Making Sense of Unstructured Data: An Experiential Learning Approach

Full Paper

SACLA 2019

S. Eybers[1[0000-0002-0545-3688]], M.J. Hattingh[2[0000-0003-1121-8892]]

[1,2] University of Pretoria, Private Bag X20, Hatfield, 0028
{sunet.eybers,marie.hattingh}@up.ac.za

Abstract. The need for competent data scientists are recognised by industry practitioners worldwide. Competent data scientists are human resources that are highly skilled, ready and able to work on industry data related projects upon graduation and the ability to work with data (to name a few). Currently tertiary education institutions focus on the teaching of concepts related to structured data (fixed format) for example database management. However, the hidden value contained in unstructured data (data with no fixed format) introduced the need to introduce students to methods for working with these data sets. As a result, an experiential learning approach was adopted to expose students to real-life unstructured data. Third year students were given an assignment whereby they could use any publicly available unstructured data set or an unstructured dataset supplied to them following a set methodology (CRISP-DM) to discover and report on the hidden meaning of the data. As part of the assignments students had to reflect on the process. Twenty student assignments were analysed in an attempt to identify the effectiveness of the experiential learning approach in the acquisition of skills pertaining to unstructured data. The findings of the study indicate that the experiential learning approach is successful in die teaching of the basic skills necessary to work with unstructured data. The positive aspects as well as challenges the students experienced are reported on. The lecturer's reflection reports on the appropriateness of the pre-scribed methodology, the students' performance and lessons learnt. The lessons learnt from this experience are offered up as recommendations to educators to improve on the learning experience associated with ELA within the context of educating future data scientists.

Keywords: Experiential Learning, Big Data, Unstructured Data, CRISP-DM Methodology, Data Scientist.

1 Introduction

Data has always been a key asset to organisations. Nowadays this asset has become even more important due to the potential value contained in Big Datasets [1],[2]. Big Data refers to data that constitute unique characteristics such as volume, velocity and variety (the three V's) [1], [3], [4]. Some scholars even include an additional characteristic namely value [1],[2], [5]. Volume refers to the size and subsequent quantity of data sets which are often estimated to be terabytes or even petabytes [1], [6], velocity refers to the continuous generation of data by applications such as social media whilst variety refers to different kinds of data such as operational data from various business systems, xml files and text messages [1], [6]. The different kinds of data is further classified as structured (fixed format), semi-structured (consisting of both fixed format and free text or no fixed format data), and unstructured (no fixed format) [1]. Value refers to the untapped, potential worth of the meaning hidden in large data sets [3], which might be of economic worth [2], [5]. Unfortunately unleashing the value contained in these data sets can be challenging due to reasons pertaining to technologies, processes and human aspects [3]. For example, working with technologies such as advanced data mining tools require specialized statistics tools; processes to combine data sources from various locations might be unclear; and data scientists, working with big datasets, require a combination of business, technical and analytical skills – a combination of skills rarely found in human resources [7].

Despite scarcity of data scientists the position remains one of the "most exciting career opportunity of the 21st century" with above average remuneration packages [8]. The demand for data scientists and data engineers is projected to grow with 39% [9]. The challenge from an educational perspective is to ensure that students studying in the area of informatics, information science and computer science are ready to meet the demands of industry practitioners upon graduation [10]. Although the majority of educational institutions currently focus on skills to work with data, the curriculum has a strong focus on working with structured data (such as databases, data marts and data warehouses). A current challenge in the curriculum, in particular, the institution under study, is to introduce students to ways and methods of working with unstructured data obtained from social media platforms. Also, students often, as part of their post-graduate research projects (or fourth year level projects), are faced with challenges to work with unstructured data – a skill set they have not been exposed to during undergraduate studies. As a result, this third year semester module aimed at introducing students to working with unstructured data from social media platforms. A set methodology was prescribed to guide students through the assignment (namely the Cross-Industry Standard Process for Data Mining – CRISP-DM explained later in the paper) and an experiential learning approach where students could select their own set of unstructured data from any social media source and subsequently any tool or technique to extract meaning from unstructured data. The aim of the research was to evaluate how effective the learning process was. The research question was: *how effective is an experiential learning approach in the teaching of basic skills to work with unstructured data.*

The paper starts with a brief introduction to the experiential learning approach followed by previous research focused on the topic of data science education. The CRISP-

DM methodology is explained followed by a description of the case study and proposed research method. The analysis and discussion section describes the findings after evaluating the student assignments in relation to the phases of the experiential learning approach in relation to the six steps of the CRISP-DM methodology. The discussion also includes a section of lecturer's reflection.

