I

NNP	3.891892
PRP	3.600000
WP$	3.600000
FW	3.313725
MD	3.027027
RB	3.016550
WRB	3.000000
VBP	2.599299
CC	2.372549
RBS	2.000000
VB	1.911364
Total#greaterthan	1.814815
RBR	1.472727
EX	1.285714
Avg#semicolon	1.190476
UH	1.000000
PRP$	1.000000
WDT	1.000000
Total#Questionmark	0.510638
VBZ	0.396088
DT	0.324675
JJR	0.219601
Avg#dots	0.195652
POS	0.000000

Figure 27: Features and their scores II

Specs	Score
lexical_diversity	8430.519187
Total#words	1622.019521
JJ	645.530770
Total#comma	625.858786
NN	532.759558
Total#dots	307.294802
Total#percentage	263.758794
#sentences	258.508304
VBG	179.376242
Total#Exclamationmark	121.124031

Figure 28: Best 10 features

	Total#words	Total#comma	JJ	Total#percentage	NN	Total#dots	#sentences	VBG	NNS	Avg#percentage
0	32	2	11	0	11	3	3	2	1	0
1	47	2	13	0	17	3	8	2	0	0
2	48	1	10	0	22	4	6	2	0	0
3	99	13	20	0	58	10	10	2	1	0
4	38	2	6	0	16	3	3	2	0	0
...	...	...	...	...	...	...	...	...	...	...
1163	43	2	7	0	22	8	10	6	0	0
1164	22	0	5	0	9	3	3	1	0	0
1165	39	7	9	0	14	3	3	0	1	0
1166	93	14	18	0	57	8	8	5	1	0
1167	85	11	17	0	56	6	6	1	1	0

1168 rows × 10 columns

Figure 29: Best features and the values

3.2.6 Model Building

3.2.6.1 Stylistic features

The datasets prepared were divided into two sections as training dataset and the testing dataset for the relevant training and testing purposes. There are 03 classes are available in the dataset and they were named 'Author' as '1', the rest of the classes 'Web' and 'RP' were named as '0'. Later the class labels were renamed as '1' for the Author class and '-1' for all other classes. The one-class SVM algorithm was used as the classification algorithms in the study separately. Hence, the labels are generated for the documents in the three documents. The testing dataset is set to 20% of the total dataset and the training set is set to 80% in the study.

The study dataset is having 845 paragraphs in the first document in the 'Author' class and 342 paragraphs in all the other classes. Document two consists of 778 paragraphs in the 'Author' class and 188 paragraphs in all the other classes while document three is having 1240 paragraphs in the 'Author' class and 342 paragraphs in all the other classes. Thus, the documents are imbalanced in nature. The class balancing is not practical as the one-class SVM algorithm is used in the study.

3.2.6.2 Bag of Words

In addition to one-class SVM algorithm used with the stylistic features, the Naive Bayes algorithm, Logistic Regression Algorthm, and SVM algorithm were used as the classification algorithms in the study with bag of words. The labels were generated for the documents in the three documents in this scenario also. The testing dataset is set to 20% of the total dataset and the training set is set to 80% in the study. Count vectorizer objects were created first and then training and validation data were transformed using the count vectorizer objects.

The Naive Bayes classifier, linear classifier – logistic regression, SVM classifier were used and the accuaracies are checked first on the counter vectors, next on the word level TF-IDF vectors, then on the N-gram level TF-IDF vectors and finally on the character level TF-IDF vectors.

3.2.6.3 Naive Bayes Classifier

Naive Bayes classifier is a machine learning model that is used for a classification task in a study. The Bayes Therom which is related with probabilities provide the basis for the classifier and this classifier is comparatively easy to implement. The research study uses this classifier in four levels as count vectors, char-level vectors, n-gram level vectors, and word-level tf-idf vectors.

3.2.6.4 Logistic Regression Classifier

Logistic regression classifier is used in the classification procedure of the machine learning and it is a supervised learning algorithm. The predictions can be done with the algorithm as a function of independent variables and can produce the dependent output variable. The research study uses this classifier in four levels as count vectors, char-level vectors, n-gram level vectors, and word-level tf-idf vectors.

3.2.6.5 Support Vector Machine Classifier

Support Vector Machine classifier is also known as SVM classifier in short form. It is a deep learning algorithm and also a supervised learning algorithm. It can solve many of the linear problems as well as non-linear problems practically. The research study uses this classifier in four levels as count vectors, char-level vectors, n-gram level vectors, and word-level tf-idf vectors.

The performance measures were calcualetd in each of the instance that the model was trained and the accuracies were compared.

Chapter 4

Evaluation

4.1 One-class SVM Classifier with Stylistic Features

4.1.1 For all the features

The performance measures in the figure 30 represent the results for one-class svm classifier for all the features. The accuracy score is 40.53% and the precision, recall, f1 score and support are 0.44, 0.75, 0.55 and 167 respectively for the class label '-1' which represent 'Non-author' or the all other classes. The precision, recall, f1 score and support are 0.22, 0.07, 0.11 and 171 respectively for class label '1' which is called 'Author' class.

```
Confusion Matrix :
[[125  42]
 [159  12]]
Accuracy Score : 0.40532544378698226
Report :
              precision    recall  f1-score   support

          -1       0.44      0.75      0.55       167
           1       0.22      0.07      0.11       171

    accuracy                           0.41       338
   macro avg       0.33      0.41      0.33       338
weighted avg       0.33      0.41      0.33       338
```

