

A CONCEPTUAL FRAMEWORK FOR MINING AND ANALYSIS OF SOCIAL MEDIA DATA

BY

SETHUNYA ROSIE JOSEPH

Reg. No: 14100067

BSc Computer Information System (UB) MLIS (UB)

Department of Computer Science and Information Systems

College of Sciences

Botswana International University of Science and Technology (BIUST)

sethunya.joseph@studentmail.biust.ac.bw

A Thesis Submitted to the College of Sciences in Partial Fulfilment of the Requirements for the
Award of the Degree of Doctor of Philosophy in Computer Science of BIUST

Supervisor : Dr Hlomani Hlomani

Department of Computer Science and

Information Systems

BIUST

hlomanihb@biust.ac.bw

Tel: +267 4931107

Signature: H. B. Hlomani.....

Date: .. 09 / 08 / 2018.....

Supervisor : Dr Keletso Letsholo

Department of Computer Science and

Information Systems

BIUST

letsholok@biust.ac.bw

Tel: +267 4931102

Signature: Keletso Letsholo.....

Date: .. 09 / 08 / 2018.....

August 2018

DECLARATION

I, **SETHUNYA ROSIE JOSEPH**, declare that this thesis is my own original work and that it has not been presented and will not be presented to any other university for a similar or any other degree award.

Signature ..S.R. JOSEPH.....

COPYRIGHT

This thesis is a copyright material protected under the Berne Convention, the Copyright Act of 1999 and other international and national enactments, in that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgement, without the written permission of the office of the DVCAA, on behalf of both the author and the BIUST.

CERTIFICATION

The undersigned certifies that he has read and hereby recommends for acceptance by the College of Sciences titled: **A CONCEPTUAL FRAMEWORK FOR MINING AND ANALYSIS OF SOCIAL MEDIA DATA** in fulfilment of the requirements for the Degree of Doctor of Philosophy (in Computer Science) of the BIUST.



Dr Hlomani Hlomani

(Supervisor)

Date: .09 / 08 / 2018.. ..



Dr Keletso Letsholo

(Co. supervisor)

Date: .09 / 08 / 2018.. ..

ACKNOWLEDGEMENTS

First of all, I would like to thank God for allowing me to carry on till the last minute, it was not an easy road. I am grateful for the support BIUST gave me to pursue this degree. I could have not reached this point without the financial support BIUST gave me.

I would like to thank my supervisors Dr Hlomani and Dr Letsholo for their constant support, guidance and the continuous interest they accorded me to complete my thesis. This task would not have been conceivable without their advice. The thanks are also passed on to the entire Department for all the support they gave for this thesis to be a success.

I would like to extend the thanks to all the year 2 students (class of 2016) at the Department of Computer Science and Information Systems who helped in making this study a success by volunteering to take part in the focus groups. Their commitment, focus, sacrifice, support and determination helped me achieve my objectives. Special thanks go to my colleagues and wonderful friends: Kutlwano Sedimo and Freeson Kaniwa, who were also part of these focus groups. I would also like to thank them for their continuous encouragement and support throughout my study.

Much gratitude and special thanks are also passed to all my friends especially Aubrey, Kago, Sithole and Candice for their continuous encouragement and support. I cannot find adequate words to express how essential their inspiration and support was, to bring me to this point. I would not be here if it was not for their continuous encouragement, support and thoughtful advice throughout my quest for PhD.

Finally, I am forever indebted to my family for their understanding, endless patience and encouragement when it was most required. I owe special thanks to my parents for their financial and moral support, continuous prayers, and endless patience.

DEDICATION

To my family

TABLE OF CONTENTS

DECLARATION	II
COPYRIGHT	III
CERTIFICATION	IV
ACKNOWLEDGEMENTS.....	V
DEDICATION.....	VI
TABLE OF CONTENTS.....	VII
LIST OF TABLES.....	IX
LIST OF FIGURES	XI
LIST OF ABBREVIATIONS.....	XII
ABSTRACT.....	XIII
CHAPTER 1	1
INTRODUCTION	1
1.1 AN OVERVIEW OF SOCIAL MEDIA	1
1.2 CHARACTERISTICS OF SOCIAL MEDIA DATA.....	3
1.3 MOTIVATION OF THE STUDY	5
1.4 STATEMENT OF THE PROBLEM	6
1.5 MAIN OBJECTIVE	6
1.6 JUSTIFICATION AND SIGNIFICANCE OF THE STUDY	7
1.7 CONTRIBUTIONS	8
1.8 RESEARCH METHODOLOGY	10
1.9 THESIS STRUCTURE.....	14
CHAPTER 2	15
BACKGROUND INFORMATION	15
2.1 INTRODUCTION	15
2.2 DATA MINING AND KNOWLEDGE DISCOVERY	15
2.3 BIG DATA	27
2.4 CHAPTER SUMMARY	29
CHAPTER 3	30
LITERATURE REVIEW	30
3.1 INTRODUCTION	30
3.2 DATA ANALYSIS.....	30
3.3 DATA MINING AND ALGORITHM ANALYSIS.....	41
3.4 FRAMEWORKS FOR SOCIAL MEDIA ANALYSIS (CONTENT ANALYSIS)	47
3.5 CHAPTER SUMMARY.....	52
CHAPTER 4	53

THE PROPOSED CONCEPTUAL FRAMEWORK.....	53
4.1 INTRODUCTION	53
4.2 THE BASE AND PREMISE OF THE FRAMEWORK	53
4.3 OVERVIEW OF THE PROPOSED CONCEPTUAL FRAMEWORK	54
4.4 THE COMPONENTS AND PHASES OF THE PROPOSED CONCEPTUAL FRAMEWORK	57
4.5 CHAPTER SUMMARY	61
CHAPTER 5	63
EVALUATION OF THE PROPOSED CONCEPTUAL FRAMEWORK.....	63
5.1 INTRODUCTION	63
5.2 DATA COLLECTION	66
5.3 DATA PREPARATION	66
5.4 INDUCTIVE CONTENT ANALYSIS	68
5.5 DATA MODELLING AND EVALUATION METRICS	84
5.6 THREATS TO VALIDITY	97
5.7 CHAPTER SUMMARY	98
CHAPTER 6	99
CONCLUSIONS AND RECOMMENDATIONS	99
6.1 INTRODUCTION	99
6.2.OVERALL VIEW OF THE STUDY AS PER RESEARCH OBJECTIVES	99
6.3 CONCLUSION OF THE STUDY	102
6.4 LIMITATIONS AND FUTURE WORKS	102
6.5 CHALLENGES	103
REFERENCES	104
APPENDICES	114
APPENDIX A: FOCUS GROUPS RESULTS FOR INDUCTIVE CONTENT ANALYSIS 114	
APPENDIX B: SAMPLE OF CONSENT FORM USED.....	123
APPENDIX C: THEMATIC CODING STEPS MANUAL.....	124
APPENDIX D: THE ICA USABILITY EVALUATION QUESTIONNAIRE	128
APPENDIX E: SCREENSHOTS ILLUSTRATION OF DATA CLASSIFICATION THROUGH THE MODELS (DATA MINING ALGORITHMS IMPLEMENTED THROUGH WEKA)	131
APPENDIX F: SCREENSHOTS ILLUSTRATION SOME OF THE TEXT PRE- PROCESSING FUNCTIONS APPLIED ON THE DATA FOR MODEL CLASSIFICATION.....	139
APPENDIX G: A JAVA CODE BUILT WITH TWITTER API FOR COLLECTION OF DATA FROM TWITTER	142

LIST OF TABLES

Table 1.1: Examples of social media networks, adapted: [8,32]	2
Table 2.1: An example KDD data prediction sample from a dataset showing data and its Source: [15].....	12
Table 3.1: The strengths, limitations and features of Qualitative Analysis.....	29
Table 3.2: A comparison of approaches used in Qualitative Analysis, adapted:[199].....	31
Table 3.3: The strengths, limitations and features of Quantitative Analysis.....	33
Table 3.4: A comparison between Quantitative research approaches, Source: [201]	34
Table 3.5: A general comparison between Qualitative and Quantitative research, Source: [182].....	36
Table 3.6: A comparison between the strengths and limitation of some of the supervised and unsupervised learning algorithms used in social media data mining and analysis.....	37-39
Table 3.7: A Comparison of the existing Frameworks for Content Analysis.....	43
Table 5.1: Demographic details of the Users who participated in conducting the ICA process	69
Table 5.2: The descriptive statistics analysis of the answers provided by participants on learnability of ICA	70
Table 5.3: The descriptive statistics analysis of the answers provided by participants on Ease of Use	71
Table 5.4: The descriptive statistics analysis of the answers provided by participants on perceived Usefulness	73
Table 5.5: The descriptive statistics analysis of the answers provided by participants on Flexibility of ICA.....	74
Table 5.6: The descriptive statistics analysis of the answers provided by participants on Satisfaction of ICA	76
Table 5.7: Kappa statistics performance of the classifiers: NB; IBK; J48 and SMO attained during classification of data	78
Table 5.8: Error rates of the classifier: NB, IBK, J48 and SMO attained during classification of data.....	79
Table 5.9: Average ROC calculations of the classifier: NB, IBK, J48 and SMO attained during classification of data	80

Table 5.10: Recall values of the classifier: NB, IBK, J48 and SMO attained during classification of data.	81
Table 5.11: TP rate and FP rate of the classifier: NB, IBK, J48 and SMO attained during classification of data.	82
Table 5.12: Precision values of the classifier: NB, IBK, J48 and SMO attained during classification of data.	82
Table 5.13: F1-Measures performance of the classifier: NB, IBK, J48 and SMO attained during classification of data.....	83
Table 5.14: Classification accuracy of the classifier: NB, IBK, J48 and SMO attained during classification of data.....	84

LIST OF FIGURES

Figure 1.1: Saunders’s research process onion, Source: [124].....	10
Figure 2.1: The steps that compose the KDD process, Source: [62]	16
Figure 2.2: Supervised learning process, Source:[15].....	15
Figure 2.3: The classification hyper-plane defines a region for a class +1 and for a class -1, Source:[99].....	18
Figure 2.4: The Optimum Separation Hyperplane, Source: [99].....	22
Figure 2.5: Kernel Mapping from Input Space to Feature Space,Source: [100].....	23
Figure 2.6: A decision tree’s test node (t4) is replaced in entirety with a positive classification (assuming a binary class domain), Source: [106]	25
Figure 3.1: Qualitative and quatitative data in dynamic balance, Source: [21].....	33
Figure 4.1: A collapsed view of the proposed Conceptual Framework for data mining and analysis of the social media data.....	55
Figure 4.2: A detailed view of the proposed Conceptual Framework for data mining and analysis of the social media data.....	50
Figure 5.1: The Evaluation of the proposed Conceptual Framework. Highlighted steps carried out in the evaluation of the proposed Conceptual Framework for analysing the social media data.....	58
Figure 5.2: Model performance in terms of accuracy of data with partial preparation (pre- processing only).....	60
Figure 5.3: Model performance in terms of accuracy of data with complete preparation (pre- & post processing.....	61
Figure 5.4: Categories per each focus group and the final eight selected categories.....	67
Figure 5.5: A general look of a confusion matrix.....	85
Figure 5.6: Confusion matrix of Naïve Bayes classifier performance attained during classification of data.....	85
Figure 5.7: Confusion matrix of SMO classifier performance attained during classification of data.....	86
Figure 5.8: Confusion matrix of IBK classifier performance attained during classification of data.....	87
Figure 5.9: Confusion matrix of J48 classifier performance attained during classification of data.....	88

LIST OF ABBREVIATIONS

API	Application Program Interface.
BCP	Botswana Congress Party.
BDP	Botswana Democratic Party.
BMD	Botswana Movement for Democracy.
BNF	Botswana National Front.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise.
DCA	Deductive Content Analysis.
FN	False Negatives.
FP	False Positives.
IBK	Instance Bases learning with parameter K.
ICA	Inductive Content Analysis.
KDD	Knowledge Discovery in Databases.
K-NN	K-Nearest Neighbour.
MAE	Mean Absolute Error.
MRI	Magnetic Resonance Imaging.
NB	Naive Bayes.
RMSE	Root Mean Squared Error.
ROC	Receiver Operating Characteristic.
SMO	Sequential Minimal Optimization.
SVM	Support Vector Machine.
TN	True Negatives.
TP	True Positives.
UB	University of Botswana.
UDC	Umbrella for Democratic Change.

ABSTRACT

Data has turned to be an essential aspect of every individual, group, industry, financial system, business enterprise, commercial enterprise and society. In this era, where social networks yields inconceivable volumes of data for different purposes on daily bases, an improved approach to analysis of data is required to extract information that best matches user interest. One of the ways of achieving this is through data mining. Data mining could help any consumer or producer of information to make knowledge driven decisions. Data mining helps with the analysis of texts from different perceptions into valuable information that can be fine tuned into useful organisational solutions and help forecast the future trends. There has been some work that has been done previously in the area of social media data analysis. Most of these works place more emphasis on the development of computational methods. Literature has argued that, while computational methods are good with regards to statistical and algorithmic analytics, they are limited in that they cannot capture in-depth meanings and semantics from data. Content from social media requires human interpretation and intervention for proper analysis. Therefore, the aim of this study was to develop a Conceptual Framework that offers a generic approach to addressing the limitations of computational methods to social media data analysis. In addition, the proposed Framework provides a guide on other aspects such as data gathering and text-pre-processing. The application of the proposed Framework was then demonstrated through classification of data. Even though only one application scenario was demonstrated through classification, the framework gives other options for applications such as regression, association and clustering. The proposed Conceptual Framework was evaluated in two stages using an example case of the political landscape of Botswana data collected from Facebook and Twitter platforms. Firstly, a user study was carried through the Inductive Content Analysis (ICA) process using the collected data by the use of focus groups. Additionally, a questionnaire was rolled out to evaluate the usability of ICA as perceived by the participants. This was mainly performed to help mitigate the limitations of computational methods of failure to capture in-depth meanings from data. Secondly, an experimental study was conducted where data which was pre-processed through ICA was classified through the use of data modelling, and an evaluation of data mining algorithms was made with metrics which measured their performance and accuracy. The evaluation of the ICA process was made based on “Usability” component of the ISO 9126 model. The aspects which were evaluated on Usability are: Learnability, Ease of use, Perceived usefulness, Satisfaction and Flexibility. Each evaluation metric had five questions. The total for the positive responses of strongly

agree and agree for the five questions of each metrics are as follows: Learnability scored between 91.7 % and 100%. Ease of use scored between 95.8% and 100% with the exception of one question which asked if the overall process took a lot of time. The majority mostly disagreed with 87.5 %, Perceived ease of Use scored between 91.7% and 100%, Flexibility scored between 95.8% and 100% and Satisfaction scored between 95.8% and 100%. The scores for accuracy of the model performance in classifying the data correctly for Instance Bases learning with parameter K (IBK) classifier for K-Nearest Neighbour and Sequential Minimal Optimization (SMO) classifier for Support Vector Machine were both 94.6%, Naive Bayes (NB) classifier for Naive Bayes multinomial with 82.8% and J48 classifier for Decision tree C4.5 was with 78%. The results from the experimental study show that data mining algorithms had higher accuracy in classifying data when supplied with data from the ICA process. That is, through the application of the ICA process, data mining algorithms were able to overcome the difficulty of failure to capture in-depth meanings and semantics within data. Overall, the results of this study, including the Proposed Conceptual Framework are useful to scholars and practitioners who wish to do some researches on social media data mining and analysis. The Framework serves as a guide to the mining and analysis of social media data in a systematic manner.

CHAPTER 1

INTRODUCTION

1.1 AN OVERVIEW OF SOCIAL MEDIA

Social media, as defined by Batrinca and Treleaven [1], is a web based and mobile applications that allow the creation, access and sharing of content that is freely and readily available. On the other hand, Gundecha and Liu [2] define social media as a conglomerate of diverse types of shared media, (e.g. YouTube, Twitter, Facebook and LinkedIn). Networking sites such as Facebook, Twitter, LinkedIn and YouTube are the most popularly used sites on the Internet for all age groups [3]. Other social media platforms, which are commonly used, are WhatsApp and Skype. Since December 2011, Facebook alone has registered over 845 million active users. At present, Twitter and Facebook accumulated over 1.2 billion users [2],[4].

Social media permits the formation, interactions and exchange of content [2],[5],[6],[7]. The users of social media platforms normally are the ones who share, organise groups and provide useful information [3]. Users deliberate with each other on their everyday encounters on different social media sites. Nowadays, a majority of people rely on social media to exchange information, communicate, express and vent their feelings, express their sentiments, even ask for social support, and get comfort from others online [3],[8]. Many researchers share the same view that, the acceptance of social media continues to grow fast [2],[4], consequently leading to development of blogs, wikis, social news, micro blogs, location based social networks, social bookmarking applications, product and business review sites, media sharing (of text, photo, audio and video), etc. [2],[3],[5],[9].

Other than only chatting, social media platforms allow other activities such as online shopping, gaming, business marketing, promotions and advertisement, entertainment, politics, news exchange and analysis, health discussion and awareness forums [9]. Social networks give a premise for keeping up with others through socialising, sharing and finding content and knowledge produced by others [3]. Because social online networks attract millions of people, who share their decisions online, majority of the time, what people say online on these social media platforms influence the decision making of companies, service providers, politics,

government, etc. [9]. For example, if an employee of a certain reputable company shares negative messages or information online on these social platforms and people respond showing their dissatisfaction, the company can be influenced to take some serious measures against that particular employee. In most cases, in order to protect the brand, they would dismiss the said person.

Medicin [10] has identified five major features which constitute social media as: openness and collaboration (between both users and user groups); social networking and social interactions; participation and the use of different providers (e.g. blog spaces, search engines, etc.). In addition to Medicin 's features, Jensen [11] states other things which also generate social media data. These are: places, networks of web pages, things, data on interrelated people, events extracted from text documents and complex relational databases [11].

According to Toivo [5] social media has brought ultimate change to how people communicate, with the self-made content, and the quick pace and supple manner in which they share this content. People share and distribute content in social media in various ways. Some examples of these ways are given in Table 1.1 adapted from Toivo [5] and Gundecha and Liu [12].

Table 1.1: Examples of the Social Media Networks, adapted: [5], [12].

Social networks	Facebook, Google+, MySpace, LinkedIn, IRC
Media products community	YouTube, Flickr, Slideshare
Blog services	Wordpress, Blogger, Twitter, Posterous
Information creation community	Wikipedia, Wikispaces
Link sharing services	e.g. Digg, Diigo
Virtual community	e.g. Habbo Hotel, Second life

Apart from usage by individuals, businesses, institutions, organisations, etc., are now tapping on social media platforms to harness social media data for making the rightful decisions. Social network analysis has become popular mainly because of its use in various activities such as business advertising and marketing, search engines and administrative dynamics. However, collecting and accumulating huge amounts of data can be overwhelming and challenging to make sense of it [13]. In order to overcome this challenge, they resort to applying data mining

to the data, to extract patterns and useful ideas from it. Data mining may aid in the analysis of data through segmentation and classification [14]. As for companies, businesses, organisations, or consumers and producers of information, the success comes only if insights could be quickly discovered from the data and used to come up with better plans, decisions and actions [13]. According to Bhadani and Jothimani [15], making sense out of the vast data can help organisations to make informed decisions and to be at a competitive edge. Data is considered to be a gold mine of insights. Data mining uncovers hidden patterns in data that could be useful in predicting future behaviours. The insights discovered through data mining are transformed into decisions that turn into actions that are used in day to day business operations that need to be carried out quickly and at a large scale [13]. For example, to be at a competitive edge, with proper mining and analysis, insights found from data could help to unravel hidden profits, maximise customer loyalty, and minimise client churn. If data is left unprocessed to uncover the insights, organisations, governments and businesses would not make any informed decisions, and in most cases, their day-to day operations may suffer. As for businesses and organisations, they could even run at loss, lose clientele or even cease operating. As for the governments, the departments will perform poorly and quality of service provided to citizen will be of a low standard hence poor service delivery.

1.2 CHARACTERISTICS OF SOCIAL MEDIA DATA

On a daily basis, researchers are challenged with massive and limitless amounts of data and encounters serious limitations to establish methods of content analysis [16], [17], [18], [19]. Data which is accessible on social networking sites is noisy, distributed and dynamic [9]. Data is generally formless and presented in a diversity of ways (e.g. blog posts or tweets, web pages). Moreover, content published online is continually becoming more intricate as it is interrelated and connected through hashtags or hyperlinks [20]. Data from Social media is uncontrolled and distributed [2]. Because of that, it becomes overwhelming for researchers to understand how it flows. Social media data is most of the times formless and unstructured [21]. Uncovering hidden patterns and meanings from unstructured data from numerous data sources can be challenging. This is because, different social media sites serve different purposes and meet different needs of users, for example, people use LinkedIn mostly to professionally network in terms of jobs, projects and specialities whilst people use Facebook to mostly socialise in an uncontrolled, relaxed manner. The use of the social media sites may not necessarily be the same. These sites are constantly advancing and are also dynamic. For

instance, Facebook has invented numerous concepts e.g., numerous user privacy policy changes, user's timeline, live feeds chats and videos, and creations of in-groups and customised individual and business pages for marketing, etc. There are quite a number of motivating questions and lessons associated with human behaviour that can be explored and learnt from social media data [2]. Huge volumes of self-made content are generated over social networks daily. This inclination is possible to go on with more growth and production of content in the future [1], [2], [3], [8]. According to Bigus [21], there are some challenges which drive the growth of social media. These challenges revolve around: hearing a user's voice, improving users's experiences, and identifying the source of information best suited for users to use. As pointed out by Bigus [21], these challenges creates plenty of opportunities for data miners to come up with new methods and inventions to analyse social media data as some of the solution(s) to these challenges lay hidden in the data.

Researchers encounter varieties of information online at an increasing velocity, volume, veracity and with different value which constitutes the so called V's of Big Data [22]. Even though there are many aspects to Big Data, this study gives more emphasis to the veracity aspect. When processing online data, veracity of data should be looked at, especially when dealing with social media data. According to Sivarajah et al. [23], veracity was first invented as one of the characteristics of data by the International Business Machines Corporation (IBM) to denote untrustworthiness intrinsic in numerous sources of data (structured as well as unstructured data). Marr [24], shares the same viewpoint that veracity denotes the trustworthiness and disorderliness of the data. Saporito [25] and Sivarajah et al.,[23] argue that veracity is all about complete understanding of data, not just its quality, since there are always chances of finding inconsistencies in the data. In their definition, Akerkar [26] and Zicari [27] point out that veracity primarily is addressing inaccuracy, biasness, untruths, doubts, disorderliness and erroneous evidence in the data. The veracity feature measures how accurate data is and its worth for analysis purposes [28]. Sivarajah et al. [29], point out that, on social media networks, people's opinions differ, in most cases they are even unclear mainly because they are subjective to varying human interaction and interpretations. In addition, Sivarajah et al. point out that most of the times, people tend to post untrue fabricated stories on various platforms online, hence it is necessary to pay attention in filtering the data well when presenting quality data. This sentiment is also shared by Marr [24], that Big Data is characterised by various forms data and as such, this makes its quality and accuracy difficult to

control, especially data gathered from online platforms such as Twitter which is characterised by noise, spelling mistakes, abbreviations, hash tags, informal, unreliable and in-accurate content. Thus, according to Gandomi and Haider [30], there is need to address the ambiguousness and inaccuracy of data in another façade of Big Data, which is tackled with the help of tools and analytics for managing and mining untrustworthy data. Therefore, when users start using data, they genuinely get involved and get more volitional to invest in numerous attempts to clean data [25].

1.3 MOTIVATION OF THE STUDY

Of late, data mining community has shown a swift increase in interest to social network analysis. The basic motive is the need to make good use of knowledge from plentiful quantities of data gathered, relating to social behaviour of online users [31]. Large samples of data from social media sites have been analysed and are continually being analysed in order to get refined results, essential to assess the impact of certain cases on the communal discourse in a broader manner. Several studies have been carried out on analysis on social media data from different fields such as machine learning, natural language processing and data mining, on various discourses such as politics [32],[33],[34]; healthcare [35]; marketing [36],[37], athletics and sports [38],[39],[40]. These studies covered a variety of experiments which include: event detection [41],[42],[43],[44],[45], popularity prediction [42],[44], topic discovery [8],[46],[47]; information propagation and diffusion [48],[49],[50], and text classification [51],[52],[53],[54],[55], to name a few. These studies, in most cases have put more emphasis on statistical models and algorithms which in essence are purely computational methods. However, pure computational and algorithmic methods of analysis in their own capacity remain limited in understanding the hidden in-depth meanings or subtleties of human language. In reporting the limitations and future works of their study, Stieglitz and Dang-Xuan [34] verified this, that although computerized language measures allowed them to test the hypotheses on a large data set, these could not capture irony, sarcasm or culture specific content from data. These sentiments are also shared by other researchers (e.g., [8],[56],[57],[58]). Most of the works which have been done lack the aspect of text analysis which involves human beings. However, human beings can intervene and interpret data better as compared to computerised methods because they can get in-depth meanings from the data. Furthermore, they can better understand data and identify errors, untrustworthy, messiness and untidiness of data. Content from the social media requires human interpretation and

intervention [8]. At times, carrying out text mining can be so helpful in identifying relevant text documents from large data records and repositories in order to carry out an in-depth content analysis [16].

1.4 STATEMENT OF THE PROBLEM

The problem this study addresses is that of the difficulty in managing, processing and analysing content from social media sites. Social media possess the characteristics of Big Data such as volume, veracity, velocity, variability and value. These characteristics make its analysis more challenging than conventional data. This affects both the producers and consumers of information (e.g. researchers, organisation, governments, businesses and individuals. There has been some work that has been done previously in the area of social media data analysis. Most of these works place more emphasis on the development of computational methods. Literature has argued that, while computational methods are good with regards to statistical and algorithmic analytics, they are limited in that they cannot capture in-depth meanings and semantics from data. Content from social media requires human interpretation and intervention for proper analysis. The challenge posed by computational methods can be attributed to the other V's: veracity (accurateness of data), variability (lack of consistency or fixed pattern) and the value (worth or usefulness of data). It is therefore very difficult and challenging to mine and analyse social media data using only computational methods. This study therefore proposed a Conceptual Framework as a possible solution towards the limitations experienced with computational methods.

1.5 MAIN OBJECTIVE

The main objective of this study was to develop a Conceptual Framework for mining and analysis of social media data that offers a generic way to addressing the limitations of computational methods to social media data analysis of capturing in-depth meanings (semantics) in data.

1.5.1 Research Objectives

The following are the specific objectives geared towards the development of the Conceptual Framework

1. To identify existing concepts (i.e., theories and methods) in data analysis, which are capable of capturing in-depth meanings (semantics) in data.

2. To identify possible processes of incorporating the concepts identified in (1) with computational methods to address their limitations.
3. To construct the Conceptual Framework from the identified concepts and processes.
4. To evaluate the conceptual Framework through an application case of the political landscape of Botswana.
 - 4.1 To manually classify social media data from Facebook and Twitter on the political landscape of Botswana.
 - 4.2 To evaluate the performance of computational methods (i.e., Support Vector Machine, Naive Bayes, Decision Tree- c4.5 and K- Nearest Neighbour and) on the manually classified data.

1.5.2 Research Questions

Similarly, the following research questions helped achieve the main objective:

1. What processes and methods of data analysis in general could be amalgamated to form a Conceptual Framework?
2. What tools exist which allow online data scrapping and what are their capabilities and short-comings?
3. What steps or processes should be taken when carrying out Inductive Content Analysis approach?
4. What are the strengths and the limitations of the manual analysis approaches of analysing social media data?
5. What are the strengths and the limitations of the data mining algorithms?
6. What metrics are applicable to the training and testing of the data models for classification?
7. How feasible is the proposed Conceptual Framework in the mining and analysis of social media data?

1.6 JUSTIFICATION AND SIGNIFICANCE OF THE STUDY

The significance of this study comes with the novelty of the proposed Conceptual Framework. Its importance does not only come with the “guidance” it provides on how social media data could be mined and analysed, but also with its generic way of mining and analysis of social media data: it is multi-faceted and allows blending of various approaches and processes. The framework is not only limited to data analysis, it also gives a guide on data collection and text-

processing. Also, it can be applied to various cases e.g., classification, clustering, regression, association of rules, etc. Also the Framework can be applied to other cases of various content analyses, not only social media data.

The overall results of this study, including the Proposed Conceptual Framework could be very useful to the scholars and practitioners who wish to do research on social media data mining and analysis. These scholars and practitioners could use the Framework as a guide to the mining and analysis of social media data in a systematic manner.

1.7 CONTRIBUTIONS

Contributions are stated in two ways: to the body of knowledge of Computer Science and publications arising from conducting this research.

1.7.1 Contributions to the body of Knowledge of Computer Science

This study contributes a Conceptual Framework for mining and analysis of social media data, this extends on other researcher's works and several attempts which have been carried out in developing tools for the mining and analysis of the social media data. The Conceptual Framework presents a comprehensive workflow to aid researchers to process and analyse data in a systematic manner. The Framework does not limit the user or researcher on how they can analyse the data. It affords the user many choices of how to gather data, pre-process and analyse data.

The Conceptual Framework helps with handling the limitations of failure to capture in-depth meanings from data by computational methods through application of the various qualitative analysis approaches to the data, such as Inductive Content Analysis (ICA), Ground Theory, Narrative Analysis, Discourse Analysis, etc. Framework gives choices on the steps and process of going about in analysing data, it has to be noted that it is a guide. If the process and steps are not carried out or applied properly, they can impact on the results.

Several studies have evaluated ICA for different purposes/contexts. This study evaluated ICA for usability guided by the ISO model 9126. The information on the evaluation can add to the already existing literature and extend knowledge on ICA evaluation and can be helpful to the researchers who wish to carry it out.

1.7.2 Parts of the Thesis that have been Peer-Reviewed and Published

This section presents the relation of the thesis and the other publications by the author. This is done to show which elements of the thesis have been peer-reviewed:

1. Some parts of the Introduction in Chapter 1, related works in Chapter 3, the proposed Conceptual Framework in Chapter 4 and the evaluation of the Framework in Chapter 5 resulted in a research article published as: *Sethunya R Joseph, Keletso Letsholo and Hlomani Hlomani A Conceptual Framework for the Mining and Analysis of Social Media Data, International Journal of Database Theory and Application, 10(10) 2017, pp.11-34,*

The discussions and results on this paper basically summaries the major components of this thesis. It gives the introduction to social media network analytics and Big Data. It discusses the limitations of the manual and computational analysis of the social media data. It also presents the proposed Conceptual Framework with its components. Finally, it presents the evaluation of the Conceptual Framework.

2. Some parts of the literature review presented in Chapter 3 was published in:

a) *Sethunya R Joseph, Keletso Letsholo and Hlomani Hlomani, Social Media Data Mining: An Analysis & Overview of Social Media Networks and Political Landscape, International Journal of Database Theory and Application, 9(7), 2016.*

Our study used datasets, which were gathered on Facebook and Twitter on the political landscape of Botswana hence the need for such review. The review is an overview and the review of the analysis of social media on political landscape of Botswana and the global arena.

b) *Sethunya R Joseph, Hlomani Hlomani, Keletso Letsholo, Data Mining Algorithms: An Overview, International Journal of Computers and Technology, 15(6), 2016.*

This analysis acts as a springboard to the discussion of the computational approaches which forms part of the proposed Conceptual Framework and some of the key topics in Chapters 4 and 5 of this thesis.

1.8 RESEARCH METHODOLOGY

In order to give direction to this study, the methodology or research process, “onion” Saunders et al. [124] was adopted. The onion illustrates the range of choices, paradigms, strategies and steps followed by researchers during the research process (See Figure 1.1). The research process onion provides a summary of the important issues that need to be taken into consideration and reviewed before undertaking any research. The different layers of the onion serve as a basis from which to consider the following: the philosophical orientation of the researcher; the research approach adopted; appropriate research design and strategies; the research timelines that are under review and the data collection techniques employed by the researcher.

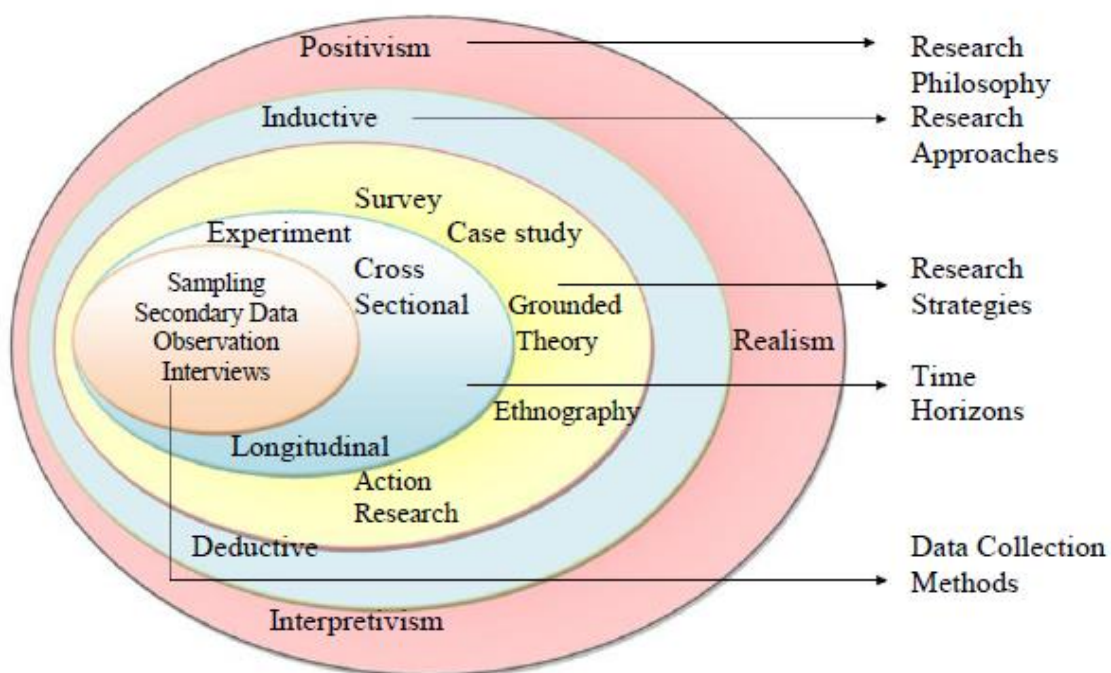


Figure 1.1: Saunders's research process onion, Source: [124]

1.8.1 Research Philosophy or Paradigm

This is the first outer layer of Saunders's onion. Here, a researcher has a choice of selecting the paradigm to follow looking at their study's objectives. The examples of the paradigms include Positivism, Realism and Interpretivist. This study followed a combination of both

Interpretive and Positivist philosophies. Interpretive paradigm involves understanding the world from an individual's point of view. The researcher attempts to observe the ongoing process to better understand a certain phenomenon. Interpretive studies usually involve human subjects. In this study, human subjects were used in focus groups to interpret and get meanings from the data that was used during the Inductive Content Analysis process as one of the approaches which were used in the evaluation and demonstration of the Proposed Conceptual Framework, hence this qualifies it to be interpretive in nature. The Positivist paradigm is characterised by understanding reality by objective testing and a single outcome of the truth. Conducting experiments and gathering quantifiable data are typical methods used in positivist paradigm and results are usually quantitatively analysed. In this study, during the evaluation of ICA, quantifiable data was processed and analysed. Also, during evaluation of the performance and accuracy of the models, the evaluation metrics quantifiably analysed the data.

1.8.2 Research Approach

This study has been shaped using Inductive research design. Saunders et al. [124] hints that the inductive approach provides the opportunity to have more clarification of what is happening (being exploratory in nature). The first objective of this study was to identify existing concepts (i.e., theories and methods) in data analysis, which are capable of capturing in-depth meanings (semantics) in data. The second objective was to identify possible processes of incorporating the concepts identified in (1) with computational methods to address their limitations. In addressing these afore-mentioned objectives, the different approaches in social media data mining and analysis were investigated, and ways in which these approaches can be improved were also explored. These were explored through the use of secondary data sources.

The formulation of the Framework was based on existing literature (through secondary sources). The research drew from the short-comings and challenges that were observed on current literature. It also involved the use of secondary sources of data for gathering information (e.g. books, journal articles, online resources, newspapers) for reviewing the literature on various areas which formed the basis of this study. In addition, in the demonstration and evaluation of the Framework through classification case, focus groups were used to explore and uncover meanings from data which was collected from online on social media platforms on the political landscape of Botswana through performing ICA. The output of the ICA was supplied to the data model algorithms to classify the data. This addressed the fourth objective of the study.

1.8.3 Research Design or Strategy

This is a plan of study that provides an overall view of how the researcher collects data for their study. The strategies may involve the use of experiments, surveys, case studies etc. In this study, apart from gathering information from secondary data sources to fulfil the first three objectives of the study, a user study was performed to evaluate the ICA process through a survey questionnaire. Experimentations were also carried out through data modelling algorithms to classify the data. This addressed the fourth objective of the study of evaluation of the Conceptual Framework.

1.8.4 Data Collection Methods

The researcher gathered data online from Facebook and Twitter through the use of digital tools (a java program built using Twitter API and Netlytic program). This data formed the data sets which were used in the ICA process and the data modelling experiments. A 30 based questionnaire was used to collect data for evaluation of ICA process from the focus group members.