2 Experiential Learning

Experiential learning theory was introduced by Kolb in 1984 [11] and widely adopted in various educational environments [12] across industries such as medical and health [13], information systems [14] and marketing [15] (to name a few). The theory postulated that learners acquire new knowledge through practically completing tasks, in other words their experience of interaction with the construct under discussion [11]. Fig. 1 illustrates how learning is perceived as a continuous process that consists of four cycles namely experiencing (i.e. interaction with the construct), reflecting (review and evaluate the experience), thinking (drawing conclusions after reflecting on the experience) and acting (apply what has been learned from the process).

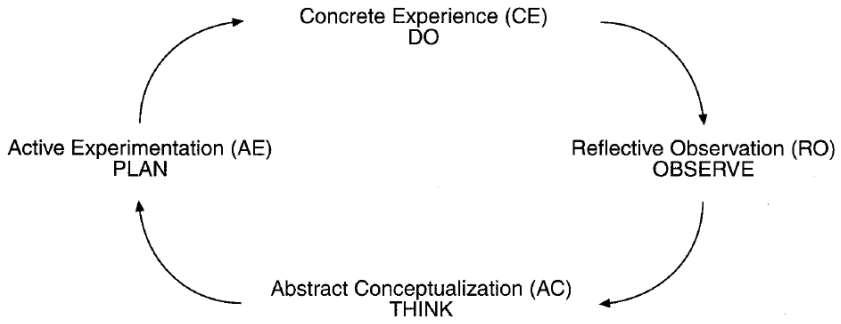

Fig. 1. Kolb's experiential learning cycle [16].

The learning can start at any point in the cycle. The benefits associated with experiential learning are: (1) increased opportunities for "analytical" reflection on tasks completed, in particular "short term experiential learning" where the task last for a short period of time [17],[18] (similar to this study); (2) "substantive" benefits which refer to the ability of students to relate theoretical constructs to the practical exposure on a deeper level than just theoretical exposure [17]; "methodological" which refers to practically apply concepts in a structured way [17]; "pedagogical" which refers to active participation of the learner into their own and peer learning [17]; and "transition" which refers to bridging the gap between applying concepts during theoretical studies into the practical requirements of industry practitioners. Lee [19] identified lower level benefits after a comparison between in-classroom learning and field-based experiential learning activities. The benefits identified included an increase in soft sills such as increased ability

to adopt to change, leadership skills and financial management skills. As a result, learners could establish their own network of practitioner contacts [15].

3 Education in the Area of Data Science

Davenport and Patil [20] describe a data scientist as a "hybrid (of) data hacker, analyst, communicator and trusted adviser" with the ability to write programming code, have contextual understanding of the environment in which they function and excellent communication skills to convey the message contained in the data to various audience levels [4], [20]. These skills can only be acquired through the exposure to real-life scenarios.

Goh and Zhang [21] acknowledged the challenge of exposing students to real-life scenarios when working with data. They referred to current educational efforts to teach students about data analytics as artificial and simplified due to the utilisation of "canned" data (a term used to refer to clean, structured data). They adopted an experiential learning approach offering students the opportunity to work on a data analytics project in partnership with a large Fortune 500 company (as a live case study). The objective of their study was to investigate the influence of the adoption of an experiential learning approach on the teaching of data analytics; to evaluate student's perceptions and attitude towards the experiential learning approach; and finally to identify challenges associated with their experience when working with big data. The findings suggested that although learning outcomes were met and student motivation increased learners were overwhelmed by the task of statistical analysis of big datasets. The project introduced additional time challenges as learners required more communication time with teachers / facilitators as well within their groups. Groups also experienced a lot of failure, and although part of the learning process, had to be explained to learners.

Serrano et al. [22] adopted experiential learning methods as part of an ongoing data science teaching project focusing on deep learning. An incremental teaching approach was used to allow students to adequately reflect on their experiences when engaging with the content presented. As a result of the lessons learned they proposed the development of a platform for experiential learning that will act as a repository for capturing and storing student experiences to be used by both students and facilitators / educators. They furthermore provided a detail list of functionalities that such a repository should offer such as a rating system to rate the difficulty of student experiences and an anonymous peer review functionality to facilitate student reflections.

Schoenherr and Speier-Pero [23] evaluated, as part of their focus on the investigation of the utilisation of predictive analytics in a supply chain management environment, the curriculum of data scientists. They found that in one particular instance, where an experiential learning approach was adopted (students had to complete a "corporate analytics project" within an organisation), students were immediately employable by organisations at above average remuneration (in line with industry requirements).

4 Cross-Industry Standard Process for Data Mining (CRISP-DM)

The CRISP-DM methodology was introduced by a consortium of both manufacturing companies, software and hardware organizations in an attempt to standardize and formalize a method for data mining [24]. The result was an independent, conceptual model proposing data as central to following a six step process during the data mining cycle. These six steps started with a clear understanding of the business under investigation, the data that are being analyzed, the preparation of the data (such and cleansing), the application of specific models to analyze the data (for example linear regression models), the evaluation of the results as a result of modelling and finally the deployment or distribution of the results and / or model to stakeholders.