Figure 30: Confusion matrix for all features

4.1.2 For the best features

The performance measures in the figure 31 represent the results for the one-class svm classifier for the best svm features. The accuracy score is 51.18% and the precision, recall, f1 score and support are 0.50, 0.77, 0.61 and 167 respectively for the class label '-1' which represent 'Non-author' or the all other classes. The precision, recall, f1 score and support are 0.54, 0.26, 0.35 and 171 respectively for class label '1' which is called 'Author' class.

```
Confusion Matrix :
[[128  39]
 [126  45]]
Accuracy Score : 0.5118343195266272
Report :
              precision    recall  f1-score   support

          -1       0.50      0.77      0.61       167
           1       0.54      0.26      0.35       171

    accuracy                           0.51       338
   macro avg       0.52      0.51      0.48       338
weighted avg       0.52      0.51      0.48       338
```

Figure 31: Confusion matrix for best features

The results should be analyzed as an overall in terms of validation accuracies, f1 score, precision and recall. The testing accuracy or the validation accuracy can be defined as the calculation of accuracy for the dataset that we did not used for the training purpose. The weighted average of precision and recall is known as the f1 score and it is the statistical measure that most of the literature sources have been used so far. Precision is also known as sensitivity which describe the fraction of positive predictive values among the all true positive and the false positive instances and the recall can be interpreted as the measure of identifying the true positive values or the correct hits. These measures can be represent in equations as follows:

Equation 1: Accuracy

Accuracy = (True Positive + True Negative) / Total

Equation 2: F1 Score

F1 score = 2 * {(Precision * Recall) / (Precision + Recall)}

Equation 3: Precision

Precision = (True Positive / (True Positive + False Positive)

Equation 4: Recall

Recall = (True Positive / (True Positive + False Negative))

4.1.2.1 F1 Score

Table 6: F1 score values

	For all features	*For best features*
-1	0.55	0.61
1	0.11	0.35

4.1.2.2 Precision values

Table 7: Precision values

	For all features	*For best features*
-1	0.44	0.50
1	0.22	0.54

4.1.2.3 Recall values

Table 8: Recall values

	For all features	*For best features*
-1	0.75	0.77
1	0.07	0.26

4.2 Naive Bayes Classifier with Bag-of-Words

The study was conducted with Naive Bayes classifier, linear classifier and SVM classifier. Classifier is an algorithm in machine learning used to allocate the class labels for the input data and these classifiers are trained by class labels.

Naive Bayes classifier is an algorithm used to classify the input data which is based on the Bayes' theorem. The main feature of this classifier is, it is assuming high independence among the features and it is also known as 'probabilistic classifier'. Figure 32 shows the confusion matrices and the accuracies for naive bayes classifier for count vectors as 0.95 and word level tf-idf as 0.76. And figure 33 shows confusion matrices and the accuracies for the n-gram vectors as 0.84 and char level vectors as 0.72.