1.8.5 Data Analysis

The analysis of data has been done through a triangulation of both the qualitative and quantitative approaches. The analysis of data that was done through the ICA process by the focus groups was qualitative. The ICA then was evaluated using a questionnaire which was analysed through the use of Statistical Package for Social Sciences (SPSS), where frequency tables were generated and summarised. The output of the ICA process which was in the form of categories were supplied to the data models for further processing through quantitative measures or metrics which evaluated the performance and accuracy. This involved quantitative analysis.

1.9 THESIS STRUCTURE

The rest of the thesis is organised as follows:

Chapter Two gives background information to some key concepts which have been covered in the thesis. The goal is to provide a context needed for the understanding of the concepts and context as used in the study.

Chapter Three discusses the literature review on the data analysis, examples of the qualitative and quantitative approaches used in data analysis, it discusses a comparison between qualitative and quantitative analysis. Data mining algorithms are also discussed. Related works on the mining and analysing the social media data are discussed with the gap analysis identified in them and a motivation for the proposed conceptual framework is also discussed. Most of the discussions made here, give an origin or set base for the premise of the Conceptual Framework proposed.

Chapter Four discusses the methodology that was applied in carrying out this study and also presents the premise of the proposed Conceptual Framework. The chapter entails research purpose, research design and research approach on the methodology part. The chapter also presents and defines the constituent components of the proposed Conceptual Framework as well as their role and relationships.

Chapter Five The chapter discusses the evaluation of the Conceptual Framework through a classification case by applying ICA and carrying out experiments on classification through data models.

Chapter Six concludes the thesis by highlighting the contributions per objective, the conclusions, limitations and future works.

CHAPTER 2

BACKGROUND INFORMATION

2.1 INTRODUCTION

This chapter gives background information of some aspects of this study. The chapter includes a discussion on: Big Data, data mining and knowledge discovery, Data representation in Knowledge Discovery in Databases (KDD) and data mining algorithms which have been used in this study.

2.2 DATA MINING AND KNOWLEDGE DISCOVERY

Data mining has been defined in different ways by numerous researchers. According to Ozer [59] and Garcia et al. [60], data mining is the uncovering of patterns or knowledge from vast quantities of data. Zaki and Meira [61] defined data mining as a way of finding out stimulating insights, unique patterns, and comprehensible descriptive and predictive models from huge amounts of data. Tan et al. [62] and Gundecha and Liu[2] defined data mining as a process of finding valuable knowledgeable bits of information from huge amounts of data. Kamruzzaman et al. [63] has pointed out that, most people regard data mining as an alternative word for Knowledge Discovery in Databases (KDD). They argued that data mining should just be taken as one of the necessary steps in the process of KDD (See KDD diagram in Figure 2.1). Han et al. [64] has shared the same sentiments that, data mining as knowledge discovery from data which describes the usual process of getting out helpful and important information from raw data.

Fayyad et al. [65] has defined KDD as a way of representing low-level data into other forms that might be more abstract, compact, or useful. KDD is a field which integrates various disciplines such as databases, machine learning, artificial intelligence and statistics [66]. Data mining is just a step in the KDD as shown in Figure 2.1. The stages that generally define KDD process are: Selection, Pre-processing, Transformation, Data mining and Interpretation/Evaluation [65]. Brief insights about some of these stages are given in the sections 2.21 – 2.23.

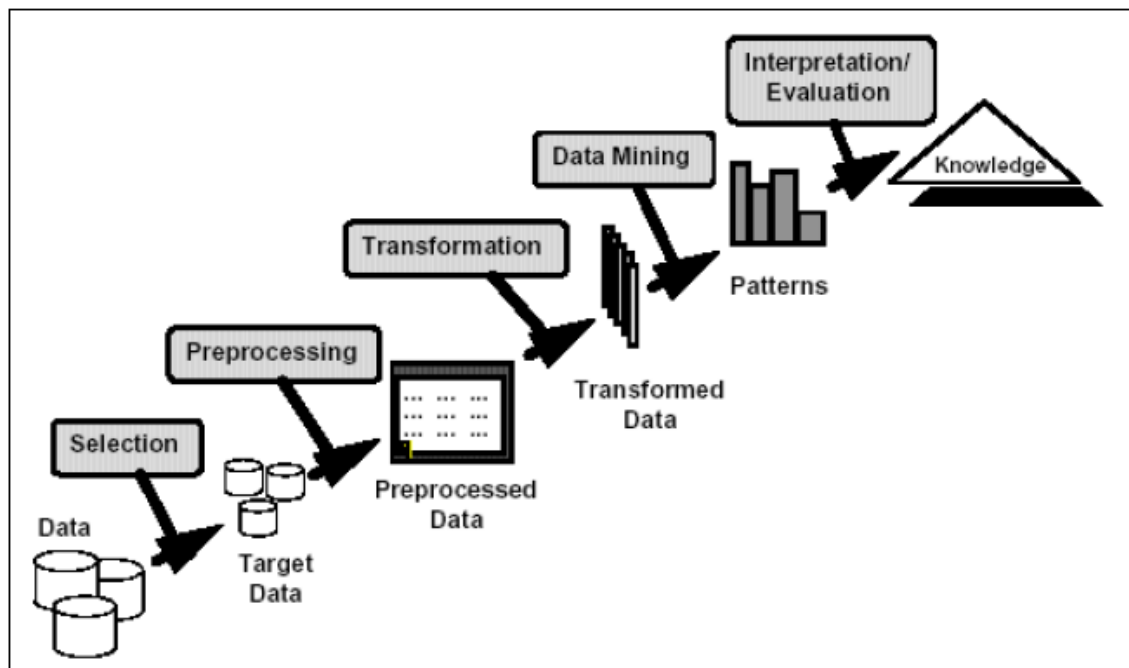


Figure 2.1: The steps that compose the KDD process, Source: [67].

2.2.1 Data Representation in KDD Process

Data is presented in a tabular format in the KDD process [67]. See the sample shown in Table 2.1 taken from Zafarani, Abbah and Liu [67]. The example shows the predictions made with regards to the person who accesses an online bookseller to purchase a particular book. These predictions have been made through studying their interest and their old purchasing history. For example, John is *likely to buy* a particular book given the history that he has previously purchased a lot from the site, has purchased resembling books before, and visited the site regularly. In this example, John is an instance. Instances are sometimes referred to data points or observations. A dataset consist of one or many data points.

Table 2.1: An example KDD data prediction sample from a dataset showing data and its properties, Source: [67].

Attributes				Class
Name	Money Spent	Bought Similar	Visits	Will Buy
John	High	Yes	Frequently	?
Mary	High	Yes	Rarely	Yes

A set of attributes are used to represent a dataset whilst values assigned to these attributes represent an instance [67]. Attributes are also known as measurements or features. In Table 2.1 example, *Name*, *Money Spent*, *Bought Similar*, and *Visits* are the features and feature values for the first instance are *John*, *High*, *Yes*, and *Frequently*. Given the feature values for one

instance, one tries to predict its class (or class attribute) value. The class attribute is *Will Buy*, and the class value prediction for first instance is *Yes*.

Psychologist Stanley Smith Stevens was first to introduce various types of features as “levels of measurement” in the theory of scales [67]. Stevens introduced four types namely: *Nominal-categorical feature* (which accepts values represented as strings); *Ordinal feature* (which place data on an ordinal scale or intrinsic order); *Interval feature* (which makes inherent ordering in data and variances are considered significant while ratios are worthless) and lastly *Ratio feature* (which add the extra multiplication and division properties).

2.2.2 Data Pre-Processing

Most of the times, data used for data mining is not instantly ready for use straight away. Several researchers share the same sentiments with regards to the quality of data used in data mining. Cleaning of data in order to reduce the noise such as spelling mistakes, abbreviations and acronyms in the data is necessary [60], [67], [68]. As pointed out by Zafarani, Abbasi and Liu [67] and corroborated by Sun [68], cleaning data for the distortions, outliers, missing values and data duplicates can reduce data noise and data size for better knowledge discovery results.

Data pre-processing prepares the data for mining [68]. Zafarani, Abbah and Liu [67] point out that aggregation; feature selection; discretization; sampling and feature extraction are typical tasks of the data pre-processing phase. The same sentiments have been shared by Gracia et al. [60], they argued that the afore-mentioned tasks are the actual typical tasks of the data preparation phase. These tasks are briefly explained as:

- a) **Aggregation:** This is done when multiple attributes are to be combined into a single one or when the scale of the attributes alters. For example, one can save image dimensions for a website by length and width of the image or by the area of the image.
- b) **Discretization:** This task involves changing continuous attributes to discrete attributes and making a decision on the range that is to be assigned to the discrete values, e.g., money.
- c) **Feature Selection:** Sometimes, not all attributes collected are useful. A few can be inappropriate, or there may lack computational strength to make use of all of the functions,

amongst many other reasons. In these instances, a subset of attributes are decided on that might ideally enhance the overall performance of the chosen data mining algorithm.

- d) **Feature Extraction:** The existing set of features or attributes get changed to a brand new set of attributes that can carry out the data mining job better. A change is done on the data, and a new set of attributes is extracted. For example, say area is a new feature being constructed, it can be made from two other attributes or features: “width” and “length”.
- e) **Sampling:** Here, a subset of attributes is chosen and processed rather than the entire set. The selected sample should ensure that it is actual representation of the entire dataset. Thus, the results acquired through on the sample data should be close to the results obtained from the whole dataset: Examples of the sampling techniques include: Random sampling; Sampling with or without replacement and Stratified sampling.

2.2.3 Data Mining and Algorithms Analysis

Several data mining techniques are being used to mine the social media data. Some these techniques are descriptive and some are predictive. The predictive techniques use a few variables to come up with unknown values of other variables, whilst the descriptive techniques discover patterns that are human-interpretable and describe the data. All the supervised learning algorithms (both classification and regression) are predictive whilst the unsupervised learning algorithms are descriptive in nature [14]. To solve a data mining task, one can apply several various techniques [66]. In data modelling, data mining algorithms are applied to find patterns in data. Data mining algorithms can be categorised as supervised and unsupervised learning. This section discusses these two well-established categories respectively:

2.2.3.1 Supervised Learning

In supervised learning algorithms, given a dataset, the values of an attribute of a class have to be known in advance prior to running an algorithm. The data in this case is referred to as training data [2]. In that dataset, instances portray the format (x, y) . In this case x represent a vector and y a class attribute. Supervised learning constructs a model that maps x to y . The task is to find a mapping $m(.)$ such that $m(x) = y$. The unlabelled set or testing set has to be also given, with a format $(x, ?)$ and y values are not known. Assumed $m(.)$ learnt from training data and x of an unlabelled instance, $m(x)$ can be worked out from the results in the prediction of the label for the unlabelled instance [67].

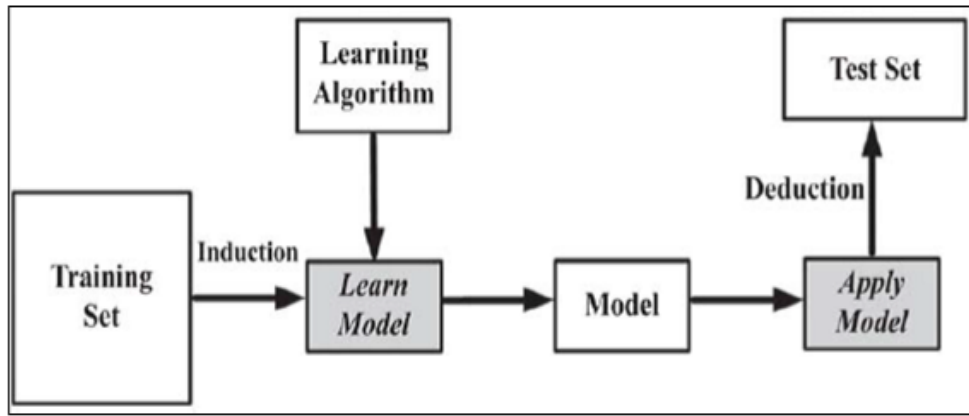


Figure 2.2: Supervised learning process, Source: [67].

Figure 2.2 shows a Supervised learning process. It begins with a set of training data where attributes are known. A supervised learning set of rules is administered on the training set in a manner known as induction. Within the induction process, the model is made. The model maps the features values to the class attribute values. The model is deployed on a testing set to guess the unknown class attribute values (deduction process). Examples of Supervised learning include: classification and regression. In classification, the class attribute is discrete whilst in regression, the class attribute is continuous [1], [67], [69].

i. Classification

According to Han, Kamber and Pei [70], classification is a two-step procedure of model development, which describes a set of already fixed classes. Each sequence is presumed to be a part of an already determined class, as predefined by the class label attribute. A set of sequences used for in model development is training set. The model is represented as decision trees, classification rules, or mathematical formulae. It is used for predicting unknown or future objects [2].

Some of the examples of classification algorithms include Decision tree learning, Naive Bayes, K-Nearest Neighbour, Support Vector Machine, etc. These are briefly explained in more detail in the sections that follow, the reason being that, they have been adopted and used in the demonstration and application of the Proposed Conceptual Framework during the evaluation in Chapter 5 of this study. However, information in relation to other algorithms can be seen on Table 3.6 of Chapter 3.

a) Naive Bayes Multinomial

The Naive Bayes multinomial classifier assigns a class to an unlabelled object based on a *maximum likelihood* principle. More precisely, the instance is allocated the class that maximizes the “*a posteriori*” probability, which is a function of the *class prior probability* and of the instance likelihood with respect to the class [71]. Bayes theorem gives a way of computing the posterior probability, $P(m|k)$, from $P(m)$, $P(k)$, and $P(k|m)$. Naive Bayes classifier presumes that the result of the value of a predictor (k) on a given class (m) is independent of the values of other predictors. This assumption is called class conditional independence [71]. This is calculated using equation (2.1):

$$P(m|k) = \frac{P(k|m)P(m)}{P(k)} \quad (2.1)$$

$P(m|k)$: is the posterior probability of *class (target)* given *predictor (attribute)* i.e. the probability of instance k being in class m (this is what was tried to be computed in this thesis).

$P(m)$: is the prior probability of *class* (probability of occurrence of class m).

$P(k|m)$: is the likelihood which is the probability of *predictor* given *class* (probability of generating instances k given class m).

$P(k)$: is the prior probability of *predictor* (probability of instance k occurring).

Naive Bayes classifiers assumes attributes have independent distributions as shown in equation (2.2) and thereby estimates

$$P(k|m) = P(k_1|m) * P(k_2|m) * ... * P(k_n|m) \quad (2.2)$$

Where $P(k|m)$: The probability of class m generating d instances, equals...

$P(k_1|m)$: The probability of class m generating the observed value for feature 1, multiplied by...

$P(k_2|m)$: The probability of class m generating the observed value for feature 2, multiplied by... (etc. until the n^{th} value

b) Support Vector Machine

Support Vector Machines (SVMs) have been generally used to address pattern classification [72], [73], [74] and non-linear regressions problems [75], [76]. The SVM creates input-output mapping functions from a set of labelled training data. In cases of classification, nonlinear kernel functions are much used in the transformation of input data to a high-dimensional feature space in which the input data become more separate in comparison to the initial input space. Maximum-margin hyper-planes are at that moment generated. The model created is contingent on only a subset of the training data closer to the class boundaries [77].

SVM learns by example to group objects into categories [78]. As pointed out by Masuyama and Nakagawa [79], the decision surface by SVM for linearly separated space is a hyper-plane and is computed by the equation (2.3):

$$w * x + b = 0 \quad (2.3)$$

Where x is an arbitrary feature vector and b and w are learned from training set linearly separable data.

The dot product $w * x$ is specified by equation (2.4):

$$w * x = \sum_{i=1}^n w_i x_i \quad (2.4)$$

SVM maps data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A training set of patterns is separable if there at least one linear classifier defined by the pair (w, b) which correctly classifies all training patterns [77],[78]. This linear classifier is represented by the hyper-plane H as shown in the previous formula (2.3) and it defines a region for class +1 patterns as shown by formula shown in (2.5) and another region for class-1 pattern as shown in formula (2.6):

$$(w * x + b > 0) \quad (2.5)$$

$$(w * x + b < 0) \quad (2.6)$$

Figure 2.3 illustrates the linear classifier defined by the hyper-plane H ($w * x + b = 0$), Source: [80].

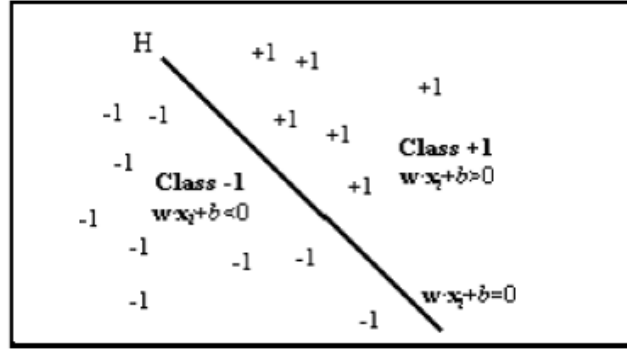


Figure 2.3: The classification hyper-plane defines a region for a class +1 and for a class -1, Source: [80].

SVM aims at make the most of the distance between two hyper-planes H1 and H2 given by the equations (2.7) and (2.8) as:

$$w \cdot x + b = 1 \quad (2.7)$$

$$w \cdot x + b = -1 \quad (2.8)$$

Where H1 is determined by equation (2.7) and H2 is also determined by equation (2.8). The distance between these two hyper-planes is $\frac{2}{\|w\|}$ [79]. Figure 2.4 illustrates the optimum separation hyper-plane.

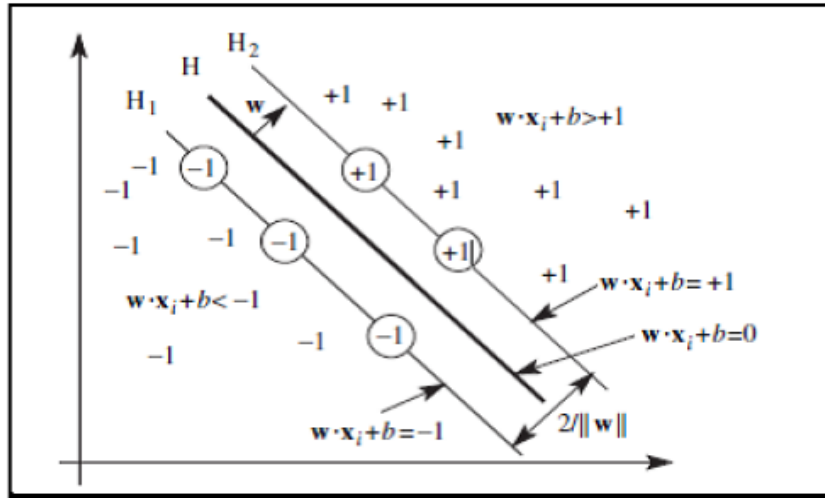


Figure 2.4: The Optimum Separation Hyper-plane, Source: [80].

The maximum separation hyper-plane is obtained by maximizing $\frac{2}{\|w\|}$ [80]. SVM makes use of kernel functions to transform the problem. Hence, linear classification technique can be carried out on non-linear classification data. According to Adu-Poku [78], the use of the kernel equations positions the data instances in such a way that, inside the multi-dimensional space, there is a hyper-plane that splits the data instances as per their types [80]. Three common

families of kernels are linear kernel, polynomial kernel and the radial basis function (RBF) kernel. The Figure 2.5 shows a simple illustration of the kernel taken from Cristianini [81].

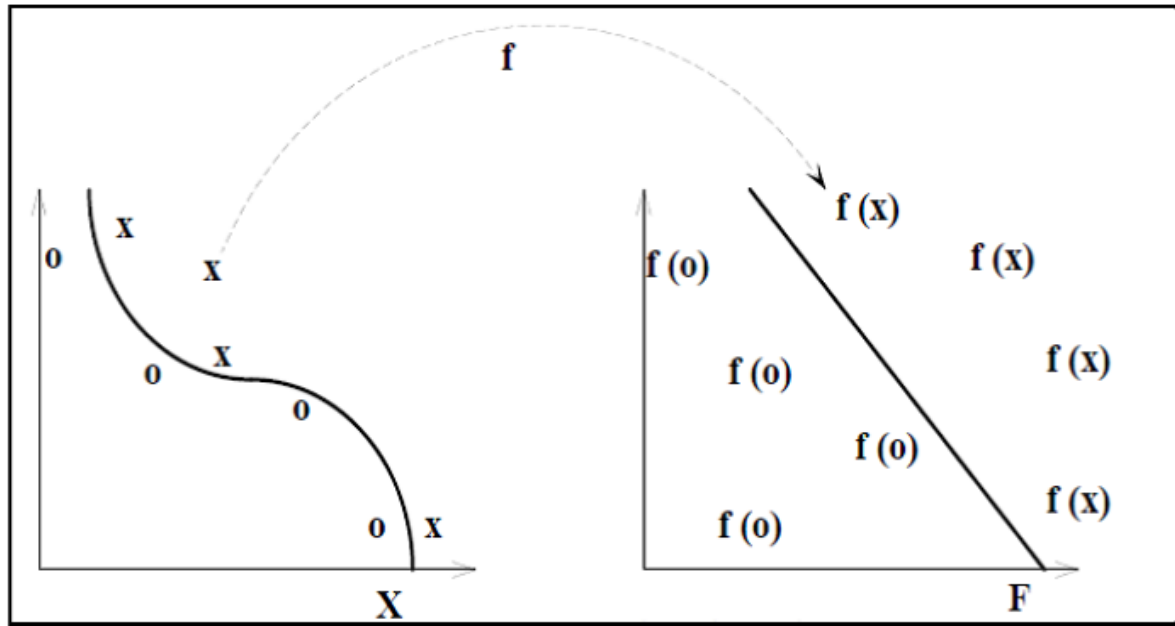


Figure 2.5: Kernel Mapping from Input Space to Feature Space, Source: [81].

c) K-Nearest Neighbour

The K-Nearest Neighbour (K-NN) algorithm classifies objects on the basis of the nearest training examples in the feature space. The classification divides data into two sets: training and test set. The K- nearest training set objects are placed in each row of the test set, and the classification is ascertained by most votes with ties broken randomly. All candidates are included in the vote if there are ties for the K_{th} nearest vector [82]. K-NN stores all existing cases and classifies new cases founded on a resemblance measure (e.g., distance functions) [83]. A case is classified by most votes of its neighbours, with the case being allocated to a class most common amongst its K-nearest neighbours measured by a distance function. In a situation whereby $K = 1$, the case is then assigned to the class of its nearest neighbour [83]. The following equations (2.9), (2.10) and (2.11) show how the distance measures functions (Euclidean, Manhattan and Minkowski) are computed:

$$\text{Euclidean: } \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (2.9)$$

$$\text{Manhattan: } \sum_{i=1}^k |x_i - y_i| \quad (2.10)$$

$$\text{Minkowski: } \left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q} \quad (2.11)$$

Where x and y are two points and the Euclidean distance between point x and y is the length of the line segment connecting them. k is the number of variables, x_i and y_i are the values of the i^{th} variable, at point x and y respectively.

All three distance measures are only usable for continuous variables. In the case of categorical variables, the hamming distance (D_H) should be applied instead (See equation (2.12)). It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

$$D_H = \sum_{i=1}^k |x_i - y_i| \quad (2.12)$$

$$x = y \implies D = 0$$

$$x \neq y \implies D = 1$$

d) Decision Tree (C4.5)

A Decision Tree is a decision modelling instrument that graphically shows the classification process of a specified input for specified output class labels [84,] [85]. In a decision Tree, a node is an attribute, and an edge between two nodes is a condition on the value of the attribute [66]. The idea of constructing a decision tree is a recursive process: firstly one has to select an attribute to place at the root of the tree and make one branch for each possible value of the attribute. Then, to classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf [66]. The priority is in the placement of the attributes, for example the Decision Tree of which an attribute is set as a root, can be decided on the basis of different measures as information gain [86] and gini index [87]. The process can be repeated for each branch which involves a subset of the original data.

According to Gula [88], a pretty straight forward way to construct a classifier is to initialize the tree with a single rule. The progression is then made down the tree by constantly making decisions to what branch to continue downwards to, eventually reaching the end node being a leaf where the final classification answer for the instance lies. Gula [88] argues that, if the solution is implemented in an algorithm that can optimize the rules at each branch by removing paths that hold none, or very little valuable information and hence maximizing the number of instances to classify in the final leaf without over fitting, then one can get pretty good generalized results [88].

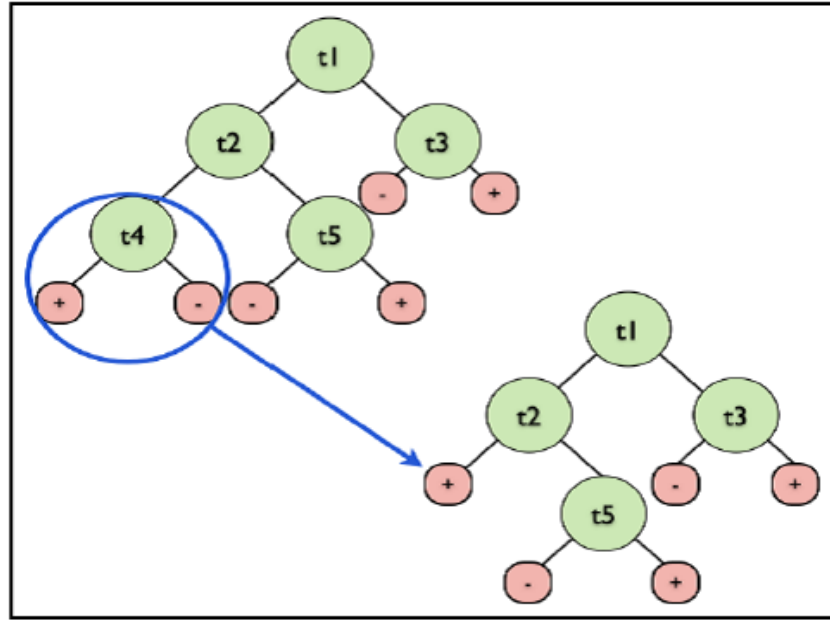


Figure 2.6: A decision tree's test node (t4) is replaced in entirety with a positive classification (assuming a binary class domain), Source: [84].

ii. Regression

Class attribute values are actual numbers in regression. For example, if one wishes to forecast the stock market value (class attribute) of a company having information about the company (features). The stock market value is continuous; hence, regression should be applied to forecast it. A dataset serves as an input to the regression method, where attributes are denoted using $x_1, x_2, \dots, x_m$ (also known as regressors) and class attribute is denoted using Y , where the class attribute is a actual number. This aids to establish the relationship between Y and the vector $X = (x_1, x_2, \dots, x_m)$ [67]. Some of the examples of regression methods include Linear regression, Logic regression, Neural Networks Multilayer perceptron.

2.2.3.2 Unsupervised Learning

This is the unsupervised division of instances into groups of the same objects. Normally when discussing the unsupervised learning, most researchers mainly discuss clustering [1], [67]. In clustering, the data is mostly unlabelled. As a result, the label for each instance is unknown to the clustering algorithm. This is the key difference between supervised and unsupervised learning. Any clustering algorithm involves a distance measure. Instances are positioned into different clusters depending on their distance to other instances [89]. The most common distance measure for continuous features is the Euclidean distance as shown in (2.13):

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \quad (2.13)$$

$$=\sqrt{\sum_{e=1}^n (x_e - y_e)^2}$$

Where $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ are n -dimensional feature vectors in R^n . After a distance measure is carefully chosen, instances are grouped using it. Clusters are generally denoted by compacted and abstract notations. “Cluster centroids” are one common example of this abstract notation. Lastly, clusters are assessed. There is still a large debate on the subject of cluster evaluation because of the non-existence of cluster labels in unsupervised learning. Other proximity measures which can be used when measuring the similarity or dissimilarity between objects include Minkowski distance and Mahalanobis distance, simple matching coefficient, Jaccard, cosine similarity and Pearson’s correlation [2]. Examples of some of the unsupervised clustering coefficient algorithms include K-means, Density based and Apriori.

2.2.3.3 Handling Random Variables

Basic probability theory normally is used to deal with random variables, and distributions [90]. A random variable is a random number (value) determined by chance, or more formally, drawn according to a probability distribution [71]. The probability distribution can be estimated from the observed data (e.g., throwing dice). The probability distribution can be synthetic. Random variables can be discrete or continuous variables. The typical random variables in machine learning problems are associated with the input data, output data and the noise [84], [85]. Random variables help predict target variables where true (presumably deterministic) function is not given but rather, only observations are provided [78], for example, one can observe the number of times a dice lands on 4, can estimate the probability, given the input that the dice will land on 4, but one does not know where the dice will land and can only make a guess to the most likely value of the dice, given the input. The random variables only take on discrete values, e.g., throwing a dice: possible value: $V_i \in \{1, 2, 3, 4, 5, 6\}$. The probable chances sum up to 1 [90].

Discrete distributions are particularly important in classification. There are categorized as: Classic discrete distributions I, II, and III. On the Classic distribution I, there is Bernoulli distribution. In Bernoulli distribution, a Bernoulli random variable takes on only two values, i.e., 0 and 1.

$p(1) = p$ and $p(0) = 1 - p$ or in compact notation:

$$p(x) = \{p^x(1 - p)^{1-x}\} \quad (2.14)$$

If $x = 0$ or $x = 1$ or 0, otherwise.

The performance of a fixed number of trials with fixed probability of success (p) on each trial is known as Bernoulli trial.

As for classic discrete distribution II, like Bernoulli distribution: binary input variables is 0 or 1 and probability $p(1) = p$ and $p(0) = 1 - p$. e.g., what is the probability of k successes, $p(k)$, in a series of n independent trials? ($n \geq k$)

$p(k)$ is a binomial random variable:

$$p(k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (2.15)$$

Where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Bernoulli distribution is a subset of the binomial distribution (i.e., $n = 1$). For classic discrete distribution III there is multinomial distribution. Here, a generalization of the binomial distribution to multiple output is made (i.e., multiple classes can be categorized instead of just one class. n independent trials can result in one or r types of outcomes, where each outcome c_r has a probability $p(c_r) = p_r$ ($p_r = 1$)

What is the probability $p(n_1, n_2, \dots, n_r)$, i.e., the probability that in n trials, the frequency of the r classes is $(n_1, n_2, \dots, n_r)$?, this is a multinomial random variables:

$$p(n_1, n_2, \dots, n_r) = \binom{n}{n_1 n_2 \dots n_r} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$$

Where $\binom{n}{n_1 n_2 \dots n_r} = \frac{n!}{n_1! n_2! \dots n_r!}$

2.3 BIG DATA

Social media data is part of the Big Data. This section therefore gives a brief description of what Big Data is in relation to the social media data as well as the properties of Big Data which translates to social media data. Large or Big data refers to huge, composite data sets continually growing with several, separate sources [91]. This is data whose scale, range, and intricacy require new structure, strategies, algorithms and analytics to manipulate it and draw out value and concealed knowledge from it [92]. The concept of Big Data has been around for quite a long time. Majority corporations recognise that in the event that they seize all of the information that streams into their organizations, they can analyse it and get considerable value from it [93]. Nevertheless, long back in the 1950s, long before the

inception of the term “big data”, corporations were performing simple analytics, basically numbers in spread-sheets that had been manually tested to realise insights and trends [93].

Big Data might encompass data accumulated and shared through the internet of things. This could include content from various sources such social media sites, search engines, blogs, network activity reports, business and organizational data such emails, reports and survey responses, web server logs and internet clickstream records, cellular-telephone calls and sms’s detail records and system data captured through sensors, etc. [94]. A few people solely associate Big Data with formless data but consulting firms like Gartner Inc. and Forrester Research Inc. also take into account transactions and other structured records to be valid components of Big Data analytics applications [94].

2.3.1 Properties of Big Data

Often, Big Data is defined using the V’s: Volume, Velocity, Variety, Veracity and some often add the fifth V of value. The following sections define these V’s.

Volume: Marr [95] defines volume as the huge quantities of data generated every second whereas Williamson [96] highlights that the key feature that makes data “Big” is the sheer volume. It is pointless to cognize on least storage units due to the fact the total quantity of information is growing at a fast pace yearly. Nobody is sincerely aware of how much new data is being generated, but the amount of data being gathered is massive [96].

Velocity: According to Williamson [96] velocity is the occurrence and speed that the data is being generated at. He gives an example of voluminous phone short messages (SMS), Facebook status updates, or credit card swipes being sent on a particular telecom service daily [96]. On the other hand, Marr [95] refers to velocity as the pace at which new data is produced and the pace at which it travels around. One example, is how social media posts can circulate and go viral in short period of time [95].

Variety: Marr [95] states that variety refers to the distinctive sorts of data which can be used. Previously, the focus was mainly on structured data that was carefully built-in into tables or relational databases. Most of online data is unstructured (text, images, video, voice, etc.) but with advancement of technology, it can now be analysed and put together even though its types varies [95]. Williamson [96] defines variety as one the most interesting developments in technology as more and more information is digitized.

Veracity: According to Marr [95] veracity refers to the disorderliness or trustworthiness of the data. With data assuming many varieties of forms, it is difficult to control the accuracy and quality of data gathered online. For examples posts from social media sites such Facebook, Twitter, etc., are full of mistakes, abbreviations, typos and informal speech. Even Williamson [96] shares the same sentiments that veracity refers to the trustworthiness of the data.

Value: Marr [95] points out that the online applications yield unprecedented amounts of data in terms of Volume, Velocity, Variety and Veracity. However, the latest technologies give us some chances to analyse all kinds of data to uncover hidden meanings from data and add value to it or even get value from it [95]. Williamson [96] argues that at the end of the each analysis, one should be able to gain insights and get value out of the data which they have analysed.

2.4 CHAPTER SUMMARY

This chapter gave a background summary of some concepts which are the springboard of this study. These include an overview of the Data mining and Knowledge Discovery; Data representation in Knowledge Discovery in Databases; Data processing; Data Mining and Algorithm analysis. Data mining and algorithm analysis covered concepts of the Supervised and Unsupervised learning algorithms with their examples. Some of the Supervised learning algorithms explained in this chapter have been used in the experimentations which were carried out in this study. How random variables are handled during classification was also explained. The chapter ends with a discussion on Big Data and the properties of Big Data being the five V's: Value, Veracity, Variety, Volume and Velocity. These are explained as social media data, being a constituent of Big Data, inherits the properties of Big Data and this study dealt with analysis of social media data.

CHAPTER 3

LITERATURE REVIEW

3.1 INTRODUCTION

This chapter discusses data analysis, with respect to the type (qualitative and quantitative) of data used for analysis, the type of analysis: qualitative and quantitative analysis, their strengths and limitations, their examples and comparison between them. The chapter also discusses data mining and algorithm analysis with regards to social media analytics. Related works on Frameworks for social media data analysis have been discussed mainly to establish the gaps which exist in the previous works and to give the premise for the Conceptual Framework of this study. The afore-mentioned topics which are discussed in this chapter give the base and premise of the proposed Conceptual Framework to be presented in Chapter 4 because some of them have been adopted to construct the Conceptual Framework.

3.2 DATA ANALYSIS

Bohm [97] describes data analysis as a process of splitting data into bits and pieces, and then splicing them together [97], while Dey [98] describes it as a way of analysing data into clear-cut components, to show its typical features and structure. Dey [98] points out that, analysis is important so as to avoid relying completely on impressions and intuitions about the data. Data analysis, from the qualitative or quantitative research perspectives, includes inspection of the data with regards to the potential resolutions to the questions or problems initially set at the beginning of the research process [98], [99]. The definite chore of analysis actually depends on the data-gathering methods applied on the research process. Generally this is to make descriptive accounts founded information gathered by various data- collection techniques [99]. Sections 3.2.1 to 3.2.4 discuss the comparison between qualitative and quantitative data; qualitative and quantitative data analysis with their strengths and limitations. Examples of the approaches used in each analysis are also stated. The section on data analysis is concluded with a comparison of the qualitative and quantitative researches.

3.2.1 Comparison between Quantitative and Qualitative data

According to Berg [99], information collected with the aid of researcher(s) can probably be used to answer the questions or resolve the issues which have been pointed out. How one decides on how to collect this information is basically a researcher's preference and largely lies upon on boundaries set by the stakeholders or the nature of the issues and setting [99], [100]. Thus, as in any general methodological approach, the researcher is guided by the research questions. Some problems will direct the researcher towards carrying out interviews with relevant parties. Nevertheless, other studies may best addressed with archival data. Certainly, a few researchers may decide to triangulate their research in order to enhance their findings and probably enrich the eventual evaluation and understandings [99].

Data collected via surveys and questionnaires are analysed better through the use of statistical computational methods. However, Dey [98] argues that arithmetical measures often involve assumptions about the likelihood of attaining answers that can only be satisfactory when using random samples, naturally through the survey method. In situations where researchers have fulfilled these assumptions, they should freely adopt appropriate procedures to go about processing their data. However, in situations where researchers are unsatisfied with these assumptions, they may apply statistics to inspect the quantitative features of the data, for instance, to be more accurate and precise in distinguishing or creating classification schemes [98]. Some simple statistical procedures tend to be useful in processing even the most peculiar and unstructured data [101]. The kind of data collected is quantitative in nature. It has a lot of numeric data.

Qualitative data mainly looks at meanings uncovered from data. Data collected through interviews, focus groups, observations is mostly qualitative in nature. It is discrete hence it is best analysed using thematic and content analysis methods [102], whereas the way numbers are analysed, is through statistics and mathematics [98], [99].

