International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217121

A Framework for Social Media Data Mining and Analysis to Product and Service Development- Case of The Zimbabwean National Art Gallery

Kumbirayi Kwenda [1], Noreen Sarai*[2], Tinashe Gwendolyn Zhou[3]

[1] Systems Analyst, Kunda.org, Harare, Zimbabwe

*[2]Computer Science, Midlands State University, Gweru, Zimbabwe

[3]Information Systems, Midlands State University, Gweru, Zimbabwe

Article Info
Volume 7, Issue 1
Page Number: 194-209
Publication Issue :
January-February-2021

ABSTRACT

Social networks have become a vital component in personal life. People are addicted to social network features, updating their profile page and collaborating virtually with other members have become daily routines. Web data mining is a new trend in current research studies. This study sought to develop a framework for social media data mining and analysis for the betterment and advancement of products (art effects) and services (exhibitions) with in the contemporary art industry of Zimbabwe through a case study of the Zimbabwean National Art Gallery. The information for this paper was gathered through the use of in-depth interviews, questionnaires, key-word search and API from the organization's social media twitter handle were used to get the data for analysis. Focus group participants were chosen from the National art gallery and a matching number was selected from the artist who makes the artworks which will be soles or exhibited through the national art gallery. The findings suggested that, yes it is possible to inform the next batch of products with information mined from analyzing the sentiments or reviews of the previous set of artworks. Hence with this in mind the researchers managed to develop a framework which can be used to implement social media mining in the art sector of Zimbabwe. The proposed framework tried to handle the major limitations in current web mining frameworks by handling challenges such as special symbols, slang use, Data Validity analysis, time frame, and methodologies.

Article History
Accepted : 05 Feb 2021
Published : 14 Feb 2021

Keywords : Social Media, Data Mining, Sentiment Analysis, Knowledge Data Discovery

I. INTRODUCTION

This study interrogated the importance of focusing on social media data mining and provided a holistic approach that will likely bring about change in market-driven products when the framework is implemented effectively.

Data mining concepts and researches have successfully produced a variety of methods, tools, techniques, and algorithms for handling, manipulating and deducing meaning from data to

solve real-world problems and help companies create products that are socially acceptable. Business intelligence, data warehousing, bioinformatics, predictive analytics, and decision support systems are some of the application domain for traditional data mining. The Principal objective of data mining is to rummage around for any relationships and informative data within enormous posed data which is unrelated, mixed, not trained and direction-less.

Social media has changed our discussions about products and services yet not the business exercises fundamental underneath according to Fan and Gordon, (2014). The same social media information has been utilized by organizations to understand progressively about their business environment, competitors and suppliers. Pattern recognition and other social media mining instruments, for example, analytics help to call attention to any critical or minor changes in people groups' assessments and sentiments, general vibe, behaviors, and perceptions that can affect product design and development as stated by Fan and Gordon (2014).

These other studies have looked into text mining from websites and others have looked into a framework to discover potential ideas of new product development. According to Kurniawati, Nargiza, and Graeme, (2013) their proposed text mining method needed more rules in order to extract useful information including the relevance score more accurately. With the Zimbabwean situation, it is cheap for products and services users to go on social media than any other online website to posts their findings, which conclude the need for a study about product information through social media.

1.1 PROBLEM STATEMENT

Organizations are unable to read the information being posed to them by their clients on social media to better develop their products and realize profits and market share. Challenges such as special symbols, slang use, Data Validity analysis, time frame, and methodologies have affected dearly on organizations (Holsapple, Hsiao, and Pakath, 2014). The type of artworks and exhibitions which are being shown in Zimbabwe are not fetching more tourists locally, regionally and internationally. If the framework which is going to be designed by the end of this study is used effectively, the National Gallery of Zimbabwe will be in a position to curate shows which suit the customer needs and will attract more viewers from across the globe.

1.2 RESEARCH OBJECTIVES

In this study, the researchers assessed the contribution made by mining social media on the development of new products and service provision.

- To develop a framework for social media data mining and analysis for product and service development
- To assess if social media data can be manipulated (mined, analyzed and be interpreted) to give meaningful conclusions.
- To identify the models which can be used to mine social media data
- Identify opportunities that can arise by mining social media content.
- To come up with ways in which the model can be used in the Contemporary Arts industry in Zimbabwe.

Organizations find it difficult to implement and exploit social media in their respective companies(Bertoni and Chirumalla, 2011). The study examined and explored the different techniques which can be used for social media data exploration. The researchers came up with a framework that can be implemented in trying to get the best out of the people's posts on social media, which can be used to inform in the product and/or service development processes.

1.3 RESEARCH QUESTIONS

The main question of the study was how to come up with a framework for social media data mining for better development of a product within the contemporary Arts industry in Zimbabwe?

The sub-research questions were as follows:-

- Can businesses/companies in the arts industry depend on social media data insights?
- Are there any other models that have been developed that mitigate the need for the Arts industry?
- What business opportunities and challenges are brought by the analysis of social media data?
- Will the developed framework be in a position to improve decision making in the art Industry of Zimbabwe?

1.4 RESEARCH HYPOTHESIS

H_0: Social media data mining and analysis have an impact on the acceptance of developed products and services.

H_1: Social media data mining has no relationship with the acceptance of developed products and services.

1.5 DELIMITATIONS

The study looked into the Contemporary and visual Art Industry (to be specific National Gallery of Zimbabwe). The study only looked at social media marketing as the only affecting factor holding all else equal as suggested by Widener and Li, (2014).

II. LITERATURE REVIEW

2.1 DATA MINING KNOWLEDGE DISCOVERY (DMKD)

The first Knowledge Data Discovery (KDD) conference was a one day conference that was done at the International Joint Conference on Artificial Intelligence (IJCAI) in Detroit with less than ten publications being reviewed. In 2007, there were 17 papers presented at the data mining conferences 17 years later which marked a significant improvement. The DMKD has attracted attention from various fields of technology, ranging from databases, pattern recognition, machine learning, statistics, AI, Mathematics, data visualization, arithmetic and optimization, and economics. Despite the clear-cut increase in development, the DMKD field is still ambiguously defined (Peng et al., 2006). Usually, it is perceived by outsiders as a collection of fairly related algorithms, implements, and tools.

