Journal of Information
2021 Vol. 6, No.1, pp. 1-14
ISSN(e): 2520-7652
DOI: 10.18488/journal.104.2021.61.1.14

BIG DATA FRAMEWORKS FOR SITES AND PRODUCTS RECOMMENDATION

Ogbuju, E.[1+]
Ejiofor, V.[2]
Okonkwo, O.[3]
Onyesolu, M.[4]

[1]*Department of Computer Science, Federal University Lokoja; Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria.*
[1]*Email: emeka.ogbuju@fulokoja.edu.ng Tel: 2348032618951*
[2,3,4]*Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria.*
[5]*Email: ve.ejiofor@unizik.edu.ng Tel: 2348035611654*
[3]*Email: to.okonkwo@unizik.edu.ng Tel: 2348036739089*
[4]*Email: mo.onyesolu@unizik.edu.ng Tel: 2348039536257*


(+ Corresponding author)

Article History
Received: 22 March 2021
Revised: 26 April 2021
Accepted: 14 May 2021
Published: 4 June 2021

Keywords
Sentiment analysis
Collaborative filtering
Big data
CRISP-DM
Recommender
e-commerce.

ABSTRACT

The improvement of the IT infrastructure in an e-commerce platform is essential for both customer satisfaction and increased revenue. While different techniques had been applied towards achieving this, there is still need to engage customer feedbacks in providing an all-inclusive solution to the recommendation systems available in the e-commerce domain. The motivation is on making a more exact recommendation with the traditional collaborative system by mining the feedbacks and uncovering their sentiments using big data analytic systems. This paper describes the design of a big data framework that may be used for shopping sites recommendations and another that may be used for product(s) recommendations to prospecting customers. The use of the cross industry standard process for data mining is applied in proposing the new system. Although the techniques of Hadoop/MongoDB tools are described within the proposed designs, it concentrates mainly on the architecture and algorithm of the system in a holistic approach to enable the platform providers, e-commerce merchants and practitioners find a guided implementation of it using any tool of choice.

Contribution/Originality: This study documents the use of the cross-industry standard process for data mining methodology to design big data frameworks that implement the application of sentiment analysis of review texts as preprocessed input into the collaborative filtering algorithm for both sites and product recommendation.

1. INTRODUCTION

There is a global transformation in the economic sector across nations. Electronic or Digital Commerce, commonly known as e-commerce, is thriving globally. The e-commerce businesses have the ability to digitally go outside the country to draw upon customer bases which are definitely larger in the e-commerce context than in a physical retail context. However, setting up a new and compliant IT infrastructure that would effectively handle the demands of the (current and future) e-commerce is always a herculean task. The IT infrastructure is perhaps the most important component in setting up an e-commerce business. The other important components are the merchants of the products/services, the consumers of those products/services and the service providers who deliver the products/services. The IT infrastructure which connects the consumers with the merchants of the products/services must be built on a set of objectives that would meet the needs of the consumers. Top among the needs of the consumers include reducing customer navigation time by offering a smart product recommendation system, providing additional value by profiling, customizing or personalizing offerings, increasing customer loyalty by incentive offerings and receiving customer feedback on the general products/service presentations, image

quality, updated/timely offering, courier services and packaging. The IT framework for an e-commerce company must be based on these set of objectives. Applying requirement engineering procedures, these objectives must be translated into requirements for the IT infrastructure. The goal of the IT infrastructure would now be to collect a history of these requirements from their various locations online into one type of storage and build Machine Learning and analytics algorithms on top of them to bring solutions that would meet the objectives.

