

## BWIN815: Industry Integration Project

### PROJECT REPORT

Enhancing and Assessment of the Forensic Linguistic Software Tool  
(2020)

|             |  |
|-------------|--|
| Prepared by | Gervian Du Preez, Charles Henry White, Wian Spamer |
| Date        | May 2020   |

|                     |  |
|---------------------|--|
| Client              | Mr. Zander Janse van Rensburg/Prof Marlene Verhoef |
| Client Organisation | NWU Writing Center                                 |
| Project Team        | Gervian Du Preez, Charles Henry White, Wian Spamer |
| Project Supervisor  | Mr. Robert Maxwell                                 |
| Project Number      | Project 3  |

Report submitted in partial fulfilment of the  
requirements for the degree M.Sc. (BMI) at  
the North-West University.



Centre for BMI  
Senteramo ya DII  
Sentrum vir BWI



## APPROVAL SHEET

**TITLE:** Enhancing and Assessment of the Forensic Linguistic Software Tool (2020)

**CLIENT:** Mr. Zander Janse van Rensburg/Prof Marlene Verhoef

**ORGANISATION:** NWU Writing Centre

**DOCUMENT NUMBER:** Project 3

**CLASSIFICATION:** Restricted

**SYNOPSIS:** The NWU Writing Centre has a forensic linguistic tool "PlagR" which they would like to enhance by improving the user interface and ensuring the program is glitch free. They would also like to assess the accuracy of the metrics provided by the assessment of the tool. Lastly, they also want to have the confidence to present this tool and its findings to an external forensic investigation committee.

**KEYWORDS:** Forensic Linguistic, Enhance, Asses

**PREPARED BY:** Charles Henry White, Gervian Du Preez, Wian Spamer

**DOCUMENT VERSION:** FINAL

**COURSE:** BWN815: Industry Integration Project

---

**APPROVED BY:** **DIRECTOR: CENTRE FOR BMI**

Robert Maxwell

**DATE:** May 2020

**DISTRIBUTION:**

???



***This page has been left blank to group the different sections of the report***



## EVALUATION FORM

### BWN815: Industry Integration Project

This work has been done as part of the requirements for the degree MSc (BMI).

| Evaluation             | Responsibility* | Weight | Mark (%) |
|------------------------|-----------------|--------|----------|
| <b>Academic</b>        |                 |        |          |
| Report quality         | APS, BCD        | 40     |          |
| <b>Business</b>        |                 |        |          |
| Project Management     | CPO             | 10     |          |
| Business Applicability | CPM, CPS        | 50     |          |
| <b>Overall</b>         |                 |        |          |

**CPM: Client Project Manager:**

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Name: \_\_\_\_\_

**CPS: Client Project Sponsor:**

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Name: \_\_\_\_\_

**APS: Academic Project Supervisor:**

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Name: \_\_\_\_\_

**BCD: BMI Centre Director:**

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Name: \_\_\_\_\_

EVALUATION FORM



***This page has been left blank in order to group the different sections of the report***





## ACKNOWLEDGEMENTS

We would like to express gratitude towards Mr. Robert Maxwell, our project supervisor, for his guidance, encouragementsa and constructive recommendations throughout this project. We would also like to thank Mr. Zander Janse van Rensburg for his willingness to consistently engage with the project team and provide useful comments during the project. Lastly, we would like to thank all parties for their continuous empathetic cooperation and engagement during the Covid-19 pandemic.



***This page has been left blank in order to group the different sections of the report***

## TABLE OF CONTENTS





## TABLE OF CONTENTS

|  |    |
|--|----|
| ACKNOWLEDGEMENTS .....                     | 1  |
| TABLE OF CONTENTS .....                    | 3  |
| LIST OF ABBREVIATIONS .....                | 4  |
| EXECUTIVE SUMMARY .....                    | 5  |
| 1 INTRODUCTION .....                       | 6  |
| 1.1 Tool Enhancement .....                 | 6  |
| 1.2 Tool Assessment .....                  | 6  |
| 2 PROJECT DESCRIPTION AND BACKGROUND ..... | 7  |
| 3 METHODOLOGY .....                        | 12 |
| 3.1 Diagnostics .....                      | 12 |
| 3.2 Reporting and Documentation .....      | 13 |
| 3.3 Credibility .....                      | 13 |
| 3.4 Usability .....                        | 14 |
| 4 RESULTS .....                            | 15 |
| 4.1 Diagnostics .....                      | 15 |
| 4.2 Reporting and Documentation .....      | 16 |
| 4.3 Part 3: Methodology Report .....       | 16 |
| 4.4 Usability .....                        | 16 |
| 5 Conclusions and Recommendations .....    | 18 |
| 6 APPENDIX A: Code .....                   | 1  |
| 7 APPENDIX B: Results Output .....         | 2  |
| 2  |    |