The absence of coordinated definitions and strong views of the knowledge discovery field causes troubles in knowledge discovery and obstructs the long haul progression of the field. To advance toward this edge here in Africa, it is important to create theoretic structures that portray the general subjects and give coordinated perspectives on the assessment of this relatively new and fast-growing business area.

Twitter and Facebook have been the most researched and mined social media pages recent studies have confirmed. Numerous researches, however, confirm that content which had been broadly analyzed or used in these studies mostly comes from Twitter rather than Facebook (Joseph, Letsholo and Hlomani, 2017a). Joseph, Letsholo and Hlomani,(2017a) completed an investigation on Iran elections by dissecting the online networking information vastly available on Twitter. In their research, Joseph and the team used user grids, histograms, and frequency of top key phrases to evaluate social activism. Different research was internet-based or social media substance was used, especially Twitter data, is the investigation on the examination of political conclusions of the Germans government polls that was seen over and led by Tumasjan.

Ediger et al.,(2010) utilized and broke down Twitter information using Cray XMT, a diagram portrayal toolbox for immense charts exhibiting interpersonal social data information. They utilized the GraphCT to inspect and process information from Twitter. Since

Twitter's messages systems and hubs show up essentially as tree-organized as news dispersal frameworks, inside the online information as bunches of discussions, they utilized the GraphCT to do positioning of entertainers inside these discussions and help investigators to focus on impressive decreased information subsets.

Hong et al.,(2013) States that, in internet-based life i.e. social media, for example, Twitter, data which is viewed as imperative by the general populous regularly spreads through re-tweets. By inspecting the qualities of such information which is spreading across the board, this can help in significant various tasks (e.g., headlining stories recognition, customized messages proposals, viral showcasing and others). All things considered, Hong et al.,(2013) did an examination to research how the fame of messages could be anticipated through the number of future re-tweets. They likewise got some understanding of what variables lead to data spreading on Twitter.

2.2 DATA MINGING METHODS

Besides clustering, classification is the most common method in Data Mining. Classification is the process of grouping data into categories so that the data can be understood but the next user. When using this supervised learning technique, the data objects will be grouped into pre-labeled classes of data, this will be done using the likeliness of the objects and how similar the data is to the present conditions of the class. For instance, when classifying, you take keywords from the articles, then determine the number from which the given keywords belong to.

Clustering is now the process of segmenting the look-alike groups of information into clusters. It can also be called unsupervised learning since the data objects are being placed into undefined groups. Unlike classification, clustering has no pre-determined groups but instead, groups are made through self-similarities within the population. Elements of groups share a common set of properties.

2.2.1 DATA CLUSTERING

Data clustering is, as mentioned above, a way to divide a population into subgroups in order to simplify the location of the data required. There are several methods to accomplish this task and some of those methods are presented here below.

2.2.2 RANDOM FOREST MODEL (RFM)

As the name of the mining model sagest, it consists of many individual trees operation in the same domain trying to classify information. The given class with unanimous votes from the decision trees working single-handedly in the random forest will become the predicted class. While each tree dribbles towards predicting a model the one with the largest amount of votes becomes the selected model.

2.2.3 HIDDEN MARKOV MODEL

The Hidden Markov Model (HMM) is a mathematical algorithm with its system modeled from the Markov process. Zhao and Ohsawa,(2018) pointed out that, Markov process has so a hidden layer with quit a number of states. In most cases it utilizes three parameters to give description to a relationship between the visible series and the hidden pairs of the transition matrix. The hidden layer is influenced by the visible layer as stated by Zhao and Ohsawa (2018). This HMM if still upcoming in the field of social media mining, that why publications in connection to the HMM and how it helps on sentiment analysis is still limited.

2.3 SENTIMENT ANALYSIS

Sentiment analysis is the computational recognition and analysis of suppositions, sentiments, feelings, and subjectivities in posts and text materials. As the extraordinary utilizer of social media data mining, sentiment analysis will be mainly focusing on the

programmed retention of positive or negative suppositions from the mined data. It is mainly noted that texts frequently contain the blend of positive and negative sentiments, so it's regularly valuable to distinguish the extremity of sentiment in texts (the good, the bad or impartial) and sometimes it is also important to check the quality of the sentiments being communicated (He et al., 2015). Support Vector Machine (SVM), Naive Bayes, Maximum Entropy and Matrix Factorization, are the machine learning strategies that are mainly used to arrange and order data into positive or negative classes.

2.4 CHALLENGES OF SOCIAL MEDIA MINING

Just like any other technology which deals with big data, social media data mining has its own problems. These problems have led to many kinds of research being done so as to improve ant the writer of this paper hopes to solve some if not all by implementing a framework that minimizes these challenges. The challenges include, the data being in fluid form, the volume of the content, use of special characters, jargon, shorthand, abbreviations, validity, the sensitivity of the available content, typing errors, legitimacy extraction problems, variability of the sources.

Considering the amount of data being posted on the social platform, this is a major concern on itself. If the mining algorithms are not up to date, old or expired information might be used to make decisions that will lead to fatal losses. The data from social media comes in large amounts and the volume will be rapidly increasing, this proves the dynamic nature of the data as suggested by Dang-xuan (2012).

2.5 SOCIAL MEDIA IN ZIMBABWE

Social media in Zimbabwe has been the form of communication in Zimbabwe for a while now. With the usage of internet and other platforms which do support social media in Zimbabwe, maces have now shifted from using traditional forms of communications, traded for social media. Grownups from 15 years and older from Zimbabwe covers a population of 6,776,000. It was noted that approximately one in three device owners uses the Internet (25.6%), accessed twitter, Facebook or any other social media. The use of internet has since increased in Zimbabwe and as said by Sengweni et al.,(2015) social platforms have not been fully utilized by industries be it advertising or data mining although some improvements have been seen all around from the past periods.