To achieve these, Big Data technology is the answer. The web holds a large number of user-generated contents which accounts for the big volume characteristics of Big Data. These contents address issues on products, people and political opinions on various blogs and social media channels. These contents are generated at a very high speed and in diverse formats like texts, videos and images thereby accounting for the high velocity and variety characteristics of Big Data. This paper is motivated by the enormous availability of text-based data on the Internet especially on shopping sites where product reviews is commonly provided. These unstructured texts are readily available and call for intelligent approaches to derive insights from them in order to enrich the e-commerce business. In the e-commerce world today, decisions about purchasing a new product need to be made very carefully. Due to the enormous availability of customers' reviews available for just a product, it is a very tedious task for a prospective customer to traverse this huge number of reviews in order to make just a purchasing decision. In this vein, product manufacturers who wish to receive and analyze their customer feedbacks in order to improve on quality, determine market trends or predict sales, are faced with these huge number of reviews in an overwhelming manner. The product sellers themselves also face the same challenge because it is easier for them to gain a competitive advantage when they present goods with positive customer goodwill for marketing.

These problems in all their unstructured nature are embedded in textual documents. The solution lies in the application of sentiment analysis. Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, 2015). Customer reviews are presented in texts. Hence, it requires the application of text mining to make automated meaning out of it. Text mining involves the pre-processing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyse these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results (Feldman and Sanger, 2007). The relevance of existing customer feedbacks is that it helps prospective or potential customers to measure product quality, durability and usability before they could commit to a buy. Sentiment analysis classifies these reviews into positive, negative or neutral.

Almost every website that provides sales and services do so through one form of recommendation or another to improve its service provision. Table 1 outlines some major websites using recommender services.

Table-1. Applications of recommender online.

S/N	Website	URL (www.)	Service
1	Book Crossing	bookcrossing.com	Books
2	Last FM	last.fm	Music
3	Netflix	netflix.com/ng	Movie
4	Movie Lens	movielens.org	Movie
5	Book Lens	grouplens.org	Books
6	Jester Joke	eigentaste.berkeley.edu	Jokes
7	Facebook	web.facebook.com	Friends
8	Twitter	twitter.com	Friends
9	Yahoo News	yahoo.com/news	News
10	Google News	news.google.com/news	News

These system does not make provisions for recommending a shopping site to a customer. Actually, no organization in business would want to recommend the services of other organizations at the expense of its own.

The system mainly operates a recommender service for goods/services. Therefore, choosing a shopping site for a customer is purely a manual process. The customer has to surf through the huge customer reviews for an informed decision making. If the customer is not satisfied with the feedback presented by previous customers, the customer may visit another e-commerce site and continue the search. This cycle continues until the customer eventually makes a decision for a purchase.

The current e-commerce system provides a platform for a customer to search for products, select the product of his/her choice and add same to a shopping cart. In a scenario where a customer wants to purchase more than one product, he/she continues with this search-and-select activity until the products had been added to the customer's shopping cart. The customer proceeds with the purchase of the product(s) by paying the total amount accrued in the shopping cart. While the search-and-select activity is ongoing, the e-commerce system may provide the customer with relevant recommendations of similar products that may be of interest to the customer. The customer may select a product from these recommendations thereby shortening his/her search time.

After making a purchase, the customer may return to the shopping site to write a review of his/her experiences with the product(s) and/or the site generally. A typical review contains the name of the reviewer (a verified customer), a text of the review, date and time of the submission of the review, a $1-5$-star rating indicating a level of total satisfaction (that is 5 stars) or complete dissatisfaction (that is 1 star) with the product. Note that the star rating may range from $0-10$ depending on the site. A screenshot of a typical review page is presented in Figure 1. The screenshot shows the feedback of two (2) customers; one (1) of which was totally satisfied with the product thereby rating it 5 stars while one (1) of the customers was not completely satisfied with the product thereby rating it 3 stars. A customer rating of 4 or 5 stars may also indicate satisfaction, 1 or 2 stars may indicate dissatisfaction while a rating of 3 stars may be classified as neutral or objective. In essence, the value of the rating is a function of the customer's experience with the product. This experience is also expressed in the review texts. Hence, the rating is the numerical value of the review text. So, some customers may quantify their experience well while others may not.