## LIST OF ABBREVIATIONS

|     |                               |
|-----|-------------------------------|
| FL  | : Forensic Linguistic         |
| NWU | : North West University       |
| SPT | : Student Project Team        |
| PTC | : Pairwise Textual Comparison |
| SC  | : Stylometric Clustering      |
| NLP | : Natural Language Processing |



## EXECUTIVE SUMMARY

The project provided by the NWU Writing Centre concerns the use of a web-based application to aid in investigations of Academic Misconduct, specifically misconduct in the form of plagiarism. The writing Centre already possess such an application called PlagR, which allow them do identify and analyze the similarity between documents, through the Pairwise Comparison feature and to identify the different writing styles in a document with the Stylometric analysis feature.

The current problem presented to the SPT is that the client finds the existing application confusing to use and that not all its functions seem to work properly. Furthermore, the client is not confident enough in the application that they can present it and its findings to an external forensic investigation committee. Lastly, they are unsure about the accuracy of the metrics provided by the application.

To complete the project successfully the SPT determined that they had to do the following. They had to analyze the code of the application to detect and fix any errors that may cause certain functions to work properly, improve the user interface so that it is easier to use and to provide a report on the methodology used in the application along with an analysis on the accuracy of the tool.

In completion of the project the SPT discovered the following:

They found that the code was undocumented and not logically organized making it difficult to understand and improve upon. This made the application unsustainable since any changes in the packages used will require the code to be drastically altered if the code needs to be worked on. This was the case when the SPT tried to analyze the code but found that certain features that worked on the implemented code on the Shiny server did not work when used on a local computer. This is because the 'quanteda' package was updated causing errors in the calculation of the similarity metrics in the pairwise feature. Furthermore, while the SPT struggled to make changes to the UI due to difficulties in understanding the code enough to alter it, they did manage to fix the error in exporting the similarity results.

In response to the difficulties surrounding the code the SPT determined it would be better to provide the client with a user guide on how to navigate the UI instead of changing it along with a methodology report on the packages used as a source on the credibility of the applications results. This is to give the client confidence in its accuracy and in presenting the application and its findings to an external forensic investigation committee. Furthermore, they concluded it would be beneficial for future projects to organize and document the code to the best of their ability.

This report summarizes the steps taken by the SPT through the course of the project.



# 1 INTRODUCTION

To give the reader, the necessary background, provided is an overview of the clients request from the NWU Writing Centre. The Writing Centre has a forensic linguistic tool called “PlagR” and they wish to:

- Enhance this tool.
- Asses the accuracy of the tool

## 1.1 Tool Enhancement

The tool provided does not operate perfectly, it had a few major glitches that prevents the Writing Centre from successfully implementing it. This reduced the confidence the user has in the tool to present the results provided to the user. The client was not completely satisfied with the User-Interface (UI) as it was difficult to use and did not provide any external guide to facilitate this process. He wanted it to be redesigned to provide an improved understanding of the tool to the user, ultimately improving the User-Experience (UX). The tool provides the user with various metrics that have been implemented for analysis. These metrics are of no relevance as the user a) does not understand metrics and b) does not know how to interpret them. The client also requested an extractable report which incorporates the findings in a fathomable manner.

## 1.2 Tool Assessment

The client requested that the metrics deployed be assed for accuracy and reliability. He requested the assessment be done in such a manner to provide him with the confidence to present the findings in-front of a forensic linguistic external panel. A thorough diagnostic to be done on the code to ensure reliability of the program.

To conclude the client had the following issues:

- Potential errors in the program
- Difficult to use
- Question credibility
- Ineffective input and output

In the next section we will give the entire problem context by providing a description of background. We will describe the applications in detail, by discussing each feature in detail.

## 2 PROJECT DESCRIPTION AND BACKGROUND

In this section the phases of the implementation of the application will be discussed. Commencing with academic misconduct at NWU. Then describing the macro process of the investigation and lastly ending off with a detailed description of how to use the application.

We start of by explaining the process of academic misconduct at the NWU. The process consists of a five-phase process as is depicted in figure 1.

