International Journal of Database Theory and Application
Vol. 10, No.10 (2017), pp. 11-34
hhtp://dx.doi.org/10.14257/ijdta.2017.10.10.02

A Conceptual Framework for the Mining and Analysis of the Social Media Data

Sethunya R Joseph[1*], Keletso Letsholo[2] and Hlomani Hlomani[3]

[1,2,3] *Computer Science Department, Botswana International University of Science and Technology, Palapye, Botswana*
[1]*Sethunya.joseph@studentmail.biust.ac.bw*, [2]*letsholok@biust.ac.bw*
[3]*hlomanihb@biust.ac.bw*

Abstract

Social media data possess the characteristics of Big Data such as volume, veracity, velocity, variability and value. These characteristics make its analysis a bit more challenging than conventional data. Manual analysis approaches are unable to cope with the fast pace at which data is being generated. Processing data manually is also time consuming and requires a lot of effort as compared to using computational methods. However, computational analysis methods usually cannot capture in-depth meanings (semantics) within data. On their individual capacity, each approach is insufficient. As a solution, we propose a Conceptual Framework, which integrates both the traditional approaches and computational approaches to the mining and analysis of social media data. This allows us to leverage the strengths of traditional content analysis, with its regular meticulousness and relative understanding, whilst exploiting the extensive capacity of Big Data analytics and accuracy of computational methods. 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. Additionally, a questionnaire was conducted to evaluate the usability of ICA as perceived by the participants. Secondly, an experimental study was conducted to evaluate the performance of data mining algorithms on the data from the ICA process. The results, from the user study, showed that the ICA process is flexible and systematic in terms of allowing the users to analyse social media data, hence reducing the time and effort required to manually analyse data. The users' perception in terms of 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 produced higher accurate results in classifying data when supplied with data from the ICA process. That is, when data mining algorithms are integrated with the ICA process, they are able to overcome the difficulty they face to capture 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 the social media data in a systematic manner.

Keywords*: Social media, Big Data, Inductive Content Analysis, Conceptual Framework*

Received (May 5, 2017), Review Result (September 28, 2017), Accepted (October 10, 2017)

ISSN: 2005-4270 IJDTA

1. Introduction

Gundecha and Liu [1] portray the social media as a combination of diverse kinds of shared media which incorporate the traditional media (e.g., daily papers, radio and TV) and non-traditional media (e.g., Twitter, Facebook, LinkedIn, and YouTube). The networking sites such as Facebook, Twitter, and YouTube are the most used sites for all age groups [2]. Other common social media platforms which are normally used are WhatsApp and Skype. According to Gundecha and Liu [1], since December 2011, Facebook alone has registered over 845 million active users. At present, Twitter and Facebook accumulated over 1.2 billion clientele [3]. Unlike in the traditional media, Social media give its users an easy way to convey and connect amongst themselves in a convenient way [1]. It permits the creation, dealings and exchanges of user made content [4]. The users of the online networking platforms typically are the ones who share, arrange groups and give helpful information [2]. That is to say, on different online networking sites, users examine and share their ordinary experiences. Nowadays, a larger part of individuals rely upon online networking to impart, express their sentiments, share data and encounters, vent feelings and stress, look for social support and so on [2], [5]. Social media platforms permit different activities, for example, web based shopping, gaming, business showcasing, advancements and commercial, stimulation, entertainment, politics, news exchange and analysis, health discussion and awareness forums [6]. Numerous researchers have a similar view that, the acceptance of social media continually grows at a faster rate, bringing about an advancement of informal communities, web journals, social news, smaller scale online journals, area based interpersonal organizations, social bookmarking applications, media content sharing (photograph, sound and video), business survey destinations, wikis, and so forth [1]; [2]; [6]; [7].

Traditional research methods of empirical research face a huge challenge with this era of Big Data [61, 62, 63]. On daily basis, researchers are challenged with massive amounts of data and encounters serious limitations to establish methods of traditional content analysis [8]. The researchers encounter variety of information at an increasing velocity, volume, and with different value as well as veracity which constitutes the so called Vs of the Big Data [60]. The information which is accessible on social networking sites is noisy, distributed and dynamic [6]. Data is generally formless and presented in a diversity of ways (*e.g.*, blog posts or tweets, web pages). It is gradually not easy to make use of the well-known methods of manual content analysis alone for this kind of web data, because of the speed and size this data grows at [8]. Evans [9] points out that even though the conservative qualitative methods for analysing text are valued, they require much more effort and are labour-intensive. One of their limitations is failure to scale well when presented with massive volume of texts. Take for instance, manual content coding and close reading, they are inadequate on their own for analysing the large textual datasets that are emerging such as newspapers, journal articles, books archives, media transcripts assuming digital form as well as social media data [9]. Some researchers overcome this challenge by using computational methods. However, pure computational automated methods usually cannot capture in-depth meanings within data since content from the social media requires human interpretation and intervention [5]. Large samples of data can be easily analysed through the use of the computational methods in order to get the contrast data essential to assess the impact of certain cases on the communal discourse on a broader manner. However, the computational and algorithmic methods of analysis in their capacity remain limited in understanding the hidden meanings or subtleties of human language [52], [53]. At times, carrying out text mining can be so helpful in identifying relevant text documents from large data records and depositories in order to carry out an in-depth manual content analysis [8].

In summary, it is very difficult and challenging to mine and analyse social media data using either traditional methods alone or computational automated methods on their

individual capacity [59]. This is because there is a challenge of analysing data, and also availability of tools and guides that can help achieve this task. These tools would include Frameworks, models and workflows which have been developed explicitly to analyse online social media data.