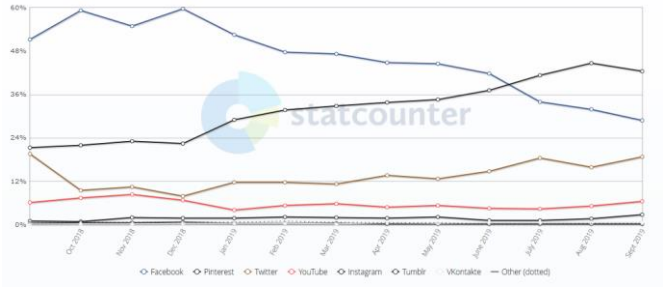


Figure 2. 1 : Social Media Stats Zimbabwe

The mobile networks companies in Zimbabwe offer social media bundle which enables users to use their smart devices wherever they will be. The fig 2.3 below showed that:

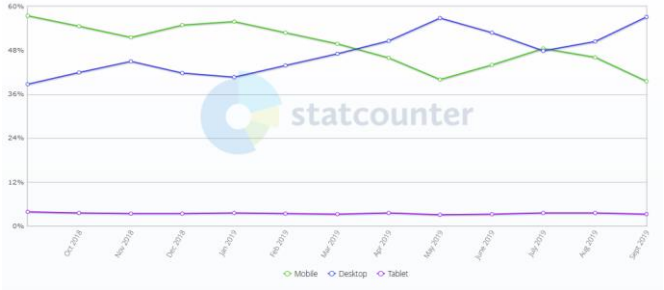


Figure 2. 2 : Desktop vs Mobile vs Tablet Market Share Worldwide

2.6 CONCEPTUAL FRAMEWORK

Fernando, Gapar and mdJohar(2015) noted that product quality does improve if the web mining algorithms are used to get information from websites and come up with better ideas for the next product.

The three colleagues did not look in the functionality of these web algorithms on social media data which is unstructured and is of streaming nature.

Assuming that other things are constant during the performance of the study, it will be possible to relate these two variables and confirm that indeed, artworks and exhibitions acceptance in the market depends somehow on the use of customer reviews on previous shows and other works. Frameworks such as figure 2.4 has worked for some nations who speak one language. However In countries like Zimbabwe, it is really hard to get a tweet or a post with only one language hence language manipulation and text structure will be not easy to manipulate.

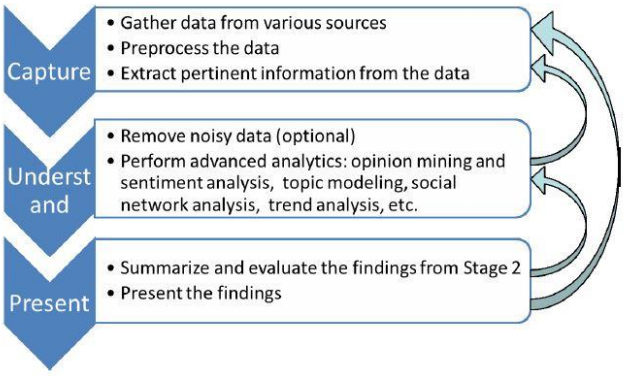


Figure 2. 3 : Social Media Analytics Process 1

Social media data have the qualities of Big Data, for example, volume, veracity, speed, fluctuation, and value (Holsapple, Hsiao, and Pakath, 2014). These qualities make their analysis more testing than customary data. Other researches have proven that data from review pages and websites can inform product development. Challenges such as special symbols, slang use, Data Validity analysis, time frame, and methodologies have affected dearly on organizations (Holsapple, Hsiao, and Pakath, 2014). Since most of the data on social media is not directed to one person. It is hard for organization to pick the relevance of some other information linking it to their world Social media data mining being the independent variable and the dependent variable being the quality and or acceptance of the product or service developed.

During the study by Fernando, Gapar and MdJohar,(2015), a conclusion that many algorithms are developed and deployed in the web structure mining area was established. Interesting patens can be identified from these web content mining, which is similar to traditional data mining, hence it was beneficial to use both mining algorithms together, to mine data from social media. The framework was going to be designed to mitigate the disadvantages of the existing algorithms single-handedly and to give birth to social media mining techniques in the arts industry of Zimbabwe.

III. METHODOLOGY

3.1 METHODOLOGICAL FRAMEWORK

The author suggested distinguishing influential organizations and corporations in specific sectors, such as the arts, contemporary art, to compare their social media citations for competitive analysis and effectiveness of the found information for product development.

3.2 RESEARCH APPROACH

The researchers used the mixed approach on the choices, cross-sectional time horizon with a deductive approach under the positivist philosophy according to Sahay, (2016). Given the data obtained for this research, the deductive approach was used to identify the challenges of why social media data mining has not been adopted and utilized in the arts sector in Zimbabwe and to develop a framework and guide its use. Data collection methods include simply using the APIs which are accessible from the organization under study. The APIs obtained from the social platforms sometimes when it is not viable to obtain the API security tokens, crawling and parsing HTML can be the viable way to obtain the much-needed review, remarks.

3.2.1 RESEARCH STRATEGY

The research strategies which were utilized by the researchers consisted of survey experiments and case study. Outlined here is the detailed procedure of how: - interviews were conducted on the sample focus group, with results being recorded. Data collected from social media was analyzed to check for most searched for items and come up with a word cloud. The word cloud was the bases for the questionnaire which was developed and distributed.

3.3 DATA COLLECTION INSTRUMENTS AND TECHNIQUES

3.3.1 PRIMARY DATA

The primary data was data collected by the author himself for the purposes of the study at hand. (Of, In and Systems, 2008) the researchers used APIs to collect data from Facebook and Twitter. Also as part of the data collection, the researchers also interviewed employees at the national gallery of Zimbabwe and lastly the researchers conducted interviews with some of the participants of the research.