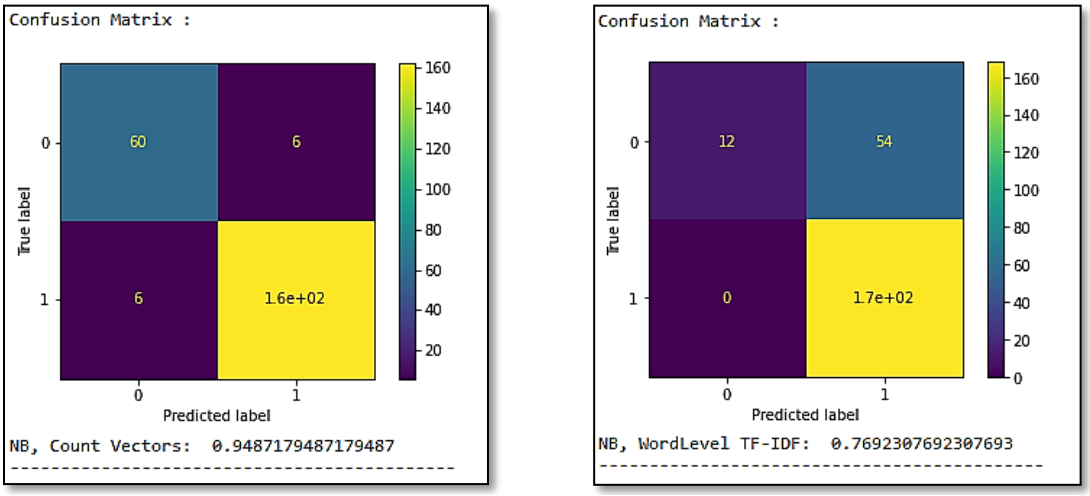


Figure 32: Naive Bayes - Count vectors and Word level TF-IDF

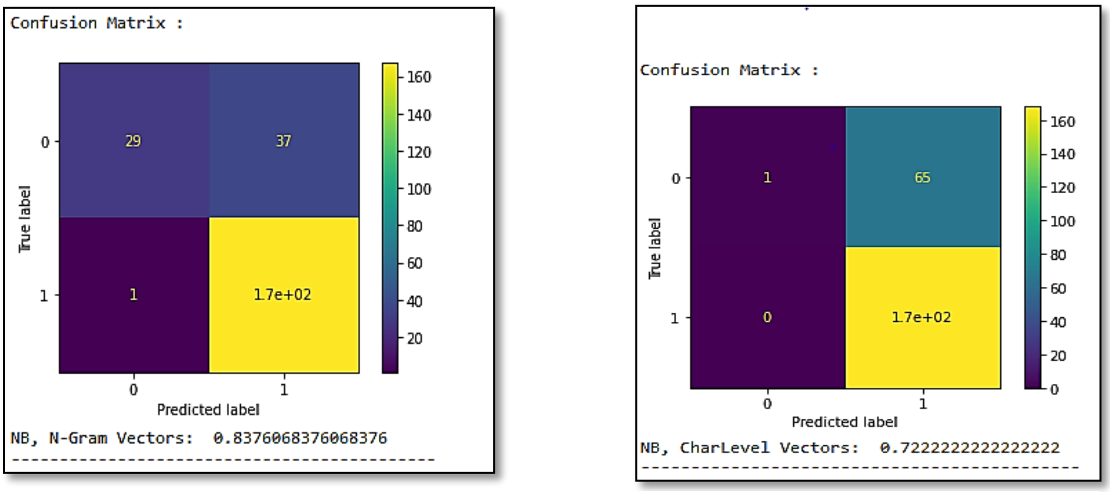


Figure 33: Naive Bayes N-Gram vectors and Char level vectors

4.3 Logistic Regression Classifier with Bag-of-Words

Linear classifier use linear combination of features to classify the labels of the input data and mostly linear conmbination of features are used as inputs under the linear logistic regression.

According to the figure 34, the accuracy of count vectors on linear classifier, logistic regression is 0.94 and the confusion matrix is given. Also, the word level tf-idf has accuracy of 0.81 and the confusion matrix is given. Accuracy for the n-gram vectors is 0.78 and char level vectors has 0.82 of accuracy.

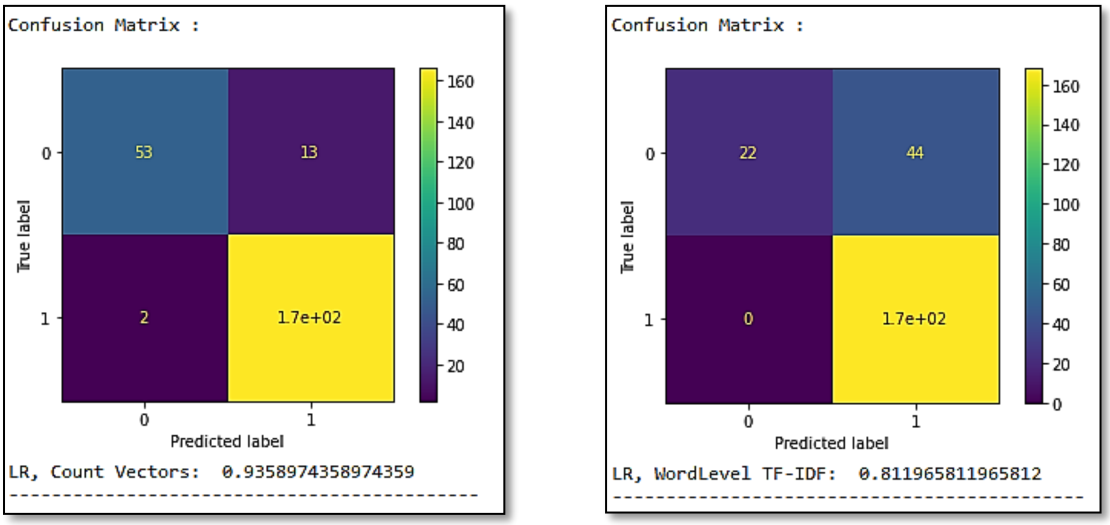


Figure 34: Linear Classifier, Count vectors and Word level TF-IDF

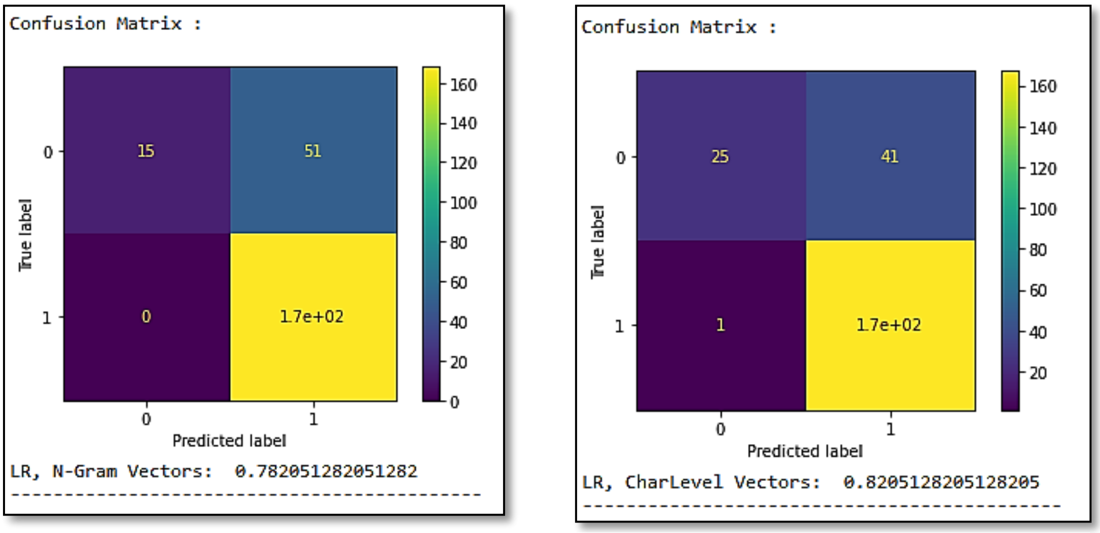


Figure 35: Linear Classifier, N-Gram vectors and Char level vectors

4.4 Support Vector Machine Classifier with Bag-of-Words

Support vector machine (SVM) is a machine learning algorithm used for supervised learning and also it is mostly used for classification problems. SVM classifier is used in the study to check for the accuracy value of prediction.