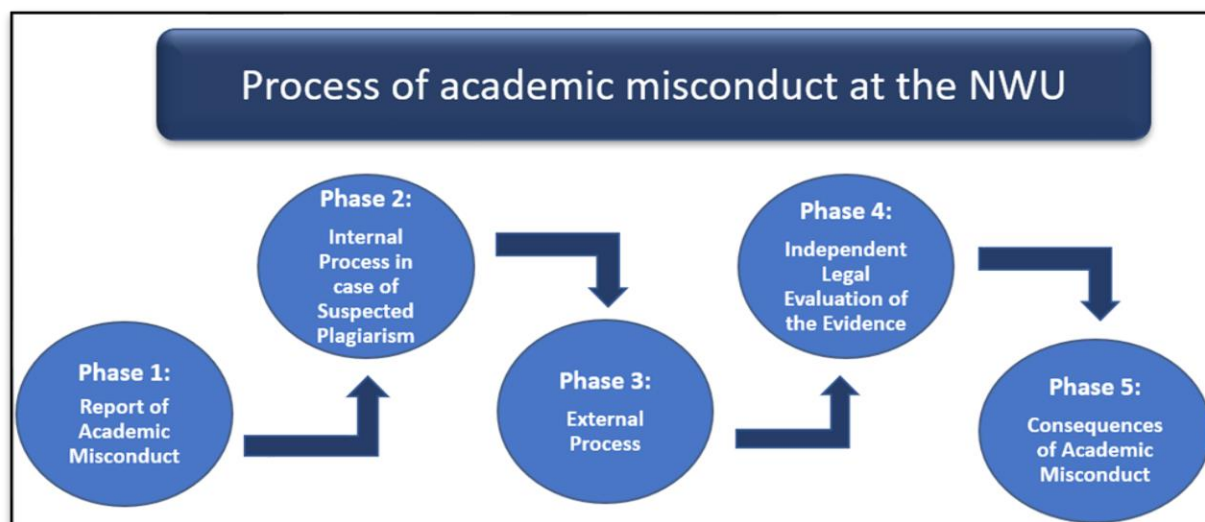


Figure 1: Process of Academic Misconduct at the NWU

Where we will focus on phase 2 as our client is an internal forensic linguistic investigator and is responsible for determining plagiarism and authorship in suspected text. The internal process goes as follows; a forensic linguistic expert first takes the alleged plagiarised text and submits it to “Turnitin” (A commercial, internet-based plagiarism detection service). To obtain a list of documents that might have been plagiarised from. The client then does further analysis on the “PlagR” tool. This entire process is depicted in figure 2.

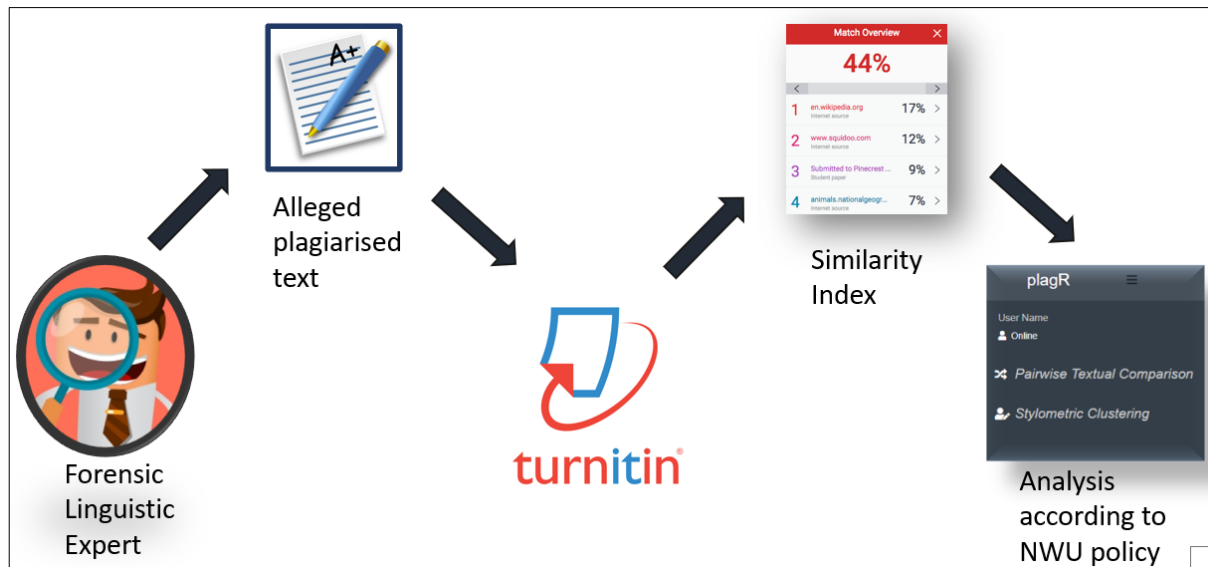


Figure 2: Internal Process in the Case of Suspected Plagiarism

The report proceeds by explaining the tool “PlagR” in detail. The tool has two main features, “Pairwise Textual Comparison” PTC and the “Stylometric Clustering” SC feature. The PTC is used for comparing textual similarity between suspected text and evidence text. The second feature is the SC which is used to identify authorship in a suspected text.