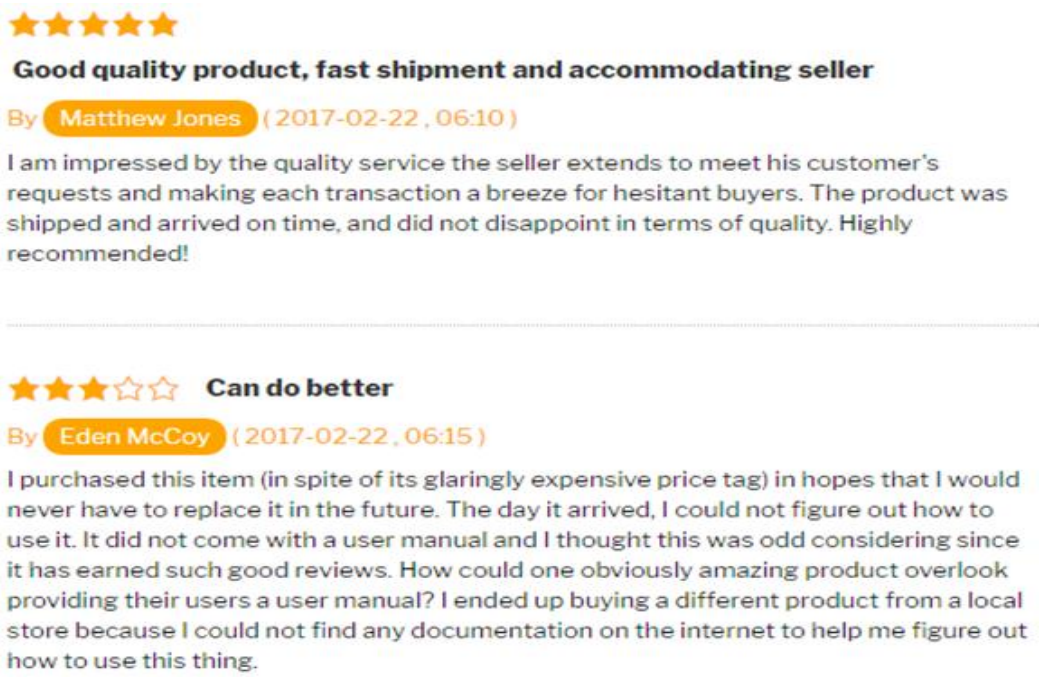

Figure-1. A Typical review page.
Source: https://www.cminds.com/wordpress-plugins-library/customer-reviews-plugin-wordpress/

The ratings are recorded by the e-commerce system and may further be used by the recommendation service to make recommendations to prospective customers through an algorithm known as Collaborative Filtering (CF). In

the current e-commerce system, a typical recommendation engine employs either an Item-Based Collaborative Filtering (IBCF) or a User-Based Collaborative Filtering (UBCF). The CF employs user-item preferences to offer information recommendation to customers.

In line with meeting the needs of the consumers in the e-commerce context, collaborative filtering had been used to recommend products and services to customers without due consideration to the customer feedbacks. Though collaborative filtering is one of the most successful methods to build recommender systems (Wang & Tang, 2015) it has its drawbacks in predicting a customer's preferences without consideration to the sentiments expressed by other customers on the product. It is against this background that this paper sets out to explore novel ways of integrating customer feedbacks and its sentiments into the recommendations provided. Hence, the aim of this work is to enhance the e-commerce IT infrastructure by proposing Big Data Analytics frameworks that would assist the recommendation of both shopping sites and products to prospective customers. The specific objectives are to (i). design a hybrid recommendation framework using sentiment analysis and collaborative filtering techniques, and (ii). build a smart product recommendation engine.