The figure 36 shows the confusion matrix and the accuracy for the SVM classifier in n-gram vectors as 0.81 and the accuracy for the count vectors as 0.89. Meanwhile, figure 37 shows the confusion matrix and the accuracy for the word level tf-idf as 0.88 and accuracy for char level vectors as 0.90 for SVM classifier.

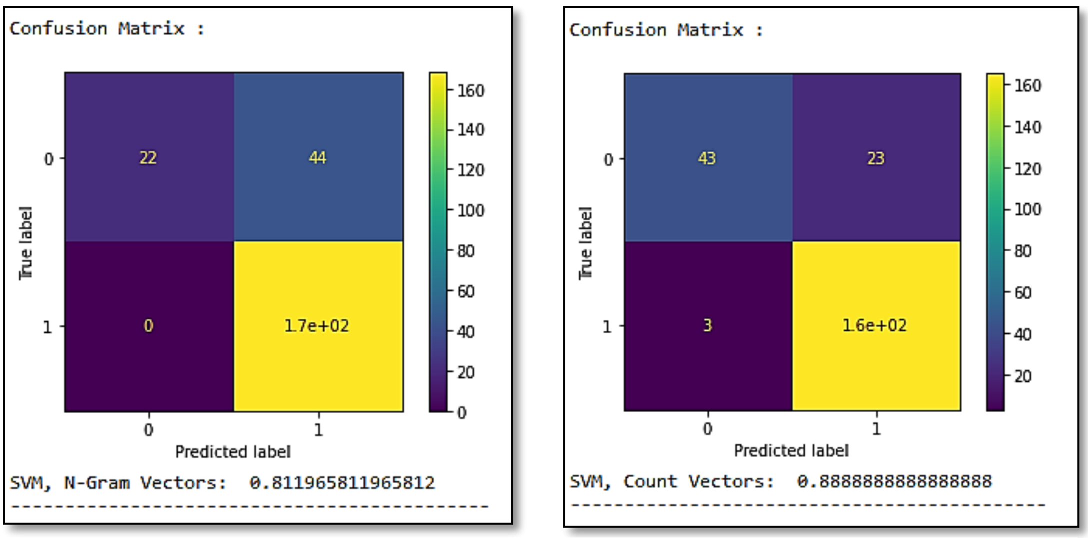


Figure 36: SVM Classifier, N-Gram vectors, Count vectors

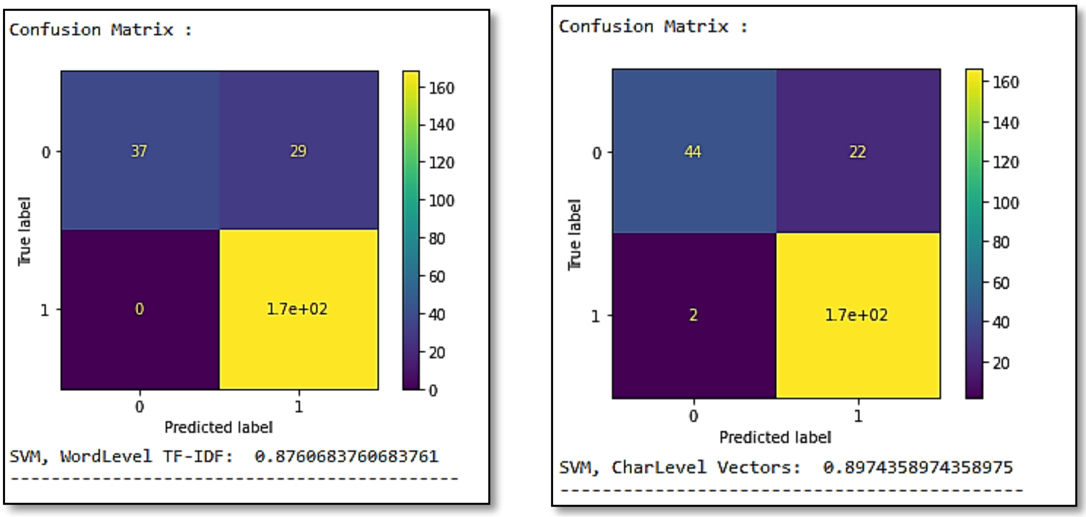


Figure 37: SVM classifier, Word level TF-IDF, Char level vectors

According to the results received, in the context of one-class svm classifier with stylistic features, the validation accuracy for the model built with the best features is 51.18%. In the context of bag-of-words, the highest validation accuracy for Naive Bayes classifier was obtained for the count vectors and the value was 94.87%. The highest validation accuracy was retrieved as 93.59% for the count vectors in logistic regression classifier and regarding the svm classifier also, counter vectors showed 88.89% of highest validation accuracy.

Chapter 5

Conclusion

5.1 Conclusion of the study

The main aim of the research study is towards the intrinsic plagiarism detection to identify the deviations in the writing styles of the authors and thus the outlier detection is needed. The features were extracted from the documents and after the process of preprocessing, a machine learning model was built. The imbalanced nature of the text sources were considered and avoided using data augmentation techniques such as SMOTE technique, as it breaks the practicality of the real world problem.The model was built using the features extracted and the one class SVM algorithm was used as the classifier. The training data portion was 0.8 and the validation data portion was 0.2 from the total. The results were obtained using the trained model and the model performance measures were taken. The model accuarcies are compared among the classifiers one-class svm classifier, logistic regression classifier, naive bayes classifier and svm classifier in the context of bag-of-words.