3.3.2 SECONDARY DATA

Social media accounts have since been long created by the national gallery of Zimbabwe and some quality data has been kept as reference points. The researchers managed to go through achieved information and asses

3.3.3 QUESTIONNAIRES

J. Knapp Kenneth,(2006) reiterated that, questions of the survey include numerous investigations with coordinated response categorizations; some open-ended investigations may also be included. Problems are evaluated (now and then overwhelmingly) for inclination, arrangement, lucidity, and facial legitimacy. For the most part, researchers try to utilize an enormous number of members to evaluate the research from both execution and exactness. The researchers distributed the questionnaire to every participant of the survey

3.3.4 HOW THE QUESTIONNAIRES WERE DISTRIBUTED

The questionnaire was distributed using the internal electronic mail to senior management and some were given to individuals (artists) when they visited the gallery. The researchers took the time to explain to every individual how the questioner was to be completed. The researchers followed the questioners and collected them for data processing.

3.3.5 INDEPTH INTERVIEWS

In-depth interview are defined as part of the qualitative research technique where, intensive individual interviews are conducted. Interviews were conducted on the employees of the national gallery. The researchers wanted to get to know how exactly the organization was being affected by not implementing social media mining

3.3.6 CASE STUDY

A case study was defined by Saunders (2009) and Robson (2002) as “a strategy for doing research which involves an empirical investigation of a particular contemporary phenomenon within its real life context using multiple sources of evidence”. “A case is a bounded entity such as an organization which acts as a unit of analysis”. “The purpose of using many case study method is to be able to relate findings from one case with findings from another and decide whether they are the same and be able to generalize from the results” Saunders (2009). Information is present on 3 regional galleries and the artists who supply them with the works though approaches such as thorough interviews and questionnaires observation and questionnaires. It is essential in that, a case study technique can be used to assess and analyze diverse occurrences. Another benefit as a result of using case study to gather information is its capability to use information from dissimilar bases to make it stress free to analyze.

3.4 DEPENDABILITY OF FINDINGS

It is of significant importance to check whether the information received is worthy or qualify to be utilized as the trusted responses of the research. They are of the utmost importance throughout the analysis, as they measure just how good the research results are. The researchers used face validity. This face validity is a two-step process which involves having your survey reviewed by two different parties. The first was the researchers were familiar with the topic at hand, and evaluated if the questions successfully captured the research understudy. Secondly the researchers cleaned the collected data. They entered the collected responses into a spreadsheet to clean the data. Having one person read the values aloud and another entering them into the spreadsheet did very well in reducing the risk of error.

3.5 SYSTEM FOR SAMPLING

The technique used to choose members to partake in the investigation is called the focus groups as specified by Broglio et al., (2014). It characterizes the sample as a specific group of people involved or taking part in a given study to be a focus group. To respond to the exploration questions, a purposive determination methodology is utilized to decide on the number of members to be linked with the study at hand and the location where the research is going to take place. Focus groups and in-depth interviews were used to collect data for this research.

3.5.1 TARGETED POPULACE

This denotes the group of people or the number of participants who were asked to take part in the project at hand. This study is targeting 3 decision making groups within the National Arts Gallery of Zimbabwe, the Artists, the ICT department and the senior management. National Gallery of Zimbabwe (NGZ) is the largest contemporary art organization in Zimbabwe and is entirely run by the government of Zimbabwe, and decisions on the type of art to be produced, sold and displayed are made locally by both parties, i.e. senior management and artists.

3.6 SAMPLE SCOPE

Targeted populace	Questionnaire sent	Responses
Senior management	5	5
ICT department	5	5
Marketing and sales department	5	5
Artist (Photographers)	10	7
Artist (sculptors)	10	8
Artist (painters)	10	9
Totals	45	39

Table 3.1: Focus groups

Too small a sample will yield scant information; but ethics, economics, time and other constraints require that a sample size not be too large. The researchers reached the scope based on three main factors budget access and time. Constraint regarding the timeline and monetary budget for the work to be done was faced by the author. The NGZ template in which the research conducted included 5 staff members in the Information Communications and Technology Department, the whole department consists of 5 people and all were chosen to participate in the study, another 5 people were selected from the marketing and sale department, and 5 people were then selected from the senior management group, the group consisting of decision-makers show creators and curators as well involved strategy formulation. 24 participants from the group of artists were also selected to take part in the survey. The artist's group was subdivided into 3 visual arts storytellers that is, painters (9) sculptors (8) and photographers (7). Lastly, a survey was created on Facebook to get the reviews from artists whom the author could not

reach but who still followers the organization in question. Of all the galleries in Zimbabwe national gallery was selected because it is the biggest of all the galleries in Zimbabwe and all artists who need to be recognized abroad need to be having a certificate from the national gallery of Zimbabwe.
The sample size was then classified into only 2 categorical groups (Management and Artist) so that a Chi-square test could fit. A Run test was used on half of the sample (the management group) testing for variability in perception over the use of social media mining this would answer some of the research questions.

3.7 ETHICS AND VALUES

Voluntary responses were sought from clients, the researchers produced the consent form from Midlands State University to prove that the study was academic. Consent forms (clarifies the objectives of the study and guidelines on how to complete the questionnaires) were handed to show the respondent's consent to take part in the study. Emphasis was given on confidentiality, names were not to be disclosed on the questionnaires by the participants. To ensure that research ethics are observed the researchers the help of another observer to conduct the second survey.

The researchers went on to ask for permission to use Twitter and Facebook APIs from the same organization's national gallery of Zimbabwe. Social media posts where mined and usernames and personal names were removed from the client's posts.

IV. DATA ANALYSIS AND DISCUSSION

This research study was mainly focused on coming up with a framework for social media mining and the researchers has tried in the previous chapters to explain how the framework would help organizations to come up with meaning full conclusions from analyzing their organizational responses from social media. From the questionnaire, we do find out that:-

4.1 SOCIAL MEDIA PRESENCE ON PERSONAL LEVELS

The national gallery of Zimbabwe employees, all have personal social media accounts which they use for personal updates, thus 100% of the sample being on social media. While the artist group had 82% of the selected sample on social media with the remaining 18% claiming that social media is distractive and they prefer to work uninterrupted.