2. Related Works

Various studies have been carried out on mining and analysis of the social media content from Twitter and Facebook sites. However, content which had been broadly analysed or used in these studies mostly comes from Twitter. Gaffney [10] carried out a study on Iran elections by analysing the social media data accessible on Twitter. In his analysis, Gaffney made use of user grids, histograms, and frequencies of top keywords to quantify online activism. Another study in which social media content was utilized, particularly Twitter information, is the study on the analysis of political sentiments of Germans federal elections which was conducted by Tumasjan [11].

Yang and Counts [12] carried out a study of network analyses of information diffusion on Twitter. They captured the three main attributes of information dissemination: scale, speed, and range. They found out that certain attributes of the tweets can forecast more information broadcast. They also established that attributes of the users such as the frequency at with which a user is cited generally in a particular tweet, tend to be a strong predictor.

Ediger *et al.*, [13] used and analysed Twitter data through the use of Cray XMT, a graph characterization toolkit for huge graphs demonstrating social networks data. They used the GraphCT to examine and process data from Twitter. Since Twitter's messages networks and nodes appears mainly as tree-structured as news dissemination systems, within the online data as clusters of conversations, they used the GraphCT to do ranking of actors within these conversations and assist analysts to pay attention to considerable reduced data subsets.

Sakati *et al.*, [14] designed a system which reported earthquakes which were occurring in Japan based on the Japanese tweets posted on Twitter. According to them, Twitter has numerous users who are geographically dispersed throughout Japan. Also, Japan has numerous earthquakes hence it is best to detect the earthquake by monitoring the tweets. Their system detected an earthquake occurrence and sent an email, probably before a probable location was hit by an earthquake.

Hong *et al.*, [15] points out that, in social media such as Twitter, information which is considered vital by the public normally spreads through re-tweets. By examining the characteristics of such widespread messages, this can help in important numerous tasks (*e.g.*, breaking news detection, personalized messages recommendations, viral marketing and others). As such, Hong *et al.*, [15] carried out a study to investigate how the popularity of messages could be predicted through the number of future re-tweets. They also brought in some understanding on what factors lead to information dissemination in Twitter.

A few studies have been done on mining the social media data on healthcare *e.g.*, Jamison-Powel et al. [16] did a study by mining and analysing data from Twitter to discover how individuals are utilizing social media to deliberate on their mental health conditions, in particular reference to insomnia. Social media content has also been mined and analysed to promote marketing and businesses [17]. Cheong and Cheong [18] mined and analysed social media data from Tweeter with regards to the 2010-2011 Australian floods. Social media data mining has also been applied to sporting activities such as athletics *e.g.*, Frederick *et al.*, [19] carried out an Internet survey on Tweeter and Facebook to decide the sort of adherents the competitors have. Social media data mining has been used to analyse the sports and athletics events and activities [20]; [21]; [22].

Another study where social media data have been used has been done by Bandari *et al.*, [23] by predicting the social popularity of news articles without using early popularity measurement. They considered features of a news article prior to its publication in order to discover if any predictors relevant only to the content exist. They wanted to establish if it is possible to make reasonable forecast of the spread of an article based on the content features. Their data of study was collected from Feedzilla and the measurements of the spread were performed on Twitter. Chen *et al.*, [5] did a study by mining data from Twitter to comprehend the students' learning experiences and encounters at Purdue University who are doing engineering courses.

A limitation or gap which is observed in these previous studies is that in almost all of these discussed studies, it is clear that data was just analysed. There were no standard procedures or guides such as Frameworks, models and workflows which guided the mining and analysis of the social media data in a systematic manner. Also, most of these studies used only one approach (computational methods) of which if they could have blended the approaches (both traditional and computational) may be their results could have been more improved or better than they are. This view is shared by researchers such as Lewi and Saunders [59], that an approach blending computational and manual methods could yield fruitful results. Hence, the main aim of this study is to develop a Conceptual Framework which integrates both the traditional approaches (methods) with the computational approaches to mine and analyse the social media data in a systematic manner.

3. The Proposed Conceptual Framework

The proposed Framework shown in Figure 1 shows three iterative phases. The Framework phases begin with data collection, trailed by text pre-processing and the data analysis phases. In between the phases, data is passed to the storage after collection and to knowledge repository after analysis. The large solid arrows show connections between the phases denoting process flow whilst the dotted arrows show the data flow. All these components and phases which make up the entire Framework are clarified and elaborated more on sections which follow.

3.1. Data Collection Phase

Data collection phase involves gathering data from different online social media sites. The proposed Conceptual Framework demonstrates that data can be collected through the use of traditional methods and computational methods. Some of examples of these traditional methods of gathering data include: surveys, interviews, questionnaires, checklist, observations, focus groups, etc. These methods can also be automated, instead of being paper based. Computational methods incorporate computerized tools, algorithms and programs. These are implemented to assist with data collection from online platforms. These methods are computerized and automatic.

3.2. Text Pre-processing Phase

The second phase of the Framework is the text pre-processing. The data which has been collected online and kept in data storage is actually processed at this phase. Before the traditional data analysis could take place, data is examined to see if it is in a better state of elucidation and better understanding. Data is properly transcribed and transformed into written text before analysis can begin [24]. Data is properly cleaned by: noisy text removal, corrections of abbreviations, acronyms and short language. This is done because it is said that pre-processing affects the quality of classification [25]. The pre-processing of text is done during the initial Qualitative analysis process which is the traditional approach. The other text pre-processing is also conducted just before model

implementation in data modelling stage and it includes word parsing and tokenization, attribute selection, feature extraction, lemmatization and stemming.

3.3. Data Analysis Stage

The last phase of the Framework is the data analysis phase. This is where data is processed to make it meaningful and comprehensible. Processed information normally helps in making good informed decisions. This phase consists of the qualitative and quantitative analysis approaches. The qualitative analysis is the traditional approach (manual analysis) and the quantitative analysis is the computational approach.