Qualitative data is rich and diverse, as it includes almost any type of data: text, sounds, pictures, videos, etc. [102]. Textual content is never the only, neither is constantly the most effective approach of communicating qualitative information. In a digital era, artwork and layout have turned out to be powerful media tools [98]. The afore-mentioned kinds of data do bring meaningful information in a form different or aside from numbers. Nevertheless, it has to be emphasised that numbers too occasionally can give most effective meanings. For instance,

numbers on football sweaters, car registration numbers, plot numbers, etc. Treating these numbers as numerical data to be subject to mathematical manipulations (e.g., addition, subtraction, etc.) would be so silly and absurd. However, sometimes it is difficult to differentiate between the use of a number as a descriptor of quality or its use as an extent of quantity. A typical example is when numbers are used to represent names for manipulation purpose, for instance when coding the response categories in an interview for analysis purposes numbers are used, especially when dealing with cases where one uses Likert scales.

In differentiating quantitative data from qualitative data, in terms of numbers and meanings, researchers need to be careful not to treat numbers as numbers, where they have been used to convey meaning [98]. It might be wrong to anticipate that quantitative data should take priority over qualitative data honestly because it includes numbers. Let us take an example of checking body weight, one may use the scales and check how much they weigh. This is a quantitative measure, however, it does not show how the weight is distributed, or how a particular point in the scale translates into overall appearance. One might choose to rely on how they look, whether they look ‘fat’ or ‘thin’ or maybe ‘just right’. These are qualitative judgements, but in a social context these may be the judgements that count.

Qualitative data is sometimes considered to be richer and more valid than quantitative data [103]. However, it is sometimes said to be too subjective since judgements are not made in terms of conventional standards. Therefore it is necessary for researchers to be careful in consideration of the reliability and validity of the methods they choose. Often qualitative data has been intently linked with research approaches that emphasise unstructured methods of attaining data [104].

Numbers and meanings can be said to be mutually dependent. A number can determine meaning and vice versa [98]. However, this kind of relationship alters as one ascends the measurement hierarchy. If the meanings are more fixed and stable, it can be easy to apply the mathematical manipulations, however, if the concepts are ambiguous and elastic, it is highly likely impossible to quantify the data in a meaningful manner.

The T'ai-chi T'u diagram shown in Figure 3.1 can be used to represent this relationship. The Figure shows a dynamic balance of two opposing forces: qualitative represented by meaning and quantitative represented by number [105]. The diagram reveals the mutual dependence of both kinds of data. Each component complements each other, although at lower levels of

measurement questions of meaning are uppermost, while at higher levels of measurement, questions of number loom largest.

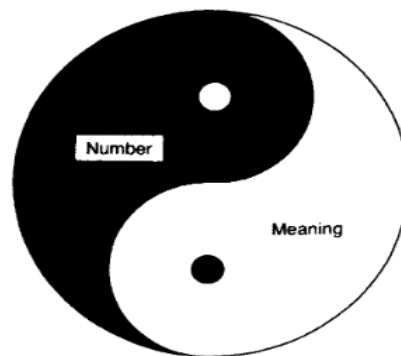


Figure 3.1: Qualitative and Quantitative data in dynamic balance, Source: [98].

3.2.2 Qualitative Analysis

Qualitative research is gives a larger, more close-up idea of the subject in order to have a better understanding of something. It is often considered appropriate for attainment of in-depth meanings and understanding of causal reasons and motivations [106]. It uncovers hidden meanings and perceptions regarding a certain phenomenon. Qualitative research may be a prior step to the quantitative analysis for creating ideas [107]. It often creates ideas and hypotheses which can later be used in quantitative research. The features of Qualitative analysis, the strengths as well as limitation are briefly discussed in Table 3.1.

Table 3.1: The strengths, limitations and features of Qualitative Analysis

Defining Features	Strengths	Limitations
- Qualitative research is interpretive in nature and infinitely creative. The researcher makes some interpretations from the data they have gathered [99],[108]. - In trying to makes sense of the data, numerous techniques can be used to make sense of the data (e.g., grounded theory [109], content analysis, thematic analysis [100] or discourse analysis, etc.) - Researchers immerse themselves in the field and their natural surroundings in order to fully understand the events, since events can be easily comprehended only if they are realised in context [99]. - The contexts of inquiry are natural, not contrived. Nothing is predefined or taken lightly [98],[108]. - Qualitative researchers let the participants speak for themselves, to provide their perspectives in words and other actions [103],[107]. Hence, qualitative research is an interactive process in which the persons studied teach the researcher about their lives [110],[111]. - The qualitative researcher is part and parcel of the data, without the active participation of the researcher, no data exists [112]. - The design of the study evolves during the research, and can be adjusted or changed as it progresses [99],[113]. - For the qualitative researcher, there is no single reality, it is subjective and exist only in reference to the observer [103],[110],[114],[108]. - Theory is data driven, and emerges as part of the research process, evolving from the data as they are collected [115].	- A researcher gets a closer involvement and gains an insider's view of the field [116,117]. - Because of the researcher's involvement, they can easily discover or even uncovers issues that are often missed (such as subtleties and complexities) by the other enquiries such as the positivistic inquiries [111]. - Descriptions in Qualitative analysis often play an important part of proposing possible relationships, causes, effects and dynamic processes [118],[119]. - Qualitative analysis gives a reflection of social reality because it allows for ambiguities/contradictions in the data [120],[107]. - The descriptive and narrative style of qualitative analysis might be of a particular benefit to the practitioners as they could turn to qualitative reports in order to examine forms of knowledge that might otherwise be unavailable, thereby gaining new insight [112],[121].	- Qualitative designs do not generally draw samples from large-scale data sets because of the time and costs constraints [114], [115]. - Qualitative analysis is mostly criticized with regards to adequate validity or reliability. This is because of the nature of qualitative data: it is subjective and origins in single contexts, thus, making it difficult to apply conventional standards of reliability and validity. - It is difficult to repeat the qualitative studies because of the central role played by the researcher in the generation of data [113]. - Also, contexts, situations, events, conditions, and interactions cannot be repeated to any extent nor can generalizations be made to a wider context than the one studied with any confidence [114],[115],[108] - It takes a lot of time to gather data, interpreting and analyse it qualitatively. These processes are lengthy. Also, Analysis of qualitative data is difficult and requires expert knowledge on the subject area necessary to make some interpretations. Therefore, great care must be taken when doing so, for example, if looking for symptoms of mental illness [114].

In summary, what is communicated in Table 3.1 is that, researchers immerse themselves in data so as to interpret it, in order to uncover or discover hidden meanings which explain certain phenomena. In analysing qualitative data, researchers seek to uncover hidden patterns and links in the data. In Qualitative analysis, the data collection and analysis are lengthy and the analyses

are done on the natural surroundings or context. The relevance of this subject discussion to this study is that, this process of data analysis had been adopted and applied as one of the components in the construction of the Proposed Conceptual Framework. It had been part of those approaches, processes and theories which formed the base and premise of the Proposed Conceptual Framework.

3.2.2.1 Examples of Qualitative Approaches

Researchers distinguish varieties of qualitative analysis. According to Braun and Clarke [100] and Dey [98] qualitative data analysis varies, there is no one clear definite way of analysis, but rather, various approaches, linked to the diverse viewpoints and aims of researchers. Researchers's viewpoints and aims differ, and to accomplish these, may follow different kinds of analysis. Some of these approaches used for analysis in Qualitative researches are briefly discussed and compared in terms of data collection implications in Table 3.2

Table 3.2: A comparison of approaches used in Qualitative Analysis, adapted: [122]

Type of Approach	Defining Features	Data Collection Implications
Phenomenology	-Focuses on individual experiences, beliefs, and perceptions. -Text used as a proxy for human experience.	- Questions and observations are aimed at drawing out individual experiences and perceptions. -In focus groups, group experiences and normative perceptions are typically sought out. -In-depth interviews and focus groups are ideal methods for collecting phenomenological data.
Ethnography	-Oriented toward studying shared meanings and practices (i.e., culture). -Emphasizes the emic perspective. -Can have a contemporary or historical focus.	-Questions and observations are generally related to social and cultural processes and shared meanings within a given group of people. -Traditionally, it is associated with long-term fieldwork, but some aspects are employed in applied settings. -Participant observation is well suited to ethnographic inquiry.
Inductive Thematic Analysis OR Inductive Content Analysis	-Draws on inductive analytic methods (this would be same for Grounded Theory). -Starts with data gathering -Involves identifying and coding emergent themes within data in to categories. -Leads to theory development -Most common analytic approach used in qualitative inquiry. -Starts from a specific level of focus to a more general focus	-ITA/ICA requires generation of free-flowing data. -In-depth interviews and focus groups are the most common data collection techniques associated with ITA/ICA. -Notes from participant observation activities can be analysed using ITA/ICA, but interview/focus group data are better.
Narrative Analysis	-Narratives (storytelling) used as source of data. -Narratives from one or more sources (e.g., interviews, literature, letters, diaries).	-If generating narratives (through in-depth interviews), then questions/ tasks need to be aimed at eliciting stories and the importance those stories, hold for participants, as well as larger cultural meaning.
Grounded Theory	-Inductive data collection and analytic methods. -Uses systematic and exhaustive comparison of text segments to build thematic structure and theory from a body of text. -Common analytic approach in qualitative studies.	-In-depth interviews and focus groups are the most common data collection techniques associated with GT. -Sample sizes for grounded theory are more limited than for ITA because the analytic process is more intensive and time consuming.
Case Study	-Analysis of one to several cases that are unique with respect to the research topic -Analysis primarily focused on exploring the unique quality.	-Cases are selected based on a unique (often rarely observed) quality. -Questions and observations should focus on, and delve deeply into, the unique feature of interest.
Discourse/ Conversation Analysis	-Study of “naturally occurring” discourse -Can range from conversation to public events to existing documents. -Text and structures within discourse used as objects of analysis.	-These linguistically focused methods often use existing documents as data. -Conversations between individuals that spontaneously emerge within group interviews or focus groups may be studied but are not preferred. -Participant observation is conducive to discourse analysis if narratives from public events can be recorded.

Table 3.2 briefly shows a comparison has been made based on the type of data collections employed. The relevance of this subject discussion to this study is that, some of the approaches mentioned here, had been adopted and applied as some of the components in the construction of

the Proposed Conceptual Framework. It had been part of those approaches, processes and theories which formed the base and premise of the Proposed Conceptual Framework. Additionally, ICA was used in the application case and evaluation of the Framework in Chapter 5. Reasons for this selection are also stated in Chapter 5 where it has been applied.

3.2.3 Quantitative Analysis

The main aim of quantitative analysis is to measure the data and evaluate it from the angle of statistics and other normally accepted metrics [98]. This provides the basis for generalizability of the results, from a sample to the entire population of what will be studied. Often, in some cases, quantitative analysis is succeeded by the qualitative analysis for reviewing the results more carefully [123]. Quantitative analysis is briefly discussed in Table 3.3 with its features, strengths and limitations.

Table 3.3: The strengths, limitations and features of Quantitative Analysis

Defining features	Strengths	Limitations
- Quantitative research aims for objectivity (i.e., without bias), and is separated from the data and is capable of being seen by anyone [124]. - Normally, the design of the study is sort and established before it begins [113]. - Quantitative analysis is carried out in testing some theories and eventually supporting or rejecting those theories [98],[110],[124]. - Statistical data analysis helps turn quantitative data into useful information to help with decision making [124]. - Statistics can used to summarise data, describing patterns, relationships, and connections. - Statistics can also be descriptive or inferential [108]. Descriptive statistics help summarise data whereas inferential statistics are used to identify statistically significant differences between groups of data [124]. - Quantitative researchers try to control extraneous variables by conducting their studies in the lab [118].	- Quantitative approach is viewed as scientifically objective and rational as data can be interpreted with statistical analysis, (since statistics are based on the principles of mathematics) [107],[123]. - Deemed very helpful and useful in testing and validating already constructed theories. - The use of Sophisticated softwares lead to a fast analysis: removes much of the need for prolonged data analysis, especially with large volumes of data involved [125]. - Experiments and studies could be easily repeated and also tested since Quantitative data is based on measured values and can be numerical data is less open to ambiguities of interpretation [125].	- Quantitative experiments do not take place in natural settings. - It does not give the participants to express themselves or the choices they make, also to explain the meaning they give to a certain phenomenon or context being researched. [123] - Requires a researcher with good command and application of the statistical analysis. Poor knowledge of the application of statistical analysis may negatively affect analysis and subsequent interpretation [126] - It needs large sample sizes for more accurate analysis, good reliability and generalizability purposes [107]. - It has an element of confirmation bias as the researcher might miss observing phenomena because of focus on theory or hypothesis testing rather than on the theory of hypothesis generation [110,124].

Table 3.3 basically points out that Quantitative analysis's main objective is to measure the data and evaluate it from the angle of statistics and other mathematically accepted metrics which can be used to manipulate arithmetic problems. The researcher ought to carefully manipulate data with the aid of comprehensive set of rules, guiding the whole process. The results of quantitative analysis should normally be able to be generalised and applied to a larger scale or to the entire population of study. These results should permit easy duplication. The relevance of this subject discussion to this study is that, this process of data analysis had been

adopted and applied as one of the components in the construction of the Proposed Conceptual Framework.

3.2.3.1 Examples of Quantitative Approaches

There are four main types of quantitative research designs namely: descriptive, correlational, quasi-experimental and experimental. The distinction between these approaches often pertains to the level the researcher plans for manipulation of the variables in their experimentations. A brief description of each design is detailed in the sections that follow. A comparison between them is also made in Table 3.4.

A **Descriptive Design** tends to define the recent status of a variable. The researcher do not immediately starts with hypothesis generation. The hypothesis after formulated after data had been gathered. Data gathering is generally through observations.

A **Correlational Design** searches for the connection between variables through statistical analyses. But, it does not check for cause and effect. Data gathering is also through observations.

Quasi-Experimental Design tends to look for a cause-effect connection between variables. It is sometimes denoted as causal-comparative. Groups used for control are named and exposed to the variable. Results are compared against the results of groups that were not exposed to the variable.

Experimental Designs Sometimes is referred to as true experimentation. It makes use of a scientific method to find the cause-effect connection among a set of variables in a study. During the experimentation, researchers try by all means to control all variables except the one being computed (the independent variable). The outcomes of the impartial variable on the dependent variable are gathered and analysed for a relationship.

Table 3.4: A comparison between Quantitative research approaches, Source: [106]

Type of Design	Key-Focus & Control of Variables	Intervention Applied	Example	Common Study Designs
Descriptive	Observational; Describe "What is"; Variables not controlled	No	A description of Teenagers' attitudes Towards smoking	Comparative Descriptive design; Cross-sectional Designs, Longitudinal designs
Correlational	Explores and Observes Relationships Among variables; Variable not controlled	No	A study of the Relationship Between IQ and Clinical depression	Descriptive Correlation Designs; and Model-testing designs
Quasi-Experimental	Test for causality With Suboptimal Variable control; Independent variable not manipulated	Yes	A study of the effect Of an after school Physical activity Program on Childhood obesity rates	Pre- & post-test designs; Post test only designs; Interrupted times- series designs
Experimental	Tests causality With optimal Variable control; Independent Variable is manipulated	Yes	Effects of new diet Treatment plan on Insulin levels in diabetics	Classic Experimental Designs; Randomized Designs, Crossover Designs, Nested designs

The relevance of this subject discussion to this study is that, some of the approaches mentioned here, had been adopted and applied as some of the components in the construction of the Proposed Conceptual Framework. It had been part of those approaches, processes and theories which formed the base and premise of the Proposed Conceptual Framework.

3.2.4 Comparison between Qualitative and Quantitative research

The section on Data analysis is concluded by making a contrast between the Qualitative and Quantitative research in general. This is summed up in Table 3.5

Table 3.5: A general comparison between Qualitative and Quantitative research, Source: [127]

Criteria	Qualitative	Quantitative
Purpose	Understand and Interpret Social interactions	Test hypotheses, check the cause and effect. Develop predictions for the future
Studies group	Small, selected intentionally	Larger and selected randomly
Data type	Words, images, objects	Numbers and statistics
Data form	Open ended responses, interviews, participants observations, fields notes	Precise measurements using structures and validated instruments for data collection
Types of data analysis	Patterns, features, themes identification	Statistical relationships identification
Researchers' role	Researchers may be known to participants in the study and participants' characteristics may be known to the researchers	Researchers and their biases are not known to the participants in the study Participants characteristics are hidden
Results	Particular findings, less generalizable	Generalizable findings, can be applied to the other populations

3.3 DATA MINING AND ALGORITHM ANALYSIS

Several data mining techniques are being used to mine social media data, amongst these are the supervised and unsupervised learning algorithms. Some of these techniques are descriptive while some are predictive. For predictive techniques, some variables are used to predict the unknown or future values of other variables whilst for the descriptive techniques, human-interpretable patterns that describe data are established. All the supervised learning algorithms are predictive whilst the unsupervised learning algorithms are descriptive in nature [14]. Examples of some of the commonly used data mining algorithms for social media data are discussed with their strengths and limitations in Table 3.6.

Table 3.6: A Summary of the strengths and Limitations of some of the Supervised & Un-Supervised Learning algorithms

ALGORITHM	HOW IT WORKS	STRENGTH	LIMITATIONS
1. Decision Trees (C4.5, ID3, CART CHAID, C5.0)	-Decision is learnt from training the data set. -Each non-leaf node in a tree represent a feature and each branch represent a value that the feature can take -Instances are classified by following a path that starts at the root node and ends at a leaf by following branches based on instance feature values [67]. -Construction of decision trees is based on heuristics -Can be used for both classification or regression [128].	-Multiple Decision trees can be learned from the same data set -They can both correctly predict the class attribute values for all instances in the dataset -Implicitly perform variables screening or feature selection. -Requires relatively little effort from users for data preparation -Non-linear relationships between parameters do not affect tree performance. -Easy to interpret and explain. -Allow the addition of new possible scenarios [67].	-Decision trees are constructed recursively from training data using a top down greedy approach in which features are sequentially selected [67]. -Without proper pruning or limiting tree growth, they tend to over-fit the training data, making them somewhat poor predictors. -Low Performance since you need to 'redraw the tree' every time you wish to update your CART model and poor resolution on data with complex relationships among the variables. -They are only two possibilities (left-right) at each node hence there are some variable relationships that decision trees just cannot learn. -Practically limited to classification.
Bayes (Naïve Bayes, Naïve Bayes multi-nomial, etc.)	-Bayesian network is a model that encodes probabilistic relationships among variables of interest [129]. -In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features [130]. -Naïve Bayesian technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data [131]. -Used for classification or regression [128].	-Easy to use. -Effective if the training set is large enough. -Ability to learn; as the training set gets larger, the results get more and more accurate (intelligence) [67]. -Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. -If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. - Naïve Bayesian classifiers simplify the computations and exhibit high accuracy and speed when applied to large databases. -Bayesian classifiers give satisfactory results because	-Does not consider the sequence of words(Non-relevant word feature) -It cannot learn interactions between features. -Is based on the so Bayesian theorem and is particularly suited when the dimensionality of the inputs is high -Considerably higher computational effort is required [132]. -In Naïve Bayes approach it is assumed that the data attributes are conditionally independent [129] which is not always so.

		focus is on identifying the classes for the instances, not the exact probabilities [130].	
K-Nearest Neighbour	-K-NN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g. distance functions). KNN has been used in statistical estimation and pattern recognition as a non-parametric technique. -A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If $K = 1$, then the case is simply assigned to the class of its nearest neighbour [133]. -Used for classification or regression [128].	K-NN algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours. -It is analytically tractable. -KNN is simple in implementation and it uses local information, which can yield highly adaptive behaviour. -A major strength of the KNN algorithm is that it lends itself very easily to parallel implementations [134].	- It is a type of lazy learning where the function is only approximated locally and all computations are deferred classification until [135]. -One of the weaknesses of the K-Nearest Neighbours Algorithm as a classifier for an IDS is its large storage requirements. -KNNs are also known to be highly susceptible to the curse of dimensionality and slow in classifying test tuples [135].
Support Vector Machine	-Structural risk minimization seeks to find a hypothesis (h) for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set [136]. -An SVM maps input (real valued) feature vectors into a higher dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [134]. -Used for classification or regression [128].	-High accuracy [136]. -SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space [137] -have the ability to update the training patterns dynamically whenever there is a new pattern during classification. -Nice theoretical guarantees regarding over-fitting (Support Vector Machines are able to model complex nonlinear decision boundaries and are less prone to over fitting than other methods). -Used in text classification problems where very high-dimensional spaces are the norm [130].	-Memory-intensive, hard to interpret, and kind of annoying to run and tune -You do not have a nice probabilistic interpretation, -You cannot easily update your model to take in new data (using an online gradient descent method). -A disadvantage of SVMs as a classifier is its high algorithmic complexity and extensive memory requirements [138]. This consequently makes the speed both in training and testing slow.
Linear Regression	-Linear Regression is a statistical procedure for predicting the value of a dependent variable from an independent variable when the relationship between the variables can be described with a linear model [139]. -Used for regression [128].	-Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results. - Linear regression is often used to model non-linear	-Linear regression is limited to predicting numeric output. - A lack of explanation about what has been learned can be a problem -Does not work well for data with continuous or binary outcomes

		relationships [140]. -Linear regression is useful for data with linear relations or applications for which a first-order approximation is adequate [141].	
Logistic Regression	Logistic regression is classification analog of linear regression. It is preferable to trees in the same situations that linear regression is preferable to regression trees-when effects are small and when predictors contribute additively (no interactions) [142]. -Used for classification [128].	-Lots of ways to regularize the model, and one do not have to worry about the features being correlated [143]. -One can easily update the model to take in new data (using an online gradient descent method). -Use it if one wants a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're unsure, or to get confidence intervals) or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model [67]. -Logistic regression is intrinsically simple, it has low variance and so is less prone to over-fitting. -logistic regression is faster and more reliable when the dimension gets large. -Logistic regression is less complex and easier to inspect	-Text classification is a classic problem [139]. -It is unstable when one predictor could almost explain the response variable, because the coefficient of this variable will be boosted to as high as possible, here is a case when people turn to discriminant analysis. -Requires more assumptions and is sensitive to outliers [144].
Neural Networks (Multilayer Perceptron)	Can be used for both classification and regression purposes [128].	Powerful, robust and generally more accurate [128].	More complex and difficult to evaluate results [128].
UNSUPERVISED LEARNING ALGORITHMS			
Distance Based(K-Means)	K-Means methodology is a commonly used clustering technique. In this analysis the user starts with a collection of samples and attempts to group them into 'k' Number of Clusters based on certain specific distance measurements. -K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters [140]. -Pattern discovery where there are no data labels [128]. -Used for clustering	-With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). -K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular [60],[145]. -It is well suited to generating globular clusters. -The K-Means method is numerical, non-deterministic and iterative.	-The K-Means algorithm is repeated a number of times to obtain an optimal clustering solution, every time starting with a random set of initial clusters [140]. -Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome). -Fixed number of clusters can make it difficult to predict what K should be. -Does not work well with

			non-globular cluster [144].
Density Based(DBScan)	Density based clustering algorithm has played a vital role in finding non-linear shapes structure based on the density. -It uses the concept of density reachability and density connectivity [146]. -Builds clusters based on the density of data points -Used for clustering	-Does not require a prior specification of number of clusters. -Able to identify noise data while clustering. -Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is able to find arbitrarily size and arbitrarily shaped clusters [146].	-DBSCAN algorithm fails in case of varying density clusters. -Fails in case of neck type of dataset [146].
Association (Apriori)	-Used rules defining a pattern where condition A implies B [128]. -Used mining frequent item sets for boolean association rules [144]. -Uses a "bottom up" approach, where frequent subsets are extended.	-Allow the pruning of many associations -Uses large item-set property -Easily parallelized -Easy to implement [144].	-More search space is needed and I/O cost will increase [144]. -Number of database scan is increased thus candidate generation will increase hence increase in computational cost -Normally discover a huge quantity of rules, some being irrelevant.
Ensemble (Bootstrapping, Random Forest AdaBoost)	Can be used for both classification and regression	Uses several algorithms for more accuracy [128].	Interpretations of the results from multiple models can be more challenging [128].

3.3.1 Summary of Table 3.6

In summary of the supervised machine algorithms, Linear Regression algorithms' benefits are that, they are simple, fast and require little data and are widely used. However, their limitations are that they work with linear model assumptions. They are also suitable for numeric data only. As for decision trees, they are easy to understand, fast. However, there is a risk of over-fitting when working with them. As for the K-nearest neighbour, it is simple and allows use of the entire training set, however, it may have to experiment with K values to optimize. It works well with small numbers of inputs and with no missing data. It may need to normalize the data. Support Vector Machine is powerful, handles high dimensional data and is used for linear and non-linear models. However, they use only numeric or dummy coded categorical data and decision boundary can be difficult to interpret. Naive Bayes is fast and often considered as the baseline since it uses probability. However, it assumes input to be independent. Input is categorical or binary, but numeric for Gaussian. Neural Networks are powerful, robust and generally more accurate. However, they are more complex and difficult to evaluate results.

For the un-supervised Machine Learning algorithms, Distance based learning clustering algorithms are easy to use and to implement, number of groups must be selected ahead of time. However, they do not handle concave shapes well. Density based clustering algorithms can find clusters with concave shapes. However, a spatial index is required to run efficiently. Association algorithms, set rules defining pattern where condition A implies condition B and are simple. However, they can be memory and processor intensive. Ensemble algorithms use several algorithms are required for more accuracy. However, Interpretation of the results from multiple models can be challenging.

3.3.1.1 Justification for using Supervised Learning Algorithms

Based on the afore-mentioned information, the supervised learning algorithms were chosen to be used in the evaluation demonstration of the Framework through a Classification case application. These algorithms are: Naïve Bayes, Support Vector Machines, Decision Tree and K- Nearest Neighbour. The reason being that, they are able to work with categorical data as in this study, categories of data were deduced from the online data and were used for classification by the models. As pointed out by Hoyt et al.[128], these models work well with target variables which are categorical and mostly used for classification.

3.4 FRAMEWORKS FOR SOCIAL MEDIA ANALYSIS (CONTENT ANALYSIS)

Of late, the data mining community has shown much more interest in analysis of social networks. The simple enthusiasm is the demand to draw knowledge from ample amounts of data gathered, relating to social behaviour of online users [31]. This interest has led to the development of several computational methods to help mine and analyse social media data content. This section gives a brief summary of some of those works which have been developed as Frameworks for the mining and analysis of social media data. The summary is followed by a table making a comparison of these Frameworks.

i) Agichtein et al. 's classification Framework [147]:

This framework works by uniting evidence from diverse sources of information that can be adjusted for a given social media type, and quality definition by manipulating features of social media that are instinctively interrelated with quality. The classifier is trained to correctly choose and assess the features for each particular kind of item, task, and quality definition. The framework uses several classification algorithms for experimentation (e.g., SVM and log-linear classifiers).

ii) Stieglitz and Dang-Xuan's Social Media Analytics Framework [58]:

Stieglitz and Dang-Xuan came up with a methodological framework for social media analytics in political context. Precisely, the framework summarises most important politically applicable issues from the angle of political establishments and equivalent methodologies from distinct scientific disciplines. The framework was used to investigate whether Twitter messages have an impact on re-tweet behaviour with regards to re-tweet quantity as well as re-tweet speed as a vital feature for encouraging information diffusion in social networks. Computerised measures of sentiment analysis and regression analysis have been applied in the experimentations.

iii) Supervised Learning In Quest (SLIQ) [148]:

SLIQ has been developed for use in data mining applications which makes use of very large training datasets. SLIQ elegantly manages disk-resident data that is excessively huge to fit in memory. It is a decision tree classifier which has the capability of handling numeric and categorical attributes. It makes use of a unique technique for pre-sorting in the tree-growth phase, joined with a breadth-first tree growing strategy to allow classification of disk-resident datasets. In addition, SLIQ also uses a new tree-pruning algorithm that is reasonable in terms

of cost, and results in solid and precise trees. The blend of these techniques allows SLIQ to gauge for huge data sets and classify data sets regardless of the number of classes, attributes, and records.

iv) Scalable Parallelizable Induction of Decision Tree (SPRINT)[149]:

SPRINT is a data mining algorithm designed to remove the memory restrictions, speed and scalability issues. SPRINT was developed in order to overcome the short comings of a scalable algorithm called SLIQ which had been proposed by Mehta et al. In SLIQ, the in-memory data structure size grows directly proportional to the input records number, hence, limiting the amount of data that can be classified. Therefore, in SPRINT, the algorithm has been made to be parallelized easily, permitting numerous processors to work together to create a single reliable model.

v) DisCo: Distributed Co-clustering Framework [150]:

This framework introduced applied methods for distributed data pre-processing, and co-clustering by using Hadoop, an open source Map-Reduce implementation. DisCo scales perfectly well when dealing with large datasets. In addition, it processes and analyse very large datasets efficiently (up to quite a hundreds of gigabytes) by extracting and storing data in an appropriate representation.

vi) Luo' et al.' s Simple Multiplicative Algorithm (SMA) [151]:

SMA has been developed to address numerous, challenging and problematic data mining tasks, by parallelizing a successions of the data mining algorithms. The series of algorithms which have been parallelized are: Support Vector Machines, the nonnegative least square problems, and the LASSO (the ℓ_1 regularized regression). They have been transferred into multiplicative matrix processes that can be set up into numerous constituents cores or nodes. In these algorithms used, the processes involved only multiplication of a long vector that kept on changing with an unvarying huge matrix. That is how parallelization has been deployed, which required just less memory copy processing amongst computational nodes.

vii) Wu et al.'s Big Data Processing Framework [152]:

This is a data-driven framework that encompasses demand-driven accumulation of information sources, mining and analysis, user interest modelling, security and privacy conditions in relation to big data analysis. It serves as a guide to that. The Framework comprises of the three tiers from back to front. Tier I considers data accessing and computing, Tier II takes care of

data privacy and domain knowledge and Tier III is about Big Data mining algorithms. Wu et al, argue that, in normal data mining systems, the processing and operations require intensive computing units for data analysis and comparisons.

viii) HACE Theorem [152]:

The HACE theorem suggests that Big Data starts with large-volume, which comes from heterogeneous and autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. The name HACE comes from that explained notion. According to Wu et al. [152], one of the key features of the Big Data is the massive volume of data characterised by heterogeneous and various dimensionalities. This is so, because each information gatherer chooses their own schemata or protocols for data recording, and varying applications also results in dissimilar data representations. Wu et al., point out that, being independent, each data source has the ability to produce and gather information without getting help from any centralized control. In addition, as the volume of Big Data increases, the complexity and the relationships underneath the data also increases and gets complicated. Wu et al., argues that, these joint features indicate that Big Data needs a “big mind” to combine data for maximum values.

Table 3.7: A comparison of the existing Frameworks for Content Analysis

Name of the Framework	Approach type	Application tasks	Content	Content Analysis Techniques deployed?	Text Pre- processing -Human intervention & interpretation [A] -Computerised tools[B]	Text Post- processing -Human intervention & interpretation [A] -Computerised tools[B]
Agichtein et al 2008	Algorithms	Classification	Q & A Posts	Yes	[A] [Yes] [B] [No]	[A] [Unclear] [B] [Unclear]
Stieglits & Dang-Xuan 2013	Guide	Classification Regression	Posts(senti ments)	Only sentiment analysis	[A] [No] [B] [Possibly]	[A] [No] [B] [Possibly]
SLIQ	Algorithm	Classification	Posts/Mess ages	No	[A] [No] [B] [No]	[A] [No] [B] [Possibly]
SPRINT	Algorithm	Classification	Posts	No	[A] [No] [B] [No]	[A] [No] [B] [Possibly]
DISCO	Algorithms	Co-clustering	Posts	No	[A] [No] [B] [Possibly]	[A] [No] [B] [No]
Luo's SMA	Algorithms	Classification Linear regression	text	No	[A] [No] [B] [Possibly]	[A] [No] [B] [No]
Wu et al Big Data Processing Framework	Model guide	Guide for Big data analysis	Text	N/A	[A] [N/A] [B] [N/A]	[A] [N/A] [B] [N/A]
HACE Theorem	Theory/ guide	Guide for big data management, storage and analysis	Text	N/A	[A] [N/A] [B] [N/A]	[A] [N/A] [B] [N/A]

3.4.1 Summary of the gaps analysis of the discussed frameworks

Based on the criteria used to evaluate and compare these Frameworks in Table 3.7, the gap analysis which comes out of this evaluation show that, these frameworks have put more emphasis on statistical models and algorithms only. They are implemented through algorithms. Some are serial while others work in parallel. Most of these algorithms are leaning towards only classification of data, except for only two which addresses co-clustering and linear regression. Since most of them have put much emphasis on algorithms, content analysis is not carried out, if it is, then it is not documented.

Furthermore, when it comes to text pre-processing, which according to the criteria is either through human intervention and interpretation or through computational tools, it is clear that almost all of them except one do not actually involve the human intervention and interpretation in pre-processing of the text. It is not clear whether they pre-process the data using the computational tools either. On the issue of post processing of text, almost all the stated Frameworks do not carry it out. It has to be noted however that as other researchers have pointed out (e.g., [8, 34, and 58]), computational methods have a limitation with understanding

the in-depth meanings and semantics in data, which requires human intervention and interpretation. This gap has also been pointed out by Stieglitz and Dang-Xuan [34] in their limitation and future works of their framework that, even-though the computerized language measures allowed them to test the hypotheses on a large data set, the computerized language measures cannot capture irony, sarcasm, or culture-specific contexts. In their future researches, they suggest that they need to support and enhance their results by performing qualitative analyses. This clearly shows how qualitative analysis is necessary when dealing with social media data.

3.4.2 Motivation for the proposed Conceptual Framework based on gap analysis

The gap identified in the above frameworks of non-inclusion of human interpretation and intervention in pre-processing the data to capture in-depth meanings and semantics in data motivates the purpose for proposing this study's Conceptual Framework. The Conceptual Framework brings in the idea of content analysis which involves human interpretation and intervention in analysing the data in order to fine tune it for further analysis through computational methods. As pointed by Papadimitriou and Sun [150], many researchers commence with believe that the input data is always in a fitting and appropriate form for use and yet it is not. The Conceptual Framework does not put much emphasis only on the statistical and model algorithms for analysis of data, it allows both the content analysis (manual human intervention as well as the computational methods), this has also been highlighted by Chen et al. [8] and Stieglitz Dang-Xuan [34] that most research put emphasis on statistical models and algorithms, the same sentiments also has been pointed out by Stieglitz et al. [58] that most researches collect large data sets during a specific time frame on specific subjects and analyse it quantitatively. The application tasks of the proposed Conceptual Framework also are not only limited to one case, unlike in the discussed Frameworks, the Conceptual Framework is not only limited to classification of data only. Other tasks such clustering, regression, mining with association rules, etc., could be carried out.

The proposed Conceptual Framework also is not limited to data analysis only, it also gives a guide to the data gathering as well as data pre-processing tasks. This tries to address the gap as an observation made by Stieglitz et al. [58] that some Framework focus on data analysis only and yet social media analytics consists of several steps, of which data analysis is only one of. Before the data can be analysed, it has to be discovered, collected and be prepared. As pointed

out by Stieglitz et al. [58], the field of social media analytics aims to combine, extend, and adapt approaches for the analysis of social media data. This has gained considerable attention and subsequent acceptance in academic research. However, as pointed out by Stieglitz et al. [58], there is still lack of comprehensive discussions of social media analytics, and of general models. Even though existing frameworks are useful to decide what to research, and to locate individual projects within a larger context, they do not offer guidance on how to carry out the research, and which challenges might arise.

3.5 CHAPTER SUMMARY

This chapter discussed literature review on data analysis: the qualitative and quantitative analysis as well as approaches for each. A contrast was also made between qualitative and quantitative data. Data mining algorithms have also been discussed with regards to their strengths and limitations. These afore-mentioned concepts formed the basis and premise of the proposed Conceptual Framework. Some of the processes and concepts have been adopted to be used in the construction of the proposed Conceptual Framework. The related works have been discussed, with the existing frameworks that have been developed for social media content analysis. A gap analysis had been identified and discussed from these frameworks and it motivated the statement of the problem of this study. The limitations observed from these frameworks were that, they put much emphasis only on the statistical and model algorithms for analysis of data, and also, most of them only focused on data analysis aspect only. In order to address these limitations this study proposed a Conceptual Framework that offers a generic way to addressing the limitations of computational methods to social media data analysis of capturing in-depth meanings (semantics) in data.

CHAPTER 4

THE PROPOSED CONCEPTUAL FRAMEWORK

4.1 INTRODUCTION

This chapter presents the proposed Conceptual Framework and its components. The proposed Conceptual Framework is only discussed in relation to its components. Its application is discussed in Chapter 5.