The CRISP-DM methodology is similar to other data mining methodological approach such as Knowledge Discovery Databases (KDD) Process model and the Sample, Explore, Modify, Model, Assess (SEMMA) model [25]. Although the number of steps to be executed to perform data mining differ amongst the methodology (nine in KDD, six in CRISP-DM and five in SEMMA) the meaning of the steps are similar. For example, the data understanding step (step 2) in the CRISP-DM methodology corresponds to the selection of a data sample set and the exploration of the data sample set in the SEMMA methodology and the selection and preprocessing of data in the KDD model [26].

The CRISP-DM methodology was furthermore selected for the purpose of this experiential learning exercise for the following reasons: (1) it was conceptual and therefore applicable to any scenario when working towards understanding data; (2) it is a complete, workable methodology [25], [27]; (3) the methodology was prescribed in post-graduate studies (i.e. fourth year studies) and therefore seemed applicable as an introduction to subsequent studies; (4) the SEMMA method is linked to the SAS enterprise software suite and therefore software tool specific [25].

5 Case Study Description

A total of 123 students enrolled with the third year, second semester course entitled "Trends in Information Systems". The course offered an introduction to a variety of novel concepts such as IS security and bitcoin technologies (to name a few). As part of the course students were exposed to the concept of Big Data and the subsequent challenge of working with unstructured data. The learning outcomes for the session were:

- Understand the data lifecycle as part of the Software Development Lifecycle (SDLC);
- Describe the characteristics of unstructured data;
- Overcome challenges associated with unstructured data;
- Understand and practically implement the six steps of the CRISP-DM methodology.

At the end of the class students were given a practical assignment to complete. The main objective of the practical assignment was to use the current concepts explained during the session and apply it in order to work with unstructured data. The task instruction was to follow the CRISP-DM methodology to identify, clean and interpret data from any publically available, unstructured social media platform (for example Twitter or Facebook) or select one of the unstructured datasets supplied to them. Students who selected their own datasets were free to choose any, publically available, unstructured dataset and no further information was supplied to them (for example how to use Twitter, what topic or threads to use). Students were allowed to use any free tool to assist them in the process of data acquisition, data cleansing, analysis and presentation. To assist them in the process, a practical example of what the intended outcome should look like was presented in class.

As part of the assignments students had to write a report using the six steps of the CRISP-DM methodology (Table 1 outlines the details of each step). Students were also instructed to include copies of the screen(s) where the actual process of data preparation, modelling and evaluation was followed. There was no need to deploy or implement their proposed solutions but learners had to make suggestions how an organisation can use the outcome of their analysis process.

A conclusion section was added for students to conclude and reflect on what they have learned. Marks were allocated to each one of the sections. Table 1 contain a summary of the rubric used for evaluating the assignments.

Table 1. Practical Assignment Rubric.

CRISP-DM step	
1.	**Business understanding** What type of business will benefit from this analysis? What are the goals of the business, i.e. what do they want to achieve as a business? What question would you like to answer with the exercise? How will you go about to answer the business question (high level plan).
2.	**Data understanding** Collect unstructured data (from any source). Tip: Twitter might be the easiest. Describe, explore, verify data.
3.	**Data preparation** ETL: extract, transform (i.e. clean), load data into another structure (flat file, table, etc.). Include copies of your screen where you prepared your data.
4.	**Data modelling** Decide what you are going to do with the data – apply complex statistical algorithms or do basic modelling, for example categorisation. Include copies of your screen where you modelled your data.
5.	**Evaluation** What does the results mean? Do I need to repeat the analysis? Include copies of your screen where you show your results.

6.	**Deployment** If you were to share your results with the rest of your students – how would you do that? What was your experience working with the datasets (good or bad?). Was the dataset appropriate for what you wanted to achieve? What challenges did you face?
7.	**Conclusion** Did you enjoy the assignment? What did you like? What did you dislike? What did you learn?

6 Research Method

The authors followed an interpretive approach to analyse a sample of 20 from the 123 assignment submissions. Saturation was reached after 15 assignments (i.e. no new concepts emerged), but another 5 assignments were analysed to confirm saturation. All data sets used by the students made use of publically available data that were not password protected. Although, the individual data records, in the data sets used by the students, did not disclose any identifying attributes the organisations associated with the data sets were in some instances revealed. The authors followed the ethical procedure recommended by Langer and Beckman [28] who prescribed that if public data, that is not password protected is used, researchers do not need to obtain permission to use the data. However, the anonymity and privacy of the users were respected as the authors anonymised the organisation names by grouping them into industries and thereby insuring the privacy of the organisation.