We describe the PTC which is as follows:

1. First submit the suspected text

The tool offers the user two methods of doing this, either the user can upload a suspected document, or the user can paste in suspected text in provided text box. By pressing calculate the user then proceeds to macro analyses which displays the submitted text and highlights the suspected plagiarised sections which are also hyperlinked. The Turnitin button refers the user to the “Turnitin” website. This can all be seen in figure 3.1

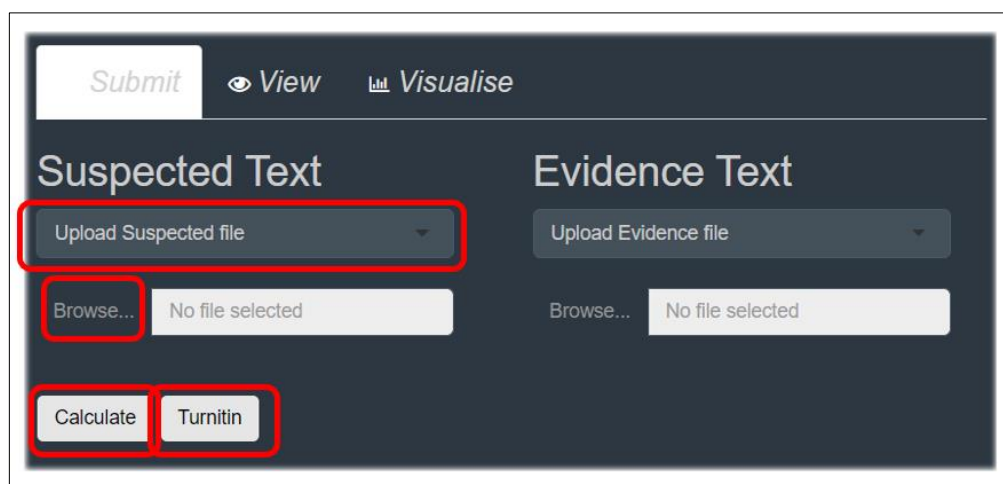


Figure 3: Pairwise Textual Comparison Submit Tab

## 2. Paste suspected specific text

Under the view tab the user is presented with two columns the suspected and evidence piece. The highlighted text is suspected plagiarized text and the user is required pasting this text in the empty text box. Pasted text can then be further analysed to determine the extent of similarity between suspected and evidence text. The user then clicks on the compare button to do micro analysis. Figure 3.2 illustrates this process.