3.3.1. Traditional Approach

Content from social media needs to be explored to get patterns, meanings and make some understandings of what the online users are trying to impart or share. Qualitative content analysis is a process intended to deduce categories or themes from raw data relying on reliable effective interpretation as well as inference [26]. These can be achieved through Inductive Content Analysis or Deductive Content Analysis approaches as shown in the proposed Framework as shown in Figure 1.

Qualitative analysis is performed on the data mainly because data gathered from social media is much unstructured, ambiguous, full of sarcasm, acronyms and mis-spellings hence it requires human interpretation [5]. When dealing with huge amounts of data gathered online, assumptions which are faulty are likely to occur in the analysis of data, if automatic algorithms are used without qualitatively inspecting the data [5].

3.3.2. Computational Approach

Other than analysing data quantitatively, quantitative data analysis can be carried out to get to understand results easily [27]. Even though data could be analysed quantitatively in a traditional approach, in the proposed Conceptual Framework the quantitative analysis is mainly the computational approach. As shown in the Figure 1, quantitative data analysis can be performed through various methods or procedures. Some of these procedures which are shown in the proposed Conceptual framework include: *Data tabulation* (frequency distributions & percentage distributions, *Data disaggregation* (data disaggregation across different variables and subcategories of variables), *Data descriptives* (Mean, Minimum and maximum values, Median, Mode) and Moderate *and advanced analytical methods* (*e.g.*, data modelling -applying machine learning and data mining algorithms). The modelling sub-phase is basically about the use of data mining algorithms to uncover patterns in the data. This is characterized by several steps such model implementation, model training and testing. After the model is said to be ready and working well it is deployed to a large scale data to analyse it. The models could be implemented through supervised or unsupervised algorithms to perform some analysis of the data. This analysis could either be classification, regression, clustering etc.

3.4. Data Storage

The proposed Conceptual Framework has data storage and knowledge discovery repository components. Data storage is where data is kept and saved immediately after it has been collected before it can be processed and analysed. On the other hand, the knowledge repository acts as storage of information which had been processed and analysed for further/future use (*e.g.*, online databases). Some of the data from the repository could be passed to data storage as shown in Figure 1. The data which is normally passed to the data storage is the one which will be re-used for various analyses of different aspects of research *e.g.*, fraud detection, political landscapes, *etc*.

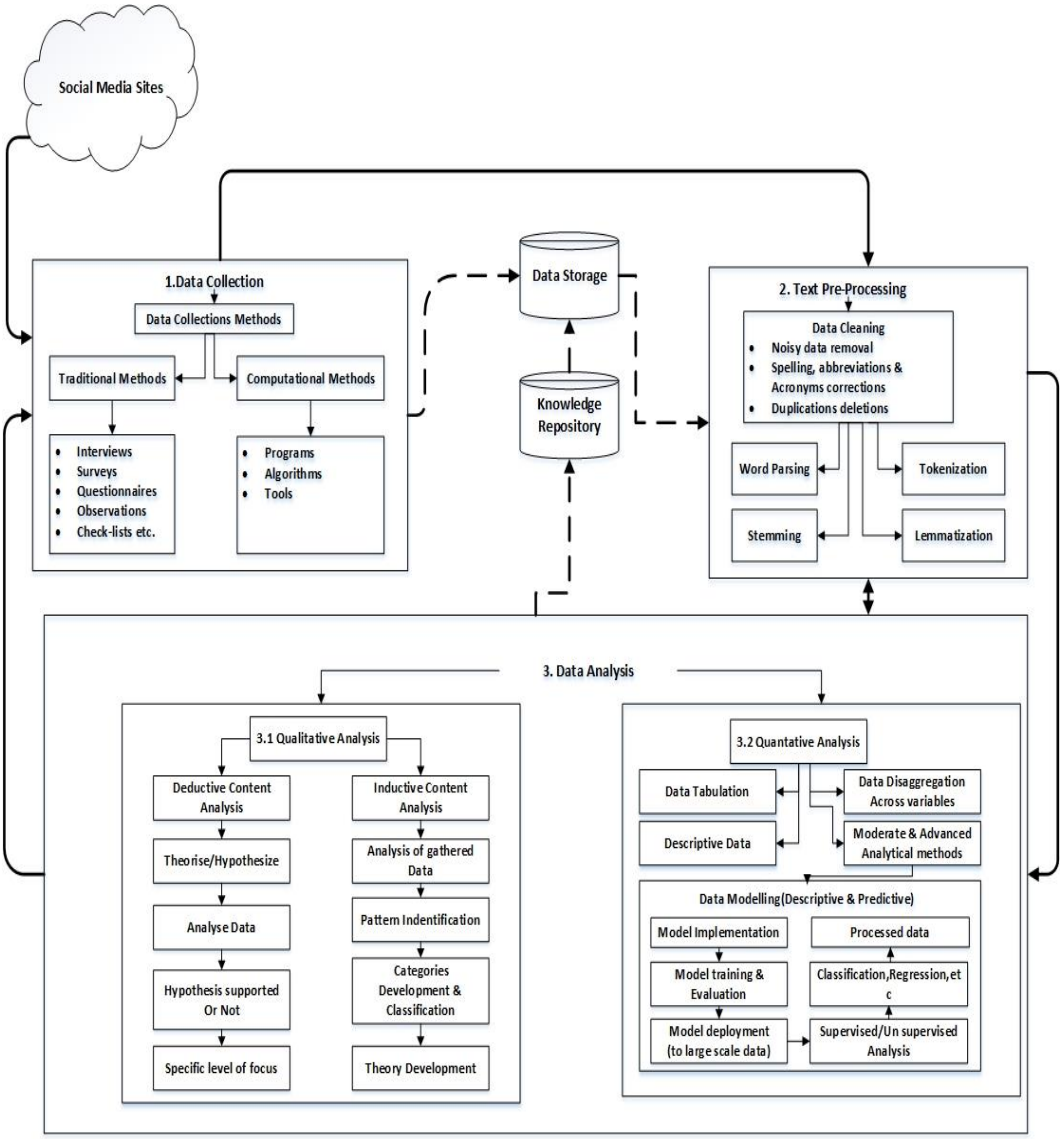

Figure 1. A Detailed View of the Proposed Conceptual Framework for Data Mining and Analysis of Social Media Data