According to the results received, the validation accuracy for the model built with the best features is 51.18% reagrding one-class svm classifier with stylistic features. In the context of bag-of-words, the highest validation accuracy for Naive Bayes classifier was obtained for the count vectors and the value was 94.87%. The highest validation accuracy was retrieved as 93.59% for the count vectors in logistic regression classifier and regarding the svm classifier also, counter vectors showed 88.89% of highest validation accuracy.

The model built with stylistic features provides the f1 score values as follows: the f1-score values for the testing datset were higher for the Non-author class, '-1' which represent the all the other classes with out the 'Author' class. The values were 0.61 and 0.55 respectively for the dataset with full features and dataset with best features. The f1-score for the class '1', which is known as 'Author' class in full feature dataset is 0.11 and 0.35 for the dataset with best features. The precision values in the results are also higher for the best feature set as the values are 0.50 for class '-1' and 0.54 for class '1'. The recall values are also considerably higher for the class '-1' which mention the values as 0.75 for all the features and 0.77 for best features.

The performance measures of the model built in the study with stylistic features show lower values than the other classifiers used with bag-of-words. For example, the naive bayes classifier shows 94.87% of accuracy for count vectors, 76.92% of accuracy for word level tf-idf, 83.76% of accuracy for n-gram vectors and 72.23% of accuracy for char level vectors. Meanwhile, logistic regression classifier shows 93.58% of accuracy for count vectors, 81.19% of accuracy for word level tf-idf, 78.20% of accuracy for n-gram vectors and 82.05% of accuracy for char level vectors. In addition, support vector machine classifier shows 88.89% of accuracy for count vectors, 87.60% of accuracy for word level tf-idf, 76.06% of accuracy for n-gram vectors and 89.74% of accuracy for char level vectors.

5.2 Future Work

As the model built with stylistic features with the usage of one-class svm algorithm shows a lesser value compared to the algorithms used with bag-of-words. Therefore, the validation accuracies can be improved with feeding more data to the model and also with more number of stylistic features in order to increase the performance measure values. The different types of classifiers also can be used for the study instead of One Class SVM algorithm used in the research study.

Appendix I - Codes

```python
# -*- coding: utf-8 -*-
"""
"""

import nltk
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer,PorterStemmer
from nltk.corpus import stopwords
import re
import numpy as np
import pandas as pd

from sklearn import model_selection, preprocessing, linear_model, naive_bayes, metrics, svm
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn import decomposition, ensemble

import matplotlib.pyplot as plt
from sklearn.metrics import plot_confusion_matrix

nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

Figure 38: Importing libraries for the program

```python
df = pd.read_csv(r'D:\MBA\Research Project\Last\After Interim\Code\
                 Document_1\Combined1.csv',encoding='mac_roman',header=None,names=['class','text'])
df
```

Figure 39: Read the .csv file prepared with text data

```python
"""# Pre-processing"""

lemmatizer = WordNetLemmatizer()
stemmer = PorterStemmer()

def preprocess(sentence):
    sentence=str(sentence)
    sentence = sentence.lower()
    rem_num = re.sub('[0-9]+', '', sentence)
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(rem_num)
    filtered_words = [w for w in tokens if len(w) > 2 if not w in stopwords.words('english')]
    stem_words=[stemmer.stem(w) for w in filtered_words]
    lemma_words=[lemmatizer.lemmatize(w) for w in filtered_words]
    return " ".join(lemma_words)
    # return " ".join(tokens)

df['cleaned_text'] = df.apply(lambda x: preprocess(x.text),axis=1)
df
```

Figure 40: Preprocessing for other mechanisms

```python
"""# Feature Extraction"""

def cnt_sentence(x):
  cnt = x.count('.')+x.count('?')+x.count('!')
  if cnt ==0:
    cnt =1
  return cnt


def tot_words(x):
  cnt = len(x.split(' '))
  return cnt

def avg_words(x,sent_cnt):
  cnt = round(tot_words(x)/sent_cnt)
  return cnt

def lexical_diversity(x):
  ld = len(set((x).split(' '))) / tot_words(x)
  return ld

#stopwords
def count_stop_words(x):
  cnt = 0
  for w in x.split(' '):
    if w in stopwords.words('english'):
      cnt=cnt+1
  return cnt
```

Figure 41: Defining features to be extracted

```python
df['#sentences'] =  df['text'].apply(lambda x: cnt_sentence(x))
# df['#sentences'] = np.where(df['#sentences']==0,1,df['#sentences'])

df['Total#words'] =  df['cleaned_text'].apply(lambda x: tot_words(x) )
df['Avg#words'] =  df[['cleaned_text','#sentences']].apply(lambda x: avg_words(x['cleaned_text'],x['#sentences'])

```

Figure 42: Extracting features I

```python
df['lexical_diversity'] =  df['cleaned_text'].apply(lambda x:  lexical_diversity(x)  )

df['Total#dots'] =  df['text'].apply(lambda x: x.count('.'))
df['Total#comma'] =  df['text'].apply(lambda x: x.count(','))
df['Total#semicolon'] =  df['text'].apply(lambda x: x.count(';'))
df['Total#colon'] =  df['text'].apply(lambda x: x.count(':'))
df['Total#Exclamationmark'] =  df['text'].apply(lambda x: x.count('!'))
df['Total#Questionmark'] =  df['text'].apply(lambda x: x.count('?'))
df['Total#Hyphens'] =  df['text'].apply(lambda x: x.count('-'))
df['Total#percentage'] =  df['text'].apply(lambda x: x.count('%'))
df['Total#lessthan'] =  df['text'].apply(lambda x: x.count('<'))
df['Total#greaterthan'] =  df['text'].apply(lambda x: x.count('>'))
```

Figure 43: Extracting features II