4.2 ACCESS TO COMPANY ACCOUNTS AND TYPE OF POSTS

The organization (NGZ) being analyzed have so many social media accounts but the most popular ones are Twitter and Facebook. 33% of the management sample size had access to the company accounts and confirmed that their postings consisted of a variety of media format i.e. the post could be in various forms such as video images and blog posts. This means also the replies for this post on a company account will be in all formats (memes, pictures gif, posts, videos, etc). The whole artist group confirmed that they had no access to the NGZ social media accounts (back end access), but had access to posting on the company social media fun pages and groups.

4.3 KNOWLEDGE OF SOCIAL MEDIA DATA MINING

The section had two parts of the same questioners i.e. before and after hence a non-parametric test was used to test the significance of the change in opinion. This test was done on only the employees of the gallery. The employees were asked if they knew anything about data mining on a scale of 1-10 they had to rate their experience the results were recorded in the pre % column they were shown an example of

text mining (sentiment mining) and they were given the same questionnaire to check if their opinion has changed and the results were recorded in the post % column. The total number of people who participated in this exercise was 15 and the results were as follows.

Id	Pre %	Post %
1	20	70
2	30	70
3	90	90
4	10	10
5	40	60
6	20	80
7	40	90
8	50	50
9	0	60
10	20	20
11	10	40
12	60	80
13	90	100
14	100	100
15	50	60

Table 4.1 : Participants responses toward data mining

On the collation of results, there were 2 categories namely, the pre explanation and the post explanation as to assess the perception, understanding of data mining and analysis. The interview and questionnaire questions examined the knowledge and understanding of data mining and social media. The results, both pre and post explanation were anonymous, but each respondent was assigned a peculiar number to allow pairing of results. Cohen's difference test (Cohen's d) was first used to investigate the variances of the means of the 2 categories i.e. pre and post explanation. (Salkind, 2010) suggests that Cohen's d has 3 categories in interpreting investigations which are

- Small effect (p-value= 0.2)
- Medium effect (p-value =0.5)
- Large effect (p-value =0.8).

```
> cohen.d(pre, pos, na.rm = TRUE, pooled = TRUE, paired =TRUE)

Cohen's d

d estimate: -0.786355 (medium)
95 percent confidence interval:
     lower      upper
-1.2594423 -0.3132677
```

Figure 4.1 : Cohen results for pre and post-test, from R

Armed with a p-value =|0.786355|, the results depict that the explanation which was done before the post explanation test had a significant influence on the understanding of data mining at a 95% confidence interval.

Also question 12 and 13 were analyzed using a distribution-free test called the sign test. With this none parametric test, the researchers hoped to assess the effects of the mining analysis which was done.

H_0: the social media mining example did not change the views of the participants
H_1: the social media mining example used had an impact on the participant's views.

```
Console  Terminal  Jobs
~/
  Length of x must equal length of y
> y <- c(-1,-1,1,-1,-1,-1,-1,1,-1,-1,-1,1,1,1,1)
> z <- c(1,1,1,-1,1,1,1,1,1,1,-1,-1,1,1,1,1)
> SIGN.test(z,y=NULL, md=0, alternative = "two.sided", conf.level = 0.95)

        One-sample Sign-Test

data:  z
s = 12, p-value = 0.03516
alternative hypothesis: true median is not equal to 0
95 percent confidence interval:
 1 1
sample estimates:
median of x
         1

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.8815      1      1
Interpolated CI       0.9500      1      1
Upper Achieved CI     0.9648      1      1
```

Figure 4.5 : Sign test results from R

Conclusion

From the results obtained from R, the statistical package used to analyze the sign test. Test statistic s = 12 with a P-value of 0.03516 tested at 95% confidence level. We fail to accept H0 and conclude that the social media mining example had an impact on the participant's view.

4.4 WORD CLOUD

The researchers used API Keys and security keys from tweeter to come up with sentiment analysis of the most talked-about words. The process included setting twitter authorization, collection of 10,000 tweets, conversion of data into a corpus, cleaning, removal of stop words, and displaying the results. The key search word which was used was "National Gallery of Zimbabwe" and figure 4.6 shows the results which were obtained from text mining. Word cloud implement mining algorithms such as Random forest, Spiral, Quadtree data structure algorithm, etc. on this search random forest was implemented.

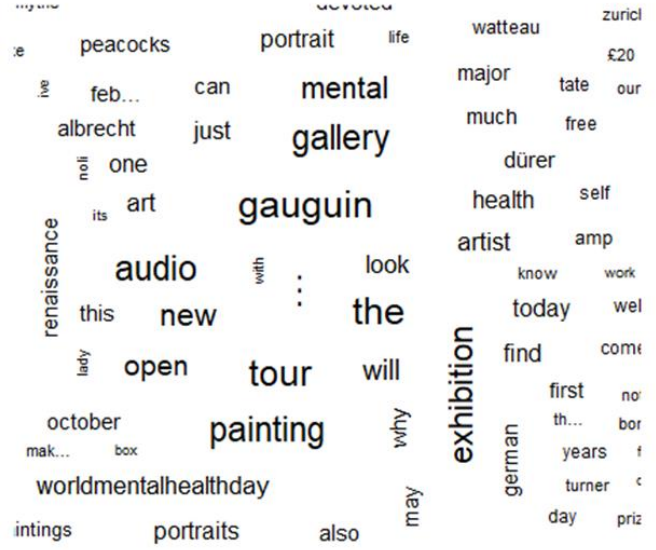


Figure 4.6 : word cloud: National Gallery of Zimbabwe from R

From figure 4.6 we see that the most talked-about words Gauguin, Audio, exhibition, painting, tour, gallery, mental, and so on. Reflecting on these words we can see that Gauguin is an English painter with works in the permanent collections in the Zimbabwean gallery. There are a few audio exhibitions. Paintings are the most exhibited. School tours are done regularly. This information was obtained by interviewing the conservation manager questions about the words themselves.

This proves that text can be mined and better information can be obtained from the social media pages. With this use of what has been mined will further the organization exhibition and quality of artworks being produced.

4.5 MODELS AND ALGORITHMS

The researchers also noted that there are quite a number of algorithms that could be used to mine social media data. Depending on what the end result of your data mining quest the explorer will have to choose the most reliable mining algorithm. Models such as Spiral algorithm, quadtree, decision tree, random forest, hidden Markov model were the most mentioned.