4.2 THE BASE AND PREMISE OF THE FRAMEWORK

The main objective of this study was to develop a Conceptual Framework for mining and analysis of social media data that offers a generic way to addressing the limitations of computational methods to social media data analysis of capturing in-depth meanings (semantics) in data, which had been articulated in the statement of the problem. In designing the Conceptual Framework, this study was guided by the principles of Business Process Modelling (BPM) [153]. In business process management and systems engineering, BPM is the activity of representing processes of an enterprise, so that the current process may be analysed, improved, and automated. BPM has various techniques which are the simplest and most primary representation of the processes. The processes represented can be anything (e.g., a manufacturing process, an administrative or service process, a project plan, etc). In this study's context, there were data management and analysis process. This is a general tool that can be adapted for variety of purposes [153]. It guided the design, mapping and modelling of the proposed Conceptual Framework. In modelling and conceptualisation of the Framework, extensive review of the literature was made with regards to the previous works in relations to Frameworks, models and approaches which have been made in the domain of social media data analysis or content analysis. A summary of each of those works have been made in section 3.4 of Chapter 3. Furthermore, an articulation of the limitations/gaps of the other Frameworks, models, etc., proposed by other researchers have been made in Chapter 3, in sections 3.4 and section 3.4.1) and Table 3.7. A discussion on how this study (proposed Framework) will narrow those gaps was made in section 3.4.2. In addition to the afore mentioned, conclusions have been made under some of the processes of data analysis after literature review

discussion of each, conclusion makes it clear and indicate if the process will be adopted in development of the proposed conceptual Framework. All these aforementioned formed the base and premise of the proposed Conceptual Framework. Hence, the components which are used in the formulation of the Framework draws from the literature review discussions.

4.3 OVERVIEW OF THE PROPOSED CONCEPTUAL FRAMEWORK

This conceptual framework is generic to the task of mining and analysis of social media data. This section gives explanations and descriptions of the components of the Framework only. The application of the Framework is demonstrated in Chapter 5 through one application case of data classification. However, as shown in the Framework, it can be applied not only to classification, it can be applied into cases such as clustering, regression, association of rules, etc.

The collapsed view of the proposed Conceptual Framework shown in Figure 4.1 shows three iterative phases. The Framework phases start with data collection, followed by the text pre-processing and the data analysis phases. In between these phases, data is passed to the storage after collection and to knowledge repository after analysis. The large solid arrows show the connections between the phases denoting process flow whilst the dotted arrows show the data flow. A more detailed view of the actual steps and processes occurring in each phase are illustrated in Figure 4.2.

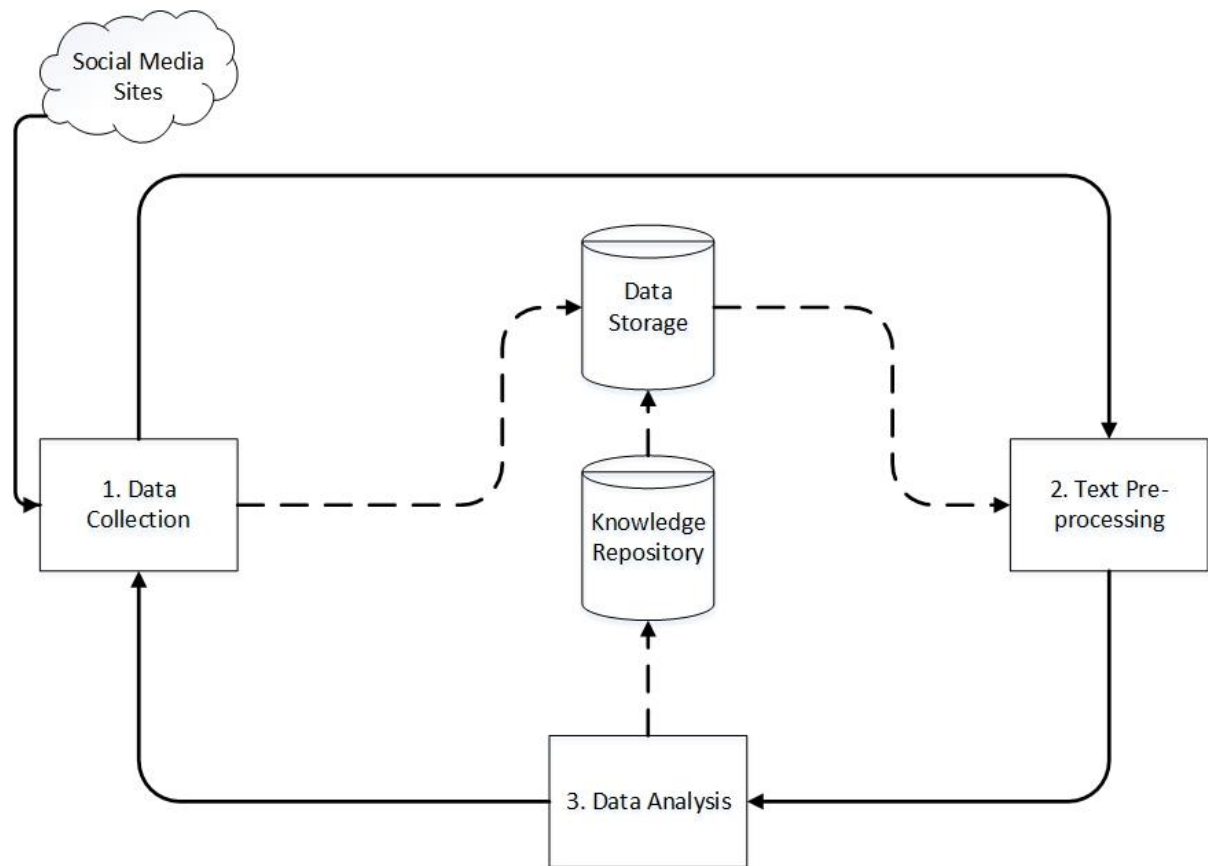


Figure 4.1: A collapsed view of the proposed Conceptual Framework for data mining and analysis of the social media data.

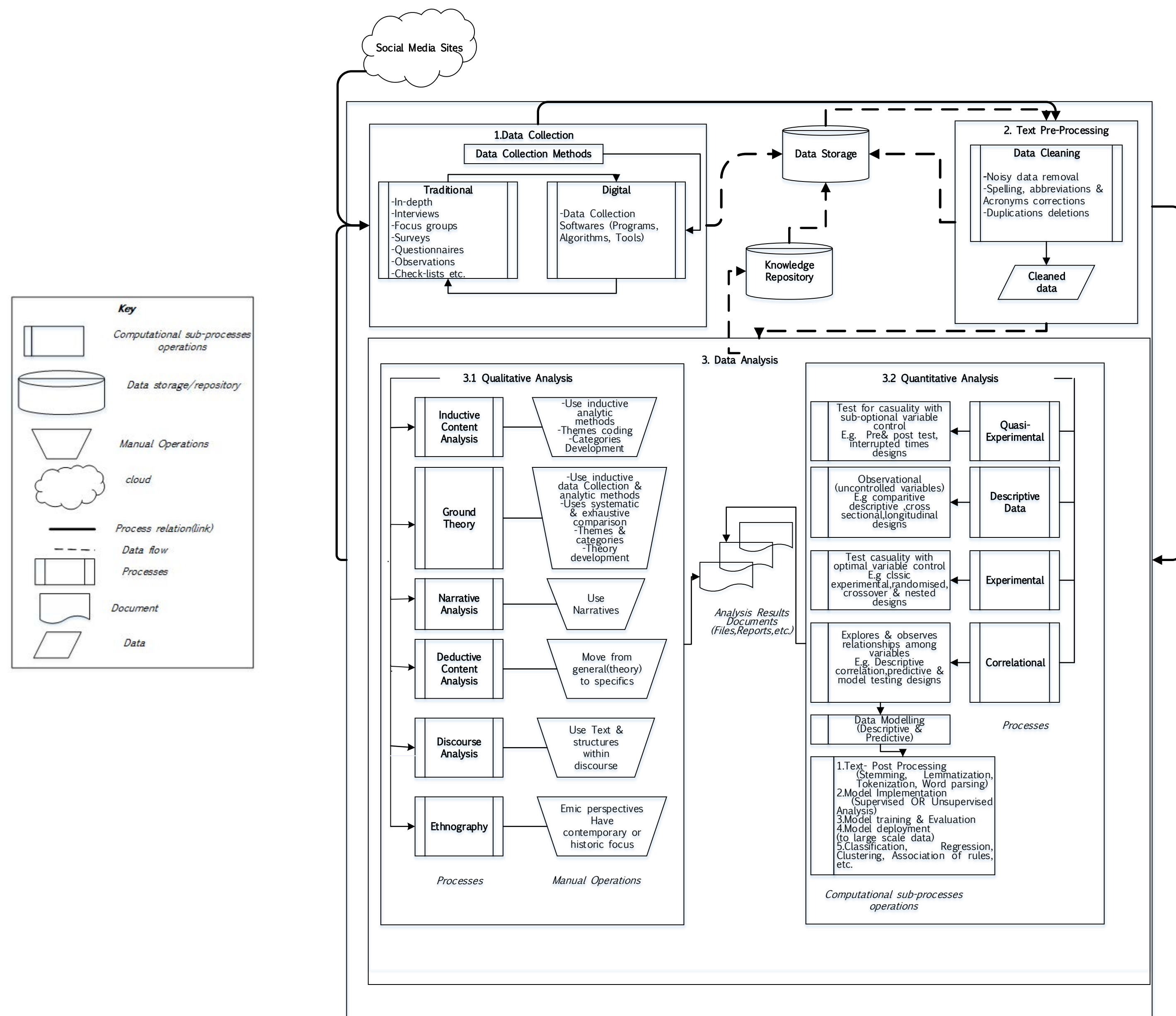


Figure 4.2: A detailed view of the proposed Conceptual Framework for data mining and analysis of social media data.

A more detailed view of the proposed conceptual Framework shown in Figure 4.2 shows all the three phases in depth: data collection, text -pre -processing and the data analysis with the steps and processes involved in each phase. It shows the data storage and knowledge discovery repository components. All these components and phases, which make up the entire Framework are explained and discussed more on sections which follow.

4.4 THE COMPONENTS AND PHASES OF THE PROPOSED CONCEPTUAL FRAMEWORK

The following sections discuss the phases and components of the proposed Conceptual framework. The three phases: Data collection, Text pre-processing and Data analysis phases are discussed respectively, together with the processes or steps taken in each phase. Other components of the Framework: data storage and knowledge discovery repository are also discussed.

4.4.1 Data Collection phase

The first phase of the proposed Conceptual Framework involves gathering data from different social media sites. Data collection is mainly a methodical approach to accumulating and assessing data from a diversity of sources to attain a comprehensive and precise picture of a phenomenon. Data collection allows an individual, a group, society, association, etc., to be able to give solutions to problems, answers important questions, assess results and come up with forecasts about future probabilities and trends [154]. Precise data collection is vital to preserving the integrity of research, generation of well-versed business decisions and safeguarding quality.

4.4.1.1 Data Collection methods

In the proposed Conceptual Framework, social media data can be gathered through the use of the digital methods. Furthermore, traditional methods could be incorporated in situations where a researcher does not understand certain things or aspects of a phenomenon in the information they already gathered through digital methods. In this case, a researcher has the option of visiting the “subjects case” and getting clarification through the use of the in-depth interviews as well as focus groups (in case of conducting qualitative research). This argument is supported by researchers such Boitnott [155]. Boitnott argues that, the evidence for digital data collection is pretty significant because it is cheaper, easier and can reach more people and tends to be more reliable. However, there may still be some places where traditional data

collection techniques make sense. Boitnott cite focus groups and in-depth interviews as some examples, which can allow one to go beyond quantitative data and into more qualitative discussions of trying to explore certain phenomena and getting an understanding and meanings from data. This is also supported by Hawe et al. [156], that Focus groups or group discussions are suitable in exploring a topic further, getting a wider understanding of why the target group behaves or thinks in a certain manner, and help in ascertaining the reason for such behaviour.

These methods are explained in the sections that follow.

- i) Digital methods: As shown in the Conceptual Framework, data could be gathered through the use of the digital methods such computerised tools, algorithms and programs. These tools tend to be fast and can collect vast volumes of data within a short period of time. There are several of these tools, some are open source, commercial off the shelf, educational tools and some are user generated customised programs.
- ii) Traditional methods: This is where information is gathered from people or sources through the use of the traditional ways as shown in Figure 4.2. This normally involves human beings. Some examples of these traditional methods of gathering data include: surveys, interviews, questionnaires, checklist, observations, and focus groups. These techniques can also be automated instead of being carried out and applied in a traditional manual way. They can be electronic and rolled out in an online setting shared through emails or links. These methods are very slow and take time in collecting data [8],[155].

4.4.2 Text Pre-Processing phase

The second phase of the Framework is text pre-processing. The data which has been gathered online and kept in data storages is actually processed here. It is cross checked if it is in a good state of interpretation and better understanding. Data is properly transcribed and transformed into written text before analysis can begin [110]. This makes it easy to read and interpret. Interpretations and corrections are made for abbreviations, acronyms and short language commonly known or referred to as slang. Cleaning of data involves removal of noisy texts, corrections of mis-spelled words, and discarding of repeated words. This is done because it is said that pre-processing operations affect the quality of the classification [157]. These text-pre-processing steps are performed during the initial stage of the qualitative analysis process. The pre-processing could be done by humans or through computational tools.

4.4.3 Data Analysis phase

The last phase of the Conceptual Framework is data analysis. This is where data is processed to make it meaningful and easy for understanding. Normally, information which gets processed is used to help in making informed decisions. This phase has **Qualitative** and **Quantitative** analysis sub-phases. In this Framework, social media data has to be analysed qualitatively before it could be analysed quantitatively. This is to overcome the limitation of the computational methods in analysing social media content. **Qualitative** and **Quantitative** analysis have been discussed extensively in the literature review chapter in section 3.2

Qualitative Analysis: Qualitative research produces rich, complete and usable processed data that helps in getting in-depth understanding of a perspective. It affects a repeated interaction amid theory and analysis. Qualitative research uses a descriptive and narrative style. In analysing qualitative data, researchers try to find to uncover patterns and probable fundamental links among variables. Qualitative interpretations are constructed, and numerous methods can be applied to analyse data. As shown in the Framework, qualitative analysis could be carried out through either one of the following approaches: grounded theory [109], Inductive Content/thematic analysis [122], discourse analysis, narrative analysis, ethnography, phenomenology, case study, etc. Qualitative analysis gives understandings into context investigated. Also, it often generates ideas and hypotheses for further use in quantitative research [123]. And in the context of this study's Framework, it prepares data for further analysis through computational methods quantitatively.

The reason for performing the qualitative analysis in the context of this study is that, social media data content is usually unstructured and ambiguous hence it requires human intervention and interpretation [8],[34]. There is high possibility of making incorrect assumptions if automated algorithms are involved in the analysis of large scale social media data without involving a qualitative intervention in the data [8],[34].

Content analysis usually starts amid the early phases of information gathering. This early immersion in the analysis stage assist researchers move forward and backward between idea advancement (concept development) and information gathering. It may coordinate their ensuing information gathering to sources that are extra valuable for tackling the research questions and objectives [116].

Quantitative Analysis: is useful in assessment because it offers quantifiable and easy to comprehend results [158]. As shown in the Framework, it can be performed through various approaches such as Descriptive design, Correlational design, Quasi-experimental and Experimental designs [125]. Experimental design includes moderate and advanced analytical methods. There are a variety of analytical procedures that one can perform with the social media data such as statistical and data modelling (application of machine learning and data mining algorithms). The modelling sub-phase focuses on the application of data mining algorithms in order to establish patterns in the data. These can be characterised as Descriptive and Predictive modelling.

For Descriptive modelling, orientation to data interpretations is made, with focus on understanding the fundamental data relations, such as identifying correlations, trends, groups, clusters and anomaly detection [66]. On the other hand, the Predictive modelling is grounded on inductive learning, where a model is built by generalising a set of training records. The purpose of this modelling is to forecast the value of a specific attribute (i.e., target variable) based on the values of other attributes. The basic assumption of the inductive approach is that the trained model should be capable enough to forecast the values of the future unseen attributes accurately [66].

The algorithms used to build these models could either be supervised or un-supervised learning algorithms [159]. Naturally, a data mining problem can be solved through various techniques and methods, there is no defined one standard procedure on how a problem could be solved (for example, clustering, association rule mining and prediction). Prediction can be divided into classification and regression. For each technique, several algorithms can be applied (e.g., for classification, several algorithms can be used to classify the data like Naive Bayes, Decision trees etc.)

Some techniques have certain requirement(s) on the form of data. Hence, a considerable attention is often needed to be paid to the data so that those requirements are met [66]. When implementing the algorithms as shown in the Conceptual Framework. This is normally called text post processing. Examples of some of the tasks which are carried out in post processing in order to prepare the data for modelling includes word parsing and tokenization, attribute selection, feature extraction, lemmatization and stemming. These are carried out before data

could be supplied to the models for further analysis. This helps improve accuracy and performance of the models.

Evaluation of the models before proceeding to the final deployment is very important and necessary [66]. The models are tested after training, based on the evaluation measures which would have been proposed to be used at that particular instance. The models have to be reviewed and evaluated to ensure that they achieve the intended purpose accurately and that their performances are optimal. The data models are deployed into usage if there are considered to be operating well. They are deployed to perform various tasks of analysis such as classification, regression, clustering, etc.

Finally the processed and analysed data that comes as a result of the data analysis phase is passed to the knowledge repository for storage and use for future projects. The afore-stated phases (Data collection, Text-pre-processing and Data analysis) of this proposed Conceptual Framework are iterative.

4.4.4 Data Storage and knowledge Repository

The proposed Conceptual Framework has data storage and knowledge discovery repositories components. Data storage is where data is kept and saved immediately after it has been collected. On the other hand, the knowledge repository acts as storage of information which had been processed and analysed for further/future use. As described by Techopedia [160], data storage is a broad term referring to the process of storing data in electromagnetic or other forms for use. There are various kinds of data storages and these play diverse roles in a computing environment. Data could be stored in hard storages (external devices) or soft storages such as cloud computing [160]. Well, the data storage and repository in the proposed Conceptual Framework could be any of kind, which the researchers will be comfortable with using and also be able to afford. The format of data stored will also depend on the kind of storage facilities infrastructure and its requirements on how it should be.

4.5 CHAPTER SUMMARY

This chapter discussed the methodology that was applied in carrying out this study. It described the research purpose, research design and research approach that were applied in carrying out this study. The chapter also discussed the overview of the proposed Conceptual Framework and its components. The proposed Conceptual Framework has been illustrated on two views: the collapsed view and a more detailed view. The collapsed view showed the

phases and components of the Framework without the steps and process whilst a more detailed view showed the phases and the components together with the processes and steps involved in each phase. The components of the Conceptual Framework include data collection phase where data is gathered from various online sources. In this stage, it is indicated that data could be gathered through traditional methods or computational methods. The second stage of the Framework is text pre-processing. In this stage data is cleaned and made ready for analysis. The third stage is of data analysis which had two sub-phases of the qualitative and the quantitative approaches. The Conceptual Framework has data storage and repository components which stores data before and after analysis. The next chapter discusses the evaluation of the proposed Conceptual Framework.

phases and components of the Framework without the steps and process whilst a more detailed view showed the phases and the components together with the processes and steps involved in each phase. The components of the Conceptual Framework include data collection phase where data is gathered from various online sources. In this stage, it is indicated that data could be gathered through traditional methods or computational methods. The second stage of the Framework is text pre-processing. In this stage data is cleaned and made ready for analysis. The third stage is of data analysis which had two sub-phases of the qualitative and the quantitative approaches. The Conceptual Framework has data storage and repository components which stores data before and after analysis. The next chapter discusses the evaluation of the proposed Conceptual Framework.

CHAPTER 5

EVALUATION OF THE PROPOSED CONCEPTUAL FRAMEWORK

5.1 INTRODUCTION

The proposed Conceptual Framework gives generic ways to the mining and analysis of data. It could be applied to various cases. However, in this study, the application of the Framework has been demonstrated on the Classification case as one of the many examples which could be applied in the analysis of data. Other examples of analysis which could be performed in the data are shown in the Framework, (i.e. Clustering, Association and Regression).

The proposed Conceptual Framework has been evaluated in twofold: 1) by classifying data by applying the Inductive Content Analysis to clean, pre-process and uncover patterns and themes to classify data into categories. 2) by automatically classifying data through data mining algorithms: Naïve Bayes, K-Nearest Neighbour, Support Vector Machine and C4.5 Decision Tree, using the categories which have been developed in ICA.

After completion of the ICA process, it was evaluated for usability as per the ISO 9126 model. The evaluation of ICA was conducted through a user study, using real data and human participants in the form of focus groups who carried out the ICA. The data models were evaluated too, through the use of metrics to measure their accuracy and performance.

This evaluation also involved gathering data online and cleaning the data (pre-processing). These are the first two stages depicted in the Framework. These are other cases in which the proposed Conceptual Framework could be applied to, one can apply it to the task they are interested in, e.g. if they only want to collect data, or only want to perform text-processing or want to performed data collection, text pre-processing as well as analysis, they can choose to do so. This proposed Framework gives them that flexibility. For demonstration of this Framework, data was collected from the social media sites: Facebook and Tweeter, it was pre-processed both manually and automatically and it was finally analysed through ICA and data modelling. This is indicated through the highlighted components of the Conceptual Framework in Figure 5.1. The ICA process has been applied to help mitigate the shortfalls of the computational methods of failure to interpret and get-in-depth meanings from social

media data, which normally results in poor classification. As mentioned in the introduction chapter and on the purpose of the study, the computational approaches have limitations of failure to understand semantics and in-depths meanings from data. The evaluation of the Framework is explained in the next sections of this chapter which follows

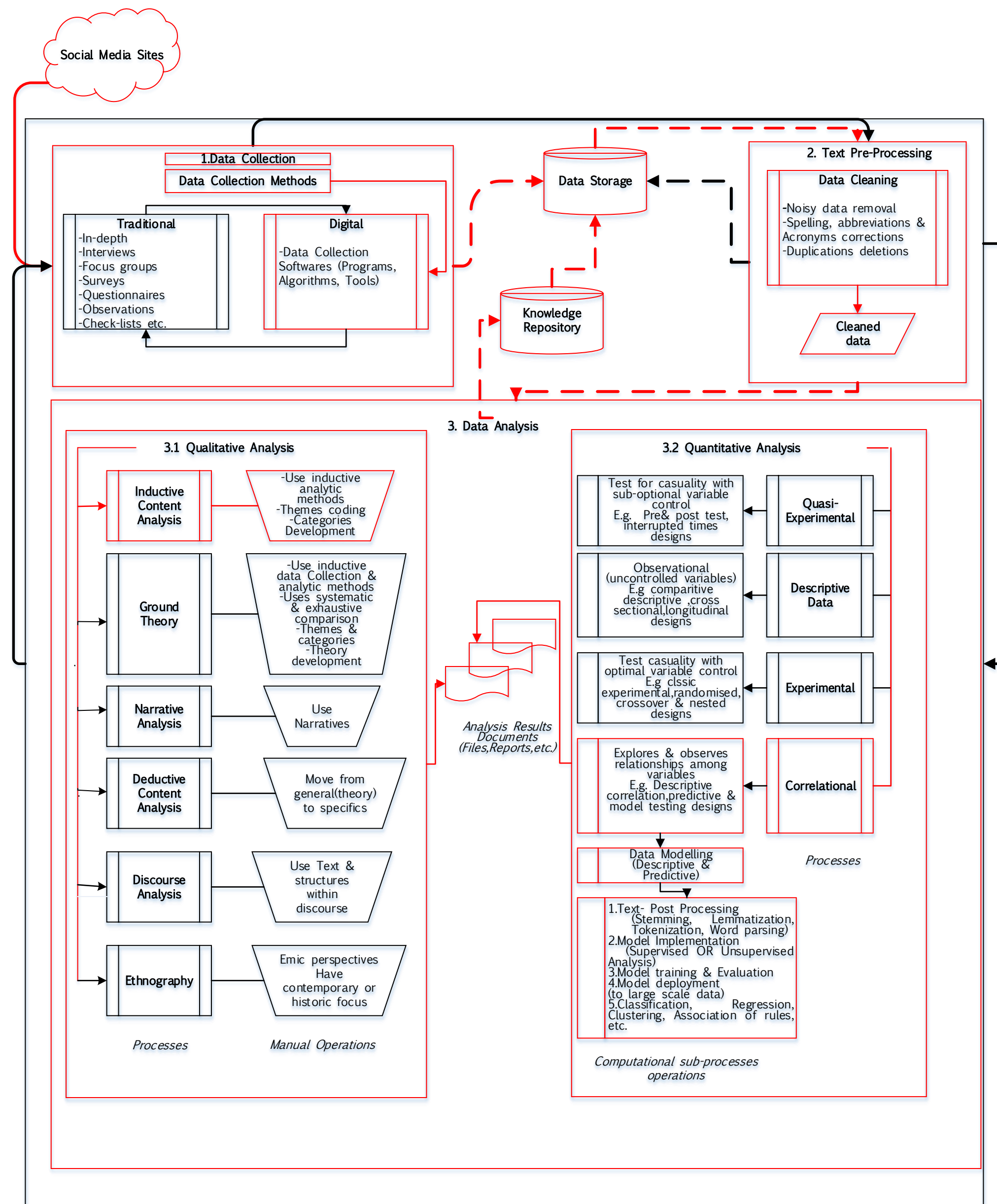


Figure 5.1: The Evaluation of the proposed Conceptual Framework: The highlighted steps were carried out as a demonstration of the application of the Framework

5.2 DATA COLLECTION

The data sets which were used for the evaluation of the Framework were gathered from Facebook and Twitter on the political landscape of Botswana (The data could have been on any topic, it is used here as a test case). Data from Facebook was gathered through the help of a social media-monitoring tool called Netlytic [161] and a user ID generating tool site called lookup-id.com [162]. The Netlytic tool allowed us to gather information from Facebook's open public groups and pages of Botswana political parties: Botswana Democratic Party (BDP), Botswana Movement for Democracy (BMD), Umbrella for Democratic Change (UDC), Botswana National Front (BNF), Botswana Congress Party (BCP) and the University of Botswana- Umbrella for Democratic Change (UB-UDC).

Facebook had a constraint of not permitting us to get data or posts from individual profiles of users even though individual users shared more content online in their timelines. It only allowed us to get posts from open public groups. Consequently, this constraint was overcome by implementing a Java program utilizing the Twitter API's to get information from individual profiles in Twitter. With regards to Twitter, searches were made with hashtags and individual names of prominent political parties' leaders.

5.3 DATA PREPARATION

Data was prepared and made ready for analysis in two ways: text pre-processing and post processing. These are discussed in the sections that follow:

5.3.1 Text Pre-Processing

Text pre-processing was performed by the focus groups on the data by cross checking if the data sets which had been collected online were in a better state to understand and interpret. Data was cleaned and interpretations were made where necessary. Translations and amendments were made for truncations, acronyms, short dialect (slang) and noisy messages were deleted, spelling mistakes were rectified, and repeating words were removed. This was done in light of the fact that it is said that pre-processing operations affect the quality of the analysis of data. This pre-processing was performed before carrying out the Inductive Content Analysis process. This was done with regards to what Patton [110] pointed out that content analysis can be applied to numerous kinds of data as the data is being converted into written text amid analysis process.

i) Text Post Processing

After the focus groups finished with pre-processing the data and carrying out the qualitative analysis, the data was post processed to make it ready and suitable for classification through data mining algorithms. This involved performing the following tasks to the data: word parsing and tokenization, feature extraction, attribute selection lemmatization and stemming (See Appendix F). In order to see how noisy or un-noisy data affected the classification through data modelling, two experimental tests were performed. The first experiment was carried out with data which was only partially pre-processed through the manual process. The second experiments were done using data which was fully pre-processed and post processed, both manually and automatically. This was done to compare if there will be any differences in the performance of the models in terms of accuracy. The first classification could be deemed as having the noisy data as the pre-processing was not fully completed when data was classified through models, and the second test could be deemed to have used un-noisy data as the pre-processing task was fully completed. The results of these experiments are captured in the histograms in Figure 5.2 and 5.3. However, it has to be noted that the manual analysis was mandatory for all the experimentation because, if the data is totally un-structured in a noisy format with wrong spellings mistakes, the data models will totally fail to even load the dataset(s).

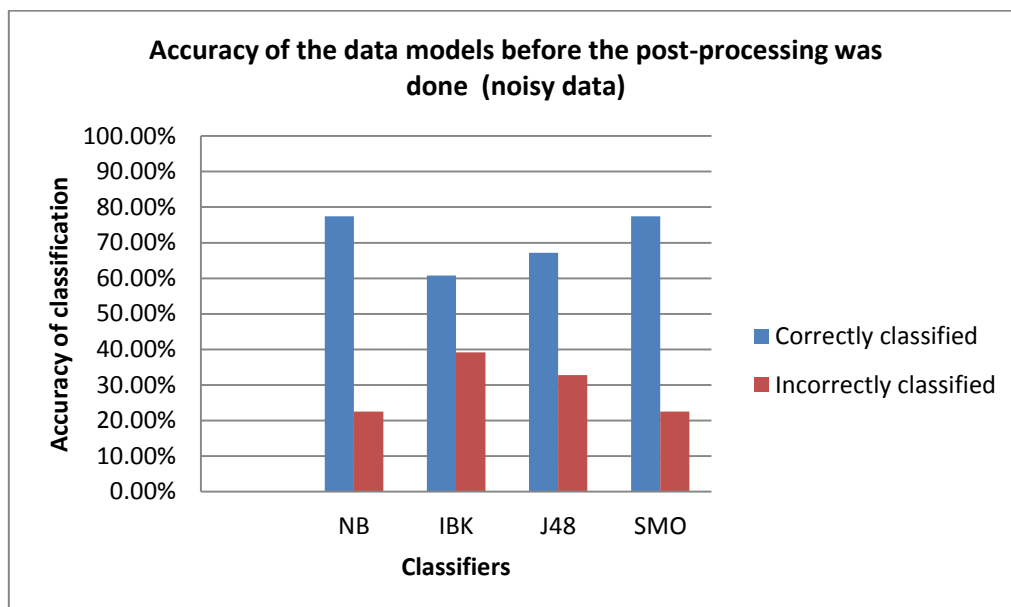


Figure 5.2: Model performance in terms of accuracy of data with partial preparation (pre-processing only)

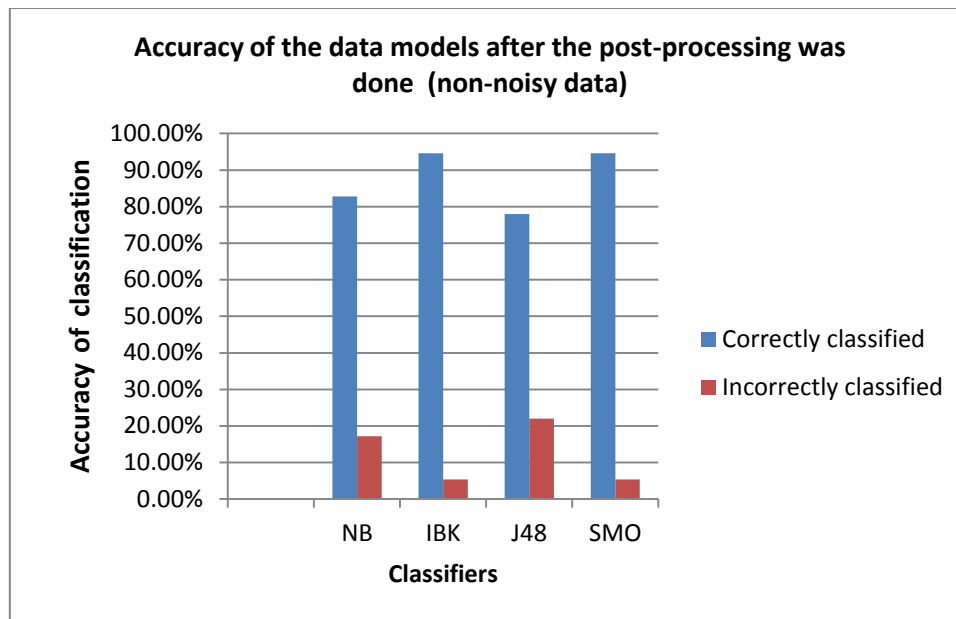


Figure 5.3: Model performance in terms of accuracy of data with complete preparation (pre & post processing)

Comparison of the two histograms shown in Figure 5.2 and 5.3, show that, data models performed better when the dataset were fully pre-processed and post processed (both manually and automatically). The readings are slightly higher than those of partial pre-processing. Even though the first experiments where data was partially pre-processed through manual way only, is lower than the second experiments, it can be observed that, still the incorrect classifications are less than 40%, which is reasonable. This is because manual pre-processing played a major role to the attainment of that, hence manual pre-processing could not be ignored as it is mandatory.

5.4 INDUCTIVE CONTENT ANALYSIS

The process of Inductive Content Analysis (ICA) was carried out by the researcher and four focus groups. These were twenty four second year Computer Science and Information Systems students and two post graduates Computer Science students. Participation in the study was on a voluntary basis, and the participants signed a consent form to indicate their agreement to participate (see Appendix B). Justification for this sample size is that, samples sizes used for qualitative studies are usually much smaller than those used in quantitative studies [102], [163]. Crouch and McKenzie [119] state that, the reason for using a small sample size is because qualitative research is aimed at getting in-depth meanings and not constructing generalised hypothesis statements. Qualitative analysis is dependent on Saturation. Finding most or all of the insights will result in the accomplishment of

saturation. Saturation happens when addition of more participants to the study no longer result in additional perspectives or facts. Glaser and Strauss [109] endorse the idea of saturation for attaining the right sample size in qualitative studies. Other guidelines have also been suggested: Morse [115] suggests about 30-50 participants for ethnography; Morse also has suggested 30-50 for grounded theory, at the same time, Creswell [113] suggests only 20-30 participants. And for phenomenological studies, Creswell recommends 5-25 participants while Morse suggests at least six. According to Mason [102], there are no exact rules when determining a suitable sample size in qualitative research. Qualitative sample size may best be determined based on the time allocated to the study, availability of resources and objectives of a study [114]. Twenty four participants were therefore deemed appropriate for this study.

Before the focus groups could actually start, the researcher trained them on what ICA is and how it is carried out. They were also provided with guidelines to follow. These guidelines also contained the manual coding guide for classification (see Appendix C for these guidelines). These guidelines were informed by various qualitative analysis research ideas especially on the Inductive Content Analysis subject [110], [116], [157], [164]. This was done in support of the argument that, in order to produce valid and reliable inferences there has to be some established clear rules and procedures for processing data in qualitative content analysis [103]. Content analysis can be flexible or more standardized depending on the goals of the study. However, it can begin with data preparations, text- proceeding, text analysis up through to reporting of findings in a study [118].

After cleaning the data, the focus groups carried out the following steps of Inductive Content Analysis:

Step 1: Defining the Unit of Analysis

In this Inductive Content Analysis process, the unit of analysis was defined as single words and some short expressions. Each sentence/post or tweet was viewed as a code fragment. Each code portion was comprised of a few words. This was done because in content analysis, researchers need to define the unit of analysis. The unit of analysis is the simple piece of text to be analysed during content analysis. Messages have to be broken down into units prior to coding data, and much consideration has to be made as variances in the unit definition can impact coding decisions and the comparability of results with other related studies [165].

Hence, as pointed out by Weber [165], defining the coding unit is one of the most fundamental and important decisions.

In Content analysis, separate themes are generally used as the unit for analysis. An illustration of a theme could be conveyed through a single word, a phrase, a sentence, a paragraph, or a whole document [166]. When using a theme as the coding unit, a researcher is mainly trying to find out some expressions of a concept. Thus, they might assign a code to a text chunk of any size, as long as that chunk represents a single theme or issue of relevance to context of study under investigation [166]. This was taken into consideration when carrying out this step of ICA in this study. From the data which was collected from Facebook and Twitter on political landscape of Botswana, a lot of these posts were a huge chunk of paragraphs. The themes or unit of analysis were therefore made of a single word, in some cases small phrases were used to code the themes (not sentences though, just a few words, see Appendix A).

Step 2: Developing Categories and a Coding Scheme

When it came to categories development of this study, the focus groups applied inductive reasoning; they carefully examined the data through constant comparison and immersed themselves within the data so as to deduce themes and categories. Literature emphasises that coding schemes could be designed either inductively or deductively [118]. During this process, the focus groups members allowed themes to emerge directly from the data. They also used the summative content analysis approach (see appendix A). This in agreement with what Zhang and Wildermuth [118] have highlighted that categories and a coding scheme can be derived from three sources: the data, previous related studies, and theories. In this study, the categories came from data which was being analysed.

This study involved the use of focus groups, so multiple coders were involved hence on ensuring consistency of coding, they were trained and given some guidelines on how to code and categorise the data. Content analysis permits assign a theme to more than one category concurrently [103]. Training was done because it is said that when more coders are involved, consistency of coding ought to be ensured. A coding manual had to be designed, which typically contains descriptions or rules for coding the data, and examples have to be given also [167]. Often some manuals have an extra field for capturing of notes as coding proceeds. The coding manual usually used in this study contained descriptions and

instructions on how to carry out the ICA process (see Appendix C). Also, to facilitate the whole process additional documents were provided to help with the analysis process.

Step 3: Testing Your Coding Scheme on a Sample of Text

It was ensured that clarity and consistency of the category definitions were taken into consideration following the coding guidelines provided in the manual by coding one of the data set(s) by passing on to another sub set of the focus groups participants for further verification. Then during the focus groups meetings, each member of each focus group discussed and agreed on the coding of the categories they selected. Then, afterwards all the focus groups participants agreed collectively on the final categories which should be selected. All this occurred because as pointed by Zhang [118], a better way of checking clarity and uniformity of the category definitions is to code a sample of the data.

After coding a sample, the coding consistency has to be examined, usually by careful assessment between the coders thorough what is called of inter-coder agreement. The coding rules should be reviewed if the level of consistency happens to be low. Also, as Schilling [117] postulates, any arising problems and doubts with regards to definitions of categories, coding rules, or categorization of specific cases, need to be conversed and resolutions should be made collectively by the research team working on the study. All afore- mentioned steps are to be repeated until sufficient coding consistency is achieved [167]. This was kept in mind throughout the entire ICA process until its completion for this study.