All assignments followed the structure as outlined in the rubric in section five (Table 1). The authors obtained the consent from the students to use their work as part of a research paper and has obtained permission from the Faculty ethical committee to conduct the study.

Thematic content analysis was used to analyse the data. The researchers followed the six steps proposed by Braun and Clarke [29]. The first author was the examiner of the assignment and was therefore familiar with the content of the assignment. The co-author familiarised herself with the content by reading the first two assignments before analysis started. Initial codes were captured as they emerged and recorded under every step of the CRISP-DM methodology. As the analysis continued and themes emerged/reviewed/named, it was easy to see how the six steps of the CRISP-DM methodology mapped to the four stages of the experiential learning approach. The following section presents the analysis and discussion of the findings.

7 Analysis and Discussion

The analysis and discussion outline section followed the phases of the experiential learning approach namely abstract conceptualisation, concrete experience, reflective observation, and the lecturer's reflection. Each of the six CRISP-DM steps could be related to the four phases of the experiential learning approach.

7.1 Abstract Conceptualisation - Business Understanding

The abstract conceptualisation stage is concerned with learners trying to make sense of the problem at hand. For the particular assignment students could use any publicly available data set from social media or a variety of unstructured datasets supplied to them. The objective was to understand the message(s) the data can communicate to various audiences and questions that can be formulated which can be answered through the data. The majority of students selected data from Twitter from diverse industries namely: Gaming, Financial Industry, music industry, Government Activism, Government (Treasury), App Store, Fast-moving consumer goods (FMCG - Beverage Company), Communications Company.

The data revealed that the students were able to contextualize the data, identifying the parties that would be interested in the answers to questions that relate to the data. For example, Participant 5 stated that, *"These businesses want to know their market/customers better while situating themselves to a favorable position in their operating markets*". Contextualizing results is a very important skill of a data scientist [4], [20], [30]. A study by Kennan [30] reported that in a business environment context, it is important for a data scientist to know what the organization does, who the customers are and what the operating environment is. She further reported that the context is different for different countries due to regulations and in the government requires graduates not to know all the contextual knowledge but to be aware that "*the context in which data and information are used are highly varied, understand examples, and where to look for specific contexts and be prepared to continue learning on the job*".

In order to contextualize the potential results, the students were required as part of the CRISP-DM methodology, to research the companies, understanding their mission, vision and goals. This aspect of the assignment was very important, as the students were exposed to real-life companies, and had to make sense of how the data set support the organisation's mission, vision and goals. A few students did this exceptionally well, studying the business, understanding the goals of the business and how the social media data relates to those goals. This exposure to real-life scenarios is important in delivering industry ready graduates [10], [21]. Participant 3 indicated that "…*it is very important that the academy has a strong social media presence to attract potential donors, spread the word about the music programs on offer and to promote any upcoming events*." Kennan [30] states that in order to understand data within its context one has to understand the intended audience of the information.

7.2 Active Experimentation - Data Understanding and Preparation

The active experimentation stage is concerned with planning the "forthcoming experience" [16]. In this instance the students had to plan, in accordance with the CRISP-DM methodology, how to approach the collecting, extracting, cleaning, loading and storing of their chosen data set. Part of this extract, transform and load (ETL) process was the verification of the source data.

The students used a number of techniques to clean the data sets. Such as splitting datasets into smaller parts according to the original date into day, month and year then

combining those attributes into a new column. The columns were furthermore labelled using meaningful names, classified type of Tweet (for example RT for retweet, or an original tweet), classify according to keywords and removing hyperlinks. All of these activities are essential as it point to their ability to work with data [4], [20].

Furthermore, students understood the importance of recognising missing data and the potential implications it might have on the results, or that it might not affect the result depending on the way the data is analysed. Through the ETL process the students were also able to identify and discard redundant data as they recognised it would have no purpose in the analysis of the data.

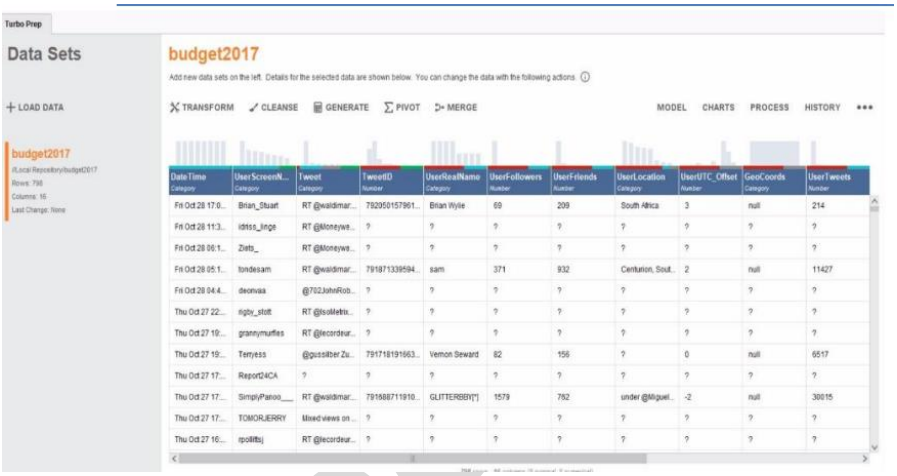

Fig. 2. Example of data set before the cleansing process.