2. LITERATURE REVIEW

Esfandiari, Honarvar, and Aghamirzadeh (2016) recommended a system using a model-based collaborative filtering refinement model that makes a more exact recommendation from customers' feedback. The system runs with Spark processing model in Hadoop context to provide a very fast speed processing. This novel work functions with a combination of the collaborative-based algorithm and sentiment analysis of users' reviews. First, the collaborative filtering algorithm makes the recommendation. Second, the users' reviews were analysed and products with negative sentiment were removed from the result of the collaborative filtering thereby leaving the products with positive results to be the final recommended items. This approach is laudable as novel due to its application on Spark and its hybrid approach with both traditional collaborative filtering and Machine Learning. The data used in this study is Amazon Reviews for movies which have both the product information and the reviews for the products. However, the Machine Learning method applied in the sentiment analysis is the dictionary-based approach with SentiWordNet which gave the sentiment results at the document level. This approach had been noted for its lower accuracy in Mali, Abhyankar, Bhavarthi, Gaidhar, and Bangare (2016); Singh, Piryani, Uddin, and Marisha (2013) and Denecke (2008).

Gupta, Kumar, and Gopal (2015) worked on sentiment analysis using Hadoop streaming with Python. The researchers proposed sentiment analysis architecture and used the HDFS for data storage and Map Reduce program for processing. The research used 1680 clothing products reviews from Amazon and achieved 88% positive on the training dataset and 31% negative reviews on test datasets. This work used the traditional method of sentiment detection with lexicons. The Hadoop framework allowed it to operate in two phases where the Mapper will parse the given input file in the first phase and the Reducer will calculate the sentiments in the second phase. Other works that followed this approach was done by Alenezi and Mesbah (2015) and Batool, Khattak, Maqbool, and Sungyoung (2013). These works, however, did not proceed to make recommendations from the sentiments. Using the sentiments discovered from tweets or reviews to improve business situations is a more worthwhile attempt in the field of Big Data.

Luo, Miao, and Xiaoxia (2011) presented an approach called Feature-Grading which is a far-reaching algorithm used to make a recommendation of products in the e-commerce business. This procedure depends on the incorporation of feature mining, sentiment analysis and customers' browsing history. The procedure of Feature-Grading was divided into five key steps: extracting a general list of the feature set for a category of products; extracting modifier and negative words set; acquiring a particular list of features and grading; acquiring particular product weight set; and acquiring item weight set. In the end, recommendations are made by presenting the item

with the best position. The researchers used the genuine data of mobiles and their audits from the well-known e-commerce site Amazon.cn as an exploratory data.

Kumar, Sachdeva, Mahajan, Pande, and Sharma (2014) automated the process of collecting user reviews for any products or services online and analysing their sentiments. It shifts off irrelevant reviews and measures the sentiment score for the relevant ones. It presents the results to the business owners in an organized graph and charts expressing the actual sentiments in order to help improve the business experience with good customer satisfaction. Another work along this line was done by Song, Fan, Liu, and Tao (2011) from online review datasets. Song et al. (2011) proposed an approach to automatically extract features from product reviews and use it for sentiment analysis. The researchers deployed a pattern matching algorithm to extract the keywords. These works were limited to providing insight into the customer experiences. Hence, they were not extended to attempt recommendation with the sentiment result.

Flesch (2014) designed and deployed a Big Data Analytics dashboard that could provide a visual representation of analytics based on freely available open-source tools. The work was in the garment industry using Bangladesh Factory Disasters' data retrieved from Facebook Social Data Analytics Tool (SODATO). The developed tool could perform language analysis and visualization, word frequency analysis and visualization and sentiment analysis and visualization. This, however, cannot offer any form of recommendation for its users.

Salvi, Pawar, Kadu, and Dike (2016) proposed a shopping site that recommends best shopping site for product purchase based a sentiment analysis results. The system performs its analysis using the reviews of the product(s) from different sites. The usefulness of this system is to help customers save the searching time spent on different sites in search of a quality product. The methodology employed in this system is a categorization of the sentiments using the positive and negative words dictionary. This work and its approach, though similar to the work in this paper, lack the ability to categorise sentiments from other standardized lexicons like Finn and NRC lexicons. Lexicon approaches had been applied in Ray and Chakrabarti (2017) and in Ogbuju, Ejiofor, Ihinkalu, and Ajulo (2017). The reusability of the reviews for other applications or ratings is questionable because the system lacks the presence of a Data Lake.