Step 4: Category development

Schilling [117] suggests that, when adequate consistency has been attained, the coding rules can be used to code the whole dataset(s). While coding, one has to keep on checking the coding closely, to avoid “drifting into an idiosyncratic” and ignoring what the codes mean. Hence, all the data for this study was carefully examined and the irrelevant ones were discarded. Relevant data was all coded in the relevant categories. The relevant data was the one containing content on the political landscape of Botswana (as this was the subject matter of the test case).

The data sets which were selected for qualitative analysis contained 2037 code segments. However, 1483 segments were discarded because there were irrelevant and not containing any information related to politics. Only 654 segments which contained information on the political landscape of Botswana were used for the manual analysis categorization. This

clearly corroborated what Thomas [157] has pointed out in his qualitative analysis steps, that amongst the generally presumed rules that inspire qualitative coding, a substantial quantity of text might not be allocated to any category, because the text may be not related to the study objectives. Refer to Appendix A, to see how each focus group categorised the data.

The focus groups then collaborated as one forum to further collapse the categories into important ones. In this exercise, twenty eight important categories were identified to have emerged from the data collectively from all the four focus groups. These categories are: Economic scandals, Corruptive governance, Incompetent governance, Judiciary agitation, Un-transparent governance, Unfair rigged elections, Abuse of power, Poor workers welfare (conditions), Lack of proper benchmarking (judiciary, political & economic), Partisanship, Cronyism, Disenchantment of rightful cadres, Poor education delivery, Poor education mandates, Poor education system, Un-consultative governance, Workers oppression, Unbeneficial economic programmes, Political propagandas, Political dynasties and elites, Draconian labour laws, Forceful and immoral enact of labour laws, Workers exploitation, Violation of rights and freedom, Disrespect to rule of law, Extra judiciary killings wastages of tax money, and mis-use of public funds (see Figure 5.4 showing each of the four focus group's categorization).

During the process of analysis, there was an overlap of some themes (i.e. some themes were belonging or featuring in more than one category). Thus, one segment could belong to several categories (this is commonly referred to as multi label classification). It was noted also that some categories overlapped. This is what Thomas [157] described as “links” in his qualitative analysis steps, that one of the commonly assumed rule of qualitative analysis is that one section of the text assigned to multiple categories.

Even though several themes emerged from the data which led to creation of a number of categories during the Inductive Content Analysis phase (see Figure 5.4 and also refer to the Appendix A which show the themes and categories), only eight prominent categories were finally selected. This was consistent with the Qualitative Inductive Analysis rule that a good exercise should end up with a range of three (3) to eight (8) final prominent categories [157], [168]. The general rule is that an inductive coding that ends up with major categories exceeding eight may seem as incomplete. If this happens, the researcher (coder) needs to

combine some categories which have a link or have to carefully decide which themes and categories are most important and so relevant.

The final eight prominent categories, which emerged from this study are: Economic and political scandals, Corruptive governance, In-competent governance, Judiciary agitation, Un-transparent and Un-consultative governance, Partisanship and cronyism, Poor education system and Oppressive governance (See the centre of Figure 5.4 also Step 6). Most of the data published online came from opposition political parties. This is because the majority of members of these parties are youth who are apt with technology. This then should not be construed to portray the overall political landscape of Botswana as such generalizations cannot be made from our datasets to represent the entire population of Botswana. However, the datasets served the purpose of achieving the goal of this research as stated in the study's objectives. The political landscape datasets were only used for the purpose of classification as an example of social media data being part of Big Data, not merely for understanding and analysing the political landscape of Botswana.

Step 5: Assessing the Coding Consistency

To ensure that the coding was consistent for this study, after every exercise before the break, the coders stopped and made a quick check and some brief discussions to ensure that they do not drift. It is always advisable for researchers to recheck the consistency of the coding at the end of coding the whole dataset. Zhang [118] cautions about the danger of assuming that, the whole coding of the entire corpus was consistent just because one sample was coded in a consistent and reliable matter. Because of fatigue, human coders are probable to making some more mistakes as the coding continues. The ICA process of our study was carried out by human coders hence a precaution was taken to avoid being trapped in this situation. It is cautioned that, from the initial consistency inspection, there could have been new more additional codes which had been added, and also the coders' interpretations of the categories and coding rules may change slightly over time, which may cause more irregularities [116],[167]. Hence, each time after coding new categories, the code books were updated to ensure that there was consistency throughout the entire ICA process.

Step 6: Drawing Conclusions from the Coded Data

At this point, the focus groups discussed and made some conclusions with regards to the final themes and categories collectively in an open discussion. This has been echoed by Zhang and

Wildermuth [118] that researchers have to make extrapolations and report their interpretations of meanings derived from the data. The focus group re-looked at the twenty eight categories carefully to identify relationships and similar patterns amongst them. This is what Bradley [169] advises on activities such as discovering the properties and features of the categories, finding associations between categories, discovering patterns, and testing categories alongside the full range of data should be done in order to finalise the selection of the final prominent categories. This view is also shared by Zhang and Wildermuth [118] that in this step, a researcher has to make sense of the themes or categories identified, and their properties. According to Zhang and Wildermuth [118], this is a vital step in the analysis process, and its accomplishment is dependent on one's reasoning capabilities.

In this study, Inductive Content Analysis only covered up to presenting the selected eight categories (see the centre part of Figure 5.4). The ICA was performed in this study in order to come up with themes or categories which were used for further classification of data through the use of data mining algorithms. Anything beyond this step was outside the scope and objectives of this study. Figure 5.4 shows categories as per each of the four focus groups. The centre part of Figure 5.4 shows the final collapsed eight categories which were finalised by the whole focus groups collectively and used for further classification using the data models.

5.4.1 Justification for using ICA over other Qualitative Approaches

ICA was selected to be used to qualitatively analyse the data because, firstly there were no pre-defined themes to start with, so ICA plays a very vital role in determining the themes because researchers immerse themselves with the data, interpret it to uncover hidden in-depth meanings which could explain certain phenomena. Again, since the application case of the Framework was classification, ICA was the most suited approach to use because the aim was to create some categories which could be supplied to the models. Also, the classification models (NB, SVM and K-NN, C4.5) are mostly used for classification because they are supervised learning and work with the target variables which are categorical (nominal). In addition, the datasets also had variable instances that had to be made categorical hence the use of ICA.

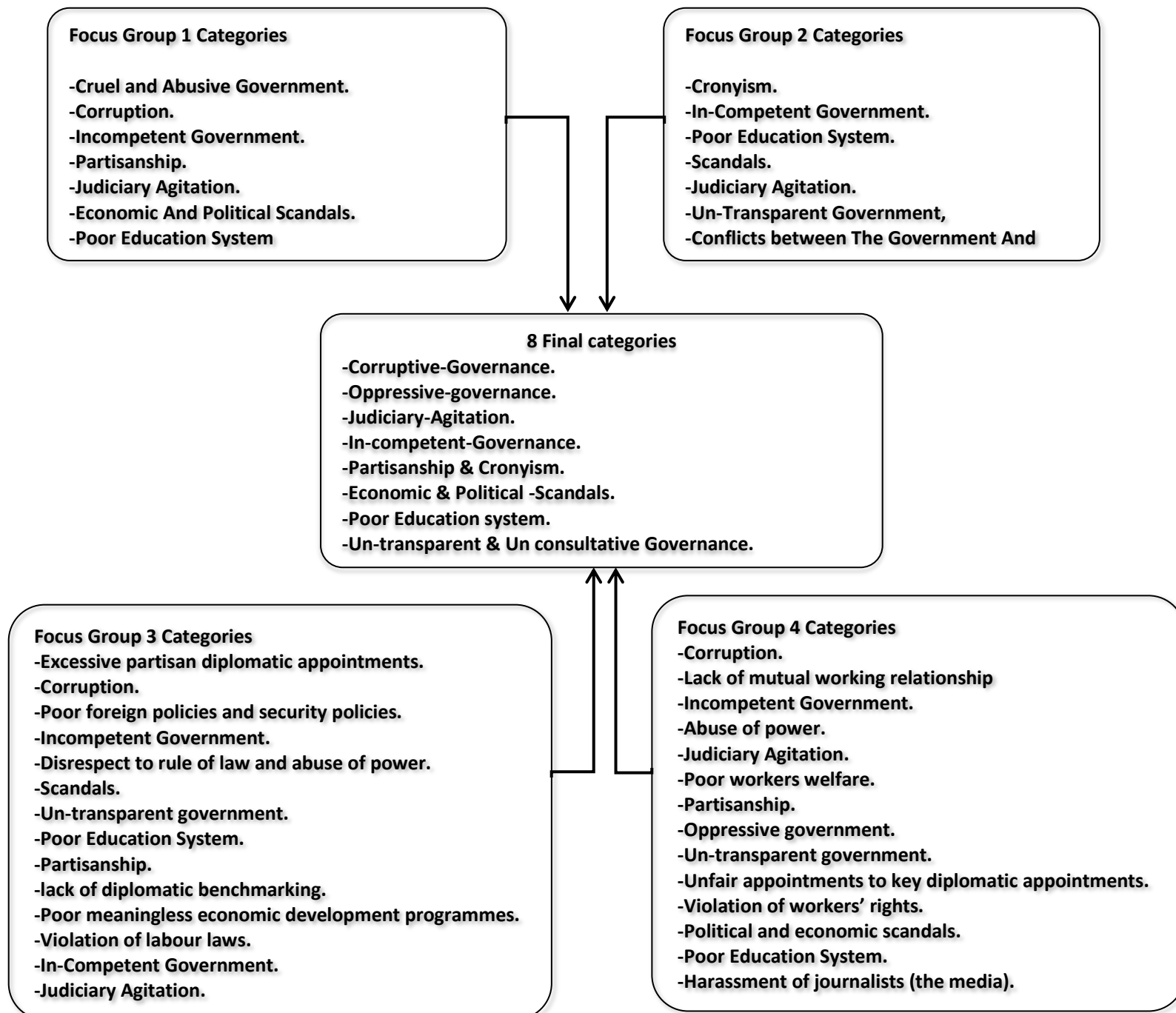


Figure 5.4: Categories per each focus group and the final eight selected categories

5.4.2 The Usability Evaluation of the Inductive Content Analysis

Upon completion of the Inductive Content Analysis (ICA), an evaluation of ICA was made. The ICA process was evaluated based on “Usability”, one of the components of the ISO 9126 model [170],[171],[172]. The ISO 9126 model is commonly used to evaluate software and systems. The purpose of this evaluation was for validity and reliability purposes to check if the focus groups were not biased and carried the ICA steps accordingly. For this study, a 30 question based usability evaluation questionnaire was formulated. The questions which have been asked in the aforementioned questionnaire have been adapted from eight usability testing questionnaires from various researchers which are widely and commonly used for usability testing of websites and software systems. The usability testing questionnaires which formed our sources are: Questionnaire for User Interface (QUIS) by Chin et al. [173], Perceived Usefulness and Ease of Use (PUEU) by Davis [174], Nielsen’ Attributes of Usability (NAU) by Nielsen [175], Nielsen’ heuristic evaluation (NHE) by Nielsen [176], USE Questionnaire: Usefulness, Satisfaction and Ease of use (USE) by Lund [177], After Scenario Questionnaire (ASQ) by Lewis [178], Practical Heuristics for Usability Evaluation (PHUE) by Perlman [179] and Purdue Usability Testing questionnaire (PUTQ) by Lin et al [180].

Questions that addressed learnability, perceived ease of use, perceived usefulness, flexibility and satisfaction were targeted. These are some of the aspects that build up to the “usability” characteristic/component as per the ISO 9126 model. The questionnaire also captured the demographic details of the participants. The questionnaire was rolled out to the participants who took part in carrying out the ICA for this study. From the twenty six participants who took part in carrying out the ICA, those who later on were able to take part in evaluating the ICA were twenty four. The following sections show the results and the discussion of the evaluation of the ICA process.

The results in Table 5.1 show that the participants who took part in evaluating the ICA were twenty two year two students and two post graduates students. There were fourteen (14) male and eight (8) females who answered the questionnaire. These results show that there were more males than the females participants. There was no one who was aged below twenty (20). From the total number of those participants who answered the questionnaire, three (3) females were doing Computer Science (CS) whilst five (5) were doing Information Systems (IS) and twelve (12) males were doing CS and four (4) were doing IS.

The majority of the participants were proficient on both Setswana and English languages. Only three (3) participants indicated that they were not proficient with Setswana. These demographics give an impression that when conducting the ICA, it is possible to apply or carry it out across any discipline, like in this case it has been carried out or applied by participants with Computer Science and Information Systems backgrounds. Furthermore, it shows that one had to be proficient in English in-order to carry out the ICA. The instructions of ICA are laid out in English and even data sets have to be properly transcribed to a standard language which could be easily understood by everyone which was English.

Table 5.1: Demographic details of the Users who participated in conducting the ICA process

	Age			Program of Study			Year of Study			Language Proficiency (English & Setswana)		
Gender	20-25	=<31	Totals	CS	IS	Totals	Year 2	Post- Grad	Totals	English Only	Both Languages	Totals
Females	8	0	8	3	5	8	8	0	8	0	8	8
Males	14	2	16	12	4	16	14	2	16	3	13	16
Totals	22	2	24	15	9	24	22	2	24	3	21	24

Table 5.2: The descriptive statistics analysis of the answers provided by participants on learnability of ICA

	Learnability Questions									
	I easily remember the ICA process		ICA process steps are clear and understandable		It was easy to learn ICA steps and apply them		I performed the ICA process quickly		The supplement reference materials for ICA were clear and helpful	
Values	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Strongly disagree	1	4.2	0	0.0	0	0.0	0	0.0	0	0.0
Disagree	0	0.0	0	0.0	0	0.0	1	4.2	0	0.0
Neutral	1	4.2	2	8.3	1	4.2	0	0.0	0	0.0
Agree	10	41.7	12	50.0	12	50.0	17	70.8	5	20.8
Strongly agree	12	50.0	10	41.7	11	45.8	6	25.0	19	79.2
Totals	24	100.0	24	100.0	24	100.0	24	100.0	24	100.0

The results from Table 5.2 show the participants' perception on the learnability of ICA. The majority of the participants pointed out that it was easy to remember the ICA process. It can be realised that, from Table 5.2, 50% of the participants strongly agreed and 41.7% agreed to this statement. The number of those who strongly agreed and those who agreed adds up to 91.7% which clearly indicates that the ICA steps were easy to remember. When asked if the ICA steps were clear and understandable 50% of the participants agreed to this and 41.7% strongly agreed, this put together out-performed other values of neutral, disagree and strongly disagree. It shows that the participants understood the ICA steps clearly.

According to the participants, it was easy to learn and apply the ICA steps, this is shown by the results recorded on the agree value of 50 % and strongly agree value of 45.8%. Majority of the participants agreed that they performed the ICA process quickly, the value of this is 70%. The participants strongly agreed by 25 % hence the two values when added show that the ICA process is quick to perform. The last aspect mentioned in learnability is about the supplementary materials provided for ICA being clear and helpful. The participants strongly agreed with a value of 79.2% that the materials were helpful to them. 20 % of the participants

agreed that the supplementary materials were helpful. This show that 99.2% of the participants reckoned that the supplementary materials which were provided were clear and helpful. Overall, one can conclude that the ICA process was easy to learn.

Table 5.3: The descriptive statistics analysis of the answers provided by participants on Ease of Use

	Ease of Use Questions									
	The overall process of ICA took a lot time to be completed		The ICA process was not demanding and required less effort		After training I took less time to perform the ICA process		It required the fewest steps possible to accomplish what I wanted to do		I did not notice any inconsistencies as I carried out the steps of ICA	
Values	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Strongly disagree	2	8.3	0	0.0	0	0.0	0	0.0	0	0.0
Disagree	9	37.5	0	0.0	0	0.0	10	41.7	0	0.0
Neutral	10	41.7	1	4.2	0	0.0	9	37.5	0	0.0
Agree	1	4.2	13	54.2	18	75.0	1	4.2	10	41.7
Strongly agree	2	8.3	10	41.7	6	25.0	4	16.7	14	58.3
Totals	24	100.0	24	100.0	24	100.0	24	100.0	24	100.0

The perceived ease of use of the ICA by the participants is shown in Table 5.3. When asked if the ICA took a lot of time to carry out, the majority of participants were neutral with 41.7%, whilst 37.5% disagreed that it took a lot of time to complete it. 8.3% of the participants also strongly disagreed. The total aforementioned values when added together outnumber the values of those who agreed of 4.2% and those who strongly agreed with 8.3% that the ICA took a lot of time to complete.

The participants had indicated that the ICA process was not demanding and did not require a lot of time by agreeing with a value of 54.2% and strongly agreeing with a value of 41.7 %. Those who were neutral were only 4.2%. There were no participants who neither disagreed nor strongly disagreed with this statement. This shows that the ICA as perceived by the participants, does not require much effort and is not demanding. The participants had shown

that after training, when carrying out the ICA process they took less time, 75 % of the participants agreed to this statement whilst 25% strongly agreed. There was no one who disagreed with this statement.

When asked if the ICA process required fewest steps possible to complete it, majority of the participants disagreed with this with 41.7%. It was followed by 37.5% of the participants who were neutral implying that they felt the steps were just moderate (neither few nor many). The participants who agreed to this statement stood at 4.2% whilst 16.7% strongly agreed to this statement. What could be drawn from this statement according to the participants is that the ICA process does not require fewest steps to be completed. Majority of the participants indicated that they did not notice any inconsistencies when carrying out the ICA steps with 58.3 % of the participants strongly agreeing and 41.7% agreeing. There were no objections to this statement. Overall, the results show that it easy to use ICA process to analyse data. The results are positive on almost all the questions which have been asked on the perceived ease of use aspect.

Table 5.4: The descriptive statistics analysis of the answers provided by participants on Perceived Usefulness

	Perceived Usefulness Questions									
	I found using the ICA process useful in my task of analysing the social media data		Using the ICA process made it easier to help analyse online social media data		Using the ICA process in analysing the social media content enabled me to accomplish tasks more quickly		ICA helped me be more productive in analysing the social media		ICA met the task of analysing the social media data	
Values	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Strongly disagree	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Disagree	0	0.0	0	0.0	1	4.2	1	4.2	0	0.0
Neutral	0	0.0	0	0.0	0	0.0	1	4.2	0	0.0
Agree	7	29.2	11	45.8	12	50.0	6	25.0	8	33.3
Strongly agree	17	70.8	13	54.2	11	45.8	16	66.7	16	66.7
Totals	24	100.0	24	100.0	24	100.0	24	100.0	24	100.0

Table 5.4 shows how participants perceive the usefulness of the ICA process in analysing social media data. The participants were asked if they found the ICA process useful in the task of analysing the social media data. In response of this, 70.8 % of them strongly agreed and 29. 2% agreed. There were no objections to this question. There were asked further if using the ICA process made it easier and helped them analyse the social media data. 54.2 % participants strongly agreed and 45.8% agreed to that question.

The participants indicated that using the ICA process in analysing social media content enabled them to accomplish tasks more quickly by agreeing with a 50% and strongly agreeing with 45.8 %, leading to 95.8 % in total. Only 4.2% disagreed. 66.7 % of the participants strongly agreed that ICA helped them to be more productive in analysing social media and 25% agreed to this. This in total out-numbered the 4.2 % of those who were

neutral and 4.2 % of those who disagreed. The final statement on the perceived usefulness was stating that the participants using ICA met their tasks of analysing the social media data. 66.7 % of the participants strongly agreed to this statement while 33.3% agreed. Overall, the participants perceived the ICA useful in analysing the social media data.

Table 5.5: The descriptive statistics analysis of the answers provided by participants on Flexibility of ICA

	Flexibility Questions									
	Performing the ICA steps was not rigid		Users performed the ICA process according to their discretion		The ICA steps and tasks are dependent on the context of the text		The ICA steps were not flexible enough to be carried out simultaneously		The ICA steps followed one after another in a chronological order making it easy and clear to carry out	
Values	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Strongly disagree	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Disagree	0	0.0	0	0.0	0	0.0	1	4.2	0	0.0
Neutral	1	4.2	0	0.0	0	0.0	19	79.2	0	0.0
Agree	3	12.5	7	29.2	8	33.3	0	0	8	33.3
Strongly agree	20	83.3	17	70.8	16	66.7	4	16.7	16	66.7
Totals	24	100.0	24	100.0	24	100.0	24	100.0	24	100.0

Users' perception on flexibility of the ICA has been shown in Table 5.5. The majority of the participants with a value of 83.3% indicated that they strongly agree that performing the ICA steps was not rigid. 12.5 % agreed to this and only 4.2 % were neutral. There were no objections to this statement. With regards to the statement which stated that the participants performed the ICA process according to their discretion, the participants strongly agreed with 70.8% and agreed with 29.2%. There were no objections to this statement.

The participants strongly agreed with 66.7 % that the ICA steps were dependent on the context of the text while 33.3% agreed to this statement. Similarly, there were no objections to this statement. On the statement which stated that the ICA steps were not flexible enough to be carried out simultaneously, majority of the participants with a 79.2% were neutral, 4.2 % disagreed whilst 16.7% strongly agreed with the statement. The majority of the participants with 66.7% strongly agreed and 33.3 % agreed that that ICA steps followed one

after another in a chronological order making it easy and clear to carry out. Still, there were no objections to this statement. Overall, one can conclude that there is room for flexibility when carrying out the ICA process in analysing the social media data.

Table 5.6 shows the overall satisfaction of the participants with the ICA process. Most of the participants indicated that they were pleased with the overall steps they carried out to complete the ICA process. They strongly agreed with a value of 58.3 % and those who agreed were 41.7%. There were no objections to this statement. When asked if they could recommend ICA to other researchers, the majority of the participants with 79.2 % strongly agreed and only 20.8% agreed. There were no objections.

The participants indicated that they were satisfied with the ease of completing the tasks in ICA process. 50% of them strongly agreed and 50% agreed. With satisfaction of the amount of time they took to complete the whole process, 75% strongly agreed that they were satisfied and 20.8% agreed that they were satisfied while 4.2% were neutral. The participants indicated that they were overall satisfied with the output of the ICA process they carried out. 75% strongly agreed that they were satisfied while 25.0 agreed

Table 5.6: The descriptive statistics analysis of the answers provided by participants on Satisfaction of ICA

	Satisfaction Questions									
	I am satisfied with the overall steps I carried out to fully complete the Inductive Content Analysis activity I participated in		I would recommend ICA to other researchers doing data analysis in their researches		Overall, I am satisfied with ease of completing the tasks in this Inductive Content analysis activity		Overall, I am satisfied with the amount of time it took to complete the tasks in this activity		Overall, I am satisfied with the output of the ICA I carried out	
Values	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Strongly disagree	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Disagree	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Neutral	0	0.0	0	0.0	0	0.0	1	4.2	0	0.0
Agree	10	41.7	5	20.8	12	50.0	5	20.8	6	25.0
Strongly agree	14	58.3	19	79.2	12	50.0	18	75.0	18	75.0
Totals	24	100.0	24	100.0	24	100.0	24	100.0	24	100.0

In conclusion of evaluation of the ICA, one could argue that when it comes to usability, ICA is easy to carry out, useful and flexible enough to be applied to analyse the social media data. It is also easy to learn. ICA is one of the qualitative approaches of analysing data in the Framework. Therefore, since ICA is one of the components of the Framework, it could be said that, by extension the proposed Conceptual Framework is easy to learn and use. It is also flexible in allowing the analysis of social media data. It is useful to those who wish to analyse social media data.

5.5 DATA MODELLING AND EVALUATION METRICS

After completing the ICA process, the second evaluation was carried out. The concluded and confirmed eight categories from the ICA were used to perform text classification through data modelling. Four data mining algorithms for classification were implemented through

WEKA: Naive Bayes multinomial (NB), Support Vector Machine (SVM), K-Nearest Neighbour (K-NN) and C4.5 Decision Tree. The eight categories were provided to a classifier of each algorithm and the models of each were created, trained and tested with the data set(s) which had been gathered online.

The reason for using these four algorithms was for demonstration purposes and for comparisons between the algorithms, to see how they will perform in terms of accurately classifying the data, given the same data set. Another justification is that, since the demonstration or application case of the proposed Conceptual Framework is of classification of data, it was best to use these classification models (NB, SVM and K-NN, C4.5). As pointed out by Hoyt et al. [128] these models are mostly used for classification because they are supervised learning and work with the target variables which are categorical (nominal), as opposed to selecting other algorithms (e.g. regression models) which are needed when the response is numeric or continuous [128],[181]. The demonstration of the application of this study's Framework is classification and also the dataset has variable instances that are categorical.

The performance of the data mining algorithms was evaluated primarily from two perspectives: for ***predictive performance*** and for ***prediction accuracy***. The metrics that constitutes predictive performance are the Root mean squared error, Mean absolute error, and Kappa, whilst the precision, recall, F1-Measures, confusion matrix, TP and FP rates evaluate the classifiers' ***prediction accuracy*** as detailed below [182], [183].

a) Kappa statistics

The Kappa statistics of the models which have been used in this study are shown in Table 5.7. Kappa is a controlled value of agreement for chance agreement [184]. Kappa metric takes the observed accuracy and compares it with an expected accuracy. Kappa is basically, a measure of how nearly the cases grouped by the machine learning classifier coordinated the data labelled as ground truth, controlling for the exactness of a random classifier as measured by the expected [185]. The kappa statistics is computed as shown in equation (5.1).

$$K = \frac{p(x)-p(i)}{p(i)} \quad (5.1)$$

Where $p(x)$ = percentage agreement and $p(i)$ = chance agreement. If $K = 1$, agreement is perfect between the classifier and ground truth. If $K = 0$, indicates there is a chance of agreement.

Table 5.7: Kappa statistics performance of the classifiers: NB; IBK; J48 and SMO attained during classification of data

Classifier	Kappa Statistics Results
NB-multinomial	0.7926
IBK	0.9354
J48	0.7359
SMO	0.9355

The kappa readings of the classifiers range between 0.9355 and 0.7359. These values are greater than 0 and less than 1, but closer to 1. This means the classified results are substantial to almost perfect agreement to the ground truth. The values are closer to 1 which is a perfect agreement. It has to be noted that, there is no standardized interpretation of the kappa statistics [185]. This argument is supported by Landis and Koch [186] who considers 0 - 0.2 as slight, 0.21 - 0.40 as fair, 0.41 - 0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect. Righat [185]'s argument is also verified by Fleiss et al. [187]'s interpretation who considers kappa greater than 0.75 as excellent, 0.40-0.75 as fair, and less than 0.40 as poor. As hinted by Righat [185]'s careful observation shows that both these aforementioned scales are somewhat arbitrary.

b) Error rates

In any experimentation, for the execution to be exact, one needs to put into thought the mistakes or errors, which influence the whole procedure and measure their error rate. In our case, error rates have been calculated using Root Squared Mean Error and Mean Absolute Error the measure. Refer to Table 5.8 for the results.

i. Mean Absolute Error

The Mean Absolute Error (MAE) is a measure that indicates the proximity of predictions to the ultimate outcomes. MAE shows the magnitude of error(s) to expect from the forecast on average. MAE is given by equation (5.2)

$$MAE = \frac{1}{k} \sum_{j=1}^k |n_j - x_j| = \frac{1}{k} \sum_{j=1}^k |c_i| \quad (5.2)$$

Basically, Mean Absolute Error is computed through an average of the absolute errors shown in equation (5.3)

$$C_i |n_j - x_j| \quad (5.3)$$

Where n_j = prediction and x_j = true value

ii. Root Mean Squared Error

Root Mean Squared Error (RMSE) is the square root of the mean of the squares of the values. Before errors could be averaged, they are squared first. [184]. The RMSE C_j of an individual program j is evaluated by the equation (5.4):

$$C_j = \sqrt{\frac{1}{k} \sum_{i=1}^k \left(\frac{P_{(ji)} - m_i}{m_i} \right)^2} \quad (5.4)$$

Where $P_{(ji)}$ = the value predicted by the individual program, j = fitness case and m_i = the target value for fitness case i .

Table 5.8: Error rates of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	Mean Absolute Error	Root Mean Squared Error
NB-multinomial	0.0434	0.1874
IBK	0.0221	0.0843
J48	0.0737	0.1919
SMO	0.1881	0.2922

The results on Table 5.8 show both the MAE and RMSE measures of the classifiers. For MAE, value ranges between 0.02 and 0.18. SMO has the highest MAE whilst IBK has the lowest MAE as compared to all the classifiers used. The Table further shows the RMSE of the classifiers. IBK has the least value of 0.08, and SMO has the highest value of 0.29.

A comparison could be made between the RMSE and MAE to establish if the forecast contains huge but infrequent errors. If the difference between RMSE and MAE is huge, it implies more inconsistent error size [188]. When comparing our results of MAE and RMSE

for each classifier, the difference is small, this implies less inconsistent error size. As pointed out by Nau [189], Chai and Draxler [190] there is no complete criteria for good or bad values of RMSE or MAE. The RSME and MAE can be applied at the same time to analyse the deviation in the errors in a set of forecasts. MAE will always be smaller or equal to RMSE. The larger the variance between them, the greater the difference in the individual errors in the sample. All the errors are of a similar extent if the $MAE=RSME$ [191].

c) Receiver Operating Characteristic

Receiver Operating Characteristic (ROC) curves are the best way to compare diagnostic tests. Thus, the area under the ROC curve usually has a range of 0.5 to 1. The interpretation of this is: 0.5 means a worthless test whilst 1 is a perfect test. So the area under the ROC curve should be more than 0.5 and less than or equal to 1 for it to be the perfect test [192]. The true positive rate (Sensitivity) and the positive rate (100-Specificity) are plotted against each other for different cut-off points in a ROC curve. On the ROC curve, each point denotes a sensitivity/specificity pair matching to a specific decision threshold. A test with a complete discrimination (no overlap in the two distributions) has a ROC curve that comes about through the upper left corner (100% sensitivity, 100% specificity) [193].

Table 5.9: Average ROC calculations of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	Average ROC
NB-multinomial	0.975
IBK	0.999
J48	0.967
SMO	0.992

As shown in Table 5.9, IBK had the largest value of 0.999 and J48 had the least value of 0.967. The readings ranges from 0.96- 0.99, this means that the classifiers' ROC readings are more than 0.5 and closer to 1 which is a perfect test. This falls within the limits that a perfect test should be greater than 0.5 and closer to 1 [192].

d) Recall (True Positive rate or Sensitivity)

Recall is the measure which shows a classification model's ability of selecting instances of a particular class from the dataset. It is the extent of actual positive which are predicted positive [79]. The True Positive (TP) rate is the extent of instances that were classified correctly as k in the relevant class k , TP is the same as Recall. To find it in a confusion matrix, the diagonal elements are divided by the sum over the relevant row (See Figure 5.5). Given the confusion matrix C and a set of labels $\{L_j\}$, the standard definition of recall is calculated as shown in equation (5.5).

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{c_{ii}}{\sum_{j=1}^{f+1} c_{ij}} = c_{ii}/ATotal_i \quad (5.5)$$

Where: f is the size of the sample, C_{ii} is the predicted class for a multi class case and $ATotal_i$ is the actual total for the class. Please refer to Table 5.10 for Recall and Table 5.11 for the TP rate values of the classifiers which have been used in this study.

Table 5.10: Recall values of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	Recall
NB-multinomial	0.828
IBK	0.946
J48	0.780
SMO	0.946

SMO and IBK classifiers had the highest recall values of 0.946 and J48 had the least value of 0.780. All the classifiers scored a value greater than 0.5. The readings imply that a reasonable proportion of instances were classified in the categories which they truly belong to.

e) The False Positive Rate

The False Positive (FP) rate is the extent of instances that had been incorrectly classified as class k , whilst belonging to another different class not k . To find it in a confusion matrix (See Figure 5.5), the diagonal elements are subtracted from the sum of the column of a class x , then divided by the sum of the rows of other classes.

Table 5.11: TP rate and FP rate of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	TP Rate	FP Rate
NB-multinomial	0.828	0.034
IBK	0.946	0.009
J48	0.780	0.040
SMO	0.946	0.008

The TP rate of all the classifiers are greater than 0.5. They range from 0.7 to 0.95. This means that proportion of the instances which were classified according to the classes among all the examples which truly have the said classes is good. It means most instances were correctly classified to their relevant classes. For the FP rate, the readings are small, it means that the proportion in which the instances were wrongly classified to different classes other than the actual classes is less. The results show that there were only few cases of incorrectly classified instances. It can be argued that the classifiers are more accurate.

f) Precision (Confidence)

The Precision is the extent of the instances that actually have class k amongst all those that were categorized as class k (predicted positives which are actually positive). To calculate it in a matrix, the diagonal elements are divided by the sum over the relevant column. It checks the accuracy of a predicted class [79]. Precision is defined by the formula (5.6).

$$p_i = \frac{TP}{TP+FP} = \frac{c_{ii}}{\sum_{k=1}^n c_{ki}} = \frac{c_{ii}}{PTotal_i} \quad (5.6)$$

Where : C_{ii} is the predicted class for a multi class problem and $PTotal_i$ is the predicted total for the class

Table 5.12: Precision values of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	Precision
NB-multinomial	0.809
IBK	0.947
J48	0.777
SMO	0.953

The classifiers precision values range from the highest scored by SMO with a value is 0.953, and the least being J48 with 0.777. All the classifiers recorded a value greater than 0.5. With a reading greater than 0.5, it simply means that the predicted instances are positive [79]. Hence our readings are all greater than 0.5 which implies there are positive and have met the threshold of 0.5.

g) F1-Measure (F1- Scores)

Precision and recall scores are usually discussed together and a single measure can be derived by combining both measures into a new measure called the F1-measure [79]. The F1- Scores is the adjusted consonant mean of recall and precision. It is defined as shown in equation (5.7).

$$F = 2 \left(\frac{P \cdot R}{P + R} \right) \quad (5.7)$$

Where P is the precision and R is the Recall. A higher F1-measure indicates a better classifier.

Table 5.13: F1-Measures performance of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	F1-measures
NB-multinomial	0.815
IBK	0.944
J48	0.768
SMO	0.946

The SMO had the highest F1- measure value score of 0.946 and J48 scored 0.768 which was the least score. These results are greater than 0 and closer to 1. These are good because when F1-score at 1, it is said to have reached its best value and when 0 it is at its worst value.

h) Classification Accuracy

Classification accuracy when the instances are correctly categorized in their respective classes in a model. It is computed by the sum of correctly classified instances divided by the total number of samples [79]. The formula for accuracy calculations is shown in (5.8).

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + TN + FP + FN} \right) \quad (5.8)$$

Where FP is incorrect prediction of positive instances (False positives), FN is incorrect prediction of negative instances (False negatives), TP is correct prediction of positive instances (True positives) and TN is correct prediction of negative instances (True negatives).

Table 5.14: Classification accuracy of the classifier: NB, IBK, J48 and SMO attained during classification of data.

Classifier	Correctly classified	Incorrectly Classified
NB	154 (82.8%)	32 (17.2%)
IBK	176 (94.6%)	10 (5.4%)
J48	145 (77.95%)	41 (22.0%)
SMO	176 (94.6%)	10 (5.4%)

Table 5.14 displays the accuracy of the models or classifiers. It shows how the models have correctly and incorrectly classified the instances. The results generally show that all the classifiers are optimal because the incorrectly classified instances are less in number as compared to the correctly classified ones. The classification accuracy is normally well explained with the use of the confusion matrix.

i) Confusion Matrix

A confusion matrix is more generally named a contingency table. A confusion matrix has k dimensions where k is the number of the classes. Figure 5.4 shows a confusion matrix for a data mining model for two possible outcomes **a** (positives) and **b** (negatives). The confusion matrix shown in Figure 5.4 shows how model predictions are made. The rows resemble the actual classes in a dataset and the columns represent the predicted classes. The diagonal elements indicate the number of correctly classified instance for each class. The size of matrix can be large. To find the total number of correctly classified instances in a matrix, calculate the sum of the numbers in diagonal (from top left to the bottom right) and all others are incorrectly classified. The confusion matrix displays the true positives, the false positives, the true negatives and the false negatives.