4.6 ATTITUDES TOWARDS SOCIAL MEDIA MINING

The section was targeting Section D (Attitudes towards social media mining) of questionnaires. A parametric test was used to test the dependence level of the 2 variables social media mining and acceptance of new products and services in the Arts market. This test was done on all the employees of the gallery and the artists who participated in the survey. The total number of people who participated was 39 and the results varied from (Don't agree, Fair and Agree). The questions were

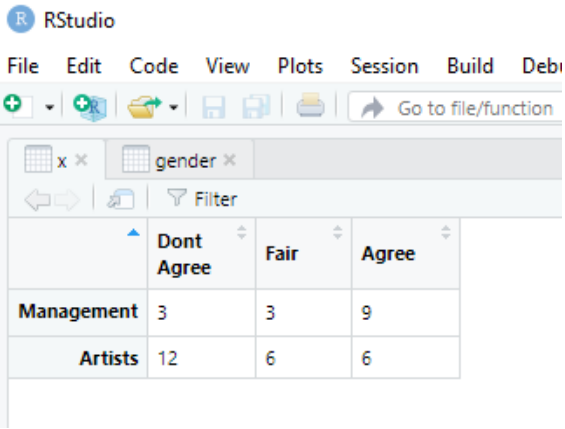
- Does social media mining improve the quality of products being developed?
- Does social media mining improve decision making?
- Social media mining makes it easy to see what's good for the brand and what's bad?
- Using social media mining tools complicates my work?
- Using social media mined results will make my products more acceptable in the market?
- I want to make other products based on the information from the social media mined information?

Stating the null Hypothesis and Alternative Hypothesis

H_0: Social media data mining has no effects on products and services being developed

H_1: Social media data mining has effects on products and services being developed.

Figure 4.7 below illustrate the results obtained by the questionnaire on the attitude of people towards social media data mining. A total of 39 people responded with 15 coming from the national gallery management and the remainder coming from the artist who exhibits at the national art gallery. Most of the artists didn't agree with the idea of data mining being used to further the development of their artworks but also a large number from the management did confirm that the mining techniques can be implemented to get better products.



	Dont Agree	Fair	Agree
Management	3	3	9
Artists	12	6	6

Figure 4.7 : results obtained from the questionnaire

A chi-square distribution test was used to test if the nature of the relationship between social media data collected and products being developed and if the chi-square value is greater than the test statistic the initial hypothesis will not be accepted hence at some level the alternative hypothesis will be true. The test statistic from the table is 4.605 which was obtained from the chi-square distribution tables (Wuensch, 2011)

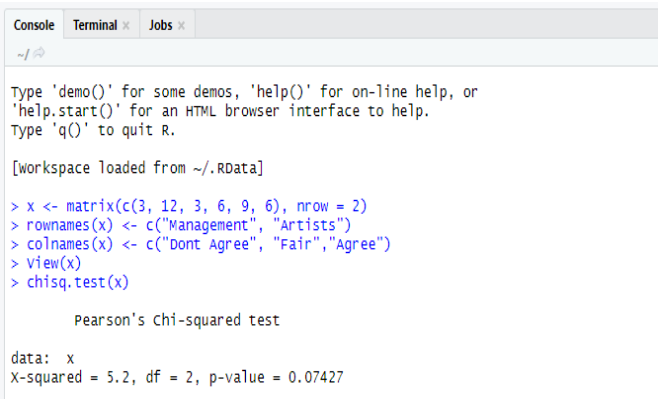


Figure 4.8 : Chi-square results from R

From the results in Figure 4.9, we get to see that the chi-square value is 5.2 with 2 degrees of freedom tested at a p-value of 0.074.

Test statistic obtained from the Chi-square distribution tables testing at a 90% level of significance and 2 degrees of freedom is 4.605.

Conclusion

Since 5.2 > 4.605, the researchers failed to accept H0 and conclude that social media data mining has effects on products and services being developed.

With these findings, the researchers managed to answer the other research question (can companies in the arts industry depend on social media data mining for better development of their products?). The findings proved that organizations can inform their new products and services using data obtained from social media.

In this chapter it was evident enough that products quality do somehow depend on how much effort was put in place to find user reviews. The literature did state that social media data is equally as good as the website reviews when it comes to products assessment. The sample strongly agreed and noted that the use of multiple language affects the collection of data. Among the suggestion, thy pointed out formalizing social media pages and make them act as company blogs. This also proved to be a "not so good" solution since much of the social media data is not directed to any organization. Creating a frame work which could deal with language processing and

steaming nature of the social media was the main point which was being pointed out.

V. RECOMMENDATIONS AND FUTURE WORK

5.1 PROPOSED FRAMEWORK

This suggested model illustrated in figure5.1 has six stages. The framework stages start with collecting data from deferent social media sites (data gathering), the second phase consists of data manipulation so as to get the correct format for analysis (extraction and preparation). Moving from each and every stage there are storage facilities, and to data warehouses, the data can be stored for later use. The arrows in the diagram represent the flow of information from one stage to another. These stages are well explained later in the same paper.