Another work that uses sentiment analysis in information recommendation was proposed by Priyadharsini and Felciah (2017). This work demonstrated a framework construction, review collection, sentiment analysis, recommendation system and fake review monitoring. The recommendation in this framework works with analysing the reviews concerning a product and classifying same into positive or negative. The products that receive the positive result will be displayed in the recommendation panel. The work went further to detect fake reviews using the media access control (MAC) address that makes it easy to identify a review that had appeared more than once (spam). The system was experimented using a C#.Net framework as front end and SQL server as backend. Instead of collaborative filtering technique, a stochastic learning algorithm was implemented to analyse the reviews, ratings and emoticons. The novelty of this system is the identification of the unique reviews only. The unique reviews are used for sentiment analysis and further make a recommendation with them. However, this system was implemented for a particular e-commerce site whereas the new frameworks in this paper would extract reviews from multiple e-commerce sites about a product and analyse its sentiment for a recommendation of the best shopping site. Another major drawback of the framework/system is its inability to store the reviews as well as the sentiment lexicons in a Data Lake for reusability.

Nair and Sreelakshmi (2017) presented system architecture for movie recommendation using sentiment analysis approach. The architecture focuses on creating a rating for a movie using the users' comments. Using collaborative filtering, a user can suggest a movie to another. The system was based on a hybrid of collaborative filtering and sentiment analysis. They further provided an algorithm for a recommendation and another for sentiment analysis. The system, however, could not demonstrate actual recommendation with a combination of the

algorithms. In the new framework's approach, a holistic big data algorithm is implemented which will utilize the result of the sentiment analysis for a recommendation in the all-encompassing e-commerce domain.

Cheng and Lau (2015) proposed a seven-layer framework for Big Data stream analysis named Big Data Stream Analytics for online Sentiment Analysis (BDSASA). The BDSASA framework had the potential to leverage a probabilistic language model that can analyse the consumer sentiments embedded in millions of online consumer reviews. The framework holds the claim to analyse consumer sentiments in near real time. However, this claim had not been tested with empirical data for authenticity. This is a major limitation to the work. The new framework will not only be more elaborate, it will employ the use of smart methods in its analyses (instead of a probabilistic language model) and would be made easy for empirical implementation.

Wang and Tang (2015) provided an architecture for real-time processing of Twitter data for the 2012 presidential election in the U.S. It opened up the dynamics of the electoral process and public opinions towards it at real time. The system was deployed on the IBM cloud for streaming analytics and hold rules for tracking and collecting tweets for nine (9) Republican candidates. The system tokenizes the tweets, matches them to the candidates, determines its sentiments *(positive, negative, neutral, unsure)* and aggregate the results by the candidate, all at real time. Though this system was applied in the political landscape, the architecture is a generic one which can be easily adapted and extended for application in other domains. However, the architecture is unable to make conclusions about a candidate since the tempo of the tweets was still on-going. The new framework, though may employ a similar architecture, will be able to make conclusions and hence the recommendation of a product based on the conveyed sentiment.

Murugavalli, Bagirathan, Saiprassanth, and Arvindkumar (2017) proposed a system that uses sentiment embeddings to perform sentiment analysis on customer reviews. The aim is to allow an e-commerce site administrator to manage the inventory of goods available for sale on the site. That is, the sales of goods with bad reviews need to be discontinued, so the system would calculate the sentiment polarity of the goods using their reviews and allow the administrator to do the needful. The system architecture has modules for user details, administrator functions, sentiment analysis calculation and inventory. Though this system is laudable because it will allow only goods with positive sentiments to be promoted, it does not provide facilities for the recommendation of the good products and the discontinuation of the bad products are done manually. This is a major limitation because a lot of time would be needed by the administrator to traverse the results of the system before taking an action.

Kim, Park, Oh, and Yu (2017) proposed a context-aware hybrid system that integrates a convolution neural network into a probabilistic matrix factorization with statistics of product items. The system captures contextual information and considers Gaussian noise differently. The methodology also compares with non-deep learning-based approaches such as matrix factorization and collaborative topic regression. A Collaborative Deep Learning approach was also compared. The dataset used in the evaluation was two movie lens datasets and one Amazon instant video dataset. The context-aware system showed slightly better performance in comparison with the other systems compared. This system demonstrated the possibility of using a document context like reviews in a real-world scenario to solve information recommendation problem. However, the system was not extended to sentiment analysis for recommendation purposes.