Predicted			Actual
a	b		
TP	FN	a	
FN	TP	b	

Figure 5.4: A general look of a confusion matrix

In our study 8 classes were used, and therefore an 8x8 confusion matrix for each classifier as shown from Figure 5.5-5.8 respectively.

a	B	c	d	E	f	G	H	← Classified as
21	0	0	0	2	1	0	0	a = corruptive-governance
0	32	1	1	3	1	0	0	b = In-competent- governance
0	0	33	4	0	0	0	0	c = Judiciary-agitation
0	1	2	38	0	1	0	0	d = Oppressive-governance
2	1	0	0	14	0	0	0	e = Partisanship & cronyism
1	1	0	0	4	8	0	0	f = Economic & political scandals
0	0	0	0	0	0	8	0	g = Poor education system
0	1	0	5	0	0	0	0	h = Un-consultative & un-transparent governance

Figure 5.5: Confusion matrix of Naïve Bayes classifier performance attained during classification of data

The matrix shows how the NB classifier has classified the instances. The totals of correctly classified instances are in a diagonal line of the matrix from the top left corner to the bottom right corner. The total number of all correctly classified instances as per the correct classes is 154 in the diagonal line. The remaining numbers in the matrix show the instance that had been incorrectly classified and their total number is of 32. These numbers tally with the results shown in Table 5.14. The numbers in the diagonal line from the top left corner to the bottom right corner actually shows the true positives of each class (category) and any other numbers which are not in the diagonal are the false negatives.

a	B	c	d	E	f	g	h	← Classified as
23	0	0	0	1	0	0	0	a = corruptive-governance
0	36	0	0	1	1	0	0	b = In-competent- governance
0	0	37	0	0	0	0	0	c = Judiciary-agitation
0	0	1	41	0	0	0	0	d = Oppressive-governance
0	0	0	0	17	0	0	0	e = Partisanship & cronyism
0	1	0	0	4	9	0	0	f = Economic & political scandals
0	0	0	0	0	0	8	0	g = Poor education system
0	0	1	5	0	0	0	5	h = Un-consultative & un-transparent governance

Figure 5.6: Confusion matrix of SMO classifier performance attained during classification of data

The matrix shows how the SMO classifier has classified the instances. The totals of correctly classified instances are in a diagonal line of the matrix from the top left corner to the bottom right corner. The total number of all the correctly classified instances as per the correct classes in the diagonal line is 176. The remaining numbers in the matrix show instance that had been incorrectly classified and their total number is 10. These numbers tally with the results shown in Table 5.14. The numbers in the diagonal line from the upper left corner to the bottom right actually shows the true positives of each class (category) and any other numbers which are not in the diagonal are the false negatives.

a	B	c	d	E	f	g	h	← Classified as
24	0	0	0	0	0	0	0	a = corruptive-governance
0	37	0	0	0	1	0	0	b = In-competent- governance
0	0	37	0	0	0	0	0	c = Judiciary-agitation
0	0	1	41	0	0	0	0	d = Oppressive-governance
1	1	0	0	15	0	0	0	e = Partisanship & cronyism
1	1	0	0	3	9	0	0	f = Economic & political scandals
0	0	0	0	0	0	8	0	g = Poor education system
0	0	1	0	0	0	0	5	h = Un-consultative & un-transparent governance

Figure 5.7: Confusion matrix of IBK classifier performance attained during classification of data

The matrix shows how the IBK classifier has correctly classified the instances. The totals of correctly classified instances are in a diagonal line of the matrix from the top left corner to the bottom right corner. The overall total number of all the correctly classified instances as per the correct classes in the diagonal line is 176. The remaining numbers in the matrix show instance that had been incorrectly classified and their total number is 10. These numbers tally with the results shown in Table 5.14. The numbers in the diagonal line from the upper left corner to the bottom right actually shows the true positives of each class (category) and any other numbers which are not in the diagonal are the false negatives.

a	B	c	d	E	f	g	h	← Classified as
22	0	0	1	1	0	0	0	a = corruptive-governance
0	30	2	2	2	1	1	0	b = In-competent- governance
3	1	28	4	0	0	0	1	c = Judiciary-agitation
0	0	4	36	0	0	0	2	d = Oppressive-governance
1	0	1	0	15	0	0	0	e = Partisanship & cronyism
3	2	0	1	4	3	1	0	f = Economic & political scandals
0	1	0	0	0	0	7	0	g = Poor education system
0	0	1	0	0	1	0	4	h = Un-consultative & un-transparent governance

Figure 5.8: Confusion matrix of J48 classifier performance attained during classification of data

The matrix shows how the J48 classifier has classified the instances. The totals of correctly classified instances are in a diagonal line of the matrix from the top left corner to the bottom right corner. The total number of all the correctly classified instances as per the correct classes in the diagonal line is 145. The remaining numbers in the matrix show instance that had been incorrectly classified and their total number is 41. These numbers tally with the results shown in Table 5.14. The numbers in the diagonal line from the upper left corner to the bottom right actually shows the true positives of each class (category) and any other numbers which are not in the diagonal are the false negatives.

In conclusion of the evaluation of data models through metrics measuring accuracy and performance, the results show that all the classifiers met the threshold in terms of correctly classifying the instances to the correct class labels or categories scoring the highest percentatges. This implies that the ICA output was of high quality because it was used to further classify the data through the models. The pre-processing of data together with qualitative analysis of ICA which involved human influenced these results. This proved that indeed when human intervension and interpretation is applied in data analysis, it helps refine data well for easy analysis through computational methods as the experiments on this study has shown. The models's performance and accuracy met the standard threshold(s). The ICA process indeed helped with fine tuning the data and the semantics which made it easier for computational analysis This is what has been echoed by other researchers such as Chung and Gray [194] that, the manner in which a model's performance is measured is very crucial and

requires special consideration. One model could perform otherwise in different fields because of the quality of the data, prescriptive measures, and the influence from decision makers. This corroborates with what Zafarani et al. [67] states that in social media data analysis, after having a depiction, measures have to be taken care of in the data with regards to quality issues. The text- preprocessing steps need to be performed data could be analysed further. Quality measures comprise removing noise from data, detecting outliers, handling of missing values and removing duplicates [67].

In conclusion of the two evaluations in the demonstration of the application of the proposed Conceptual Framework through Classification, the results, from the user study, show that the ICA process is flexible and systematic in terms of allowing the users to analyse social media data. The users' perception with regards to ease of use and usefulness of the ICA on analysing social media data is positive. The results from the experimental study show that data mining algorithms performed above the standard threshold of the metrics which evaluated the performance and accuracy in classifying data when supplied with data from the ICA process. It should be recalled that ICA is one of the qualitative approaches in which researchers immerse themselves in the data, in order to get the in-depth meanings and understanding of people's perceptions in a certain phenomenon, through which they are able to put bits and puzzles of information together to form themes, and identify similar themes and collapse them further into categories.

Based on the results of the optimum performance of the data models in classifying data correctly and meeting the standard threshold, it could be concluded that the limitations of failure to capture in-depth meanings from data were mitigated through the application of ICA hence the attainment of the main objective of this study.

5.6 THREATS TO VALIDITY

For the ICA process, the guidelines which were used were informed by various qualitative research ideas [110], [116], [157], [164]. This was done based on the support of the argument that, in order to produce valid and reliable inferences there has to be established methodical and clear measures for handling data [102]. Also, as highlighted by Struwig and Stead [195], it is important during text analysis, to have two or more groups of raters' to categorise the data independently, and then compare and agree on the categories to use. This measure was applied during the ICA, by engaging four focus groups. This ensured the reliability and

validity of the categories selected. Again, this involved an element of triangulation, because the focus groups were formed by different participants who independently analysed and made interpretation the data, and determined if it was in agreement with the coding [195]. This is supported by Schilling [117] that, any arising problems and doubts regarding: defining of categories, rules for coding, or categorization require to be considered and be resolved.

The validity and reliability of the questionnaires was ensured by re-checking the questions and passing them to some people to check if there were easy to understand and also if they have any mistakes. All mistakes were corrected and they were made easy to understand before they could be rolled out to the participants. Questions were asked in a manner which eliminated biasness and leading questions to avoid influencing the participants on how and what to answer. The questionnaire used the Likert scale questions. Overall this study triangulated both the qualitative (use of the traditional approach) and quantitative (use of data mining algorithms evaluation metrics in Section 5.5) methods to evaluate the Conceptual Framework.

5.7 CHAPTER SUMMARY

This chapter discussed an evaluation of the Conceptual Framework. The evaluation involved gathering data online, about the political landscape of Botswana from Facebook and Twitter. Data was gathered using the computational automated methods (Netlytic tool and a Java program). This is stage one of the Conceptual Framework. Data was then pre-processed and made ready for analysis and this is stage two, of the Conceptual Framework. The process of Inductive Content Analysis (ICA) was carried out to deduce themes and categories from the data (data was manually classified into related themes and collapsed into categories). This was the qualitative analysis of data that involved the human intervention and interpretation. The Inductive Content Analysis was evaluated through a questionnaire guided by “Usability” component of the ISO 9126 model. The results of the evaluation of ICA showed that the ICA process was flexible and systematic in terms of allowing the users to analyse social media data, the users’ perception in terms of ease of use and usefulness of the ICA on analysing social media data was positive. After completion of the ICA process, the categories were supplied to the classifiers of the models to further classify the data. The results from the experimental study show that data mining algorithms produced accurate results in classifying the data and the performance was above the threshold data when supplied with data from the ICA process.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 INTRODUCTION

This chapter discusses the overall view of the study, conclusions of the study, limitations and recommendations for future works. The challenges which were encountered during the execution of the entire study are also stated.

6.2 OVERALL VIEW OF THE STUDY AS PER RESEARCH OBJECTIVES

This study gave a brief introduction to social media, the characteristics of social media data and gaps that exist in social media data analysis. Social media data possess the characteristics of Big Data such as volume, veracity, velocity, variability and value. These characteristics make its analysis more challenging than conventional data. Computational analysis methods usually cannot capture in-depth meanings (semantics) within data. As a solution, this study proposed a Conceptual Framework for mining and analysis of social media data by offering a generic approach to addressing the limitations of computational methods to social media data analysis. The overall view of this study is summed and concluded per objectives.

Objective 1

The first objective was to identify existing concepts (i.e., theories and methods) in data mining and analysis, which are capable of capturing in-depth meanings (semantics). This was achieved by gathering information from various literature sources on the subject of data analysis. Information gathered here, addressed concepts of data analysis (qualitative and quantitative as well as approaches, their strengths and limitations) discussed in Section 3.2 of Chapter 3.

Objective 2

The second objective was to identify possible processes of incorporating concepts identified in (1) with computational methods to address their limitations, this was achieved by gathering information from the literature review on related works which have carried out content

analysis. Information on existing frameworks was also reviewed to find possible processes of incorporating the concepts identified in (1). This was done in Chapter 2 and 3.

Objective 3

The third objective was to construct a Conceptual Framework from the identified concepts and processes. A Conceptual Framework was constructed in Chapter 4 drawing its origin from the processes and components addressed in (1) and (2). The Conceptual Framework was designed to have three stages or phases with various components or processes. The first phase was Data collection. The proposed Conceptual Framework shows that social media data can be gathered through the use of digital methods. It further shows that traditional methods could be incorporated in situations where more data is needed with more insights. A researcher can have the option of visiting the “subjects case” and getting more clarification through the use of the traditional methods (e.g., in-depth, interviews focus groups), in case a researcher is carrying out qualitative research.

The second phase of the Framework was Text-pre-processing. This stage has the manual pre-text processing component of data cleaning which include noisy data removal, spelling checking, abbreviations and acronyms corrections and duplication deletions. Secondly, data was prepared and made suitable for analysis. This could be achieved through text pre-processing (just after data has been gathered) and post processing (performed before data could be automatically classified through data models). Post processing involves steps such as: word parsing, tokenization, stemming and lemmatization while pre-processing involves cleaning the noise from the data, formatting it and correcting mistakes and mis-spelled words.

The third stage is of data analysis which has the qualitative and quantitative analysis processes. Examples of qualitative and quantitative approaches have been stated in chapter 3 on Tables 3.1-3.5. Data modelling which is part of the Experimental approach in Quantitative analysis is also shown as the component of the Framework. The data modelling could be descriptive or predictive. The data modelling could have the supervised and unsupervised algorithms for classification, regression, etc.

The other components of the Conceptual Framework which were created were of data storage and knowledge repository. These are the components where data is stored after collection and even after analysis. This is evidenced by section 4.3 and 4.4 of Chapter 4. The proposed Conceptual Framework was designed as shown in Figures 4.1 and 4.2.

Objective 4

The fourth objective was to evaluate the conceptual framework through a political landscape case of Botswana. In demonstration and application of the Framework, data gathered from Facebook and Twitter was cleaned and made easy to understand through text-pre-processing and manual classification. Classification was done through Inductive Content Analysis. Data was manually classified through identification of similar patterns and themes. The themes were further collapsed into categories (See Appendix A-C). Inductive Content analysis was evaluated with a 30 Usability based questionnaire which was guided by the “Usability” component of the ISO 9126 model (See Appendix D). The aspects which were evaluated on Usability are: Learnability, Ease of use, Perceived usefulness, Satisfaction and Flexibility. Each evaluation metric had five questions. The total for the positive responses of strongly agree and agree for the five questions of each metrics are as follows: Learnability scored between 91.7 % and 100%. Ease of use scored between 95.8% and 100% with the exception of one question which asked if the overall process took a lot of time and majority mostly disagreed with 87.5 %. Perceived ease of Use scored between 91.7% and 100%, Flexibility scored between 95.8% and 100% and Satisfaction scored between 95.8% and 100%.

The finalised eight categories were used further to evaluate the performance of the computational methods (i.e., Naive Bayes, K-Nearest Neighbour, Decision Tree- C4.5 and Support Vector Machine,) in classifying the data. The metrics which measured the performance and accuracy of the models were applied. The scores for accuracy of the model performance in classifying the data correctly were: IBK and SMO classifiers with 94.6%, NB classifier with 82.8% and J48 classifier with 78% (As shown in Figure 5.3 in Chapter 5). Most of the evidence of this objective was presented in Chapter 5 and Appendix E.

General Objective

The main objective of this study was to develop a Conceptual Framework for mining and analysis of social media data by offering a generic approach to addressing the limitations of computational methods to social media data analysis. This main objective was carried out through performing all the objectives of the study which have been mentioned above. A conceptual Framework was proposed and formulated as per objective 3 (Chapter 4). This Conceptual Framework was evaluated as per objective four in Chapter 5. In conclusion, the main objective has been achieved of developing and evaluating a Conceptual Framework.

6.3 CONCLUSION OF THE STUDY

Based on the results of the metric evaluation of the models, performing content analysis through human intervention and interpretation helps to overcome the limitation of the computational methods of understanding the in-depth meanings and semantics on data. Inductive Content Analysis which had been applied in the application case of the Framework helped with preparing data for classification experimentation carried throughout data modelling. The evaluation of the ICA as well as the data models yielded positive outcomes. Hence, since these are some components of the Conceptual Framework, it could be extended to be the results of the Conceptual Framework.

As a measure of caution, it should be noted that the accuracy and the performance of the Conceptual Framework depends on the type of data and the quality measures which include proper pre-processing steps such as noise removal, outliers detections, missing value handling and duplicates removal. If these quality measures are not adhered to, the accuracy and performance could be affected negatively. If the pre-processing steps are carried out well, the Framework will perform well and yield better results which are accurate as revealed in our data modelling classification experiments. It should also be noted that this study tested or evaluated the proposed Conceptual Framework with classification only but it can also be applied on other data mining tasks (e.g regression, clustering, etc). Therefore the application of this Framework will depend on the tasks a researcher will be addressing at that particular moment. This is the flexibility aspect of this Conceptual Framework. It is not rigid and does not only provide one option or way of analysing data. It allows one to analyse the social media data in a systematic manner.

6.4 LIMITATIONS AND FUTURE WORKS

In conducting Inductive Content Analysis (qualitative content analysis), this study went only up to categories development. The content analysis did not go up to the last stage of theory development because that was outside the scope of this study. This was because this study required only the categories, so classification of data was made in order to deduce categories so that they got supplied to the algorithms in models implementation. Mostly qualitative content analysis also end up with some reports where meanings and interpretation coming out of data are discussed, supporting this with some direct quotations from the data. It has to be noted that this also was outside the study's scope. As stated earlier on, the main aim of conducting Inductive Content Analysis was to manually classify the data through identifying

the patterns and themes and collapsing them into categories which were used to further classify data automatically through the implemented models through data mining algorithms.

Another limitation which is worth noting is that, instead of evaluating the ICA only with questionnaires, it could have been triangulated with an interviews of the participants so as to get concrete results with certainty and assurance that the participants were being honest and not hiding anything. In addition, they could have also been observed unobtrusively to see how they were carrying out the ICA, because sometimes the participants alter their behaviour in the presence of the researcher to look good in their eyes.

Recommendations for future works on the proposed Conceptual Framework could be adding another phase of data visualizations. In the future researchers who would like to understand about the political landscape of Botswana can complete the Inductive Content Analysis which was carried out in this study. They could also build some political landscape theories after completion.

In the future, the Framework could be demonstrated through other applications such as clustering, regression, etc.

6.5 CHALLENGES

One major challenge experienced was collecting data from individual user profiles from Facebook. Facebook does not give permission to collect information from individual user profiles. It only gives permission to collect information from public open groups and pages. It also does not allow information gathering from private closed groups, regardless of membership status. This posed a great challenge because more data is posted individually on personal accounts and the study needed more data.

REFERENCES

- [1] B. Batrinca and P. C. Treleaven, "Social media analytics: A survey of techniques, tools and platforms," *AI Soc.*, vol. 30, no. 1, pp. 89–116, 2015.
- [2] P. Gundecha and L. Huan, "Mining Social Media: A Brief Introduction," 2012. [Online]. Available: <http://dx.doi.org/10.1287/edu.1120.0105>. [Accessed: 17-Jan-2016].
- [3] P. Patel and K. Mistry, "A Review: Text Classification on Social Media Data," *IOST-JCE*, vol. 17, no. 1, pp. 80–84, 2015.
- [4] Facebook Newsroom (2016). [Online]: <http://newsroom.fb.com/content/> Accessed: February 20th 2016.
- [5] S. Toivo, Social Media-The New Power of Political Influence, Version 1.0, CES, 2014.
- [6] G. Barbier, Z. Feng, P. Gundecha and H. Liu, Provenance data In Social Media, *Synthesis Lectures On data Mining and Knowledge Discovery*, 2013.
- [7] S. Esk, The Social Media Vocabulary, Sanastokeskus TSK ry: ESC 40, Helsinki, 2010.[Online]: Available: http://www.tsk.fi/tiedot/pdf/Sosiaalisen_median_sanasto, Accessed: March 9th 2016.
- [8] X. Chen, M. Vorvoreanu and K. Madhavan, Mining Social Media Data for and Understanding Students's Learning, *IEEE TOLT*, vol. 7, no. 3, July-September, 2014.
- [9] G. Nandi and A. Das, A Survey on Using Data Mining Techniques for Online Social Network Analysis, *IJCSI*, vol.10, no.2, pp. 1694-0814, 2013.
- [10] G. E. Medicin, 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness, *Journal of Medical Internet Research*, vol. 10, no. 3, pp. 22, 2008.
- [11] D. Jensen, Data Mining in Social media, Computer Science Department Faculty Publication Series Paper 67,[Online].http://scholarworks.umass.edu/cs_faculty_publications/67, 2002: Accessed: September 16th 2016.
- [12] P. Gundecha and H. Liu, Mining Social Media: A Brief Introduction. *Tutorials in Operations Research*, vol.1, no.4, pp. 1-17, 2012.
- [13] SAS (2015) Data Mining From A to Z: How to Discover Insights and Drive Better Opportunities [Online]: <https://www.sas.com>. Accessed: March 10 2016.
- [14] D. Sotiropoulos. Data Mining on Social Networks, [Online]. Available: docplayer.net. Accessed: February 10th 2016.
- [15] A. Bhadani, and D. Jothimani, D. Big data: Challenges, opportunities and realities, In Singh, M.K., & Kumar, D.G. (Eds.), *Effective Big Data Management and Opportunities for Implementation*, Pennsylvania, USA: IGI Global, pp. 1-24, 2016.
- [16] A. Waldherr, G. Heyer, P. Jahnichen, A. Niekler and G. Wiedemann, Mining Big Data with Computational methods, 2016. In G. Vowe and P. Henn (Eds.), *Political Communication in the Online World: Theoretical Approaches and Research Designs*, pp. 201-217, New York, NY, Routledge.
- [17] R. Tinati, S. Halford, L. Carr, and C. Pope, Big Data: Methodological challenges and Approaches for Sociological Analysis, *Sociology*, vol. 48, no. 4, pp. 663-681, 2014.
- [18] M. Mahrt and M. Scharkow, The value of Big Data in digital media research, *Journal of Broadcasting & Electronic Media*, vol. 57, no. 1, pp. 20-33, 2013.
- [19] D. Boyd and K. Crawford, Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, Communication & Society*, vol. 15, no. 5, pp. 662-679, 2012.
- [20] B. A. Hamilton, The Field Guide to Data Science, 2013, [Online].

- www.boozallen.com/insights/2013/data-science-filed-guide: Accessed June 12th 2016.
- [21] J. P. Bigus, Social Media mining, [Online]. Available: <https://www.slideshare.net>, Accessed: 27th February 2016.
 - [22] D. Laney, 3D Data Management: Controlling data volume, velocity, and variety, 2001. [Online]. Available: <http://blogs.gartner.com>, Accessed: March 2017.
 - [23] U. Sivarajah, M. M. Kamal, Z. Irani, V. Weerakkody, Critical analysis of Big Data challenges and analytical methods, *Journal of Business Research*, vol. 70, issue. 2017, pp. 263–286.
 - [24] B. Marr, Big Data: The 5 Vs Everyone Must Know, Published on March 6, 2014, [Online]. Available: <https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>, Accessed December 04th 2017.
 - [25] P. Saporito, 2 More Big Data V's-Value And Veracity. Published, 23-January-2014 Available at: <http://www.digitalistmag.com/technologies/big-data/2014/01/23/2-more-big-data-vs-value-and-veracity-01242817>, Accessed: December 04th 2017
 - [26] R. Akerkar, Big Data Computing. Florida, USA: CRC Press, Taylor & Francis Group, 2014.
 - [27] R. V. Zicari, Big Data: Challenges and Opportunities, In R. (Ed.), *Big Data Computing*, Florida, USA: CRC Press, Taylor & Francis Group, pp. 103-128, 2014.
 - [28] M. A. Vasarhelyi, A. Kogan, B. M. Tuttle, Big data in accounting: An overview. *Accounting Horizons*, vol.29, no.2, pp.381–396, 2015.
 - [29] U. Sivarajah, Z. Irani, Z. V. Weerakkody, Evaluating the use and impact of Web 2.0 technologies in local government, *Government Information Quarterly*, vol. 32, no. 4, pp. 473-487, 2015
 - [30] A. Gandomi and M. Haider, Beyond the hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, 2015.
 - [31] J. Srivastava, Data Mining for Social Network Analysis, *IEEE ISI 2008 Invited Talk (III)*, pp.1-125, 2008.
 - [32] D. Gaffney, #Iran Election: Quantifying Online Activism, *Proc. Extending the Frontier of Society On-Line (WebSci10)*, 2010.
 - [33] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I. M. Welp. Predicting Elections with Twitter: What 140 characters Reveal about political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.
 - [34] S. Stieglitz and L. Dang-Xuan, 'Emotions and Information Diffusion', In *Social Media-Sentiment of Microblogs and Sharing Behavior*, *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217-248, 2013.
 - [35] S. J. Jamison-Powell, C. Lineman, L. Daley, A. Garbett, S. Lawson. I can't get no sleep: Discussing #insomnia on Twitter, Session: Understanding Online Communication, CHI 2012, May 5–10, 2012, Austin, Texas, USA.
 - [36] S. F. Wambaa and L. Carter, Social Media and Networking: Concepts, Tools, and Applications, Chapter 35: Twitter adoption and use by SMEs: An empirical study, *46th Hawaii International Conference on System Sciences*, pp. 791-806, IGI Global, 2013.
 - [37] E. L. Frederick, C. H. Lim, G. Clavio and P. Walsh, Why We Follow: An Examination of Para-social Interaction and Fan Motivations for Following Athlete Archetypes on Twitter. *International Journal of Sport Communication*, vol. 5, no. 18, pp. 481-502, 2012.
 - [38] L. Wallace, J. Wilson and K. Miloch. Sporting Facebook: A Content Analysis of NCAA Organizational Sport Pages and Big 12 Conference Athletic Department Pages, *International Journal of Sport Communication*, 2011, 4, 422-444.

- [39] M. E. Hambrick and T. Q. Mahoney. It's incredible – trust me: exploring the role of celebrity athletes as marketers in online Social Networks. *Int. J. Sport Management and Marketing*, Vol. 10, Nos. 3/4, 2011.
- [40] G. Clavio, Social Media and the College Football Audience, *Journal of Issues in Intercollegiate Athletics*, 2011, vol. 4, no.15, pp. 309-325.
- [41] F. Cheong and C. Cheong. Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods (2011), *PACIS 2011 Proceedings*. Paper 46, [Online]. Available; <http://aisel.aisnet.org/pacis2011/46>, Accessed: January 30th 2016.
- [42] R. Bandari, S. Asur and B. Huberman. The Pulse of news in Social media: Forecasting Popularity, International AAAI Conference on Web and Social Media, North-America, [Online]. Available: <https://www.aaai.org/ocsarxiv.org>, Accessed: March 10 2017.
- [43] T. Sakaki, M. Okazaki and Y. Matsho. Earthquake shakes Twitter users: Real- Time Event detection by social sensors, April 26-30, Raleigh, NC, USA, 2010.
- [44] L. Hong, O. Dan and B. D Davison, Predicting Popular Messages in Twitter, WWW 2011-Poster, March 28-April1, 2011, Hyderabad, India, 2011.
- [45] H. Becker, M. Naaman, and L. Gravano, "Selecting Quality Twitter Content for Events," *Proc. Fifth Int'l AAAI Conf. Weblogs and Social Media (ICWSM '11)*, 2011.
- [46] W. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," *Proc. 33rd European Conf. Advances in Information Retrieval*, pp. 338- 349, 2011.
- [47] L. Hong and B.D. Davison, "Empirical Study of Topic Modelling in Twitter," *Proc. First Workshop Social Media Analytics*, pp. 80-88, 2010.
- [48] J. Yang and S. Counts. Predicting the speed, Scale and Range of information Diffusion in Twitter, Association of Advancement of Artificial intelligence, *Proceedings of the fourth International Conference On Weblogs And Social Media*, May 2010, vol. 1, issue. 25, pp. 355-358, 2010
- [49] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter," *Proc. 20th Int'l Conf. World Wide Web*, pp. 695-704, 2011.
- [50] M. Gomez Rodriguez, J. Leskovec, and B. Scheolkopf, "Structure and Dynamics of Information Pathways in Online Media," *Proc. Sixth ACM Int'l Conf. Web Search and Data Mining*, pp. 23-32, 2013.
- [51] D. Ediger, C. Corley, R. Farber and W.N. Reynolds, Massive Social network Analysis: Mining Twitter for Social Good, *39th International Conference on Parallel processing*, 2010.
- [52] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," *Proc. 23rd Int'l Conf. Computational Linguistics: Posters*, pp. 24249, 2010.
- [53] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification Using Distant Supervision," CS224N Project Report, Stanford pp. 1-12, 2009.
- [54] K. Nishida, R. Banno, K. Fujimura, and T. Hoshide, "Tweet Classification by Data Compression," *Proc. Int'l Workshop Detecting and Exploiting Cultural Diversity on the Social Web*, pp. 29-34, 2011.
- [55] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, Short Text Classification in Twitter to Improve Information Filtering," *Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 841-842, 2010.
- [56] M. Conway, The subjective precision of Computers. A methodological Comparison with human coding in content analysis, *Journalism and Mass Communication Quarterly*, pp. 186-200, 2006, vol. 83, issue 1, DOI 10.1177.

- [57] A. F. Simon. A unified method for analyzing media Framing: Communication in US elections: New agendas, 2001. [Online]. Available: books.google.com: Accessed March 31st 2017.
- [58] S. Stieglitz, M. Mirbabaiea, B. Rossa, C. Neubergerb, Social media analytics- Challenges in topic discovery, data collection, and data preparation, *International Journal of Information Management*, vol. 39, no.1, pp.156–168, 2018.
- [59] P. Swaroop and B. Golden, Data Mining Introduction and a Health Care Application, Robert H Smith School of Business, University of Maryland, 2009.
- [60] E. Garcia, C. Romeo, S. Ventura and T. Calders, Drawbacks and solutions of applying association rule mining in learning management systems, *Proceedings of the International Workshop on Applying Data Mining in e-Learning*, 2007.
- [61] M. J. Zaki and W. Meira Jr, Data Mining and Analysis: Fundamental Concepts and Algorithms, New York: Cambridge University Press. USA, pp. 16, 2014.
- [62] P. N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Boston: Pearson Addison Wesley, pp. 23, 2006.
- [63] S. M. Kamruzzaman, F. Haider and A. R. Hasan, Text Classification Using Data Mining, ICTM, pp. 1-19, 2005.
- [64] J. Han, M. Kamber and J. Pei. Data Mining Concepts and Techniques, San Francisco: Morgan Kaufmann, pp. 15, 2011.
- [65] U. Fayyad, G. Piatetsky-Shapiro and P. Smith, From data mining to knowledge discovery in databases. *AI magazines*, c17, pp. 37-54, 1996.
- [66] A. Zanda, Adapting autonomously classification data mining algorithms for ubiquitous devices, A thesis submitted for the degree of Doctor of Philosophy at the University of Polytechnic de Madrid, 2014.
- [67] R. Zafarani, M. A. Abbasi and H. Liu, Social Media Mining: An Introduction, New York: Cambridge University Press, 2014.
- [68] H. Sun, Implementation of a classification algorithm for institutional analysis. A Project Submitted to the School of Graduate Studies of the University of Lethbridge, 2008.
- [69] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd ed, Boston: Morgan Kaufmann Publishers, pp. 60-101, 2006.
- [70] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd ed. Chapter 8, Boston: Morgan Kaufmann Publishers, pp. 99-117, 2012.
- [71] V. Vryniotis. Machine Learning Tutorial: The Naïve Bayes text classifiers, Datumbox Machine Learning Framework, 2013.
- [72] S. Hua, Z. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular Biology*, vol. 308, no. 2, pp. 397-407, 2001.
- [73] D. J. Strauss, G. Steidl, Hybrid Wavelet-Support Vector classification of waveforms. *JComput and Appl*, vol. 148, no. 2, pp. 375- 400, 2002.
- [74] R. Kumar, A. Kulkarni, V. K. Jayaraman, B. D. Kulkarni, Symbolization assisted SVM classifier for noisy data. *Pattern Recognition Letters*, 25: pp. 495-504, 2004.
- [75] S. Mukkamala, A.H. Sung, A. Abraham, Intrusion detection using an ensemble of intelligent paradigms. *Journal of Network and Computer Applications*, 2004.
- [76] U. Norinder, Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. *Neurocomputing*, 55: pp. 337–346, 2003.
- [77] L. Wang, Support Vector Machines: Theory and Applications. Springer, Berlin, 2005.

- [78] S. Adu-Poku, Comparing Classification Algorithms in Data Mining, A thesis submitted in partial fulfilment of degree of Master of science in Data Mining at Central Connecticut State University, New Britain, Connecticut, February, 2012.
- [79] T. Masuyama, and H. Nakagawa, Applying Cascade Feature Selection to Support Vector Text Categorization. In TJOA, A.M and Wagner, R.R (eds.) *Proceedings of the 13th International Workshop on Database and Expert Systems Applications*, ISSN: 1529-4188, vol. 300, no. 6, pp. 241-245, 2002.
- [80] O. Ivanciuc, Applications of Support Vector Machines in Chemistry, *Computational Chemistry*, vol. 23. no. 6, pp. 291-400, 2007.
- [81] N. Cristianini, Support Vector and Kernel Machines, 2001.[Online]. Available: <http://www.svms.org/tutorials/Cristianini2001.pdf>, Accessed: August 12th 2016.
- [82] "Tutorial 14: Classification--K-Nearest Neighbours (KNN) and Linear Discriminant Analysis (LDA)", [Online]. Available: <http://www.ArrayTrack.com>, Accessed: August 20, 2016.
- [83] "KNN Classification", [Online]. Available: <http://www.ArrayTrack.com>, Accessed: August 17 2016.
- [84] S. Drazin and M. Montag, Decision Tress Analysis using Weka, Machine Learning-Project II, University of Miami
- [85] T. R. Patil and S.S. Sherekar, Performance Analysis of Naïve Bayes and J48 classification Algorithm for data classification, *International Journal Of Computer Science And Applications*, Vol. 6, No.2, 2013.
- [86] E.B. Hunt, J. Marin and P.J Stone, Experiments in induction. Academic Press, New York, NY, 1996.
- [87] L. Breiman, J. Friedma, R. Olshien, R. and C. Stone. Classification and Regression trees. Wadsworth and Brooks, Monterey, CA. 1984.
- [88] L. Gula, Information Data Management for the Future of Communication. A thesis for Master of Science in Engineering Technology Computer Science and Engineering, Lulea University of Technology Department of Computer Science, Electrical and Space Engineering.
- [89] T. Y. Lin, Y. Xie, A. Wasilewska, C.J. Liao, Data Mining: Foundations and Practice, Germany, Berlin: Springer, pp. 28 2008.
- [90] S. A. Macskassy, Machine Learning: Bayesian Learning, Fall 2008. [Online]. Available: <http://www-scf.usc.edu>
- [91] W. Xindong, Z. Xingquan, W. Gong-Qing, D. Wei. Data Mining with Big Data, *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, issue.1, pp. 97-107, 2014.
- [92] A. Uddin, Big Data: Its Characteristics and Architecture Capabilities, [Online]. Available: [http:// ashrafsau.blogspot.in/](http://ashrafsau.blogspot.in/) Accessed 21 November 2016.
- [93] Data Mining From A to Z: How to Discover Insights and Drive Better Opportunities [Online]: Available: <https://www.sas.com>, SAS, 2015, Accessed: March 10 2016.
- [94] "Big Data Analytics", [Online] Available: <http://search.businessanalytics.techtarget.com>. Accessed: May 2016.
- [95] B. Marr. Big Data-What is it?, Big Data in Practice, Wiley Publishers, USA, 2016.
- [96] J. Williamson, The 4V's of Big Data, [Online]..Available: <http://www.dummies.com/careers/find-a-job/the-4-vs-of-big-data/>, Accessed October 2016.
- [97] D. Bohm, Wholeness and the Implicate Order, London and New York: Ark, pp.11, 1983.
- [98] I. Dey, Qualitative Data Analysis, London: Routledge Taylor and Francis Group, pp.26, 2005.

- [99] B. L. Berg, *Qualitative Research Methods for the Social Sciences*, San Francisco: California State University, Long Beach, pp. 15, 2004.
- [100] V. Braun, and V. Clarke, Using thematic analysis in psychology, *Qualitative Research in Psychology*, vol. 3, issue 1, pp. 77–101, 2006.
- [101] H. Becker and B. Geer, 'Participant Observation: The Analysis of Qualitative Field Data' in Burgess, Robert (ed.) *Field Research: A Sourcebook and Field Manual* London: Allen & Unwin, pp. 10, 1982.
- [102] M. Mason. Sample Size and Saturation in PhD Studies Using Qualitative Interviews, *FQS*, Volume 11, No. 3, Art. 8 September 2010.
- [103] R. Tesch, *Qualitative Research: Analysis Types & Software Tools*. Bristol, PA: Famer Press, 1990.
- [104] J. Galtung, *Theory and Methods of Social Research*, London: Allen & Unwin, pp. 13 1967.
- [105] F. Capra, *The Tao of Physics: An Exploration of the Parallels Between Modern, Physics and Eastern Mysticism*, London: Fontana Paperbacks, pp. 25, 1983.
- [106] CIRT, "Quantitative Approaches", [Online]. Available: <https://cirt.gcu.edu/research/development/resources>, Accessed 07th February 2018.
- [107] M. Denscombe, *The Good Research Guide: for small-scale social research*. McGraw Hill: Sage, pp. 269–272, 2010.
- [108] S. McLeod, Research Methods: Research data-: Qualitative vs. Quantitative, [Online]. Available: <https://www.simplypsychology.org/saul-mcleod.html>: published 2008, updated 2017, Accessed: 05th February 2018.
- [109] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine, 1967.
- [110] M. Q. Patton, *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage, 2002.
- [111] N. Denzin, and Y. Lincoln, *Handbook of Qualitative Research*. Thousand Oaks, CA, US: Sage Publications Inc., 1994.
- [112] K. Punch, *Introduction to Social Research: Quantitative and Qualitative Approaches*. London: Sage, 1998.
- [113] J. W. Creswell, John W, *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage, pp. 15, 1998.
- [114] M. Q. Patton, *Qualitative evaluation and research methods*, Newbury Park, CA: Sage, pp. 9, 1990.
- [115] J. M. Morse, Designing funded qualitative research. In Norman K. Denzin & Yvonna S. Lincoln (Eds.), *Handbook of qualitative research*, Thousand Oaks, CA: Sage, pp. 220-235, 1994.
- [116] M. B. Miles, and A. M. Huberman, *Qualitative data analysis*, 2nd ed, London, Sage, 1994.
- [117] J. Schilling. On the pragmatics of qualitative assessment: Designing the process for content analysis. *European Journal of Psychological Assessment*, vol. 22, no.1, pp. 28-37, 2006.
- [118] Y. Zhang and B. M Wildemuth, *Qualitative Analysis of Content*, 2005, [online]: Available: <https://www.ischool.utexas.edu>, Accessed: Feb 16 2016.
- [119] M. Crouch and H. McKenzie, The logic of small samples in interview based qualitative research. *Social Science Information*, vol.45, no. 4, pp. 483-499, 2006.
- [120] M. A. Bright and D. O'Connor, Qualitative Data Analysis: Comparison Between Traditional and Computerized Text Analysis, *The Osprey Journal of Ideas and Inquiry*, All Volumes (2001-2008). Issue. 21, pp.1-13, 2007.
- [121] C. Seale, Using Computers to analyse qualitative data. In Silverman, D. *Doing*

- qualitative Research, London: SAGE, pp.188-201.
- [122] G. Guest, E.E Namey, and M. L. Mitchell, Chapter 1: Qualitative Research: Defining and Designing In: *Collecting Qualitative Data: Afield Manual for Applied Research*, London: SAGE, pp. 1-10, 2013.
 - [123] L. T. Carr, The strengths and weaknesses of quantitative and qualitative research: what method for nursing?, *Journal of advanced nursing*, 20(4), pp. 716-721, 1994.
 - [124] M. Saunders, P. Lewis and A. Thornhill, Research Methods for Business Students, 6th ed, Pearson Education Limited, 2012.
 - [125] R. Antonius, Interpreting quantitative data with SPSS, London: Sage Publications, 2003.
 - [126] T. R. Black, Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics, London: Sage Publications, 1999.
 - [127] ATLAS.ti, "Comparison of Qualitative and Quantitative Research", [Online], Available: atlas.com/qualitative-vs-quantitative-research.htm, Accessed: 07th February 2018.
 - [128] R.E Hoyt and A. K Yoshihashi, Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, 6th edpp.13-14, Published November 2016, Available: <http://informaticseducation.org>. Accessed November 26th 2017.
 - [129] D. Barbara, N. Wu and S. Jajodia, "Detecting Novel Network Intrusions Using Bayes Estimators", *Proceedings Of the First SIAM Int. Conference on Data Mining, (SDM 2001)*, Chicago, IL2001.
 - [130] A. Adebawale, S.A. Idowu and A. Amarachi, Comparative Study of selected data Mining Algorithms Used for Intrusion Detection, *IJSCE* , vol. 3, no.3, ISSN: 2231-2307, pp. 1-6, July 2013.
 - [131] T. Lappas, and K. Pelechris, Data Mining Techniques for (Network) Intrusion Detection Systems, Department of Computer Science and Engineering Riverside, Riverside CA, 2006.
 - [132] E. Kesavulu, V. N. Reddy, and P. G. Rajulu, "A Study of Intrusion Detection in Data Mining". *Proceedings of the World Congress on Engineering 2011*, vol III, WCE 2011, July pp. 6 - 8, 2011, London, U.K, 2011.
 - [133] S. Sayad, An Introduction to Data Mining, [Online]. Available: http://www.saedsayad.com/k_nearest_neighbors.htm, Accessed: 24 February 2016.
 - [134] W. Lee, S. J. Stolfo and K. W. Mok, "A data mining framework for building intrusion detection models," *In Proc. of the 1999 IEEE Symp. On Security and Privacy*, pp. 120-132, Oakland, CA: IEEE Computer Society Press, 1999.
 - [135] C. Yadav, S. Wang and M. Kumar, Algorithm and approaches to handle large data: A Survey, *IJCSN*, vol. 2, no. 3, pp. 2277-5420, 2013.
 - [136] S. L. Lo, R. Chiong and D. Cornforth, Using Support Vector Machine Ensembles for Target Audience Classification on Twitter, *Plos One*, vol. 10, No. 4: e0122855, [Online]. Available: <https://doi.org/10.1371/journal.pone.0122855>, 2015, Accessed: March 26th 2016.
 - [137] W. Lee, S.J. Stolfo and K.W. Mok, "Mining in a data-flow environment: Experience in network intrusion detection," (Chaudhuri, S. & Madigan, D. Eds.). *Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pp. 114-124, San Diego, CA: ACM, 1999.
 - [138] T. D. Lane, "Machine Learning Techniques for the computer security domain of anomaly detection", Ph.D. Thesis, Purdue Univ., West Lafayette, IN, 2000.
 - [139] S. Desai and S. T. Patil, Efficient Regression Algorithms for Classification of Social Media Data. *International Conference on Pervasive Computing, IEEE*, vol. 3 no. 15, pp. 1-5, 2015.