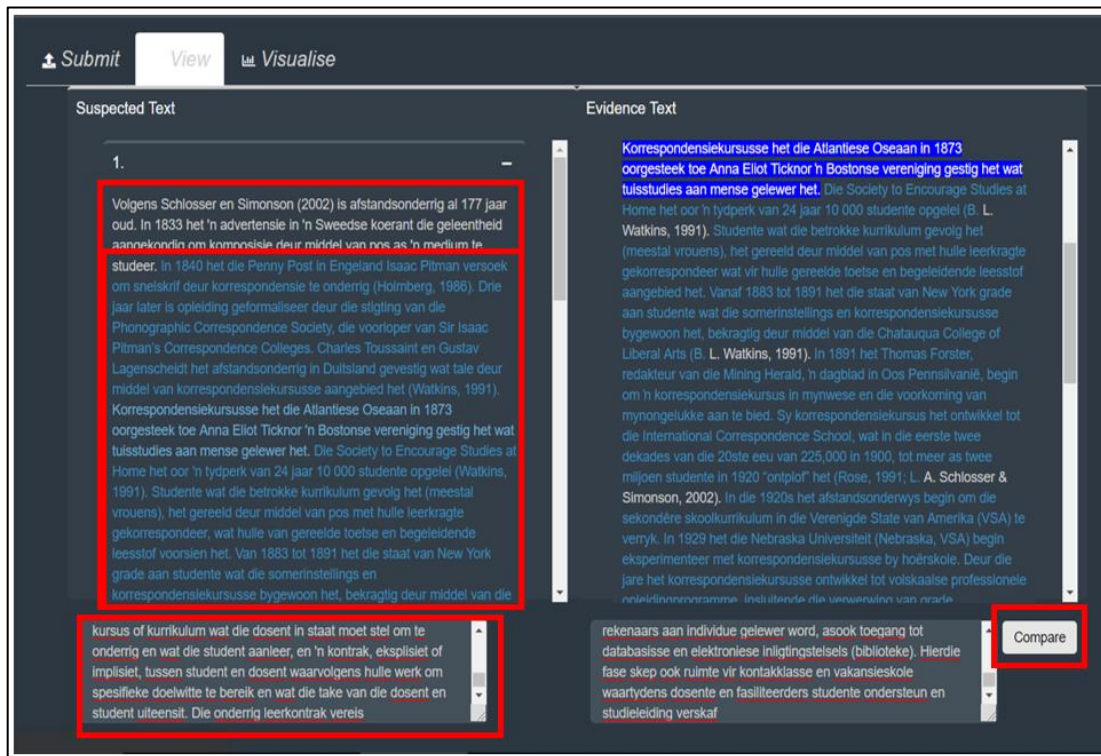


Figure 4: Pairwise Textual Comparison View Tab

## 3. Obtain similarity values

Micro analysis is then done based on selected similarity measures by clicking on the calculate button. These values determine the extent of plagiarism based on various statistical methods. Results can then be exported to a word document post micro analysis by clicking on the export button. See figure 3.3.