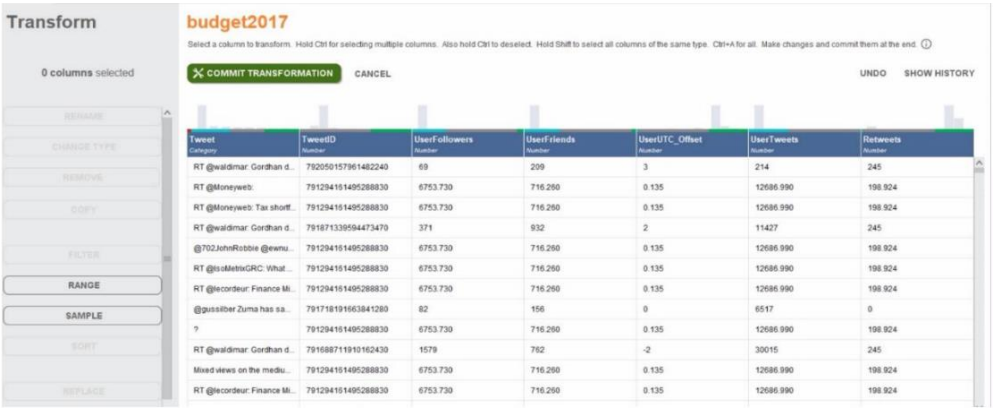

Fig. 3. Example of data set after the cleansing process.

As mentioned above, one of the challenges for a data scientist is the skill to work with technologies [3]. In this assignment students were exposed to a variety of tools to

complete the ETL tasks. For example Rapid Miner, ParallelDots AI in MS Excel (sentiment analysis), Twitter API in RapidMiner, Twitter Analytics, Zoho Reports (now Zoho Analytics), Tableau, MS Excel Azure Machine Learning add-in and Jupyter Notebook in Python. The variety of tools available to students indicated the evolving nature of data science and by selecting to use a variety of tools will add to the students' evolving skill set.

7.3 Concrete Experience – Data Modelling, Evaluation and Deployment

The concrete experience stage is concerned with the actual completion of the activity. During this assignment, this stage refers to the modelling, evaluation and deployment of the results.

During this stage, it was observed that the students used a variety of visualisation methods to communicate their results. For example, bar charts, pie charts, scatter plot, bubble chart, ring graph, line chart, and location map. One student used a histogram to indicate how brand sentiment changed over a period of time. Some students used more than one visualisation method to communicate different messages. One student used a more advance modelling technique namely the "predict" and "simulate" functions of RapidMiner to build a Deep Learning Simulator based on the data set. Kennan [30] found in her study that there is a great need for graduates to have visualization skills which would allow them to present the data in such a way that it enables decision makers to make better and quicker decisions and communicate messages to stakeholders outside the organisation.

A second observation at this stage was the students' ability to interpret the visualised results. The majority of the students were able to correctly interpret the meaning behind the visualisation. The power of big data lies in the interpretation of the results. McAfee and Brynjolfsson [4] stated that *"Big data's power does not erase the need for vision or human insight"*. The assignment was not too complex and some of the analysis were quite basic but contained powerful messages. Participant 3 found that "*the company does not have to change what it is tweeting, rather when it is tweeting. In addition, gaining more followers should increase impressions and engagement rate. If they do these two things, they should see an improved Twitter performance*." An observation from the participant illustrates that students were able to derive meaning "*the rebranding campaign was not well received (based on the tweets analysed) as the audiences did not understand the campaign concept when it came to the brand messaging and intent. Some found it offensive and insensitive whilst others either engaged positively or were indifferent to what [the company] had communicated.*"

An important component of the methodology was to evaluate the results to *verify if the results were plausible*. This requires students to critically evaluate the results. Whilst most students reported that the results were in line with their expectations, Participant 12 evaluated the results and found that the sentiment analysis done by Azure were not correct as sometimes there was a colloquial misunderstanding which skewed the results. Whilst, Participant 7 evaluated the data post modelling and concluded that the results were not accurate as the retweets skewed the data. This illustrates a level of

awareness about the nature of the data set, which is a very important skill for data scientist. Costa and Santos [31] describe data scientists as having an inquisitive mind where they interrogate the data to understand the meaning of the data.