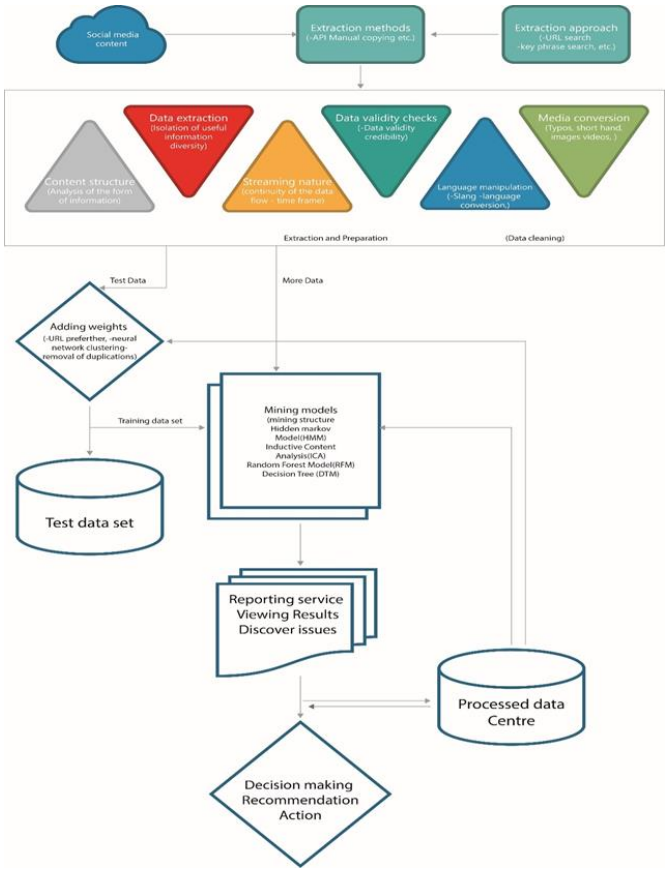


Figure 5. 1 : Proposed Framework

The framework was reviewed by two experts, one in the academic field (National University of Science and Technology) and the other from the Zimbabwean telecommunications industry (Econet Zimbabwe).

Data gathering

This stage consists of the collection of information from a selected choice of social media account. The data can be gathered using traditional and or computational systems. For example, the use of API for computational methods and the use of interviews and surveys for traditional i.e. the surveys can be done online on social media through surveys and polls.

The data can be channeled using deferent extraction approaches, such as URL search or key phrase depending on the type of result which is being searched for i.e. qualitative and quantitative data.

Extraction and preparation

After gathering your data there is a need for the data to be ready for analysis. According to (Holsapple, Hsiao, and Pakath, 2014) social media data turns to be distributed and formless. It is the duty of the mining expects to get the information together. The data has to be separated getting useful information from the free form blocks of data. It was explained by Holsapple, Hsiao, and Pakath,(2014) that information is hidden in memes, emoji's and stickers can be omitted, there is need to extract that type of information also the social media post consists of many layers hence the other information on the same particular post might not be relevant to the study in question.

Data validity checks: with this social media consisting of multiple sources of media coverage, being free form and unstructured, there is the need for a basic validity check so as to avoid spam, insensible and or drivel information. Language manipulation: not all the information on social media is made for mining some are general pear to pear conversations don in public domain hence a language of choice will be used.

There are some symbols and slang which can be used when communicating depending on location, these can only be understood by human. This process will convert this information to basic level one type of language which can be understood by the algorithms.

Not only language need to be monitored, videos images audios are posted onto social media hence there is also need for Media conversion so as to get all in one package.

Streaming nature: The content on social media is forever being generated hence there is a need to have a real-time analysis. The volume of the data is always ever-increasing which will also determine the next phase. The whole extraction and preparation must be done to only a sample of the data, which is called the training set.

Adding weights
With social media, there are so many duplications of the same post be it on one platform or different but this has an impact on the model. Having so many posts with the same message conveys the impotence of the post but some other duplications are based on error postings. Any given post will be allocated a weight based on how frequent it has appeared on a given platform (social media). With this done on the training set, it means the mining model has now the basses to work on the rest of the data using the classification models which has been used in the extraction phase.

From this stage, the training set can be stored for later use or can be parsed to the mining model of the researchers's choice.

Mining Models
Mining models will do the actual data analysis work utilizing the information obtained from the training sets. It will come up with theories and specific levels of focus. Depending on the mining model invoked, qualitative and quantitative analysis will be used. Data obtained at this stage can be considered for re-use, assessment or be used as a training set for another bigger project.

Reporting services
This stage the analyzing personnel will do documentation which will be sent to the decision-makers for consideration. Issues are discovered in this stage. The reporting service stage is linked to the data center because the data can also be stored for later use. The other link to decision-makers proves that there can be issues that want intervention from them.
Decision making

After all the stages have been completed the decision whether to use the information proposed or not.

5.2 LIMITATIONS

The limitation worth mentioning is that, instead of evaluating the focus groups participants (the management and artists) only with questionnaires, it would benefit the research if a larger sample was used including the general public, the users of the products thus concrete results would be obtained and since most of the sample population will be end-users the opinions will be honest and reflecting how the world view what they post on social media.

5.3 FUTURE WORK

The author of this paper did not go deep into the Mining models and how they will be structured. It would be of paramount impotence if the analysis of how the Analyzing models will be feeding information of this framework. Testing the system.
Tests on whether the products developed through the information obtained from mining social Medea are better or not, are yet to be done. It will be of paramount importance if these tests are done.

VI. CONCLUSION

The data mining framework was developed to assist the Arts industry of Zimbabwe, on mining data from social media and be able to use the information gathered and analyzed through the use of advanced software, to make decisions. The data mining framework can still be used to assess if a product is still viable in the market of no it gives the bases on which all language processing software s are utilized to capture all the sentiments and have a more informed decision.

VII. REFERENCES

[1]. Barbier, G. (2011) 'Social Network Data Analytics', Social Network Data Analytics. doi: 10.1007/978-1-4419-8462-3.

[2]. Bertoni, M. and Chirumalla, K. (2011) 'Leveraging web 2.0 in new product development: Lessons learned from a cross-company study', Journal of Universal Computer Science, 17(4), pp. 548–564. doi: 10.3217/jucs-017-04-0548.

[3]. Bijmolt, T. H. A. et al. (2010) 'Analytics for customer engagement', Journal of Service Research, 13(3), pp. 341–356. doi: 10.1177/1094670510375603.

[4]. Broglio, S. P. et al. (2014) 'National athletic trainers' association position statement: Management of sport concussion', Journal of Athletic Training, 49(2), pp. 245–265. doi: 10.4085/1062-6050-49.1.07.

[5]. Dang-xuan, S. S. L. (2012) 'Social media and political communication : a social media analytics framework'. doi: 10.1007/s13278-012-0079-3.

[6]. Dellarocas, C., Gao, G. and Narayan, R. (2010) 'Are Consumers More Likely to Contribute Online Reviews for Hit or Niche Products?', Journal of Management Information Systems. Routledge, 27(2), pp. 127–158. doi: 10.2753/MIS0742-1222270204.