4. Evaluation of the Proposed Conceptual Framework

The proposed Conceptual Framework has been evaluated through classifying data by integrating one of the traditional qualitative methods or approaches of data analysis: Inductive Content Analysis (ICA), together with some of the automatic computational algorithms for data mining (Naïve Bayes multi-nomial, K-Nearest Neighbour, Support Vector Machine and C4.5 decision tree). The evaluation was twofold: ICA has been carried out and upon its completion it was evaluated with a usability questionnaire guided by the ISO 9126 model. The data mining algorithms were implemented and also evaluated through the use of metric measuring accuracy and performance. The kind of analysis which had been performed in the data is of classification. Classification is one of the examples of ways in which data can be analysed. As depicted in the Framework in Figure 1, there are several techniques which can be applied to a problem in data mining such as clustering, association rule mining and prediction (which can be divided into classification and regression). This evaluation involved gathering data online and cleaning the data (pre-processing). These are the first two stages depicted in the Framework shown

in Figure 1. The evaluation has blended the components or processes in the third stage of data analysis which are Inductive Content Analysis and data modelling (use of data mining algorithms). All these tasks are explained and discussed in the sections that follow.

4.1. 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. Data from Facebook was gathered using an educational account on a social media monitoring tool called Netlytic (https://netlytic.org/) and a user ID generating tool site called lookup-id.com (https://lookup-id.com/). 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 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 but more content is shared by individual users online. 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.

4.2. Data Pre-Processing

Text pre-processing was performed on the data by cross checking if the data sets which had been collected online were in a proper condition of interpretation and a better state of understanding. Data was cleaned and interpreted justifiably. Translations and amendments were made for truncations, acronyms, short dialect (slang) and noisy messages were deleted, spelling slip-ups were rectified, and rehashing words were removed. This was done in light of the fact that it is said that pre-processing operations affect the quality of the classification. This pre-processing was performed before carrying out the Inductive Content Analysis process. The second text pre-processing was carried out to the data before it could be classified through computational methods (classification data modelling algorithms). This involved performing the following tasks to the data: word parsing and tokenization, attribute selection, feature extraction, lemmatization and stemming.

4.3. Inductive Content Analysis (ICA)

Inductive Content analysis was done by four focus groups (twenty four second year Computer Science and Information Systems students and two post graduates Computer Science students). The guidelines which were utilized amid the process of Inductive Content Analysis were informed by various qualitative analysis research ideas especially on the Inductive Content Analysis subject (*e.g.*, [25]; [48]; [49]). Inductive thinking was applied in carrying out the analysis, in-order to allow the themes and categories to emerge from the data by the focus groups and the researchers' careful examination and constant comparison. During this process, the researchers and focus groups members immersed themselves in the pre-processed data and allowed the themes to emerge directly from it.

In this Inductive Content Analysis process, the unit of analysis was defined as single words and some short expressions. Each sentence/post/tweet was viewed as a code fragment. Each code portion was comprised of a few words. 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 [25] described as "links" and as an assumed rule in qualitative analysis, that some text may be coded and classified into more than one category. All irrelevant data which

was not related to political landscape was discarded. This obviously verified what Thomas [25] hinted in his qualitative analysis steps, that among the normally assumed standards and rules that underlie qualitative coding, a lot of content may not be assigned to any class, as a great part of the content might not be important to the research objectives.

Each focus group carried out the steps of the Inductive Content Analysis in order to come up with categories and then manually analyse and classify data into those categories. Although several themes emerged from the data which led to the creation of a number of categories during the Inductive Content Analysis phase by each focus group, only eight prominent categories were finally selected. This came from the Qualitative Inductive Analysis rule that a decent practice ought to end up with a range of three (3) to eight (8) final prominent categories [25]; [28]. The general rule is that an inductive coding which ends winds up with more than eight categories, its classifications may appear as inadequate and incomplete. If this happens, the researcher (coder) needs to combine some categories which have a link or have to make hard decision with regards to the themes and categories which are important and should remain and those which should be discarded.

The finalized eight prominent categories which emerged from the data 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. 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.

4.3.1. Evaluation of the Inductive Content Analysis with a Usability questionnaire

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 [29]; [30]; [31]. The ISO 9126 model is commonly used to evaluate software and systems. 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 reason for adopting usability evaluation was, to the best of our knowledge, we have not come across any metrics which evaluates the ICA process, including even frameworks or models which analyse the social media data which could be applicable to our proposed Conceptual Framework or similar Frameworks.

The usability testing questionnaires which formed our sources are: Questionnaire for User Interface (QUIS) by Chin *et al.*, [32], Perceived Usefulness and Ease of Use (PUEU) by Davis [33] Nielsen' Attributes of Usability (NAU) by Nielsen [34] Nielsen' heuristic evaluation (NHE) by Nielsen [35], USE Questionnaire: Usefulness, Satisfaction and Ease of use (USE) by Lund [36], After Scenario Questionnaire (ASQ) by Lewis [37], Practical Heuristics for Usability Evaluation (PHUE) by Perlman[38] and Purdue Usability Testing questionnaire (PUTQ) by Lin et al.[39].

We targeted questions that addressed learnability, perceived ease of use, perceived usefulness, flexibility and satisfaction which 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 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 gives an impression that when conducting the ICA, it is possible to apply or carry it out across any discipline, like on 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 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 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 learnability results shown in Table 2 of the five asked questions, when summarized and explained in terms of the strongly agree and agree values totals (i.e. the positive perception of the participants) ranges from 91.7 % to 100%. This clearly shows a positive perception on the aspect of learnability.