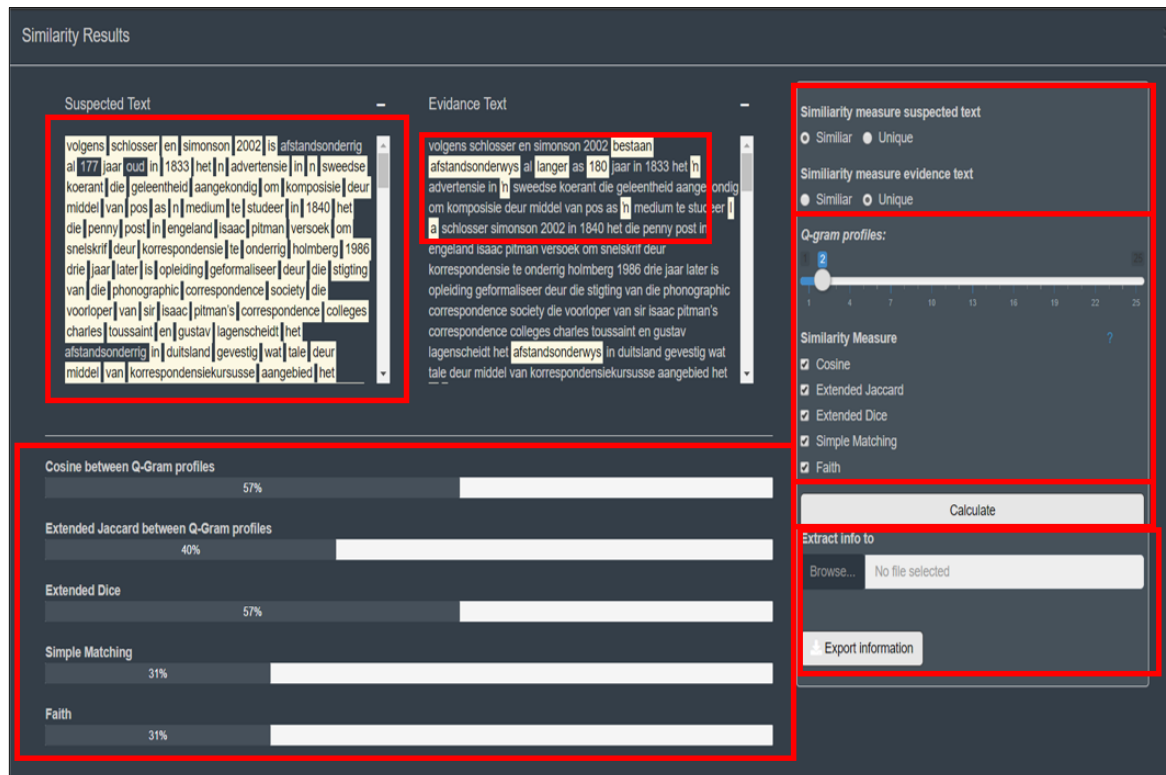


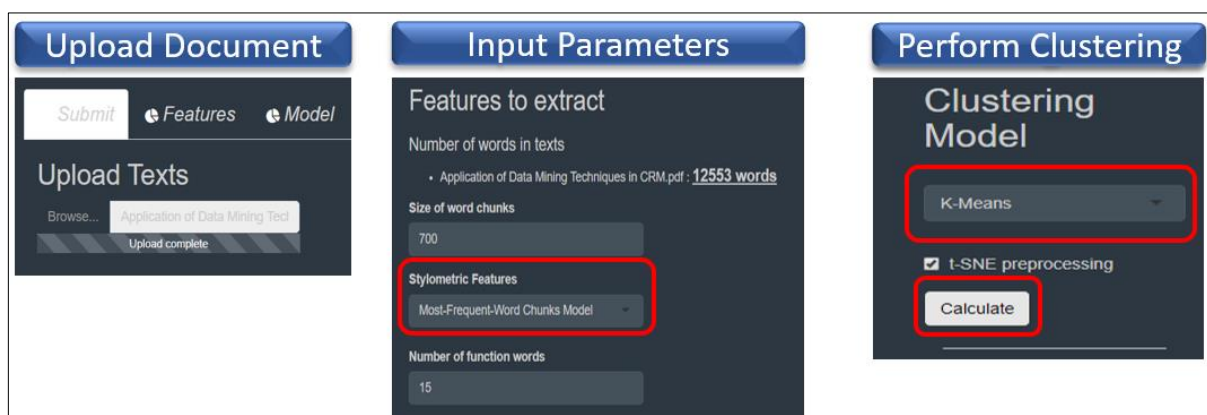
Figure 5: Pairwise Textual Comparison Similarity Results

Regarding the PTC the client was satisfied with the uploading of text, for the macro analysis section the client wanted automated text copying. For the Micro analysis, the client wanted to validate the accuracy measures, explain the functionality to the user, set a benchmark for measures and export results into a report.

Having explained the PTC feature, next we will explain the SC and how it is used to identify authorship.

The process for using the SC is as follows:

1. Upload the Document you wish to analyse
2. Input the various parameters
3. Select the clustering algorithm you wish to use and click calculate button







*Figure 6: Stylometric Clustering Process*

The clients requests for the Stylometric Clustering is to be able to determine the accuracy of the clustering algorithms implemented. The client also requested a technical report explaining these findings.

In the background we discovered that there are:

- Potential errors in the program
- That the client finds the program difficult to use
- Discontentment regarding the credibility of the program
- Ineffective input and output

This concludes the brief description of the background of the project and next the methodology will be described



### 3 METHODOLOGY

In this section a detailed description is given regarding the method of approach for enhancing and assessing the provided forensic linguistic tool. The methodology was broken down into four major components:

- Diagnostics
- Reporting and Documentation
- Credibility
- Usability

The SPT commenced by assuming that the PlagR tool was completely operational, and as a result could split up the work of the four agenda points raised by the client between the team and start solving them. This was not the case the PlagR tool had plenty of bugs which have been corrected, but this hindered the team from making effective progress. The team first had to direct all attention to ensure the program is completely operational before evaluation of metrics could be done as well as the improvement of the UI and UX.

#### 3.1 Diagnostics

The “PlagR” program was developed in a programming language called R. Upon running and testing the program the SPT discovered the code provided had certain glitches that needed fixing. To execute this the code provided needed to be understood. **As the +2000 lines of code provided was undocumented, unstructured and no guide to understand it was provided.** Receiving the code was initially a problem and post receiving after diagnosing it we realized inconsistency between the program on the shiny server and when running the program on a local server R-studio. After the discovery the team then had to wait another three weeks to gain access to the code published on the shiny server.

The SPT commenced as follows:

1. Met with the initial developer, Hugo Loubsher on two separate occasions. On the 27 of February in the SAS labs and on the 15 of March in Johannesburg. On the first occasion H. Loubsher gave the SPT a brief introduction to the code as well as optimal strategy to understanding it. |
2. **Gain knowledge** about R-shiny, a R package that enables one to develop web based interactive software.
3. The second meeting on the 15<sup>th</sup> of March H. Loubsher met with two members to systematically work through the code after having worked through the shiny package.



4. Having the R-shiny knowledge and being familiar with the code it enabled us to **sort** and **organize** the code.
5. **Gain knowledge** about other packages in R, a package is a collection of functions that increase the power of R.
6. After sorting and organizing we were able to **delete unnecessary code** that was not implemented as well as correct code that was not working due to changes in R packages.
7. While doing steps 4, 5 and 6 the SPT realized a package was not updated causing errors in the calculation of the similarity metrics in the pairwise feature
8. Lastly, we were tasked with **ensuring perfect functionality** of the tool and updating the code on the shiny server.

A snip bit of the final code can be found in Appendix A

### 3.2 Reporting and Documentation

Tasked with designing an extractable report of result out from the “PlagR” tool.