[7]. Fan, W. and Gordon, M. D. (2014) 'The power of social media analytics', Communications of the ACM, 57(6), pp. 74–81. doi: 10.1145/2602574.

[8]. Farr, B. C. (2008) 'Designing Qualitative Research', Transformation: An International Journal of Holistic Mission Studies, 25(2–3), pp. 165–166. doi: 10.1177/026537880802500310.

[9]. Fernando, G., Gapar, M. and MdJohar, M. (2015) 'Framework for Social Network Data Mining', International Journal of Computer Applications, 116(18), pp. 7–10. doi: 10.5120/20434-2765.

[10]. Gundecha, P. and Liu, H. (2012) 'Mining Social Media: A Brief Introduction', 2012 TutORials in Operations Research, (Dmml), pp. 1–17. doi: 10.1287/educ.1120.0105.

[11]. Gundechahuan, P. and Liu, H. (2014) 'INFORMS Tutorials in Operations Research Mining Social Media: A Brief Introduction', INFORM Tutorials in Operations Research., (November 2018), pp. 1–17. doi: 10.1287/educ.1120.0105.

[12]. He, W. et al. (2015) 'A novel social media competitive analytics framework with sentiment benchmarks', Information and Management, 52(7), pp. 801–812. doi: 10.1016/j.im.2015.04.006.

[13]. Hill, S. and Ready-campbell, N. (2011) 'Expert Stock Picker : The Wisdom of (Experts in) Crowds Expert Stock Picker : The Wisdom of (Experts in) Crowds', 15, pp. 73–102.

[14]. Holsapple, C., Hsiao, S.-H. and Pakath, R. (2014) 'Business social media analytics: Definition, benefits, and challenges: Completed research paper', 20th Americas Conference on Information Systems, AMCIS 2014, (2010), pp. 1–12.

[15]. J., K. K. (2006) 'Information security: management's effect on culture and policy', Information Management & Computer Security. Edited by M. T. E. Emerald Group Publishing Limited, 14(1), pp. 24–36. doi: 10.1108/09685220610648355.

[16]. Joseph, S. R., Letsholo, K. and Hlomani, H. (2017a) 'A Conceptual Framework for the Mining and Analysis of the Social Media Data', International Journal of Database Theory and Application, 10(10), pp. 11–34. doi: 10.14257/ijdta.2017.10.10.02.

[17]. Joseph, S. R., Letsholo, K. and Hlomani, H. (2017b) 'A Conceptual Framework for the Mining and Analysis of the Social Media Data', 10(10), pp. 11–34.

[18]. Kelly, B. et al. (2015) 'New Media but Same Old Tricks: Food Marketing to Children in the Digital Age', Current obesity reports, 4(1), pp. 37–45. doi: 10.1007/s13679-014-0128-5.

[19]. Kurniawati, Nargiza, B. and Graeme, S. (2013) 'The business impact of social media analytics', ECIS 2013 - Proceedings of the 21st European Conference on Information Systems.

[20]. Manuscript, A. (2015) 'Europe PMC Funders Group Global , regional and national prevalence of overweight and obesity in children and adults 1980-2013 : A systematic analysis', 384(9945), pp. 766–781. doi: 10.1016/S0140-6736(14)60460-8.Global.

[21]. Of, M., In, S. and Systems, I. (2008) 'Master of Science in Information Systems'.

[22]. Panayides, P. M. and Cullinane, K. (2002) 'Competitive Advantage in Liner Shipping: A Review and Research Agenda', International journal of maritime economics, 4(3), pp. 189–209. doi: 10.1057/palgrave.ijme.9100045.

[23]. Panisson, A. et al. (2012) 'On the dynamics of human proximity for data diffusion in ad-hoc networks', Ad Hoc Networks, 10(8), pp. 1532–1543. doi: 10.1016/j.adhoc.2011.06.003.

[24]. Peng, Y. et al. (2006) 'A systemic framework for the field of data mining and knowledge discovery', Proceedings - IEEE International Conference on Data Mining, ICDM, 7(4), pp. 395–399.

[25]. Qi, L. and Zhang, S. (2012) 'The Development of Customer Relationship Management System Based on Rough Set', Communications in Computer and Information Science, 315, pp. 328–333. doi: 10.1007/978-3-642-34240-0_43.

[26]. Sahay, A. (2016) 'Peeling Saunder's Research Onion', ResearchGate, (October), pp. 1–6.

[27]. Salkind, N. (2010) 'Encyclopedia of Research Design'. Thousand Oaks, California. doi: 10.4135/9781412961288.

[28]. Sengweni, W. et al. (2015) 'TITLE A Lotka-Volterra competition model for modelling market competition in the telecommunication industry : Case Study of Zimbabwe .'

[29]. Sharma, R. et al. (2010) 'Business analytics and competitive advantage: A review and a research agenda', Frontiers in Artificial Intelligence and Applications, 212, pp. 187–198. doi: 10.3233/978-1-60750-577-8-187.

[30]. Widener, M. J. and Li, W. (2014) 'Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US', Applied Geography. Elsevier Ltd, 54, pp. 189–197. doi: 10.1016/j.apgeog.2014.07.017.

[31]. Wuensch, K. L. (2011) 'Chi-Square Tests', International Encyclopedia of Statistical Science, pp. 252–253. doi: 10.1007/978-3-642-04898-2_173.

[32]. Zhao, X. and Ohsawa, Y. (2018) 'Sentiment Analysis on the Online Reviews Based on Hidden Markov Model', Journal of Advances in Information Technology, 9(2), pp. 33–38. doi: 10.12720/jait.9.2.33-38.

Cite this article as :

Kumbirayi Kwenda, Noreen Sarai, Tinashe Gwendolyn Zhou , "A Framework for Social Media Data Mining and Analysis to Product and Service Development- Case of The Zimbabwean National Art Gallery", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 194-209, January-February 2021. Available at
doi : https://doi.org/10.32628/CSEIT217121
Journal URL : https://ijsrcseit.com/CSEIT217121