The results shown in Table 3 on the perceived ease of Use by the participants, when summed up in terms of positive perception, the totals for the strongly agree and agree values ranges from 95.9% to 100% with exception to Q1 and Q4. These two questions asked about the negative perception with regards to the ease of use of the ICA, and majority strongly disagreed and disagreed. Their negation shows a positive perception towards the ease of use of the ICA.

As for the results shown in Table 4, the totals for the strongly agree and agree values for all the questions ranges from 91.7% to 100%. This shows that the participants have a positive perception and find the ICA very useful.

The results shown on Table 5 on the perception about the flexibility of ICA show a positive response with the totals of the strongly agree and agree values ranging from 95.8% to 100%. This is a positive response. However, Q4 stated that ICA is not flexible enough and majority of the respondents gave a neutral response, with a few disagreeing with a 4.2%.

The results shown in Table 6 with regards to the satisfaction of using the ICA show that the participants were mostly satisfied. The totals of the strongly agree and agree values ranged from 95.8 % to 100%. This is a positive impression.

Table 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**

Table 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. 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**

Table 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, very useful, flexible enough to be applied to analyse the social media data. It is also easy to learn. This could be applied generally to the proposed Conceptual Framework since ICA is one of its components. Therefore, we could say that since ICA is part of the Framework, 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.

4.4. Data Modelling and Evaluation Metrics

After completing the ICA process, the second evaluation of the proposed Conceptual Framework 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: 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 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 Mean absolute error, Root mean squared error and Kappa, whilst the precision, recall, F1-Measures, confusion matrix, TP and FP rates evaluate the classifiers' ***prediction accuracy*** [40]. These are elaborated, explained and demonstrated in the sections which follow.

a) Kappa Statistics

The Kappa statistics of the models which have been used in this study are shown in Table 7. Kappa is a controlled value of agreement for chance agreement [41]. 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 accuracy [42]. The kappa statistics is computed as shown in equation (1).

$$K = \frac{p(x) - p(i)}{p(i)} \quad (1)$$

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 7. Kappa Statistics Performance of the Classifier: 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 [42]. *This argument is supported by* Landis and Koch [54] who considers 0 - 0.2 as slight, 0.21 - 4.0 as fair, 0.41 - 0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect. Righat [42]'s *argument is also verified*

by Fleiss et al. [55]'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 [42]'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 Mean Absolute error and the Root squared mean error measure. Refer to Table 8 for the results.

i. Mean Absolute Error

The Mean Absolute Error (MAE) is a measure used to indicate how close predictions are to the eventual outcomes. MAE shows the magnitude of error(s) to expect from the forecast on average. MAE is given by equation (2)

$$\text{MAE} = \frac{1}{k} \sum_{j=1}^{k} |n_j - x_j| = \frac{1}{k} \sum_{j=1}^{k} |c_i| \quad (2)$$

Basically, Mean Absolute Error is an average of the absolute errors shown in equation (3)

$$C_i \ |n_j - x_j| \quad (3)$$

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. It squares the errors before they are averaged [41]. The RMSE C*j* of an individual program *j* is evaluated by the equation (4):

$$Cj = \sqrt{\frac{1}{k} \sum_{i=1}^{k} \left(\frac{P_{(ji)} - m_i}{m_i} \right)^2} \quad (4)$$

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

Table 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 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 [43]. When comparing our readings of MAE and RMSE for each classifier, the difference is small, this implies the less inconsistent error

size. As pointed out by Nau [44], Chai and Draxler [57] there is no complete criteria for good or bad values of RMSE or MAE.

c) Receiver Operating Characteristic (ROC)

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 interpretations of these are: 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 [45].

Table 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 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 classifier's ROC readings are more than 0.5 and closer to 1 which is a perfect test. This falls within the limits which have been laid by MedCal [45] that a perfect test should be greater than 0.5 and closer to 1.

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 [46]. The True Positive (TP) rate is the extent of examples which were classified as class *k*, among all examples which truly have class *k*. TP is equivalent to Recall. In the confusion matrix, this is the diagonal elements divided by the sum over the relevant row (See Figure 2). Given the confusion matric C and a set of labels $\{L_j\}$, the standard definition of recall is calculated as shown in equation (5).

$$\text{Recall}=\frac{TP}{TP+FN}={}^{C_n}\Big/_{\sum_{j=1}^{n+1} C_{ij}}={}^{C_{ii}}\Big/_{ATotal_i} \quad (5)$$

Where $ATotal_i$ is the actual total for the class and *Cii* is the predicted class for a multi class problem and n is the sample size. Please refer to Table 10 for Recall and Table 11 for the TP rate values of the classifiers which have been used in this study.

Table 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 which were categorized as class k, but belonging to a different class, among all examples which are not of class k. In the matrix (see Figure 2), this is the column sum of class x minus the diagonal element, divided by the rows sums of all other classes.

Table 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. We can say the classifiers are more accurate.

f) Precision (Confidence)

The Precision is the extent of the instances which actually have class x among all those which were categorized as class x (predicted positives which are actually positive). In the matrix (See Figure 2), this is the diagonal element divided by the sum over the relevant column. It is the measure of the accuracy provided that a specific class has been predicted [46]. Precision is defined by the formula (6)

$$p_i = \frac{TP}{TP + FP} = {}^{c_{ii}}/_{\sum_{k=1}^{n} c_{ki}} = {}^{c_{ii}}/_{PTotal_i} \qquad (6)$$

Where $PTotal_i$ is the predicted total for the class and *Cii* is the predicted class for a multi class problem.