The different methods and approaches discussed in this literature review had been applied in different industrial setups using recommendations. For example, the item clustering collaborative filtering technique had been applied at Amazon and YouTube (Linden, Smith, & York, 2003). Collaborative filtering is applied to Facebook, LinkedIn and MySpace to make friends recommendation (Kabiljo & Ilic, 2015). Netflix uses a hybrid system of both collaborative filtering and content-based filtering (Kismet, 2017). Other music recommender system like Pandora uses content-based approach and Last FM and Reddit uses a user-based collaborative filtering (Howe, 2009). The literature had ex-rayed related works that showed the approaches applied, the datasets used, the algorithms used,

and the frameworks employed. None of the approaches, algorithms and frameworks was holistic enough to solve the information overload challenge with a robust analytic system for both products and sites recommendation.

3. MATERIALS AND METHODS

The methodology adopted for this work is the Cross-Industry Standard Process for Data Mining (CRISP-DM). The justification for adopting this methodology is its object-oriented nature and the research poll carried out by Piatetsky (2014) which puts CRISP-DM as the leading methodology for data analytics projects. To meet the requirements of the evolving Big Data frameworks and analytics, practitioners are applying the steps of CRISP-DM with improved approaches as demonstrated by Larose and Larose (2015) and Taylor (2016). CRISP-DM is a holistic data mining life cycle methodology involving six (6) major steps/phases. It enabled the study to perform a detailed business understanding and identify the objectives of the data analytics project in the e-commerce domain. It allows data collection and review in line with the identified objectives of the work. It performs an iterative process of data preparation (or data cleansing) and modelling in order to draw conclusions or insights from the data. The next paragraphs show the phases of the methodology and how they were applied in the work.

Phase 1, Research Understanding: The primary objective of this paper is to develop analytic frameworks that would be able to use sentiments from customer feedbacks for both products and sites recommendation. The application of the frameworks should be able to reduce the collaborative filtering error and increase the recommendation precision using textual datasets. This objective was translated into a Big Data mining problem by identifying the online data sources and collecting/storing the datasets for the recommendation solution. This phase principally defined the problem and identified that the data would be sourced from the Internet by crawling the reviews from the merchant system.

Phase 2, Data Understanding: An Exploratory Data Analysis (EDA) was performed on the crawled dataset. The datasets were reviews collected from e-commerce websites and tweets collected from social media channels connected to an e-commerce website. The main fields and structure of both the review dataset and the tweet dataset were identified. The reviews would be loaded on a Hadoop Distributed File System while the tweets would be loaded into a MongoDB data store in BSON format.

Phase 3, Data Preparation: The main fields required for the task are the ReviewText of the review dataset and the text of the tweet dataset. These fields were selected specifically for sentiment analysis process. They were further cleaned and transformed for the modelling task in the next phase. The text cleansing involves turning the review into a corpus and removing odd characters like special characters, stop words, irrelevant words (e.g., will, can), and white spaces. The corpus is also lemmatized or stemmed to the root words like binding to bind, frustration to frustrate, etc. The other fields were dropped as they were not needed for further activities in the new system.

Phase 4, Modelling: This phase serves as the main analytic phase where the recommender systems would be developed. Two (2) recommender engines were proposed: the shopping recommender and the smart products recommender.