- [140] “Clustering”, [Online]. Available: <http://www.camo.com/resources/clustering.html>, Accessed: 24 February 2016.
- [141] P. Komarek, Logistic Regression for Data Mining and High-Dimensional Classification, [Online]. Available: <http://repository.cmu.edu/robotics> 2004, Accessed July 4th 2016.
- [142] T. Minka, Data Mining, Handout 24, pp. 36-350, Nov 2003, [Online]. Available: alumni.media.mit.edu, Accessed: February 26th 2016.
- [143] X. Wu et al, Top 10 algorithms in data mining, Survey paper, DOI 10.1007/s10115-007-0114, *Knowl Inf Syst*, vol. 14, no.1, pp.1-37, 2008.
- [144] G. Negandhi, Apriori Algorithm Review for Finals, SE 157B, Spring Semester, 2007, [Online]. Available: <https://www.cs.sju.edu/>, Accessed: December 12th 2016.
- [145] “Improved Outcomes Software, Performing K-Means Clustering”. [Online]. Available: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/KMeans_Clustering_Clustering_Overview.htm. Accessed: June 11th 2016.
- [146] H. Conceicao, K- Means Clustering, [Online]. Available: OnMyPhd.com, Accessed: 27th February 2016.
- [147] E. Agichtein, C. Castilo, D. Donato, Finding High Quality Content in Social Media, In Proc. WSDM’08, pp183-194, Palo Alto, California: ACM press, 2008.
- [148] M. Mehta, R. Agrawal, J. Rissanen, SLIQ: A Fast Scalable Classifier for Data Mining, In: *Proceedings of the 5th International Conference on Extending Database Technology*, pp. 18–32, Avignon, France, 1996.
- [149] M. Mehta, J. Shafer, R. Agrawal. SPRINT: A Scalable Parallel Classifier for Data Mining, *Proceedings of the 22nd VLDB Conference, Mumbai (Bombay)*, pp. 544-555, India, 1996,
- [150] S. Papadimitriou and J. Sun, DisCo: Distributed Co-clustering with Map-Reduce A Case Study Towards Petabyte-Scale End-to-End Mining, *In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, Pages 512-521, December 15 - 19, 2008, IEEE Computer Society Washington, DC, USA.
- [151] D. Luo, C. Ding, H. Huang, Parallelization with Multiplicative Algorithms for Big Data Mining, *2012 IEEE 12th International Conference on Data Mining*, pp.489-498, IEEE Computer Society, Washington, DC, USA.
- [152] X. Wu, and X. Zhu, Transactions on Knowledge and Data Engineering, *IEEE Data Mining with Big Data*, vol. 26, no.1, pp. 97-107, 2014.
- [153] “Business Process Modelling”, [Online]. Available: https://en.wikipedia.org/wiki/Business_process_modeling, Accessed July 23rd 2018
- [154] “Data Collection”, [Online]. <http://searchcio.techtarget.com/definition/data-collection> Accessed: October 2016.
- [155] J. Boitnott, Why Collecting Data Digitally is the Best Way to Go in 2016, [Online]. Available: <https://www.inc.com>, Accessed: December 05th 2017.
- [156] P. Hawe and D.D. Hall, Evaluation Health Production: A Health’s Worker Guide, Sydney: MacLennan & Petty, pp. 254.
- [157] D.R. Thomas, A general Inductive Approach for qualitative data analysis, *American Journal Of Evaluation*, vol. 27, no. 2. pp. 237, 2006.
- [158] Toolkit [Online]. Available: <http://toolkit.pellinstitute.org/evaluationguide> Accessed: August 18, 2016.
- [159] S. R. Joseph, H. Hlomani and K. Letsholo, Data Mining Algorithms: An overview, *International Journal of Computers and Technology*, vol. 15, no. 6, pp. 6806-6813, 2016.
- [160] Techopedia, Data Storage, 2017. [Online]. Available: <https://www.techopedia.com>.

Accessed: 29th May 2017.

- [161] Netlytic (2016), [Online]. Available: <https://netlytic.org>, Accessed: March- August 2016.
- [162] Look-id (2016), [Online] Available: www.look-id.com, Accessed: March- August 2016.
- [163] R. Jane; L. Jane and E. Gillian, 'Designing and selecting samples'. In Jane Ritchie Jane Lewis (Eds.), *Qualitative research practice. A guide for social science students and researchers*, Thousand Oaks, CA: Sage, 2003, pp.77-108.
- [164] S. Hsieh and S. Shannon, Three approaches to qualitative content analysis, *Qualitative H.-F Health Research*, vol. 15, issue. 9, pp. 1277-1288, 2005.
- [165] B. DeWever, T. Schellens, M. Valcke, M., and H. VanKeer, Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computer & Education*, no. 46, no. 1, pp. 6-28, 2006.
- [166] V. Minichiello, R. Aroni, E. Timewell and L. Alexander. In-Depth Interviewing: Researching People. Hong Kong: Longman Cheshire, 1990.
- [167] R. P. Weber, Basic Content Analysis. Newbury Park, CA: Sage Publications, 1990.
- [168] J. W. Creswell, Educational research: Planning, conducting, and evaluating quantitative and qualitative research. Upper Saddle River, NJ: Pearson Education, 2002.
- [169] J. Bradley, Methodological issues and practices in qualitative research. *Library Quarterly*, vol.63, no.4, pp. 431- 449, 1993.
- [170] A. Abran, A. Khelifi, A.W. Suryn and A. Seffah, Usability Meanings and Interpretations in ISO Standards, *Software Quality Journal*, vol. 11, no. 4, pp. 325-338, 2003.
- [171] S. Valenti, A. Cucchiarelli, A. and M. Panti, Computer Based Assessment Systems Evaluation via the ISO9126 Quality Model, *Journal of Information Technology Education*, vol. 1, no.3, pp. 157-175, 2002.
- [172] ISO 9126: Information Technology Software Product Evaluation-Quality characteristics and guidelines for their use, ISO 9126, 1991.
- [173] J. P. Chin, V.A. Diehl and K.L. Norman. Development of an Instrument Measuring User satisfaction of the Human-computer Interface. *ACM CHI'88 Proceedings*, pp. 213-218, 1988.
- [174] F. D. Davis. Perceived Usefulness, Perceived Ease of Use and User Acceptance of information Technology, *MIS Quarterly*, vol. 13, no.3, pp. 319-340, 1989.
- [175] J. Nielsen, Usability Engineering, Academic Press. Chapter 2.2, pp. 26, 1993.
- [176] J. Nielsen, Usability Engineering, Academic Press. Chapter 5, pp. 115, 1993.
- [177] A. M. Lund, Measuring Usability with the USE Questionnaire, STC Usability SIG Newsletters, vol. 8, no. 2, 2001. [Online]. Available: garyperlman.com/quest/.
- [178] J. R. Lewis, IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use, *International Journal of Human Computer Interaction*, vol. 7, no. 1, pp. 57-78, 1995.
- [179] G. Perlman, Practical Usability Evaluation, Based in part on Nielsen's 1993 Heuristics and Norman's 1990 principles, 1997.
- [180] X. H. Lin, Y.Y. Choong and G. Salvendy. A proposed Index of Usability: A Method for Comparing the Relative Usability of Different Software Systems, *Behaviour & Information Technology*, vol. 16, no. 4/5, pp. 267-278, 1997.
- [181] B. Deshpande, 2 main difference between classification and regression, [Online]. Available: www.simafare.com, Accessed December 2nd 2017.
- [182] N. Gayathri, S. Nickolas, A.V. Reddy and R. Chitra, Performance Analysis of Data mining Algorithms and for Software Quality Prediction, in *IEEE ARTCOM 09*

- Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing*, pp. 393-395.
- [183] S. S. Aksenova, Machine Learning with WEKA: WEKA Explorer Tutorial, pp. 1-64, 2004.
 - [184] E. Bhuvaneswari and V. R. S. Dhulipala, The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset, *Information Engineering*, vol. 2, no. 1, pp. 6-13, 2013.
 - [185] R. Righat, Kappa statistics in plain English, [Online]. Available: www.stats.stackexchange.com, Accessed: August 16 2016.
 - [186] J. R. Landis and G.G. Koch, The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, no. 1, pp. 159-174, 1977. Doi: 10.2307/2529310.
 - [187] J. L. Fleiss, J. Cohen and B.S. Everit, Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin*, vol. 72, no. 5, pp. 323- 327, 1969. Doi: 10.1037/h0028106.
 - [188] T. Wood, Using Mean Absolute Error for Forecast Accuracy, [Online]. Available: <http://canworksmart.com/using-mean-absolute-error-forecast-accuracy/>. Accessed: August 18 2016.
 - [189] R. Nau, What's a good value for R-squared? [Online]. Available: <http://people.duke.edu>, Accessed: August 2016.
 - [190] T. Chai and R. R. Draxler, Root mean square error mean absolute error-Arguments against avoiding RSME in the literature, *Geosci. Model Dev*, vol. 7, issue. 3, pp. 1247-1250, 2014.doi 10.5194/gmd-7-1247-2014.
 - [191] "Mean Absolute Error and Root Mean Squared Error", [Online]. <http://www.eumetcal.org/resources/ukmeteocal>. Accessed: August 10th 2016.
 - [192] MedCalc 1993-2016, Software bba manual. ROC curve analysis in MedCalc, [Online]. Available: <https://www.medcalc.org/manual/roc-curves.php>, 2016, Accessed: August 18 2016.
 - [193] M. Gonen, Chapter 4: Analyzing Receiver Operating Characteristic Curves [Online], <http://ebp.uga.edu/courses/Chapter%204%20%20Diagnosis%20I%208%20%20ROC%20curves>, Accessed: August 12th 2016.
 - [194] H. M. Chuang and P. Gray, Special selection: Data mining: *Journal of Management Information Systems*, vol. 16, no.1, pp. 11-16, 1999, DOI 10.1080/07421222.
 - [195] F. W. Struwig and G.B. Stead, Planning, Designing and Reporting Research, Understanding Reliability and Validity, Cape Town, South Africa : Pearson Education,. pp. 134-145, 2004.

APPENDICES

APPENDIX A: FOCUS GROUPS RESULTS FOR INDUCTIVE CONTENT ANALYSIS

Table 1: Focus Group 1 categories

Categories	Some of top 30+ words
Cruel and Abusive Government	Inconsequential ,Ministers, purports, why, would, change, status-quo, then, Poignantly, fracas, Regime, fails, yet, again, embrace, dialogue, detests, shuns, inclusive, governance, says, workers, boss, go-hang, In-fact , democracy, sold, democratic, test, Disrespect, rule, law, harassment, journalists, extra judicial, killings.
Corruption	Project,failure,abuse,stealing,government,money,unfair,corruptiveleadership,favouritism,top,jobs,delegations, appointments, extra, judicial ,killings, disrespect, to rule, law, agitation, judicial, system, abuse, power, abuse,state,media,resources, bribe, payments, high offices, illegitimate, presents, by, officials, Botswana-Railways, suspect, Sonny, Serite, set-up, those, Believe, later , other, accomplish, might, turn, out, state,witness,leaving,vulnerable,hot,headed,pro- government, judges, buy, story, because, think , go in-front,Office,President,collect,personal,file,other went, there, for, second, appointment, meeting, scheduled. Bribery,Palapye-Glass.
Partisanship	Given,scale,abuse, partisanship, cronyism, next,substantive,President, after, will, have, big, task, carved, out, for, rekindle, diplomatic- appointments, make, more, reflective,country, diversity,acolytes,appointing, authority, would, seem, like, such, appointment, form, part, patronage, largesse, gladly,spreads, around, keep circle, friends, intact. Would, seem, like, such, appointments, exist, appease,placate, tensions ,within, party more, importantly, win, control, over, many, opposing, interests,party,terrible,welter, political, abuse, successful,doling, offshore, jobs, members, establishment, many,who, have, been, discarded, actually, denies ,country,benefits ,that, could, come, bear, career, diplomats, given, due, attention, more, regrettably, culture abus, bred,even, bigger, political,administrative, fiasco.
Incompetent Government	Poor-education-system, poor-results, enslavement, public servants, sour, relations, between government, trade, unions, public ,strikes, low, salary, increments, unexplained ,deaths, high, youth, unemployment-rates, lack –of-jobs, favouritism, top, jobs, delegations ,appointments, provision, schemes, programs, eradicate, poverty, graduates, labour, exploitation, meaningless graduate, schemes, unexplained ,mysterious, deaths.
Judiciary Agitation	Disrespect,rule,law, harassment, judges, suspension, unfair, biased,judicial,judgments, presidential, dictatorship, creating, constitutional, security, tenure, infuse, laws, loop-holes, closed. Witnessed, compromised, played, botate, shamelessly, crawling, their, knees, kowtowing, secure, presidency, reduced, judiciary, shameful, low-levels, further, advocates, long, overdue, constitutional-review, bring, Botswana-constitution,up-to,standard,constitutions,best,democracies,world,irredeemably dysfunctional, corruption, punctuated, impunity, cornerstone, administration, goodwill, retains, especially as reflected, international-ratings, nothing, much, more, windfall, overhang, appropriate, term, capital accumulated, years, preceded, arrival, prove, adherence, governance, disingenuously, riding, wave, create. re-make, credibility, integrity, independent, appendage, executive, arm, worrisome,coherent, authority, strategy ,restore, literally, reduced, stature, public hold, institution, people, manning, heartfelt, reference.
Economic And Political	Mouths, stomachs, politicians,represent,ever,lasting, unquenchable, desires, purchasing, blue, train,lie, Morupule-B-project-failure, North-South-Water-carrier-project, Botswana-Railways, extra-judicial, killings,

Scandals	Political, scandals, votes, rigging, votes, fixing, exposure, youth, delinquent- acts, exposure.
Poor Education System	Poor-Education-System, leading ,disastrous, results, painful, ,whole generation, Botswana,future, leaders, our children, being, relegated, uncertain, gloomy, future ,failure ,leaders,engage, meaningfully, order , improve, education, delivery, government, never, loved, youth,education, crisis, mode, perennially ,advocating, increased, funding, resourcing , betterment, education-sector,correctly, pointed, lasting, solution, prevailing, crisis, start, engaging, teachers, improving, immediate, impact, welfare, school- results, advocating, rationalization, regulation, fees, involved, quality-Education, private, public, guiding, principle, Production, views, process, socially, laughable, comical , comrades, , kids , pathetic, public-schools, , examination, performance, lowest law, force, Legislatures, Cabinet, Ministers, public-institution, policy-makers , inspire confidence , lead, example, implementation , Revised- National-Policy.

Table 2: Focus group 2 categories

Categories	Some of top 30+ words
Cronyism	Cronyism, importantly, win, control, over, many opposing, interests, terrible, welter, political, while, successful, doling offshore, jobs, members, establishment, many, discarded, given, scale, abuse, partisanship, next, substantive, President after, have, big, task, carved, out, rekindle, diplomatic, appointments, make more, reflective, country, diversity, acolytes, appointing, authority, would, seem, like, such, appointment, form, part, patronage, largesse, gladly, spreads, around, keep, circle, friends, intact, exist, appease, placate, tensions within, party, fiasco, more, actually, denies, benefits, come, bear, career, diplomats, given, due, attention, regrettably, culture, abuse, bred, even, bigger, administrative.
Poor Education System	Correctly, pointed, lasting, solution, prevailing, crisis, start, engaging, teachers, improving, immediate, impact, welfare, school- results, advocating, Poor-Education-System, laughable, comical, comrades, kids, pathetic, public-schools, examination, performance, lowest law, force, Legislatures, Cabinet, Ministers, public-institution, policy- makers, inspire confidence, lead, example, implementation, Revised- National-Policy, leading, disastrous, results, painful, whole generation, Botswana, future, leaders, our children, being, relegated, uncertain, gloomy, future, failure, leaders, engage, meaningfully, order, improve, education, delivery, government, never, loved, youth, education, crisis, mode, perennially, advocating, increased, funding, resourcing, betterment, education-sector, rationalization, regulation, fees, involved, quality-Education, private, public, guiding, principle, Production, views, process, socially.
Judiciary Agitation	Dysfunctional, corruption, punctuated, impunity, cornerstone, administration, goodwill, retains, especially as reflected, international-ratings, nothing, much, more, windfall, overhang, appropriate, term, capital accumulated, years, preceded, arrival, prove, adherence, governance, disingenuously, riding, wave, disrespect, rule, law, harassment, judges, suspension, unfair, biased, judicial, judgments, presidential, dictatorship, creating, constitutional, security, tenure, infuse, laws, loop-holes, closed. Witnessed, compromised, played, botate, shamelessly, crawling, their, knees, kowtowing, secure, presidency, reduced, judiciary, shameful, low-levels, further, advocates, long, overdue, constitutional-review, bring, Botswana-constitution, create. re-make, credibility, integrity, independent, appendage, executive, arm, worrisome, coherent, authority, strategy, restore, literally, reduced, stature, public hold, institution, people, manning, heartfelt, reference, up-to, standard, constitutions, best, democracies, world, irredeemably.
Conflicts Between The Government And Unions	Battles, Regime, waged, against, people, most, nail-biting, one, tussle, with, BOFEPUSU, federation, public, sector, unions, Botswana.
In-Competent Government	Mis-use, public funds, untruthful, leadership, tax- payments, Selfish, blue – train, money, Transnet, reported, huge, turnover, performance, quarter,

	financial, statement, official, figure, given, excludes, bribe, payments, illegitimate, given, youth, unemployment-rates, lack –of-jobs, favouritism, top, jobs, delegations, appointments, provision, schemes, programs, eradicate, poverty, graduates, Botswana-Railways, officials, ministry, poor-education-system, poor-results, enslavement, public servants, sour, relations, between government, trade, unions, public, strikes, low, salary, increments, unexplained, deaths, high, labour, exploitation, meaningless graduate, schemes, unexplained, mysterious, deaths, board, executives, cronies, notice, embracing, changes, lead.
Scandals	Botswana-Railways, desires, purchasing, blue, train, extra-judicial, killings, mouths, stomachs, politicians, represent, ever, lasting, unquenchable, lie, Morupule-B-project-failure, North-South-Water-carrier-project, votes, rigging, votes, fixing political, scandals, , exposure, youth, delinquent- acts, exposure.
Un-Transparent Government	Doing, such, drastic, amendment, BEC- Act, without, consulting, teachers, seeks, shove, drastic, job, design, change, down, throats.

3286

Table 3: Focus Group 3 categories

Categories	Themes
Excessive partisan diplomatic appointments	Appointments, positions, often, done, wrong, reasons, party, political, considerations, tended, supersede, everything, witnessing, friendship, cronyism, rearing, their, ugly, heads, appointments, visceral, aversion, contempt, professional, career, civil, servants, links, ruling, party, behind, tragedy, indignity, aside, current, lack, strategic, grasp, diplomatic, relations.
Corruption	Government, bribe, payments, high, offices, corruption, illegitimate, presents, officials, Botswana, Railways, set-up, Believe later, other, accomplish, might, turn, out, witness, vulnerable, headed, pro- government, judges, buy, story, because, go, in-front, office, president, collect, personal, file, went, second, appointment, meeting, scheduled, bribery, Palapye, Glass, Project, failure, abuse, stealing, government, money, unfair, corruptive, leadership, favouritism, top, jobs, delegations, appointments, extra, judicial, killings, disrespect, law, agitation, judicial, system, abuse, power, abuse, state, media, resources.
Poor foreign policies and security policies	Long, Botswana, clear, up-to-date, National -Security -Strategy, country, will, continue, shooting, blanks, regard, resourcing, maintaining, security, apparatus, complimented, general, well, pointed, country, going, continue, sailing, troubled waters, regard, foreign, policy, continues, having, written, accepted, updates, legislators, went, straight, root, cause, defence, problems, intentionally, springs, continue, looting, billions -of-Pulas, annual, hide, behind, cloak, security, reasons, divulge, information, public, regarding, procurement.

Incompetent Government	Bribe, payments, illegitimate, given ,Botswana-Railways, officials, ministry, board, executives, cronies, notice, embracing, changes, lead, lack -of-jobs, favouritism, top, jobs, delegations , funds, untruthful, leadership , tax- payments, Selfish ,blue -train, money, Transnet, reported, huge, turnover, performance, quarter, financial, statement, official, figure, given, appointments, provision, schemes, programs, eradicate, poverty, graduates, labour, exploitation, meaningless graduate, schemes, unexplained ,mysterious, deaths, poor-education-system, poor-results, enslavement, public servants, sour, relations, between government, trade, unions, public ,strikes, low, salary, increments, unexplained ,deaths, high, youth, unemployment-rates, mis-use, public, excludes.
Disrespect to rule of law and abuse of power	Needless,state, courts, law,ruled, that, those, outside, scope, educators, work, want, reverse, diabolically, like,bully, man, would, battered, wife, gains, made, write, Minister, may, direct.
Scandals	Morupule-B-project, failure, North -South -Water- carrier- project, Botswana Railways, extra, judicial, killings, political, scandals, votes, rigging, vote, fixing exposure, youth, delinquent, acts, exposure, mouths, stomachs, politicians, represent, ever, lasting, unquenchable, desires, purchasing, blue, train, lies.
Un-transparent government;	Change, down, Doing, such, drastic, amendment, BEC- Act, without, consulting, teachers, seeks, shove, drastic, job, design, throats.
Poor Education System	Education, crisis, delivery , immediate, impact, improving ,welfare, school, results, advocating , rationalization , regulation, fees, involved, advocate, quality, private , public, guiding, principle, production ,views , production, process passive, acquisition, knowledge, performance, lowest ,policy makers, inspire, confidence, lead, example, implementation, revised- national policy, order, improve education, delivery, government, never, loved, youth, education, delivery, crisis mode, future, way, leaders, perennially, advocating, increased, funding, resourcing, betterment, sector , none ,correctly, pointed, out, lasting, solution, prevailing, poor-education-system, leading ,disastrous, results, painful, when, see ,whole generation, Botswana, future, leaders, children, being, relegated, uncertain, gloomy, future, failure, leaders ,engage, teachers, meaningfully.
Partisanship	Importantly, win, control, over, many opposing, interests, terrible, welter, political, while, successful, doling offshore, jobs, members, establishment , many, discarded, actually, denies, benefits, come, bear, career, diplomats, given, due, attention, regrettably, culture, abuse, bred, even, bigger, administrative, fiasco, given, scale, abuse, partisanship, cronyism, next, substantive, President after, have ,big, task, carved, out, rekindle, diplomatic, appointments, make more, reflective, country, diversity, acolytes, appointing, authority, would, seem, like, such, appointment, form, part, patronage, largesse, gladly, spreads, around, keep, circle, friends, intact, exist, appease, placate, tensions within, party, more.

lack of diplomatic benchmarking	Reason, foreign, policy, lacks, any, pivotal, diktat, almost, daily, basis, vacillates, from, one vacuum, exactly, because, anchored, shallow, intellectual, threads, ambassadors, high-commissioners, themselves, first, foremost, political rejects, home, case, made, start, appointing, more, career, diplomats, simply, start, looking, other, countries, doing, appointments, diplomatic, representatives, broadening, even, extending, appointments, even, well, known, opposition, figures.
Poor meaningless economic development programmes	Programmes, panic, wonder, why, tells, me, that, political propaganda, still, lives, will, never, die, behind, Economic, stupid, dynasties, elites.
Violation of labour laws	Government, again, trying, forcibly, immorally, enact, labour, laws, that, seek, punish, workers, take, away, their, employment, rights, trade, union, rights, regrettably, learnt, minister, labour, home- affairs, tabled, amendments, trade, dispute, act, consequently, make, entire, civil, service, essential, service, contrary, pass, defeat, freedom, association, organize, collective, bargaining.
In-Competent Government	High, youth, unemployment-rates, lack -of-jobs, favouritism, top, jobs, delegations, appointments, provision, schemes, programs, eradicate, poverty, graduates, labour, exploitation, meaningless graduate, schemes, unexplained, mysterious, deaths, mis-use, public funds, untruthful, leadership, tax- payments, Selfish, blue -train, money, Poor-education-system, poor-results, enslavement, public servants, sour, relations, between government, trade, unions, public, strikes, low, salary, increments, unexplained, deaths, Transnet, reported, huge, turnover, performance, quarter, financial, statement, official, figure, given, excludes, bribe, payments, illegitimate, given, Botswana-Railways, officials, ministry, board, executives, cronies, notice, embracing, changes, lead.
Judiciary Agitation	Loop-holes, closed. Witnessed, compromised, played, shamelessly, crawling, their, knees, kowtowing, secure, presidency, reduced, judiciary, shameful, low-levels, further, advocates, long, overdue, constitutional-review, disrespect, rule, law, harassment, judges, suspension, unfair, biased, judicial, judgments, presidential, dictatorship, creating, constitutional, security, tenure, infuse, laws bring, executive, arm, worrisome, coherent, authority, strategy, restore, literally, reduced, stature, public hold, institution, people, manning, heartfelt, reference. Botswana-constitution, standard, constitutions, best, democracies, world, irredeemably dysfunctional, corruption, punctuated, impunity, cornerstone, administration, goodwill, retains, especially as reflected, international-ratings, nothing, much, more, windfall, overhang, appropriate, term, capital accumulated, years, preceded, arrival, prove, adherence, governance, disingenuously, riding, wave, create, re-make, credibility, integrity, independent, appendage.
Violation of freedom of media	Squabbles, with, devious, regime, dates, back, when, followed, king, early, days Water-Utilities, read, between, lines, Sonny, been, thorny, issue, powers, through, Facebook, page, column, Telegraph.

Table 4: Focus Group 4 categories

Categories	Themes and thematic expressions
Corruption	Favouritism, top, jobs, delegations, appointments, bribe, payments, high, offices, corruption, illegitimate, presents, officials, Botswana, Railways, set-up, Believe later, other, accomplish, might, turn, out, witness, vulnerable, headed, pro- government, judges, buy, story, because, go, in-front, office, president, collect, personal, file, went, second, appointment, meeting, scheduled, bribery, Palapye, Glass, Project, failure, abuse, stealing, government, money, unfair, corruptive, leadership, extra, judicial, killings, disrespect, law, agitation, judicial, system, abuse, power, abuse, state, media, resources, government.
Lack of mutual working relationship	Public, servants, strike, aftermath, regime, exhibited, kinds, authoritarianism, departing, perceived, democratic, nature, looked, back, bracing, ceasefire, anytime soon, have, intensified, sharpened, advanced, their, weapon, stoking, speculations, horror, awaits, nation, workers, have, been, hardest, hit, fighting, with teachers, usual.
In-competent Government	Poverty, graduates, labour, exploitation, meaningless graduate, schemes, unexplained, mysterious, Bribe, payments, illegitimate, given, Botswana-Railways, officials, ministry, board, executives, cronies, notice, embracing, changes, lead, lack -of-jobs, favouritism, top, jobs, delegations, funds, untruthful, leadership, tax- payments, Selfish, blue -train, money, deaths, poor-education-system, poor-results, enslavement, public servants, sour, relations, between Transnet, reported, huge, turnover, performance, quarter, financial, statement, official, figure, given, appointments, provision, schemes, programs, eradicate, government, trade, unions, public, strikes, low, salary, increments, unexplained, deaths, high, youth, unemployment-rates, mis-use, public, excludes.
Abuse of power	Diabolically, like, bully, man, would, battered, wife, gains, made, write, Minister, may, direct, courts, law, ruled, that, those, outside, scope, educators, work, want, reverse, needless, state.
Judiciary Agitation	Constitutions, best, democracies, world, irredeemably dysfunctional, corruption, punctuated, impunity, cornerstone, administration, goodwill, retains, especially as reflected, international-ratings, nothing, much, more, windfall, overhang, appropriate, term, capital, disrespect, rule, law, harassment, judges, suspension, unfair, biased, judicial, judgments, presidential, dictatorship, creating, constitutional, security, tenure, infuse, laws, loop-holes, closed. Witnessed, compromised, played, botate, shamelessly, crawling, their, knees, kowtowing, secure, presidency, reduced, judiciary, shameful, low-levels, further, advocates, long, overdue, constitutional-review, bring, Botswana-constitution, up-to, standard, accumulated, years, preceded, arrival, prove, adherence, governance, disingenuously public hold, institution, people, manning, heartfelt, reference, riding, wave, create, re-make, credibility, integrity, independent, appendage, executive, arm, worrisome, coherent, authority, strategy, restore, literally, reduced, stature.

Poor workers welfare	Botswana, Examinations, Council, Amendment, Bill, Act, amend , Botswana-Examinations- Council, move, seeks, circumvent, law, coerce ,teaching, cadre, performing, duties, which, theirs, exams, invigilation, marking, tantamount, beating one's partner into submission, treats, teacher, door-mat, minister, education, tabled, parliament, week, that ,will, seek, force ,teachers, invigilate, mark, exams, part, their core, duties, constant, battles, with ,all, types ,workers, lawyers, judges, government, workers, including, nurses, doctors ,been, court, through, t representatives, trying, conditions, work. Service, provision, public, service, fallen, across, departments, like ,education , health, administration, justice, state, struggle, continues, clear, current, government, really, anti-workers, persevere , fight, rights, ensure , laws, enacted.
Partisanship	Establishment , many, discarded, actually, denies, benefits, come, bear, career, diplomats, given, due, attention, regrettably, partisanship, cronyism, Importantly, win, control, over, many opposing, interests, terrible, welter, political, while, successful, doling offshore, jobs, members, next, substantive, President after, have ,big, task, carved, out, rekindle, diplomatic, appointments, culture, abuse, bred, even, bigger, administrative, fiasco, given, scale, abuse, make more, around, keep, circle, friends, intact, exist, appease, placate, tensions within, party, more, reflective, country, diversity, acolytes, appointing, authority, would, seem, like, such, appointment, form, part, patronage, largesse, gladly, spreads.
Oppressive government	Consequential, ministers, purports, would, change, status, quo, poignantly, fracas, regime, fails, yet, again, embrace, dialogue, detests, shuns, inclusive, governance, regime, workers, boss, hang, democracy, sold, fails, democratic, test, tussle, unions revealed, true, character, regime, shed, fade, fooled, many, believe, Botswana, shining, example, democracy, Africa, such, exposure, some, demoted, government, placed, where, belongs, camouflaged, violator, human, rights, cadre.
Un-transparent government	Drastic, amendment, Change, down, Doing, such, drastic, job, design, throats ,BEC-Act, without, consulting, teachers, seeks, shove.
Unfair /unclear appointments to key diplomatic appointments	visceral, aversion, contempt, professional, career, civil, servants, links, ruling, party, behind, tragedy, Appointments, positions, often, done, wrong, reasons, party, political, considerations, tended, supersede, everything , current, lack, strategic, grasp, diplomatic, relations, witnessing, friendship, cronyism, rearing, their, ugly, heads, appointments, indignity, aside.
Violation of workers's rights	True ,Botswana, government, turn ,increasingly, becoming, enemy, workers, people, making, introducing, secretly, laws, suppose, safeguard, rights, Freedom, people, need, vigilant, freedom, loving ,persons, jealously ,guard , Freedom, coming, with secretly, laws, that ,forces, Teachers, invigilate, , solution, worse, decision, since already , failing, address,conditions, working, hours, entire ,public, service.
Un-performing government	Strikes, low, salary, increments, unexplained ,deaths, high, youth, unemployment-rates, mis-use, public, excludes, , Bribe, payments, illegitimate, given ,Botswana-Railways, officials, ministry, board, executives, cronies, notice, embracing, changes, lead, lack -of-jobs, favouritism, top, jobs, delegations , funds, untruthful, leadership , tax- payments, Selfish ,blue -train, money, deaths, poor-education-system, poor-

	results, enslavement, Poverty, graduates, labour, exploitation, meaningless graduate, schemes, unexplained ,mysterious public servants, sour, relations, between Transnet, reported, huge, turnover, performance, quarter, financial, statement, official, figure, given, appointments, provision, schemes, programs, eradicate, government, trade, unions, public.
Political and economic scandals	Political, scandals, votes, rigging, vote, fixing exposure, youth, delinquent, acts, exposure, mouths, Morupule-B-project, failure, North -South -Water- carrier-project, , represent, ever, lasting, unquenchable, desires, purchasing, blue, train, lies, Botswana Railways, extra, judicial, killings, stomachs, politicians.
Poor Education System	Perennially, advocating, increased, funding, resourcing, betterment, sector , none ,correctly, pointed, out, lasting, solution, prevailing, poor-education-system, leading ,disastrous, results, acquisition, knowledge, performance, lowest ,policy makers, inspire, confidence, lead, example, implementation, revised- national policy, order, improve education, delivery, government, never, loved, youth, education, delivery, crisis mode, future, way, leaders, painful, Education, Botswana, future, leaders, children, being, relegated, uncertain, gloomy, future, failure, leaders ,engage, teachers, meaningfully. crisis, delivery , immediate, impact, improving ,welfare, school, results, advocating , rationalization , regulation, fees, involved, advocate, quality, private , public, guiding, principle, production ,views , production, process passive, when, see ,whole, generation.
Harassment of journalists (the media)	Between, with, devious, telegraph, regime, dates, back, when, followed, king, early, days Water-Utilities, read, lines, Sonny, been, thorny, issue, powers, through, Facebook, page, column, squabbles.
Poor foreign policies and security policies	Pointed, country, going, continue, sailing, troubled waters, regard, foreign, policy, continues, having, written, accepted, updates, legislators, went, straight, root, cause, defence, problems, divulge, information, public, regarding, procurement. Long, Botswana, clear, up-to-date, National -Security -Strategy, country, will, continue, shooting, blanks, regard, resourcing, maintaining, security, apparatus, complimented, general, well, intentionally, springs, continue, looting, billions -of-Pulas, annual, hide, behind, cloak ,security, reasons.

APPENDIX B: SAMPLE OF CONSENT FORM USED

CONSENT FORM

By signing this document I agree that I have been given enough information about this Project. It has been explained to me how the information I give will be used. Furthermore I give permission for my words to be used in a report but understand that my name will not be mentioned.

#	LAST NAME	FIRSTNAME	PROGRAMME	CONTACT	SIGNATURE
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					

DATE: 24/06/2016

Name of a Researcher: Sethunya Rosie Joseph
Signature:.....