Table 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 [46]. 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 [46]. The F1- measure is the adjusted harmonic mean of precision and recall. It is defined as shown in equation (7).

$$F = 2\left(\frac{P * R}{P + R}\right) \quad (7)$$

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

Table 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 1it is said to have reached its best value and when 0 it is at its worst value.

h) Classification Accuracy

Classification accuracy refers to the instances that are correctly classified by the model. It is calculated as the sum of correct classification divided by the total number of samples [46]. It is given by the formula shown in (8).

$$\text{Accuracy} = \left(\frac{(TP+TN)}{(TP+TN+FP+FN)}\right) \quad (8)$$

Where FP is incorrect prediction of positive instance(s), FN is incorrect prediction of negative instance(s), TP is correct prediction of instance positive instance(s) and TN is correct prediction of instance negative instance(s).

Table 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	10 (5.4%)

	(94.6%)	
J48	145 (77.95%)	41 (22.0%)
SMO	176 (94.6%)	10 (5.4%)

Table 14 shows the accuracy of the models or classifiers. It shows the number of correctly classified instances and incorrectly classified instances. The results show that IBK and SMO performed better than other classifiers, followed by the NB-multinomial and lastly the J48 classifier. 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 2 shows a confusion matrix for a data mining model for two possible outcomes **a** (positives) and **b** (negatives). The confusion matrix shown in Figure 3 shows how model predictions are made. The rows resemble the actual classes in the 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. In a matrix, the number of correctly classified instances is the sum of diagonals (from top left to the bottom right); all others are incorrectly classified. The confusion matrix displays the true positives, the false positives, the true negatives and the false negatives.

Predicted

a	**b**	⁄	
TP	FN	**a**	**Actual**
FN	TP	**b**	

Figure 2. A General Look of a Confusion Matrix

In our study we had 8 classes, and therefore an 8x8 confusion matrix for each classifier as shown from Figure 3-6 respectively.

The confusion matrix shown in Figure 3-6 show the correctly classfied instances (True Positives) and incorrectly classified instances (False Poistives). In all the Figures 3-6, the correctly classified instances of each classifier are in a diagonal line of the matrix from the upper left corner to the bottom right. Any other instances not in the diagonal means they are incorrectly classified. The totals of the correctly classified instances for each classifier, shown in each matrix match or tally with the totals shown in Table 14. NB correctly classified 154 instances as per the correct classes and 32 incorrectly. SMO correctly classified 176 instances as per the correct classes and 10 incorrectly, IBK classified 176 instances as per the correct classes and 10 incorrectly and lastly J48 classified 145 instances as per the correct classes and 41 incorrectly.

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 3. Confusion Matrix of Naïve Bayes Classifier Performance Attained During Classification of Data

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 4. Confusion Matrix of SMO Classifier Performance Attained During Classification of Data

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. Confusion Matrix of IBK Classifier Performance Attained During Classification of Data

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 6. Confusion Matrix of J48 Classifier Performance Attained During Classification of Data

In conclusion of the evaluation of data models through metrics measuring accuracy and performance used in our study, 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. The same applies to all other other metrics which were used. This implies that the ICA results were of high quality because they were used for classification of data throught the models. The pre-processing steps which were performed in the ICA process influeced these results.The models 's performance and accuracy met the standard threshold(s) hence this could be applied to the proposed Conceptual Framework that if the text pre-proceesing is done well and ICA performed well, the models will ultimately perform better and accurately in terms of classying the data. This is what has been echoed by other researchers such as Chung and Gray [47] that, the manner in which a model's performance is measured is very crucial and requires special consideration. One model could perform differently in different domains because of the quality of the data, normative criteria, and the decision maker factors involved. This corroborates with what Zafarani et al. [56] states in their social media mining book that, after having a representation, quality measures have to be taken care of. The text-preprocessing steps need to be performed before processing the data. Quality measures include noise removal, outlier detection, missing values handling, and duplicate removal [56].

5. Threats to Validity

For the ICA process, the guidelines which were used were informed by various qualitative research ideas [25]; [24]; [48]; [49]. This was done based on the support of the argument that, in order to produce valid and reliable inferences there has to be a set of systematic and transparent procedures for processing data [50]. Also, as highlighted by Struwig and Stead [51], it is important during text analysis, to have two or more groups of raters' to categorize 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 [51]. This is supported by Schilling [58] that, any arising problems and doubts regarding the definitions of categories, coding rules or categorization of specific cases need to be discussed and resolved within the research team.

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) methods to evaluate the Conceptual Framework.

6. Conclusion of the Study

Overall, in conclusion of the two evaluations of the proposed Conceptual Framework through ICA and data mining algorithms, 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, hence reducing the time and effort required to manually analyse data. The users' perception in terms of 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 produced higher accurate results in classifying data when supplied with data from the ICA process.

The proposed Conceptual Framework for mining and analysing social media data presents a comprehensive workflow to aid researchers to process and analyse data in a systematic manner. This addresses the aforementioned time consumption problem with traditional approaches. If something is orderly, one takes less time to articulate and follow it, other than struggling to comprehend and finding ways to solve problems in a haphazard manner. The Framework is unique in that, it does not limit the user or researcher on how they can analyse the data. It affords the user the choice to integrate both approaches or to carry them out individually.

The Framework is also flexible since it gives users various options on how to analyse data (*e.g.*, for gathering data, it gives various options on both the traditional and computational methods. Likewise, for analysis of data it also gives various options on both the traditional and computational approaches). It helps preserve the strengths of the traditional content analysis, with its regular meticulousness and relative understanding of those researchers who advocate for qualitative studies. It is true that one can choose to only apply the traditional approaches from data collection up to the analysis stage only. The same could be said about the computational approaches, it gives the researchers that zeal to explore the extensive capacity of the Big Data and the accuracy and performance of computational methods. However, as we have stated, integrating both approaches will greatly impact on the accuracy of the data analysed as well as the performance in terms of time and effort needed. The flexibility that comes with integrating traditional approaches and computational approaches, addresses problems which were stated in relation with the traditional methods of analysis which requires much effort. This is so, because with a good use of the computational methods, more data could be analysed accurately with less effort and with an optimal performance, which could have been very difficult when doing it manually. Furthermore, with a well-performed process of text pre-pre-processing, Inductive Content Analysis (or Deductive Content Analysis, whichever is chosen) makes it easy to resolve the short comings of the computational methods of failure to capture the semantics in data.