- First, to implement the shopping recommender, the reviews from different websites (now a corpus of cleaned texts) would serve as the input to the system and an expected output would be the recommended shopping site. The corpus from the first website would be turned into a term-document matrix (TDM). The TDM shows a matrix with a column as words and a row as the corpus of the reviews. The TDM enables the system to produce a word frequency and a word cloud of all the keywords used in the review. The next step would be to perform a sentiment analysis with the corpus. Two (2) lexicon-based approaches would be employed. They are the NRC Emotion lexicon approach and the Bing lexicon approach. The NRC expresses the total counts of the sentiments in eight 8 basic emotions – anger, anticipation, disgust, fear, joy, sadness, surprise, trust, and two sentiments of negative and positive. The Bing approach holistically expresses the sentiments in three classes – positive, neutral and negative. In essence, a corpus would be classified as

positive if the total number of positive sentiments outnumbers the total number of negative sentiments while the neutral sentiments may be accounted for in the NRC classification. After completing this step, the corpus from the second website would be turned into a TDM and the same process is repeated until all the reviews from the different websites had been analyzed. The shopping site recommendation can then be performed by comparing the results of the entire analysis. The site that scored the most positive result would be recommended to the customer.

- Second, to implement the smart product recommender, product reviews and their ratings from one e-commerce site would be collected for different products. Note that this is different from the shopping site recommender where the product reviews were collected from multiple e-commerce websites. The smart product recommender is implemented for one website only though it can be used on other websites. The first step would be to mutate the raw star ratings to binary values only (i.e., 5 stars to 1 and otherwise 0 or 4-5 stars to 1 and otherwise 0). The rationale is to have the highest rating as positive and others as negative. The review texts, in particular, are cleaned as above (Phase 3) to prepare it for sentiment analysis tasks. So, only the reviews and the ratings would be selected and divided into train and test datasets. This is required because the smart product recommendation technique would be using a supervised Machine Learning approach unlike the lexicon-based approach used in the shopping site recommendation. The Machine Learning model would be performed on the train dataset and validated on the test dataset. The corpus would be tokenized to obtain the keywords and a large sparse TDM would be created. A TF-IDF model would be adopted alongside a Deep Learning algorithm (Doc2Vec) to vectorize the keywords. The model would be trained by building a logistic regression with the x-axis as the independent variable bearing the fitted TF-IDF model on the train dataset and the y-axis as the dependent variable bearing the mutated ratings as sentiments. The result of the model performance promises to be high thereby assuring a more robust system. The predicted sentiment rates would be added to the initial review dataset and the results plotted. This process would be repeated for all the reviews of the other products. The next step would then be to feed the outputs of the predicted sentiments into the collaborative filtering approach and make product recommendation from it using both IBCF and UBCF. To achieve this, two (2) smart methods were proposed:
  - The Keep Method: This approach keeps all the products with negative reviews in the system but penalizes them with a low rating of 0.
  - The Kill Method: This approach takes the mean of the sentiments for each product and removes or eliminates the product(s) with the weakest mean from ever being recommended.

In any of the methods employed, a user-item matrix would be built as it is done in the traditional collaborative filtering system. The dataset would also be split into training and evaluation sets for the collaborative filtering models.

Phase 5, Evaluation: This phase is principally responsible for evaluating the result of the modelling phase. The Receiver Operating Characteristics (ROC) method of evaluation may be adopted. ROC uses classification/decision matrix (otherwise known as confusion matrix) to evaluate models. Plotting the Sensitivity (true positive rate) and 1-Specificity (false positive rate) on the axis of the ROC plot gives clear decision assessment. The precision metric may be used to evaluate the actual recommendations while Root Mean Squared Error may be used to evaluate the predicted ratings in the collaborative filtering.

Phase 6, Deployment: The generated reports and graphs that explain the new business situations and highlight the gains of the new recommendation engine over the old system would be presented. The engine may also be taken for deployment through the use of Application Programming Interface (API) to enable others to find an active use of it. The general system flowchart which helps to achieve the deployment is presented in Figure 2. It provides the system implementation details and shows the processes taken by a system analyst to retrieve the

required datasets. Furthermore, it shows the processes taken by a data scientist to perform the analytics and recommendation tasks in the system.