Finally, students recognised that *the dataset can potentially answer different questions depending on the analysis*. Participant 7 generated nine different visualisations form the dataset which included a pie chart, five different bar charts, a scatter plot and two line charts. The scatter plot was used to indicate location. Participant 8 developed a generalised linear model between the categories. The awareness that one data set can communicate different messages is something students struggle with, however, with this assignment that prescribed the adopted CRISP-DM methodology, it allowed students to do a number of iterations of modelling whereby the result obtained after completing one cycle of the methodology introduce a new question or problem to be investigated.

7.4 Reflective Observation - Reflection

The reflective observation stage is concerned with the student's reflection on the completed activity. This stage offered students the opportunity to indicate what aspects of the assignment as well as experiential approach they enjoyed and what they found challenging. From the lecturer's perspective the reflective observation stage allowed her to make a judgement on the successfulness of the assignment. This will be discussed in section 7.5.

Some positive feedback regarding the assignment by the students included:

- *Students enjoyed the "mining" aspect of the assignment*. This refers to practically using an identified software tool, as presented in section 7.2. Kolb [11] explained that students that prefer a practical approach to solving problems refer to the converging learning style.
- The practical component of the assignment also extended to *learning new software.* Participant 10 explained that *"Overall, I enjoyed working on the assignment because I enjoy working with new software and the learning that comes with it, especially when it allows you to apply your theory work, making it interactive".* Learning how new software works whilst completing an assignment is an example of incidental learning which is imperative in assisting students to get "hands-on" experience and preparing them for the information age. Incidental learning is a "side effect" of learning whereby the learner was not aware of the fact that they would learn new software but then had to acquire new skills in order to complete the entire CRISP-DM lifecycle [32] .
- Students learnt about the potential impact of data. Participant 4 said that he could "*see the potential impact of big data"* and …"*it was best to see this when it was done practically".* Participant 2 confirmed this by stating *"...I enjoyed seeing how the steps are done practically and I learnt a great deal about how unstructured data can be used to make better business decisions"* whilst Participant 1 indicated *"...simple data can give good answers to important questions".*

- Students learnt about a variety of options to visualise data and the power of tools to manipulate data. Participant 7 stated that "*I was amazed how these tools could formulate meaningful graphs and charts based on unstructured data. Even if a [data] field was null, the tool was able to identify it without mixing it with the rest of the data"*. As stated by Wang et al. [33] the adoption of good visualisation methods can transform the challenges introduced by big data (such as the vast volumes of seemingly unrelated data) into meaningful "pictures".
- The assignment allowed for Self-Motivated Incremental Learning (SMIL) [34]. SMIL is concerned with the intrinsic motivation by a person which will allow him/her to learn a hierarchy of skills freely by repeating three phases: exploring the environment, identifying interesting situations and obtaining skills to reach these situations. Participant 15 reported *"…we were never really taught how to actually do a full data analysis. We were only ever shown the theory behind data analysis and data modelling and we had to use the information along with all the other knowledge we have from Stats, Maths, IT, etc. and figure out ourselves how to work with data and analyse the raw data. I learned that Excel is [a] much more powerful tool than I first anticipated, but it cannot compete with how strong Python is and how easily you can achieve the same results"*. The assignment furthermore introduced students to the area of data science, and as one participant explained "*It sparked an interest in Python in me and I already enrolled in two short online courses on Python and R in data analytics".*

Some challenges observed by the students:

- Data preparation took a long time due to the volume of data they had to work through. The majority of the students indicated that this was their least favourite part of the assignment. Participant 10 indicated that *"having to read through the data and generating a question or problem statement …took a while for me"*
- Some students struggled to understand the data set. Participant 3 said: *"[It was a] challenge to understand what data I was working with and the relevance it might have tot the [company]. However, after doing plenty of research I was able to overcome this*". This scenario is another example of SMIL [34] as the student has to research the environment in order to solve the data problem.
- One student indicated that he would have preferred more direction in the assignment. He said that "*[I] often felt lost and unsure of whether I was doing things correctly".*
- A few students reported that it was difficult to identify the correct technology for the task. Participant 7 stated that *"locating a data analyser tool which would best fit the dataset was challenging"* whilst Participant 20 stated that *"...being able to choose the right model and making the data fit the model was also a challenge",* Participant 16 was struggling with using the correct algorithm: *"I didn't know which algorithms to use and how to correctly use the datasets"*. It was clear that this challenge was two-fold: firstly, the students' ability in identifying the technology with the correct functionality to obtain the necessary results, and secondly the students' ability to use the chosen software as some software requires knowledge in order to use it.
- Finally, some students reported on the challenge to identify the appropriate visualisation method.

The above mentioned challenges are key competencies for data scientists [31] and exposing the students early on to these challenges will give them a chance to do SMIL in preparation for the workplace.