The SPT commenced as follows:

1. The layout of the report had to be designed in logical manner by the SPT.
2. The layout was designed to provide an extractable piece of the suspected text and the provided metrics with results.
3. The SPT was tasked with implementing the code and updating it on the shiny server.
4. Lastly the team had to ensure that the implemented code is functioning, and report is correct,

A snip bit of the report can be found in Appendix A

### 3.3 Credibility

The tool when used correctly as described in the background provides the user with certain results to make certain conclusions. These results and the entire program are derived from various R packages. R packages are collections of functions and data sets developed by a community and improve R base functionalities. To provide the user with some credibility the SPT took the following steps:

1. Do thorough **research** on all:
  - **Accuracy metrics** provided in by the PTC feature in the tool.



- Tool **functionality** available, e.g “size of word chunks” provide and user with an intuitive explanation of this and how to proceed using this function.
  - 24 R packages implemented in the program.
2. Discussed the findings with the client.
  3. **Developed a report** about the
    - Scientific methodology that the FL tool uses incorporating all aspects under the first section.
    - How to present and interpret the interpret the results to an external forensic linguistic committee.
  4. Presented this and discussed with the client.

The report has been developed and delivered to the client.

### 3.4 Usability

To give the user more comfort getting around with the tool and optimizing the user experience the SPT developed:

- Video tutorial, explaining in detail to the user how to go about using the program.

The tutorial video has been developed and delivered to the client.

In this section we described in depth the approach of the four components, Diagnostics, Reporting and Documentation, Credibility, Usability that were executed by the SPT. The next section will discuss the results obtained from the approaches followed.



## 4 RESULTS

In this section we discuss our results in the following three parts. Part 1 concerns the analysis of the code and the changes made. In Part 2 we discuss the User Interface and present an alternative to changing the UI, while in Part 3 we present our findings on the methodology of the packages used in the application.

### 4.1 Diagnostics

Through our analysis of the code we found that it was undocumented and not logically organized. This made the application unsustainable for future development since any changes in the packages used will require the code to be drastically altered, which will require a detailed understanding of the code. This will be difficult to do for anyone other than the original coder if the code is not properly documented.

This was the case during the project when we tried to run the application code on a local computer. They found that features that worked properly on the implemented code on the Shiny server, accessed through a web browser, failed to work on a local computer. The features in question include the hyperlinking of similar paragraphs between documents during Macro Analysis and the calculation of the similarity metrics in the Micro Analysis. We found that this was due to local computers always using the latest version of a package while the Shiny server used the version at the time the code was uploaded to this server and that the newest version of the 'Quanteda' package, which is the main package used by the application, changed the way it calculated the similarity metrics.

The only feature that we found that did not work on the server code was the exporting of the similarity results obtained during Micro analysis to a Micro Soft Word document.

In terms of changes to the existing code, we first altered the code so that it performed the same on a local computer as on the server and fixed the exporting of the similarity results obtained during Micro Analysis. An example of such a document can be found in appendix B.

This is only a temporary solution since any future developments to the application will require the code to be altered again for it to work on a local computer. Therefore, in terms of code analysis, our focus was on documenting and organizing the code to aid any future developments on the application by improving the understanding of the code. An example of the organized code can be found in appendix A.



## 4.2 Reporting and Documentation

This feature corresponds to the PTC feature and exports the highlighted suspected text produced by the Macro Analysis function. To make use of this feature a blank word document will be required and saved on local where it will be allocated.

- The exporting document will be a word document
- Will consist of two main sections the pasted section highlighting unique or similar words and the output results.
- The pasted text being the top part of the document and the output results being the lower section.
- The results section consisting of the five metrics.
- To execute this the SPT first had to successfully implement the code in the “PlagR” program on a local server to test it, as well as upload it onto the shiny server.

An example of such a document can be found in appendix B.

## 4.3 Part 3: Methodology Report

During the project we were unable to provide sensible interpretation for the accuracy metrics for the results produced by the application to determine plagiarism or authorship successfully and consistently. After thorough research and discussion between the SPT and client was done the two parties that with the current statistical development regarding NLP no framework can be implemented just using the metrics provided to determine whether plagiarism or identification of authorship can be determined. Therefore, we decided after communicating with the client that as an alternative we will provide a report outlining the methodology of the different packages used in the application shown in appendix. This will aid in understanding and interpreting the results produced by the application.

The methodology report will also give the client the needed confidence in the application that they can present it and its findings to an external forensic investigation committee.

## 4.4 Usability

As mentioned in section 4.1, the code that was given to us was undocumented, this made it difficult to change the current UI. Therefore, we determined that it will be better if we provided the client with a user guide, in the form of a video demonstration, on how to navigate the application. Explaining all the available features in the tool and exactly how to implement them correctly and effectively.