```
df['Avg#dots'] = round(df['Total#dots']/df['#sentences'])
df['Avg#comma'] = round( df['Total#comma']/df['#sentences'])
df['Avg#semicolon'] =  round(df['Total#semicolon']/df['#sentences'])
df['Avg#colon'] =  round(df['Total#colon']/df['#sentences'])
df['Avg#Exclamationmark'] = round( df['Total#Exclamationmark']/df['#sentences'])
df['Avg#Questionmark'] =  round(df['Total#Questionmark']/df['#sentences'])
df['Avg#Hyphens'] = round(df['Total#Hyphens']/df['#sentences'])
df['Avg#percentage'] =  round(df['Total#percentage']/df['#sentences'])
df['Avg#lessthan'] = round( df['Total#lessthan']/df['#sentences'])
df['Avg#greaterthan'] =  round(df['Total#greaterthan']/df['#sentences'])

df['Total#stop_words'] =  df['cleaned_text'].apply(lambda x:  count_stop_words(x)  )
df['Avg#stop_words'] =  round(df['Total#stop_words']/df['#sentences'])

```

Figure 44: Extracting features III

```
df

df[df['#sentences']==0]

df['text'][2]

df['No'] = df.index

dff = pd.DataFrame()

for index, row in df.iterrows():
  data_tagset = nltk.pos_tag(row['cleaned_text'].split(' '))         #pass by tokens
  df_tagset = pd.DataFrame(data_tagset, columns=['Word', 'Tag'])
  df_tagset = df_tagset.groupby(['Tag']).agg('count')
  df_tagset_piv = pd.pivot_table(df_tagset, values='Word', index=[],columns=['Tag'], aggfunc=np.sum)
  tag_df = pd.DataFrame(df_tagset_piv)


  df1 = df[df['No']==row['No']]
  tag_df['No'] =row['No']
  result = df1.merge(tag_df, how='inner',on='No')
  print(result)

  dff = dff.append(result)

dff = dff.fillna(0)

dff.to_csv(r'D:\MBA\Research Project\Last\After Interim\Code\Document_1\final_result1.csv',index=False)
```

Figure 45: Write the extracted features to a file

```
dff_copy = dff.copy()
y = dff['class']
X = dff_copy.drop(['class','text','cleaned_text'],axis=1
                  )

x_columns = X.columns
x_columns

from sklearn.metrics import confusion_matrix
from imblearn.over_sampling import SMOTE
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.svm import OneClassSVM

y = y.replace(0,-1)
y
```

Figure 46: Importing more libraries

```python
    accuracy =  metrics.accuracy_score(predictions, valid_y)
    print(i,' ',accuracy)

    if accuracy > accuracy_best:
      print('best')
      accuracy_best = accuracy
      best_no_columns = i+1


print('Best set of columns:',featureScores.nlargest(best_no_columns,'Score'))
print('Best column counts:',best_no_columns)

train_x, valid_x, train_y, valid_y = model_selection.train_test_split
    (X[featureScores.nlargest(best_no_columns,'Score')['Specs']], y,test_size=0.2, random_state=42)
classifier = OneClassSVM(gamma='auto')

# fit the training dataset on the classifier
classifier.fit(train_x, train_y)

# predict the labels on validation dataset
predictions = classifier.predict(valid_x)

accuracy =  metrics.accuracy_score(predictions, valid_y)
print(accuracy)

results = confusion_matrix(valid_y, predictions)
print('Confusion Matrix :')
print(results)
print('Accuracy Score :',metrics.accuracy_score(valid_y, predictions))
print('Report : ')
print(metrics.classification_report(valid_y, predictions))

```

Figure 47: Applying SMOTE for best features II

```python
"""### without applyting SMOTE accuracy is high()"""

train_x, valid_x, train_y, valid_y = model_selection.train_test_split
     (X, y,test_size=0.2,random_state=42)
classifier = OneClassSVM(gamma='auto')

# fit the training dataset on the classifier
classifier.fit(train_x, train_y)

# predict the labels on validation dataset
predictions = classifier.predict(valid_x)

accuracy =  metrics.accuracy_score(predictions, valid_y)
print(accuracy)

results = confusion_matrix(valid_y, predictions)
print('Confusion Matrix :')
print(results)
print('Accuracy Score :',metrics.accuracy_score(valid_y, predictions))
print('Report : ')
print(metrics.classification_report(valid_y, predictions))

```

Figure 48: Without applying SMOTE for all features

```python
"""### Feature selection"""

#apply SelectKBest class to extract top 10 best features
bestfeatures = SelectKBest(score_func=chi2, k=10)
fit = bestfeatures.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns
print(featureScores.nlargest(10,'Score'))

#select best feature set
accuracy_best = 0
best_no_columns = 0
for i in range(len(X.columns)):
  train_x, valid_x, train_y, valid_y = model_selection.train_test_split
      (X[featureScores.nlargest(i+1,'Score')['Specs']], y,test_size=0.2, random_state=42)
  classifier = OneClassSVM(gamma='auto')

  # fit the training dataset on the classifier
  classifier.fit(train_x, train_y)

  # predict the labels on validation dataset
  predictions = classifier.predict(valid_x)
```

Figure 49: Best features selection I

```python
  accuracy =  metrics.accuracy_score(predictions, valid_y)
  print(i,'  ',accuracy)

  if accuracy > accuracy_best:
    print('best')
    accuracy_best = accuracy
    best_no_columns = i+1

print('Best set of columns:',featureScores.nlargest(best_no_columns,'Score'))
print('Best column counts:',best_no_columns)

train_x, valid_x, train_y, valid_y = model_selection.train_test_split
    (X[featureScores.nlargest(best_no_columns,'Score')['Specs']], y,test_size=0.2, random_state=42)
classifier = OneClassSVM(gamma='auto')      # Assigning the model  - new born baby it won't have an
 