7.5 The Lecturer's Reflection

This section presents the reflection of the lecturer during and after the assignment. Three areas are reported on: (1) the appropriateness of the pre-scribed methodology, (2) the students' performance and (3) lessons learnt. Each of these will be briefly discussed.

The appropriateness of the CRISP-DM methodology.
Prescribing the CRISP-DM methodology was appropriate as students easily grasped the six steps of the process when working with data. The methodology guided students through the process and provided a structure for approaching an assignment that seemed daunting to some students at first. Due to its conceptual nature, students can hopefully re-use this methodology when working on similar projects in the future.

The students' performance.
Out of a total of 123 students, 11% obtained an assignment mark between 80% to 89%; whilst the majority of students (52%) obtained a mark between 70% to 79%. 26% of students obtained a mark between 60% to 69% whilst only 5% of students obtained a mark between 50% to 59%. The average for the assignment was 68% with the highest grade awarded 88%. The students who obtained a mark between 50% and 59% due to either not providing enough detail in their submission, or misinterpreting the results found. For example, two students uncovered the most prominent words associated with their data but failed to synthesise the findings to draw a meaningful conclusion from it. As a consequence, they did not answer the initial question without realising it. Overall, given the performance of the students, the assignment was successful in teaching students the basic skills necessary to work with unstructured data.

Lessons Learnt.
The following lessons were learnt

- *Working with social media data.* The students reported that they enjoyed working with unstructured social media data as they are familiar with the platforms. From this experience it is recommended that educators use data sets that students can relate to, such as social media.
- *Students found the extract, clean and transforming of data very challenging and time consuming.* This was the biggest, overwhelming tasks to students. It is recommended that students be provided with more practical demonstrations focusing on how to extract, clean and transform data. Unfortunately, the ETL process is the biggest task when working with any form of data.

- *Cater for different learning styles.* The lecturer should accommodate learners with different learning styles. The assignment leaned itself towards learners with the converging style (learners who prefer solving problems and apply their learning to practical implementations) as well as the accommodating style (learners who prefer to do things practically) [11]. Students with the diverging style (learners prefer to watch rather than do) as well as the assimilating style (strong analysts given good quality information) should be accommodated by offering them the opportunity to work in teams.
- *Offer more consultation time.* The lecturer should offer students more time to consult as students had more questions at the beginning of the assignment. This is similar to the study by Goh and Zhang [21].

8 Conclusion

This paper reports on the effectiveness of teaching basic skills to third year undergraduate students to work with unstructured data by following an experiential learning approach. It was found that the ELA followed enabled students to acquire basic skills in working with unstructured data. The students were exposed to a structured methodology that allowed them to tie the data sets to the goals of the business, they used a variety of tools and technologies to obtain, prepare, model and interpret unstructured data sets. The assignment enabled the students to experience the "nature of data", the influence of missing and redundant items have on the end-result and how one data set can provide different answers to a variety of questions. The students reported that they enjoyed the "(data) mining". The students reported that they enjoyed the acquisition of skills to work with new software and tools and realised the impact social media data has on an organisation. The challenges experienced by the students in completing this assignment, does not outweigh the benefits that students derived from it. As more organisations become data-driven graduates need to be prepared to support organisations with their data needs.

References

1. M. Tanwar, R. Duggal, and S. K. Khatri, "Unravelling Unstructured Data: A Wealth of Information in Big Data," in *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Noida, UP, 2015, pp. 1–6.
2. S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, Jul. 2015.
3. J. Gantz and D. Reinsel, "Extracting Value from Chaos," *IDC*, p. 12, 2011.
4. A. McAfee and E. Brynjolfsson, "Big Data: The Management Revolution," *Harv. Bus. Rev.*, no. October, p. 9, 2012.
5. S. Wang, W. Yeoh, G. Richards, S. F. Wong, and Y. Chang, "Harnessing business analytics value through organizational absorptive capacity," *Inf. Manage.*, 2019.