Even though this 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.

7. Limitations and Future Works

In conducting Inductive Content Analysis (the qualitative content analysis), this study went only up to categories development. The content analysis did go up to the last stage of theory development because that was outside the scope of this study. This was because we wanted the categories so that we supply them 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 our scope. As stated earlier on, our 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 classifying data automatically through the implemented models through data mining algorithms.

Another limitation which is worth noting is that, instead of evaluating the focus groups participants only with questionnaires, it could have been triangulated with an interviews of the participants so that we can get concrete results which we can be sure that they 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. Data could be modelled and visualized. In the future researchers who would like to understand about the political landscape of Botswana they can complete the Inductive Content Analysis which was carried out in this study. They could also build some political landscape theories after completion.

References

[1] P. Gundecha and H. Liu, "Mining Social Media: A Brief Introduction. Tutorials in operations research", INFORMS 2012, http://dx.doi.org/10.1287/edu.1120.0105, **(2012)**.
[2] P. Patel and K. Mistry, "A review: Text Classification on Social Media Data", IOST-JCE, vol. 17, no. 1, **(2015)**, pp. 80-84.
[3] "Facebook Newsroom", [Online]: http://newsroom.fb.com/content/default.aspx?NewsAreaId=22, Accessed February 20th 2016, **(2016)**.
[4] P. K. Pagare, "Analysing Social media Data for Understanding Students' Problem", ITCCE, **(2014)**, pp. 0975- 8887.
[5] X. Chen, M. Vorvoreanu and K. Madhavan", Mining Social Media Data for Understanding Students's Learning", IEEE TOLT, vol. 7, no. 3, **(2014)**.
[6] G. Nandi and A. Das, "A Survey on Using Data Mining Techniques for Online Social Network Analysis", IJCSI, vol. 10, no. 2, **(2013)**, pp. 1694-0814.
[7] S .Toivo, "Think Tank Social Media-The New Power of Political Influence", Version1.0, *CES*, [Online]: www.academia.edu, **(2014)**.
[8] A. Waldherr, G. Heyer, A. Jahnichen- Niekler and G. Wiedemann, "Mining Big Data with Computational Methods, Political Communication in the Online World: Theoretical Approaches and Research Designs, New York, NY, Routledge, **(2016)**, pp. 201-217.
[9] M. Evans, "A Computational Approach to Qualitative Analysis in Large Textual Datasets", PLoS ONE, vol. 9, no. 2, **(2014)**.
[10] D. Gaffney, "#Iran Election: Quantifying Online Activism, Proc. Extending the Frontier of Society On-Line", WebSci10, **(2010)**.
[11] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveral about Political Sentiment", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, **(2010)**.
[12] J. Yang and Counts, "Predicting the Speed, Scale and Range of Information Diffusion in Twitter", Association of Advancement of Artificial intelligence, **(2010)**.
[13] 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)**.
[14] T. Sakati, M. Okazaki and Y. Matsho, "Earthquake Shakes Twitter Users: Real- Time Event Detection by Social Sensors", Raleigh, NC. USA, **(2010)**.

[15] L. Hong, O. Dan and B. D. Davison, "Predicting Popular Messages in Twitter", WWW 2011-poster, (2011); Hyderabad, India.
[16] S. J. Jamison-Powell, C. Linehan, L. Daley, A. Garbett and S. Lawson, " I Can't Get No Sleep: Discussing #insomnia on Twitter, Session: Understanding Online Communication", CHI, **(2012)**; Austin, Texas, USA.
[17] S. F. Wambaa and L. Carter, "Twitter Adoption and Use by SMEs", **(2013)**.
[18] F. Cheong and C. Cheong, "Social Media Data Mining: A Social Network Analysis Of Tweets.During The 2010-2011 Australian Floods", **(2011)**, PACIS Proceedings, **(2011)**.
[19] 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, **(2012)**, pp. 481-502.
[20] L. Wallace, J. Wilson and K .Mooch, "Sporting Facebook: A Content Analysis of NCAA Organizational Sport Pages and Big 12 Conference Athletic Department Pages", International Journal of Sport Communication, vol. 4, **(2011)**, pp. 422- 444.
[21] 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, 10(¾), **(2011)**.
[22] G. Clavio, "Social Media and the College Football Audience", Journal of Issues in Intercollegiate Athletics, vol. 4, **(2011)**, pp. 309-325.
[23] R. Bandari, S. Asur and B. Huberman, "The Pulse of News in Social media: Forecasting Popularity", Computer Science Society, arXiv: 1202.033v1. Submitted 2 Feb 2012, https://arxiv.org/abs/1202.0332, (2012), [Online]: Accessed March 10 2017.
[24] M. Q. Patton, "Qualitative Research and Evaluation Methods", Thousand Oaks, CA: Sage, **(2002)**.
[25] D. R. Thomas, "A General Inductive Approach for Qualitative Data Analysis", School of Population Health, University of Auckland, August, **(2003)**.
[26] Y. Zhang and B. M. Wildemuth, "Qualitative Analysis of Content", [Online]: https://www.ischool.utexas.edu. [Accessed]: February 16 2016 study, 46th Hawaii International Conference on System Sciences, **(2005)**.
[27] "Toolkit", [Online]:http://toolkit.pellinstitute.org/evaluation-guide/analyze/analyze-quantitative-data/2016, [Accessed]: **(2016)**.
[28] J. W. Creswell, "Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative research", Upper Saddle River, NJ: Pearson Education, **(2002)**.
[29] A. Abran, A. Khelifi, W. Suryn and A. Seffah, "Usability Meanings and Interpretations in ISO Standards", Software Quality Journal, vol. 11, no. 4, **(2003)**, pp. 325-338.
[30] S. Valenti, A. Cucchiarelli and M. Panti, "Computer Based Assessment Systems Evaluation via the ISO9126 Quality Model", Journal of Information Technology Education, vol. 1, no. 3, **(2002)**, pp. 157-175,
[31] "International Organization for Standardization", ISO/IEC: 9126 Information technology-Software Product Evaluation-Quality characteristics and guidelines for their use-1991. http://www.cse.dcu.ie/essiscope/sm2/9126ref.html,Geneva, ISO, **(1991)**.
[32] J. P. Chin, A. Diehl and K. L. Norman, "Development of an Instrument Measuring User satisfaction of the Human-computer Interface", ACM CHI'88 Proceedings, **(1988)**.
[33] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use and User Acceptance of information Technology", MIS Quarterly, vol. 13, no. 3, **(1989)**, pp. 319-340.
[34] J. Nielsen, "Usability Engineering", Academic Press, Chapter 2.2, **(1993)**, p. 26.
[35] J. Nielsen, "Usability Engineering", Academic Press, Chapter 5, **(1993)**, p. 115.
[36] A. M Lund, "Measuring Usability with the USE Questionnaire", STC Usability SIG Newsletters, vol. 8, no. 2, **(2001)**, [Online]:garyperlman.com/quest/.
[37] J. R. Lewis, "IBM Computer Usabilitty Satisfaction Questionnaires: psychometric Evaluation and Instructions for Use", International Journal of Human Computer Interaction, vol. 7, no. 1, **(1995)**, pp. 57-78.
[38] G. Perlman, "Practical Usability Evaluation, Based in part on Nielsen's 1993 Heuristics and Norman's 1990 principles", **(1997)**.
[39] H. X. 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, **(1997)**, pp. 267-278.
[40] N. Gayathri, S. Nickolas, A. Reddy and R. Chitra, "Performance Analysis of Data mining Algorithms and for Software Quality Prediction", IEEE ARTCOM Proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing, **(2009)**.
[41] E. Bhuvaneswari and V. R. Sarma-Dhulipala, "The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset", Information Engineering, vol. 2, no. 1, **(2013)**, pp.6-13.
[42] R. Righat, "Kappa Statistics in Plain English", [Online]:www.stats.stackexchange.com, [Accessed]: August 16 2016, **(2016)**.