# fit the training dataset on the classifier
classifier.fit(train_x, train_y)     #building new model by feeding data    # this is classification

# predict the labels on validation dataset
predictions = classifier.predict(valid_x)

accuracy =  metrics.accuracy_score(predictions, valid_y)
print(accuracy)

len(predictions)
predictions
valid_x
valid_y
#test_x = pd.DataFrame({'lexical_diversity':[8800]})
#classifier.predict(test_x)                              #give
results = confusion_matrix(valid_y, predictions)
print('Confusion Matrix :')
print(results)
print('Accuracy Score :',metrics.accuracy_score(valid_y, predictions))
print('Report : ')
print(metrics.classification_report(valid_y, predictions))
```

Figure 50: Best feature selection II

```python
dff['class'].value_counts()

dff['class'] = dff['class'].replace('Author',1)
dff['class'] = dff['class'].replace('Web',0)
dff['class'] = dff['class'].replace('RP',0)
dff

dff = dff.drop(['No'],axis=1)
dff.columns
dff['class'].value_counts()

#!pwd

train_x, valid_x, train_y, valid_y = model_selection.train_test_split(dff['cleaned_text'],
                                                                      dff['class'],test_size=0.2,random_state=1)

```

Figure 51: Initial class value defining

```python
# create a count vectorizer object
count_vect = CountVectorizer(analyzer='word', token_pattern=r'\w{1,}')
count_vect.fit(dff['cleaned_text'])

# transform the training and validation data using count vectorizer object
xtrain_count =  count_vect.transform(train_x)
xvalid_count =  count_vect.transform(valid_x)

#see the count_vectorizer as dataframe
xtrain_count_df = pd.DataFrame(xtrain_count.todense(), columns = count_vect.get_feature_names())
xtrain_count_df
# xtrain_count_df[0:1].T[xtrain_count_df[0:1].T[0]>0]

# word level tf-idf
tfidf_vect = TfidfVectorizer(analyzer='word', token_pattern=r'\w{1,}', max_features=5000)
tfidf_vect.fit(dff['cleaned_text'])
xtrain_tfidf =  tfidf_vect.transform(train_x)
xvalid_tfidf =  tfidf_vect.transform(valid_x)

# ngram level tf-idf
tfidf_vect_ngram = TfidfVectorizer(analyzer='word', token_pattern=r'\w{1,}', ngram_range=(2,3), max_features=5000)
tfidf_vect_ngram.fit(dff['cleaned_text'])
xtrain_tfidf_ngram =  tfidf_vect_ngram.transform(train_x)
xvalid_tfidf_ngram =  tfidf_vect_ngram.transform(valid_x)

# characters level tf-idf
tfidf_vect_ngram_chars = TfidfVectorizer(analyzer='char', token_pattern=r'\w{1,}', ngram_range=(2,3),
                                         max_features=5000)
tfidf_vect_ngram_chars.fit(dff['cleaned_text'])
xtrain_tfidf_ngram_chars =  tfidf_vect_ngram_chars.transform(train_x)
xvalid_tfidf_ngram_chars =  tfidf_vect_ngram_chars.transform(valid_x)

```

Figure 52: Defining word level, n-gram level and character level tf-idf

```python
"""## Linear Classifier"""

# Linear Classifier on Count Vectors
accuracy = train_model(linear_model.LogisticRegression(), xtrain_count, train_y, xvalid_count)
print ("LR, Count Vectors: ", accuracy)
print('---------------------------------------------')


# Linear Classifier on Word Level TF IDF Vectors
accuracy = train_model(linear_model.LogisticRegression(), xtrain_tfidf, train_y, xvalid_tfidf)
print ("LR, WordLevel TF-IDF: ", accuracy)
print('---------------------------------------------')


# Linear Classifier on Ngram Level TF IDF Vectors
accuracy = train_model(linear_model.LogisticRegression(), xtrain_tfidf_ngram, train_y, xvalid_tfidf_ngram)
print ("LR, N-Gram Vectors: ", accuracy)
print('---------------------------------------------')


# Linear Classifier on Character Level TF IDF Vectors
accuracy = train_model(linear_model.LogisticRegression(), xtrain_tfidf_ngram_chars,
                       train_y, xvalid_tfidf_ngram_chars)
print ("LR, CharLevel Vectors: ", accuracy)
print('---------------------------------------------')

```

Figure 53: Linear classifier- logistic regression on vectors

```python
"""## Naive Bayes"""

# Naive Bayes on Count Vectors
accuracy = train_model(naive_bayes.MultinomialNB(), xtrain_count, train_y, xvalid_count)
print ("NB, Count Vectors: ", accuracy)
print('--------------------------------------------')

# Naive Bayes on Word Level TF IDF Vectors
accuracy = train_model(naive_bayes.MultinomialNB(), xtrain_tfidf, train_y, xvalid_tfidf)
print ("NB, WordLevel TF-IDF: ", accuracy)
print('--------------------------------------------')


# Naive Bayes on Ngram Level TF IDF Vectors
accuracy = train_model(naive_bayes.MultinomialNB(), xtrain_tfidf_ngram, train_y, xvalid_tfidf_ngram)
print ("NB, N-Gram Vectors: ", accuracy)
print('--------------------------------------------')