6. D. J. Janvrin and M. W. Watson, "'Big Data': A new twist to accounting," *J. Account. Educ.*, vol. 38, pp. 3–8, 2017.
7. A. A. Tole, "Big data challenges," *Database Syst. J.*, vol. IV, no. 3, pp. 31–40, 2013.
8. S. Baškarada and A. Koronios, "Unicorn data scientist: the rarest of breeds," *Program*, vol. 51, no. 1, pp. 65–74, Apr. 2017.
9. G. Piatetsky, "How many data scientists are there and is there a shortage?," *How many data scientists are there and is there a shortage?*, 12-Mar-2019. Online.. Available: https://www.kdnuggets.com/2018/09/how-many-data-scientists-are-there.html. Accessed: 12-Mar-2019..
10. H. Smuts and M. J. Hattingh, "Towards a Knowledge Conversion Model Enabling Programme Design in Higher Education for Shaping Industry-Ready Graduates," in *ICT Education*, Cham, 2019, pp. 124–139.
11. D. A. Kolb, *Experiential learning: experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice hall, 1984.
12. A. Y. Kolb and D. A. Kolb, "Learning Styles and Learning Spaces: Enhancing Experiential Learning in Higher Education," *Acad. Manag. Learn. Educ.*, vol. 4, no. 2, pp. 193–212, 2005.
13. S. Grace, E. Innes, N. Patton, and L. Stockhausen, "Ethical experiential learning in medical, nursing and allied health education: A narrative review," *Nurse Educ. Today*, vol. 51, pp. 23–33, Apr. 2017.
14. S. Eybers and M. J. Hattingh, "The Last Straw: Teaching Project Team Dynamics to Third-Year Students," in *ICT Education*, vol. 963, S. Kabanda, H. Suleman, and S. Gruner, Eds. Cham: Springer International Publishing, 2019, pp. 237–252.
15. L. M. Bobbitt, S. A. Inks, K. J. Kemp, and D. T. Mayo, "Integrating Marketing Courses to Enhance Team-Based Experiential Learning," *J. Mark. Educ.*, vol. 22, no. 1, pp. 15–24, Apr. 2000.
16. M. Healey and A. Jenkins, "Kolb's Experiential Learning Theory and Its Application in Geography in Higher Education," *J. Geogr.*, vol. 99, no. 5, pp. 185–195, Sep. 2000.
17. M. C. Wright, "Getting More out of Less: The Benefits of Short-Term Experiential Learning in Undergraduate Sociology Courses," *Teach. Sociol.*, vol. 28, no. 2, p. 116, Apr. 2000.
18. R. Scarce, "Field Trips as Short-Term Experiential Education," *Teach. Sociol.*, vol. 25, no. 3, p. 219, Jul. 1997.
19. S. A. Lee, "Increasing Student Learning: A Comparison of Students' Perceptions of Learning in the Classroom Environment and their Industry-Based Experiential Learning Assignments," *J. Teach. Travel Tour.*, vol. 7, no. 4, pp. 37–54, Jun. 2008.
20. T. H. Davenport and D. J. Patil, "The Sexiest Job of the 21st Century," *Harv. Bus. Rev.*, no. October, p. 8, 2012.
21. S. Goh and X. Zhang, "Incorporating Experiential Learning into Big Data Analytic Classes," in *Twenty-first Americas Conference on Information Systems*, Puerto Rico, 2015, p. 10.
22. S. Emilio, M. Martin, M. Daniel, and B. Luis, "Experiential Learning in Data Science: From the Dataset Repository to the Platform of Experiences," *Ambient Intell. Smart Environ.*, pp. 122–130, 2017.

23. T. Schoenherr and C. Speier-Pero, “Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential,” *J. Bus. Logist.*, vol. 36, no. 1, pp. 120–132, 2015.
24. M. North, *Data mining for the masses*. S.l.: CreateSpace Independent Publishing Platform, 2016.
25. U. Shafique and H. Qaiser, “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),” vol. 12, no. 1, p. 7, 2014.
26. A. Azevedo and M. F. Santos, “KDD, SEMMA and CRISP-DM: A parallel overview,” in *European Conference on Data Mining 2008*, Amsterdam, The Netherlands, 2008, p. 6.
27. R. Wirth and J. Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.
28. R. Langer and S. C. Beckman, “Sensitive research topics : netnography revisited,” 2005.
29. V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006.
30. M. A. Kennan, “‘In the eye of the beholder’: knowledge and skills requirements for data professionals,” *Inf. Res.*, vol. 22, no. 4, pp. 1–21, Dec. 2017.
31. C. Costa and M. Y. Santos, “The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age,” *Int. J. Inf. Manag.*, vol. 37, no. 6, pp. 726–734, Dec. 2017.
32. D. A. Sleight, “Incidental Learning from Computerized Job Aids,” *Michigan State University*, 1994. Online.. Available: https://msu.edu/~sleightd/inclearn.html. Accessed: 14-Mar-2019..
33. L. Wang, G. Wang, and C. A. Alexander, “Big Data and Visualization: Methods, Challenges and Technology Progress,” *Digit. Technol.*, vol. 1, no. 1, pp. 33–38, 2015.
34. A. Bonarini, A. Lazaric, and M. Restelli, “Incremental Skill Acquisition for Self-motivated Learning Animats,” in *From Animals to Animats*, vol. 4095, S. Nolfi, G. Baldassarre, R. Calabretta, J. C. T. Hallam, D. Marocco, J.-A. Meyer, O. Miglino, and D. Parisi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 357–368.