Date: 24/06/2016

Supervisors : Dr Hlomani and Dr Letsholo

APPENDIX C: THEMATIC CODING STEPS MANUAL

CODE SCHEME MANUAL GUIDE

For the study

“A Conceptual Framework For Data Mining And Analysis Of Social Media Data”

(Adapted from David Thomas’s “A general inductive approach for qualitative data analysis”)

Prepared by S. R. Joseph

This manual contains **general aspects** which are followed when performing the content analysis for the qualitative data. It contains guidance on how data has to be processed and coded into themes and categories. Therefore it is going to serve as a guide in the analysis of the data for this study of analysing Facebook and Twitter data on the political landscape of Botswana. It can be applied to any other qualitative data analysis. There are no **pre-defined** categories from the data since the intension is to explore the data and deduce the themes and categories which will be used for further classification of data.

The process of inductive coding

Inductive coding begins with close readings of text and consideration of the multiple meanings that are inherent in the text. The researcher then identifies text segments that contain meaning units, and creates a label for a new category into which the text segment is assigned. Additional text segments are added to the category where they are relevant. At some stage the researcher may develop an initial description of meaning of category and by the writing of a memo about the category (e.g., associations, links and implications). The category may also be linked to other categories in various relationships such as: a network, a hierarchy of categories or a causal sequence.

Features of categories developed from coding

The outcome from an inductive analysis is the development of categories into a model or framework that summarises the raw data and conveys key themes and processes. The categories resulting from the coding, which are the core of inductive analysis, potentially have five key features.

- 1. Label for category:* Word or short phrase used to refer to category. The label often carries inherent meanings that may not reflect the specific features of the category.
- 2. Description of category:* Description of the meaning of category including key characteristics, scope and limitations.
- 3. Text or data associated with category:* Examples of text coded into category that illustrate meanings, associations and perspectives associated with the category
- 4. Links:* Each category may have links or relationships with other categories. In a hierarchical category system (e.g., tree diagram) these links may indicate superordinate, parallel and subordinate categories (e.g., *parent*, *sibling* or *child* relationships). Links are likely to be based on commonalities in meanings between categories or assumed causal relationships.
- 5. Type of model in which category is embedded:* The category system may be incorporated in a model, theory or framework. Such frameworks include; an open network (no hierarchy or sequence), a temporal sequence (e.g., movement or time), or a causal network (one category causes changes in another). It is also possible that a category may not be embedded in any model or framework.

The following procedures are used for inductive analysis of qualitative data.

1. Preparation of raw data files ("data cleaning")

Format the raw data files in a common format (e.g., font size, margins, questions or interviewer comments highlighted) if required. Print and/or make a backup of each raw data file (e.g., each interview).

2. Close reading of text

Once text has been prepared, the raw text should be read in detail so the researcher is familiar with the content and gains an understanding of the "themes" and details in the text.

3. Creation of categories

The research identifies and defines categories or themes. The upper level or more general categories are likely to be derived from the research aims. The lower level or specific categories will be derived from multiple readings of the raw data. Categories are created from meaning units or actual phrases used in specific text segments. Several procedures for creating categories may be used.

4. Overlapping coding and un-coded text

Among the commonly assumed rules that underlie qualitative coding, two are different from the rules typically used in quantitative coding: (a) one segment of text may be coded into more than one category. (b) A considerable amount of the text may not be assigned to any category, as much of the text may not be relevant to the research objectives.

5. Continuing revision and refinement of category system

Within each category, search for subtopics, including contradictory points of view and new insights. Select appropriate quotes that convey the core theme or essence of a category. The categories may be combined or linked under a superordinate category when the meanings are similar. An overview of the coding process is shown in Table 1. The intended outcome of the process is to create three to eight summary categories, which in the coder's view captures the key aspects of the themes in the raw data and which are assessed to be the most important themes given the research objectives. NOTE: Inductive coding which finishes up with more than about eight major themes can be seen as incomplete. In this case some of the categories may need combining or the coder has not made the hard decisions about which themes or categories are most important.

Table 1: The coding process in Inductive analysis (Adapted from Creswell, 2002, figure 9.4, pp266)

1.Initial read through text data	2.Identify specific segments of information (themes)	3.Label the segments of information to create categories	4.Reduce overlap and redundancy among the categories	5. Create a model incorporating most important categories
				
Many pages of text	Many segments of text(themes)	30-40 categories	15-20 categories	3-8 categories

APPENDIX D: THE ICA USABILITY EVALUATION QUESTIONNAIRE

EVALUATION OF INDUCTIVE CONTENT ANALYSIS (ICA) PROCESS

PURPOSE: This questionnaire evaluates the Inductive Content Analysis Process (ICA) which had been carried out as part of the evaluation of the Proposed Conceptual Framework for mining and analysing social media data.

Your Participation Will Be Highly Appreciated.

Thank you very much!

SECTION A: DEMOGRAPHIC DETAILS									
Instruction: Tick on the appropriate white box (answer)									
1. Age	< 20		20-25		26-30		≤31+		
2. Gender	Female		Male						
3. Year of Study	Year 1		Year 2		Year 3		Year 4		Post Grad
4. Program Of Study	CS		IS		Other (state)				
5. Language proficiency	English		Setswana		Both				

SECTION B: For each question, tick only ONE appropriate answer on the right

QUESTIONS	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
LEARNABILITY					
6. I easily remember the ICA process	1	2	3	4	5

7. ICA process steps are clear and understandable	1	2	3	4	5
8. It was easy to learn ICA steps and apply them	1	2	3	4	5
9. I performed the ICA process quickly	1	2	3	4	5
10. The supplement reference materials for ICA were clear and helpful	1	2	3	4	5
EASE OF USE					
11. The overall process of ICA took a lot of time to be completed	1	2	3	4	5
12. The ICA process was not demanding and required less effort	1	2	3	4	5
13. After training I took less time to perform the ICA process	1	2	3	4	5
14. It required the fewest steps possible to accomplish what I wanted to do	1	2	3	4	5
15. I did not notice any inconsistencies as I carried out the steps of ICA	1	2	3	4	5
PERCEIVED USEFULNESS					
16. I found using the ICA process useful in my task of analysing the social media data	1	2	3	4	5
17. Using the ICA process made it easier to help analyse online social media data	1	2	3	4	5
18. Using ICA process in analysing the social media content enabled me to accomplish tasks more quickly	1	2	3	4	5
19. ICA helped me be more productive in analysing the social media data	1	2	3	4	5
20. ICA met the task of the analysing the social media data	1	2	3	4	5
FLEXIBILITY					
21. Performing the ICA steps was not rigid	1	2	3	4	5
22. Users performed the ICA process according to their discretion	1	2	3	4	5
23. The ICA steps and tasks are dependent on the context of the text	1	2	3	4	5

24. The ICA steps were not flexible enough to be carried out simultaneously, one had to finish one step in-order to move to the next one	1	2	3	4	5
25. The ICA steps followed one after another in a chronological order making it easy and clear to carry out	1	2	3	4	5
SATISFACTION					
26. I am satisfied with the overall steps I carried out to fully complete the Inductive Content Analysis activity I participated in	1	2	3	4	5
27. I would recommend ICA to other researchers doing data analysis in their researchers	1	2	3	4	5
28. Overall, I am satisfied with the ease of completing the tasks in this Inductive Content Analysis activity	1	2	3	4	5
29. Overall, I am satisfied with the amount of time it took to complete the tasks in this activity	1	2	3	4	5
30. Overall, I am satisfied the output of the ICA I carried out.	1	2	3	4	5

APPENDIX E: SCREENSHOTS ILLUSTRATION OF DATA CLASSIFICATION THROUGH THE MODELS (DATA MINING ALGORITHMS IMPLEMENTED THROUGH WEKA)

a). Classification process through Naïve Bayes using the Naïve Bayes Multinomial classifier

The screenshot displays the WEKA software interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'NaiveBayesMultinomial'. Under 'Test options', 'Supplied test set' is selected. The 'Classifier output' pane shows the following results:

Summary

Metric	Value	Percentage
Correctly Classified Instances	154	82.7957 %
Incorrectly Classified Instances	32	17.2043 %
Kappa statistic	0.7926	
Mean absolute error	0.0434	
Root mean squared error	0.1874	
Relative absolute error	20.7558 %	
Root relative squared error	58.0465 %	
Total Number of Instances	186	

Detailed Accuracy By Class

IP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.875	0.019	0.875	0.875	0.875	0.856	0.967	0.889	corruptive-governance
0.842	0.027	0.889	0.842	0.865	0.832	0.970	0.921	In-competent-governance
0.892	0.020	0.917	0.892	0.904	0.881	0.991	0.968	Judiciary-agitation
0.905	0.069	0.792	0.905	0.844	0.798	0.980	0.930	Oppressive-governance
0.824	0.053	0.609	0.824	0.700	0.674	0.965	0.778	PartisanshipCronysin
0.571	0.017	0.727	0.571	0.640	0.620	0.967	0.736	EconomicPolitical-scandals
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Poor-education-system
0.000	0.000	0.000	0.000	0.000	0.000	0.917	0.232	Un-ConsultativesUn-transparent-governance
Weighted Avg.	0.828	0.034	0.809	0.828	0.815	0.787	0.975	0.882

Confusion Matrix

```

a b c d e f g h <-- classified as
21 0 0 0 2 1 0 0 1 a = corruptive-governance

```

The 'Result list' on the left shows a single entry: '11:55:44 - bayes.NaiveBayesMultinomial'.

Continued...

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize

Classifier

Choose
NaiveBayesMultinomial

Test options

☐ Use training set
☒ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) class-att

Start Stop

Result list (right-click for options)

11:55:44 - bayes.NaiveBayesMultinomial

Classifier output

```

Total Number of Instances      186

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.875   0.019   0.875     0.875   0.875     0.856   0.967    0.889    corruptive-governance
          0.842   0.027   0.889     0.842   0.865     0.832   0.970    0.921    In-competent-governance
          0.892   0.020   0.917     0.892   0.904     0.881   0.991    0.968    Judiciary-agitation
          0.905   0.069   0.792     0.905   0.844     0.798   0.980    0.930    Oppressive-governance
          0.824   0.053   0.609     0.824   0.700     0.674   0.965    0.778    Partisanship&Cronysin
          0.571   0.017   0.727     0.571   0.640     0.620   0.967    0.736    Economic&Political-scandals
          1.000   0.000   1.000     1.000   1.000     1.000   1.000    1.000    Poor-education-system
          0.000   0.000   0.000     0.000   0.000     0.000   0.917    0.232    Un-Consultative&Un-transparent-governance
Weighted Avg.   0.828   0.034   0.809     0.828   0.815     0.787   0.975    0.882

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  <-- classified as
21  0  0  0  2  1  0  0 | a = corruptive-governance
 0 32  1  1  3  1  0  0 | b = In-competent-governance
 0  0 33  4  0  0  0  0 | c = Judiciary-agitation
 0  1  2 38  0  1  0  0 | d = Oppressive-governance
 2  1  0  0 14  0  0  0 | e = Partisanship&Cronysin
 1  1  0  0  4  8  0  0 | f = Economic&Political-scandals
 0  0  0  0  0  0  8  0 | g = Poor-education-system
 0  1  0  5  0  0  0  0 | h = Un-Consultative&Un-transparent-governance

```

Status

b) Classification process through Support Vector Machine using SMO classifier

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"**

Test options

☐ Use training set
☒ Supplied test set **Set...**
☐ Cross-validation Folds **10**
☐ Percentage split % **66**
More options...

(Nom) class-att

Start **Stop**

Result list (right-click for options)

11:55:44 - bayes.NaiveBayesMultinomial
11:58:50 - functions.SMO

Classifier output

```

=== Summary ===
Correctly Classified Instances      176           94.6237 %
Incorrectly Classified Instances    10            5.3763 %
Kappa statistic                    0.9355
Mean absolute error                 0.1881
Root mean squared error             0.292
Relative absolute error             90.0071 %
Root relative squared error         90.4333 %
Total Number of Instances          186

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.958	0.000	1.000	0.958	0.979	0.976	0.999	0.987	corruptive-governance
	0.947	0.007	0.973	0.947	0.960	0.950	0.988	0.946	In-competent-governance
	1.000	0.013	0.949	1.000	0.974	0.967	0.993	0.949	Judiciary-agitation
	0.976	0.000	1.000	0.976	0.988	0.985	0.997	0.988	Oppressive-governance
	1.000	0.036	0.739	1.000	0.850	0.844	0.982	0.739	Partisanship&Cronysin
	0.643	0.006	0.900	0.643	0.750	0.745	0.985	0.785	Economic&Political-scandals
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Poor-education-system
	0.833	0.000	1.000	0.833	0.909	0.910	0.995	0.905	Un-Consultative&Un-transparent-governance
Weighted Avg.	0.946	0.008	0.953	0.946	0.946	0.940	0.992	0.931	

```

=== Confusion Matrix ===
 a  b  c  d  e  f  g  h  <-- classified as

```

Status

OK

Log  x 0

Continued...

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize

Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

Test options

☐ Use training set
☒ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) class-att

Start Stop

Result list (right-click for options)

11:55:44 - bayes.NaiveBayesMultinomial
11:58:50 - functions.SMO

Classifier output

```

Total Number of Instances      186

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.958    0.000    1.000    0.958    0.979    0.976    0.999    0.987    corruptive-governance
          0.947    0.007    0.973    0.947    0.960    0.950    0.988    0.946    In-competent-governance
          1.000    0.013    0.949    1.000    0.974    0.967    0.993    0.949    Judiciary-agitation
          0.976    0.000    1.000    0.976    0.988    0.985    0.997    0.988    Oppressive-governance
          1.000    0.036    0.739    1.000    0.850    0.844    0.982    0.739    Partisanship&Cronysin
          0.643    0.006    0.900    0.643    0.750    0.745    0.985    0.785    EconomicsPolitical-scandals
          1.000    0.000    1.000    1.000    1.000    1.000    1.000    1.000    Poor-education-system
          0.833    0.000    1.000    0.833    0.909    0.910    0.995    0.905    Un-Consultative&Un-transparent-governance
Weighted Avg.   0.946    0.008    0.953    0.946    0.946    0.940    0.992    0.931

=== Confusion Matrix ===
 a b c d e f g h <-- classified as
23 0 0 0 1 0 0 0 | a = corruptive-governance
 0 36 0 0 1 1 0 0 | b = In-competent-governance
 0 0 37 0 0 0 0 0 | c = Judiciary-agitation
 0 0 1 41 0 0 0 0 | d = Oppressive-governance
 0 0 0 0 17 0 0 0 | e = Partisanship&Cronysin
 0 1 0 0 4 9 0 0 | f = EconomicsPolitical-scandals
 0 0 0 0 0 0 8 0 | g = Poor-education-system
 0 0 1 0 0 0 0 5 | h = Un-Consultative&Un-transparent-governance

```

Status

OK
Log
 x 0

c) Classification process through K-Nearest-Neighbour using IBK classifier

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose `IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last"`

Test options

☐ Use training set

☒ Supplied test set Set...

☐ Cross-validation Folds

☐ Percentage split %

More options...

(Nom) class-att ▼

Start Stop

Result list (right-click for options)

- 11:55:44 - bayes.NaiveBayesMultinomial
- 11:58:50 - functions.SMO
- 12:02:59 - lazy.IBK

Classifier output

```

=== Summary ===
Correctly Classified Instances      176           94.6237 %
Incorrectly Classified Instances    10            5.3763 %
Kappa statistic                    0.9354
Mean absolute error                 0.0221
Root mean squared error             0.0843
Relative absolute error             10.5739 %
Root relative squared error         26.1147 %
Total Number of Instances          186

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.012	0.923	1.000	0.960	0.955	1.000	0.997	corruptive-governance
	0.974	0.014	0.949	0.974	0.961	0.951	0.999	0.995	In-competent-governance
	1.000	0.013	0.949	1.000	0.974	0.967	1.000	0.999	Judiciary-agitation
	0.976	0.000	1.000	0.976	0.988	0.985	1.000	0.999	Oppressive-governance
	0.882	0.018	0.833	0.882	0.857	0.843	0.995	0.938	Partisanship&Cronysin
	0.643	0.006	0.900	0.643	0.750	0.745	0.992	0.883	Economic&Political-scandals
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Poor-education-system
	0.833	0.000	1.000	0.833	0.909	0.910	0.999	0.958	Un-Consultative&Un-transparent-governance
Weighted Avg.	0.946	0.009	0.947	0.946	0.944	0.938	0.999	0.982	

```

=== Confusion Matrix ===
 a  b  c  d  e  f  g  h  <-- classified as

```

Status

OK Log  x 0

Continued...

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A 1"weka.core.EuclideanDistance -R first-last""

Test options

☐ Use training set
 ☒ Supplied test set

Set...

☐ Cross-validation

Folds

10

☐ Percentage split

%

66

More options...

(Nom) class-att

Start

Stop

Result list (right-click for options)

11:55:44 - bayes.NaiveBayesMultinomial

11:58:50 - functions.SMO

12:02:59 - lazy.IBk

Classifier output

total number of instances

186

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.012	0.923	1.000	0.960	0.955	1.000	0.997	corruptive-governance
	0.974	0.014	0.949	0.974	0.961	0.951	0.999	0.995	In-competent-governance
	1.000	0.013	0.949	1.000	0.974	0.967	1.000	0.999	Judiciary-agitation
	0.976	0.000	1.000	0.976	0.988	0.985	1.000	0.999	Oppressive-governance
	0.882	0.018	0.833	0.882	0.857	0.843	0.995	0.938	Partisanship&Cronysin
	0.643	0.006	0.900	0.643	0.750	0.745	0.992	0.883	Economics&Political-scandals
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Poor-education-system
	0.833	0.000	1.000	0.833	0.909	0.910	0.999	0.958	Un-Consultative&Un-transparent-governance
Weighted Avg.	0.946	0.009	0.947	0.946	0.944	0.938	0.999	0.982	

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	<-- classified as
24	0	0	0	0	0	0	0	0	a = corruptive-governance
0	37	0	0	0	1	0	0	0	b = In-competent-governance
0	0	37	0	0	0	0	0	0	c = Judiciary-agitation
0	0	1	41	0	0	0	0	0	d = Oppressive-governance
1	1	0	0	15	0	0	0	0	e = Partisanship&Cronysin
1	1	0	0	3	9	0	0	0	f = Economics&Political-scandals
0	0	0	0	0	0	8	0	0	g = Poor-education-system
0	0	1	0	0	0	0	5	0	h = Un-Consultative&Un-transparent-governance

Status

OK

Log

x 0

d) Classification process through C4.5 Decision Tree using J48 classifier

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

☐ Use training set

☒ Supplied test set

☐ Cross-validation Folds 10

☐ Percentage split % 66

(Nom) class-att

Result list (right-click for options)

- 11:55:44 - bayes.NaiveBayesMultinomial
- 11:58:50 - functions.SMO
- 12:02:59 - lazy.IBk
- 12:45:01 - trees.J48

Classifier output

```

=== Summary ===
Correctly Classified Instances      145           77.957 %
Incorrectly Classified Instances    41           22.043 %
Kappa statistic                    0.7359
Mean absolute error                 0.0737
Root mean squared error             0.1919
Relative absolute error             35.2571 %
Root relative squared error         59.4352 %
Total Number of Instances          186

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.917	0.043	0.759	0.917	0.830	0.807	0.981	0.884	corruptive-governance
	0.789	0.027	0.882	0.789	0.833	0.795	0.967	0.904	In-competent-governance
	0.757	0.054	0.778	0.757	0.767	0.710	0.959	0.855	Judiciary-agitation
	0.857	0.056	0.818	0.857	0.837	0.789	0.962	0.800	Oppressive-governance
	0.882	0.041	0.682	0.882	0.769	0.750	0.983	0.781	Partisanship&Cronysin
	0.214	0.012	0.600	0.214	0.316	0.331	0.940	0.486	Economic&Political-scandals
	0.875	0.011	0.778	0.875	0.824	0.817	0.994	0.792	Poor-education-system
	0.667	0.017	0.571	0.667	0.615	0.603	0.988	0.583	Un-Consultative&Un-transparent-governance
Weighted Avg.	0.780	0.040	0.777	0.780	0.768	0.734	0.967	0.810	

```

=== Confusion Matrix ===
a b c d e f g h <-- classified as

```

Status

OK x0

Continued...

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize

Classifier

Choose
J48 -C 0.25 -M 2

Test options

☐ Use training set
☒ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) class-att

Start Stop

Result list (right-click for options)

11:55:44 - bayes.NaiveBayesMultinomial
11:58:50 - functions.SMO
12:02:59 - lazy.IBk
12:45:01 - trees.J48

Classifier output

```

Total Number of Instances      186

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.917    0.043    0.759    0.917    0.830      0.807    0.981    0.884    corruptive-governance
      0.789    0.027    0.882    0.789    0.833      0.795    0.967    0.904    In-competent-governance
      0.757    0.054    0.778    0.757    0.767      0.710    0.959    0.855    Judiciary-agitation
      0.857    0.056    0.818    0.857    0.837      0.789    0.962    0.800    Oppressive-governance
      0.882    0.041    0.682    0.882    0.769      0.750    0.983    0.781    Partisanship&Cronysin
      0.214    0.012    0.600    0.214    0.316      0.331    0.940    0.486    EconomicsPolitical-scandals
      0.875    0.011    0.778    0.875    0.824      0.817    0.994    0.792    Poor-education-system
      0.667    0.017    0.571    0.667    0.615      0.603    0.988    0.583    Un-Consultative&Un-transparent-governance
Weighted Avg.  0.780    0.040    0.777    0.780    0.768      0.734    0.967    0.810

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  <-- classified as
22  0  0  1  1  0  0  0 | a = corruptive-governance
 0 30  2  2  2  1  1  0 | b = In-competent-governance
 3  1 28  4  0  0  0  1 | c = Judiciary-agitation
 0  0  4 36  0  0  0  2 | d = Oppressive-governance
 1  0  1  0 15  0  0  0 | e = Partisanship&Cronysin
 3  2  0  1  4  3  1  0 | f = EconomicsPolitical-scandals
 0  1  0  0  0  0  7  0 | g = Poor-education-system
 0  0  1  0  0  1  0  4 | h = Un-Consultative&Un-transparent-governance

```

Status

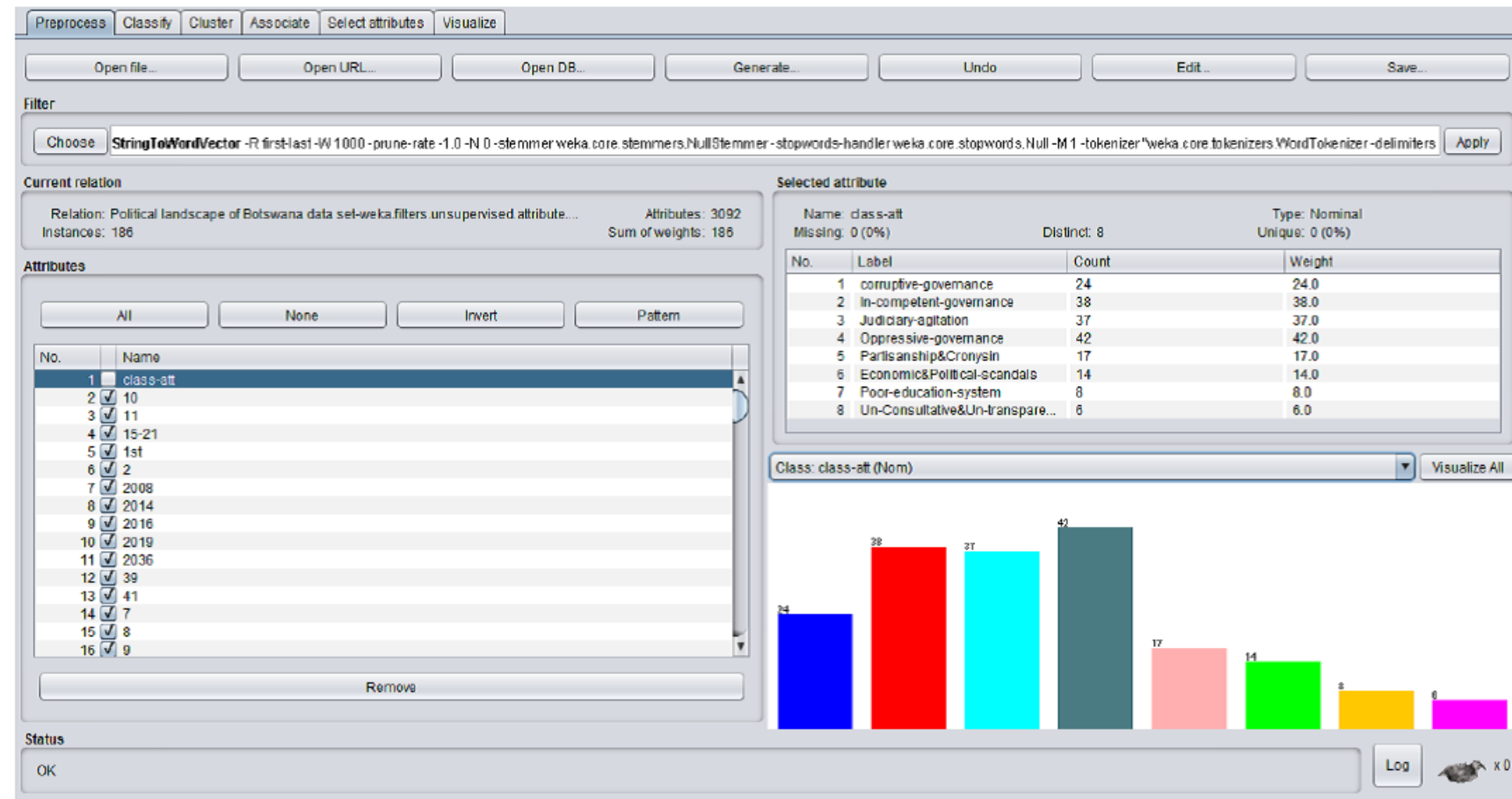
OK

Log

x 0

APPENDIX F: SCREENSHOTS ILLUSTRATION SOME OF THE TEXT PRE-PROCESSING FUNCTIONS APPLIED ON THE DATA FOR MODEL CLASSIFICATION

Figure 1: Illustration of the tokenization, lemmatization and stemming processes applied through the StringToWordVector Function



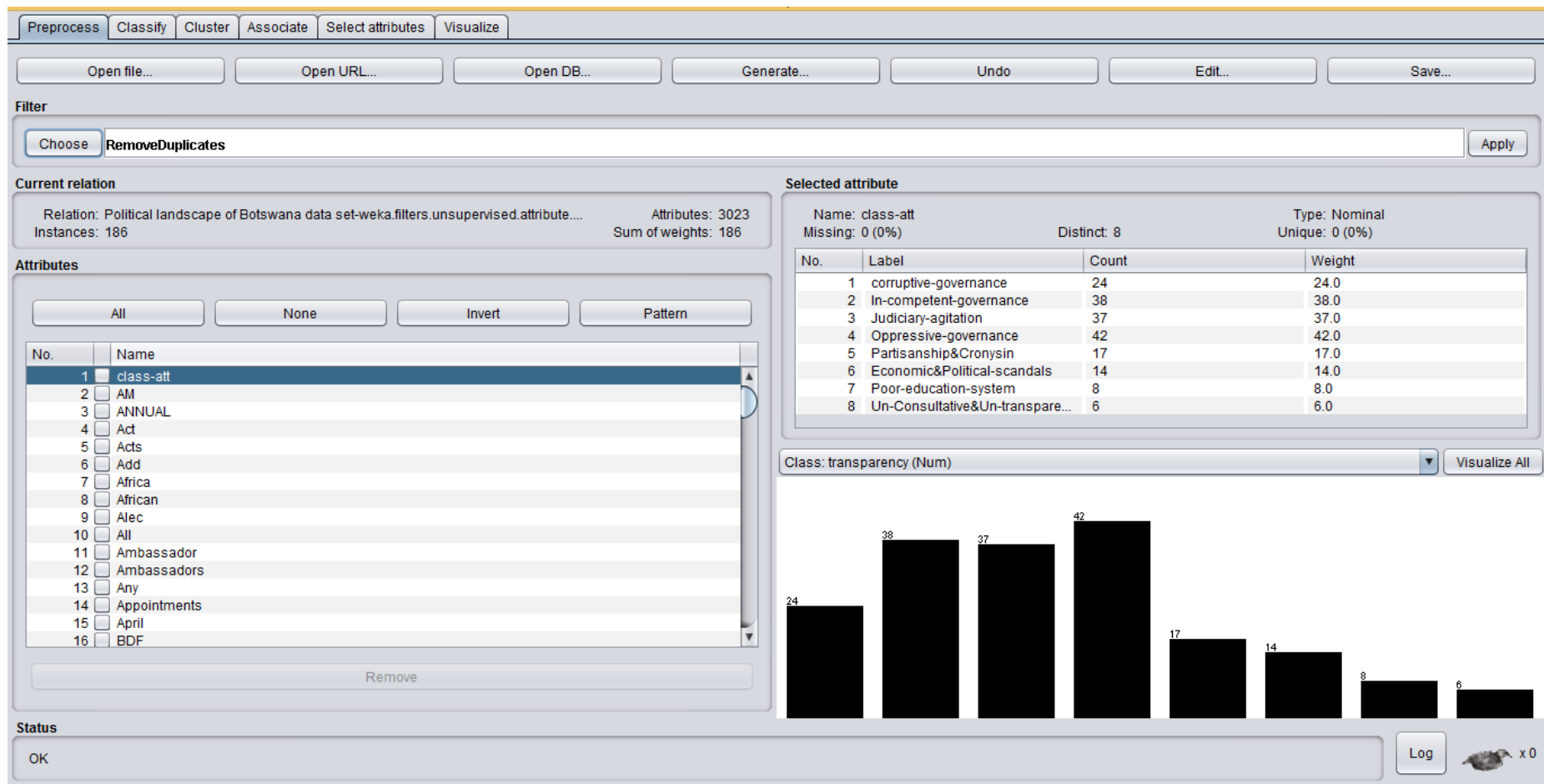


Figure 2: Illustration of removal of Duplicates and inconsistencies in the data set through application of RemoveDuplicates Function

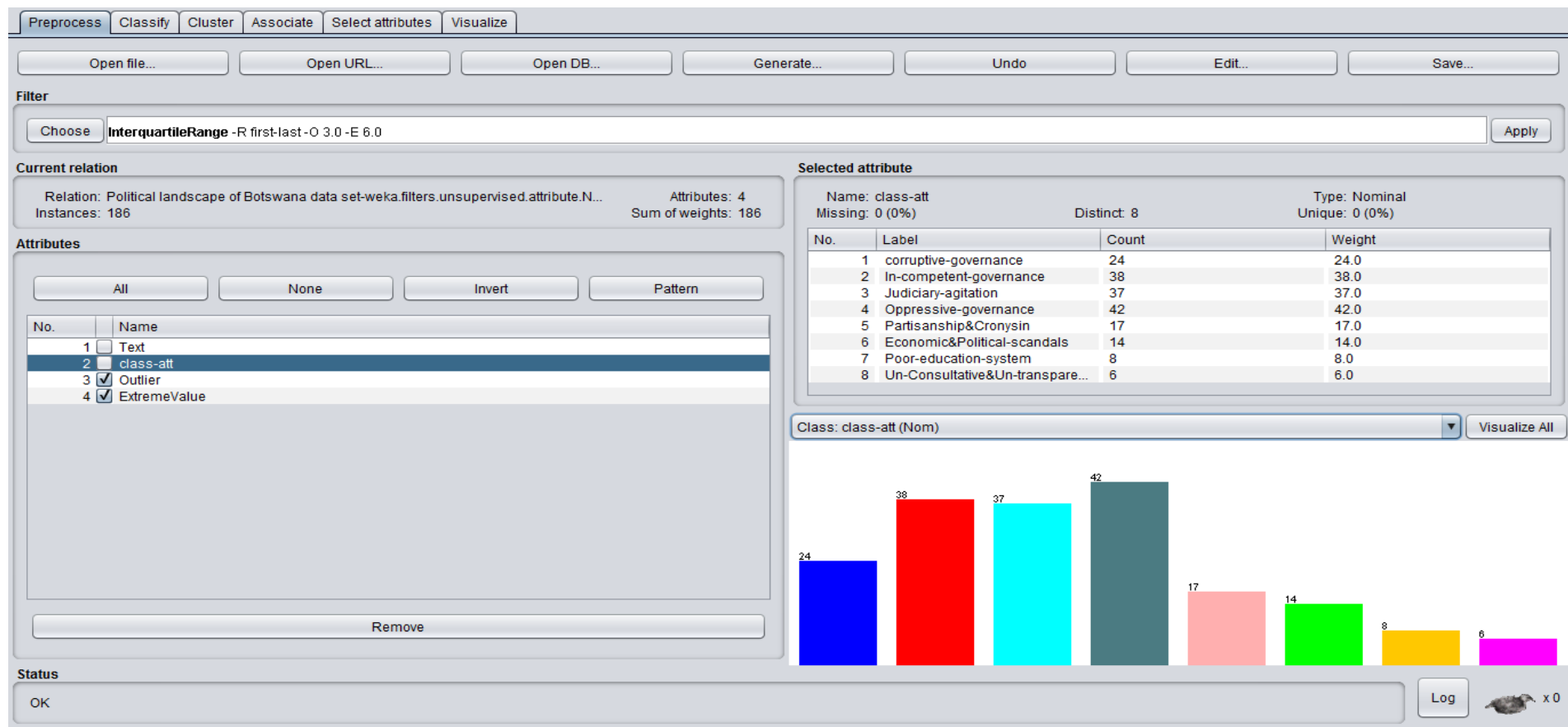


Figure 3: An illustration of dealing with outliers and extreme values through an InterquartileRange function

Continued...

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Filter

Choose

InterquartileRange -R first-last -O 3.0 -E 6.0

Apply

Current relation

Relation: Political landscape of Botswana data set-weka.filters.unsupervised.attribute.N...

Instances: 186

Attributes: 4

Sum of weights: 186

Attributes

All

None

Invert

Pattern

No.	Name
1	☐ Text
2	☐ class-att
3	☒ Outlier
4	☒ ExtremeValue

Remove

Selected attribute

Name: class-att

Missing: 0 (0%)

Distinct: 8

Type: Nominal

Unique: 0 (0%)

No.	Label	Count	Weight
1	corruptive-governance	24	24.0
2	In-competent-governance	38	38.0
3	Judiciary-agitation	37	37.0
4	Oppressive-governance	42	42.0
5	Partisanship&Cronysin	17	17.0
6	Economic&Political-scandals	14	14.0
7	Poor-education-system	8	8.0
8	Un-Consultative&Un-transpare...	6	6.0

Class: class-att (Nom)

Visualize All

Label	Count
corruptive-governance	24
In-competent-governance	38
Judiciary-agitation	37
Oppressive-governance	42
Partisanship&Cronysin	17
Economic&Political-scandals	14
Poor-education-system	8
Un-Consultative&Un-transpare...	6

Status

OK

Log

x 0

APPENDIX G: A JAVA CODE BUILT WITH TWITTER API FOR COLLECTION OF DATA FROM TWITTER

```
/*
 */

package twitterdata;

import java.util.List;
import twitter4j.*;
import twitter4j.conf.ConfigurationBuilder;
import java.io.*;

//import java.text.SimpleDateFormat;
//import java.util.Date;
import java.util.logging.Level;

/**
 *
 * @author
 */

public class TwitterData {

    /**
     * @param args the command line arguments
     */

    public static void main(String[] args)throws IOException {

        TwitterData obj = new TwitterData();

        ConfigurationBuilder cb = new ConfigurationBuilder();

        cb.setDebugEnabled(true)

        .setOAuthConsumerKey("Guz4mnHNpJb5USSAI0tch7uL3")

        .setOAuthConsumerSecret("aou57hd3BCA9RSfUleP8bWmApGeYiMiJ7IPAv5t2luH5nU97J")

        .setOAuthAccessToken("264497942-SQZ2X4Npju2LEGBAcAdvezHSg0hvtQeKQZoNkT3")

        .setOAuthAccessTokenSecret("P3nGTISqF1jgOiw5HY4fjSIWewd4knMIjV16anMNSEqkR");

        System.out.println("-----Started-----");
```

```

Twitter twitter = new TwitterFactory(cb.build()).getInstance();
try {
    Query q = new Query(query);
    //Date date = new Date();
    //String modifiedDate= new SimpleDateFormat("20140101").format(date);
    //q.setSince(modifiedDate);
    QueryResult result;
    do {
        result = twitter.search(q);
        List<Status> tweets = result.getTweets();
        for (Status tweet : tweets) {
            obj.writeToFile(tweet.getUser().getScreenName(),tweet.getText());
        }
    } while ((q = result.nextQuery()) != null);
    System.out.println("-----Done-----");
} catch (TwitterException e) {
    System.out.println(e.getMessage());
}
}

public void writeToFile(String user,String data)
{
    BufferedWriter output = null;
    try {
        File file = new File("Bushdee.csv");
        output = new BufferedWriter(new FileWriter(file,true));
        output.write(user+","+data+"\n");
    } catch ( IOException e ) {
        e.printStackTrace();
    } finally {

```

```
if ( output != null ) {  
    try {  
        output.close();  
    } catch (IOException ex) {  
        java.util.logging.Logger.getLogger (TwitterData.class.getName()).log(Level.SEVERE, null,  
ex);  
    }  
}  
}  
}  
}
```