# Naive Bayes on Character Level TF IDF Vectors
accuracy = train_model(naive_bayes.MultinomialNB(), xtrain_tfidf_ngram_chars, train_y, xvalid_tfidf_ngram_chars)
print ("NB, CharLevel Vectors: ", accuracy)
print('--------------------------------------------')
```

Figure 54: Naive Bayes Classifier on vectors

```python
"""## SVM """

# SVM Classifier on Count Vectors
accuracy = train_model(svm.SVC(), xtrain_count, train_y, xvalid_count)
print ("SVM, Count Vectors: ", accuracy)
print('--------------------------------------------')


# SVM Classifier on Word Level TF IDF Vectors
accuracy = train_model(svm.SVC(), xtrain_tfidf, train_y, xvalid_tfidf)
print ("SVM, WordLevel TF-IDF: ", accuracy)
print('--------------------------------------------')


# SVM Classifier on Ngram Level TF IDF Vectors
accuracy = train_model(svm.SVC(), xtrain_tfidf_ngram, train_y, xvalid_tfidf_ngram)
print ("SVM, N-Gram Vectors: ", accuracy)
print('--------------------------------------------')


# SVM Classifier on Character Level TF IDF Vectors
accuracy = train_model(svm.SVC(), xtrain_tfidf_ngram_chars, train_y, xvalid_tfidf_ngram_chars)
print ("SVM, CharLevel Vectors: ", accuracy)
print('--------------------------------------------')
```

Figure 55: SVM classifier on vectors

References

What is plagiarism 2020, viewed 27 August 2020, < https://plagiarism.org/article/what-is-plagiarism/ >.

Keegan, L., 2020. 79+ Staggering Plagiarism Statistics REVEALED! [2020]. [online] SkillScouter. Available at: <https://skillscouter.com/plagiarism-statistics/> [Accessed 25 November 2020].

Alzahrani, S. M. *et al.* (2012) 'Understanding Plagiarism Linguistic Patterns , Textual Features , and Detection Methods', 42(2), pp. 133–149.

Bensalem, I., Rosso, P. and Chikhi, S. (2014) 'Intrinsic plagiarism detection using N-gram classes', *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1459–1464. doi: 10.3115/v1/d14-1153.

Bensalem, I., Rosso, P. and Chikhi, S. (2019) 'On the use of character n-grams as the only intrinsic evidence of plagiarism', pp. 1–31.

Elamine, M., Mechti, S. E. and Belguith, L. H. (2017) 'Intrinsic detection of plagiarism based on writing style grouping', *CEUR Workshop Proceedings*, 1988.

Gipp, B. (2015) *Doctoral Thesis : Citation-based Plagiarism Detection : Applying Citation Pattern Analysis to Identify Currently Non-Machine-Detectable Disguised Plagiarism in Scientific Publication ... The book version of the thesis is available from Springer Vieweg Res*.

Kakkonen, T. and Mozgovoy, M. (2010) 'Hermetic and web plagiarism detection systems for student essays-an evaluation of the state-of-the-art', *Journal of Educational Computing Research*, pp. 135–159. doi: 10.2190/EC.42.2.a.

Kanjirangat, V. (2016) *J estr*. doi: 10.25103/jestr.095.02.

Kestemont, M., Luyckx, K. and Daelemans, W. (2011) 'Intrinsic plagiarism detection using character trigram distance scores notebook for PAN at CLEF 2011', *CEUR Workshop Proceedings*, 1177(January).

Kuznetsov, M. *et al.* (2016) 'Methods for intrinsic plagiarism detection and author diarization', *CEUR Workshop Proceedings*, 1609, pp. 912–919.

Liu, X. (2013) ‘Full-Text Citation Analysis : A New Method to Enhance’, *Journal of the American Society for Information Science and Technology*, 64(July), pp. 1852–1863. doi: 10.1002/asi.

Luyckx, K. *et al.* (2008) ‘<C08-1065.pdf>’, (August), pp. 513–520.

Meyer Zu Eissen, S., Stein, B. and Kulig, M. (2007) ‘Plagiarism detection without reference collections’, *Studies in Classification, Data Analysis, and Knowledge Organization*, (January 2006), pp. 359–366. doi: 10.1007/978-3-540-70981-7_40.

Oberreuter, G. and Velásquez, J. D. (2013) ‘Expert Systems with Applications Text mining applied to plagiarism detection : The use of words for detecting deviations in the writing style’, *Expert Systems With Applications*, 40(9), pp. 3756–3763. doi: 10.1016/j.eswa.2012.12.082.

Polydouri, A. *et al.* (2020) ‘An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection’, *Evolving Systems*, 11(3), pp. 503–515. doi: 10.1007/s12530-018-9232-1.

Rexha, A. *et al.* (no date) ‘Towards Authorship Attribution for Bibliometrics using Stylometric Features’.

Sari, Y., Stevenson, M. and Vlachos, A. (2018) ‘Topic or Style ? Exploring the Most Useful Features for Authorship Attribution’, pp. 343–353.

Selection, S. M., According, A. and Orientation, T. C. (2015) ‘Stylometry Metrics Selection for Creating a Model for Evaluating the Writ- ing Style of Authors According to Their Cultural Orientation’, 19(3), pp. 107–119. doi: 10.12948/issn14531305/19.3.2015.10.

Wijaya, A. (2015) ‘Two-Step Cluster based Feature Discretization of Naive Bayes for Outlier Detection in Intrinsic Plagiarism Detection’, *Journal of Intelligent Systems*, 1(1), pp. 1–8.

Zhao, Y. and Zobel, J. (2007) ‘Searching with style: Authorship attribution in classic literature’, *Conferences in Research and Practice in Information Technology Series*, 62, pp. 59–68.