[43] T. Wood, "Using Mean Absolute Error for Forecast Accuracy", [Online] http://canworksmart.com/using-mean-absolute-error-forecast-accuracy/.Accessed August 18 2016, **(2012).**
[44] R. Nau, "What's a good value for R-squared?" http://people.duke.edu/~rnau/411home.htm, May 1 2016, Accessed August 16 **(2016)**.
[45] "MedCalc", MedCalc1993-2016, Software bba manual. ROC curve analysis in MedCalc, [Online]:https://www.medcalc.org/manual/roc-curves.php [Accessed]: August 18 2016, **(2016)**.
[46] 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 Connection State University, New Britain, Connecticut, **(2012)**.
[47] H. M. Chuang and P. Gray, "Special selection: Data mining", Journal of management Information Systems, vol. 16, no. 1, **(1999)** pp. 11-16.
[48] M. B. Miles and A. M. Huberman, "Qualitative data analysis: An Expanded sourcebook", (2nd ed.) London: Sage, **(1994)**.
[49] S. Hsieh and S. Shannon, "Three Approaches to Qualitative Content Analysis", Qualitative H.-F Health Research, vol. 15, **(2005)**, pp. 1277-1288.
[50] R. Tesch, "Qualitative Research: Analysis Types & Software Tools", Bristol, PA: Famer Press, **(1990)**.
[51] F.W. Struwig and G. B. Stead, "Planning, Designing and Reporting Research, Understanding Reliability and Validity", Pearson Education, **(2004)**, pp.134-145.
[52] A. F. Simon, "A Unified Method for Analysing Media Framing: Communication in US elections", New Agendas, **(2001)**, Available at books.google.com, Accessed 31/03/2017
[53] M. Conway, "The Subjective Precision of Computers, A Methodological Comparison with Human Coding in Content Analysis", Journalism and Mass Communication Quarterly, **(2006)**, pp. 186-200.
[54] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data", Biometrics, vol. 33, no. 1, **(1977)**, pp. 159–174.
[55] J. L. Fleiss, J. Cohen and B. S. Everit, "Large Sample Standard Errors of Kappa and Weighted Kappa", Psychological Bulletin, 72, **(1969)**, pp. 323–327, 1969.
[56] R. Zafarani, M. A. Abbasi and H. Liu, "Social Media Mining", Cambridge University Press, **(2014)**.
[57] 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, **(2014)**, pp. 1247-1250.
[58] J. Schilling, "On the Pragmatics of Qualitative Assessment: Designing the Process for Content Analysis", European Journal of Psychological Assessment, vol. 22, no.1, **(2006)**, pp. 28-37.
[59] M. Saunders, P. Lewis and A. Thornhill, "Research Methods for Business Students", 6th ed, Pearson Education Limited, **(2012)**.
[60] D. Laney, "3D Data Management: Controlling data volume, velocity, and variety", [Online]: http://blogs.gartner.com, Accessed: March 2017, **(2001)**.
[61] R. Tinati, S. Halford, L. Carr and C. Pope, "Big Data: Methodological Challenges and Approaches for Sociological Analysis", Sociology, vol. 48, no. 4, **(2014)**, pp. 663-681.
[62] M. Mahrt and M. Scharkow, "The Value of Big Data in Digital Media Research", Journal of Broadcasting & Electronic Media, vol. 57, no. 1, **(2013)**, pp. 20-33.
[63] 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, **(2012)**, pp. 662-679.