The SPT created three video tutorials for the client:



1. Showing the user how to access the “PlagR” tool on the shiny server using a shiny server URL.
2. How to go about using the PTC feature in tool. Beginning with uploading a published document then showing the results output. As well as explaining each function as the tutorial proceeds ending off with showing the exporting button output.
3. Explains to user how to use the SC feature, explaining each function implemented as well as how to upload a document.

This will not only help the client to understand how the application works, but it can be distributed to others that wish to use the application in the future.

Having explained what the results produced by, as well as explaining what business value our efforts have produced. Lastly, we discuss the conclusions and recommendations to aid the client with more assurance and confidence in this tool



## 5 Conclusions and Recommendations

The document compiled has thus attempted to provide the reader with insight on the findings of the project titled 'Enhancing and Assessment of the Forensic Linguistic Tool (2020)'.

The accuracy metric research has shown that the "PlagR" can be relevant in identifying academic misconduct but that the tool still requires a lot of work. It also showed that this field NLP is an emerging field in A.I, therefore constant updates and improvements to the tool will be required.

The SPT accepts that this is an in-complete project but believes that the findings presented will assist in identifying problems that needs to be addressed. Below are some recommendations for possible future implementation to alleviate some of the problems that were detected.

The **Diagnostic** aspect of the project revealed to the SPT the difficulties in using R and its supported packages. Packages are continuously update and slightly modified which diminishes the reliability of code being successfully executed over extended periods of time. Another difficulty in connection with using R is the 'R-shiny' annual server fees which are in USD and as a result heavily dependent on the exchange rate. Due to the above-mentioned reason the SPT concluded that **the program be converted into a Python application**. The SPT team attempted to do this but server fees cost concerns were only brought during the Project Proposal meeting leaving us with insufficient time to execute.

The second recommendation would be to implement **paraphrasing identification features** into the "PlagR" tool. This recommendation is since paraphrasing is trained to predict the next word. Which provides the user with a benchmark to evaluate the other "PlagR" features on. Offering the user with more confidence in results provided.

The last recommendation the SPT suggests is to **train the language model on the boloka repository**. This will aid the model in identifying the authorship of suspected





text provided and will deeply aid the accuracy of doing this. Providing the client with the increased confidence to present his findings to an external panel.





## 6 APPENDIX A: Code

### A snippet of the structured and documented code

```
#####
# 1.1) Declare inputs as reactive #
#####

# file or word text input
susp_input_typeInput <- reactive({input$susp_input_type})
evid_input_typeInput <- reactive({input$evid_input_type})
|
#####
# 1.2) Change UI for different input types #
#####

# input method ui - Suspected text
output$susp_method <- renderUI({
  si <- susp_input_typeInput()
  # paste text
  if (si == "clip") {
    textAreaInput("susp_raw", NULL,
                  placeholder = "Suspected text...", width="100%")
  }
  # upload file
  else {
    tagList(
      fileInput("susp_files", NULL, multiple = TRUE,
                accept = c(".pdf", ".txt", ".docx", ".doc"))
    )
  }
})

# input method ui - Evidence text
output$evid_method <- renderUI({
  ei <- evid_input_typeInput()
  # paste text
```



## 7 APPENDIX B: Results Output

### Results

| Suspected.Text   | Evidence.Text   |
|--|---|
| Cosine between Q-Gram profiles   | 33.97%  |
| Extended Dice  | 32.43%  |
| Extended Jaccard between Q-Gram profiles   | 19.35%  |
| Simple Matching  | 15.72%  |
| Faith  | 15.72%  |
| Similiar   | Similiar  |
| then one wishes to maximize the a long position<br>the wheel proportion between its index and return<br>function the weighted sum of the assets expected<br>rate the index of the next long position plus the<br>proportion of of return while minimizing the risk the<br>standard deviation any enclosed short position<br>wedges is the total proportion of the portfolio rate of<br>return since this defines the level of resource<br>allocation for that holding the idea is that for of<br>uncertainty about the future payoff at a certain<br>time example the resources from a short sale of a<br>stock are used to there are also different<br>constraints depending on the type purchase<br>additional shares of the long position stock whose<br>problem to be solved | since the knapsack problem asset represented as a<br>long position the wheel proportion has been solved<br>using eas the authors adopted this encoding<br>between its index and the index of the next long<br>position plus in addition to the vector of decision<br>variables the weights the proportion of any enclosed<br>short position wedges is the so each bit from the<br>knapsack determined if an asset would total<br>proportion of resource allocation for that holding |