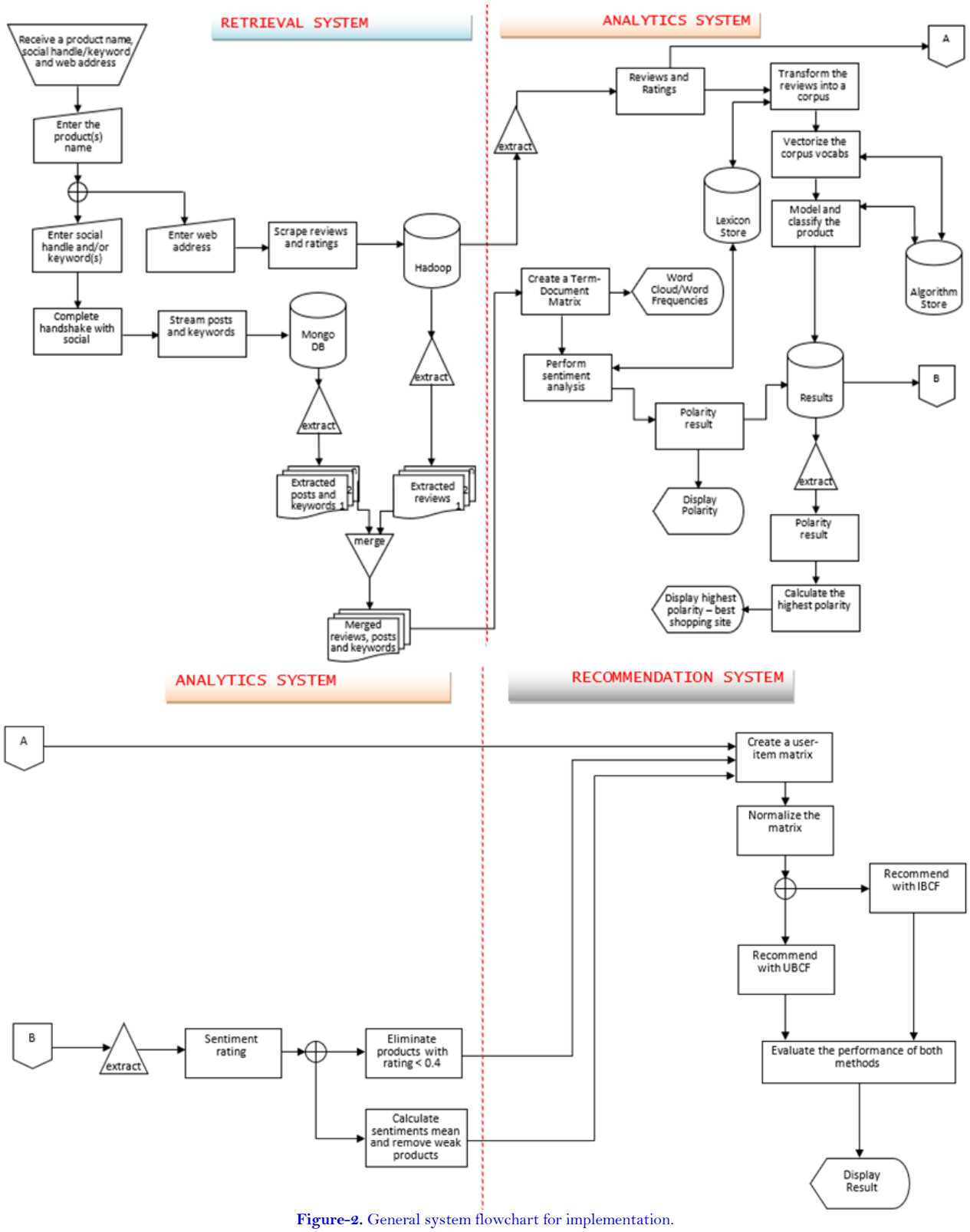

Figure-2. General system flowchart for implementation.

4. RESULTS AND DISCUSSION

The result of the proposed system is not a menu-driven system or a software application package; it is rather an analytics system that operates with subsystems using different modules. The high-level model is presented in

Figure 3 showing its operations in four (4) subsystems – acquisition subsystem, exploratory subsystem, analytics subsystem and reports subsystem.

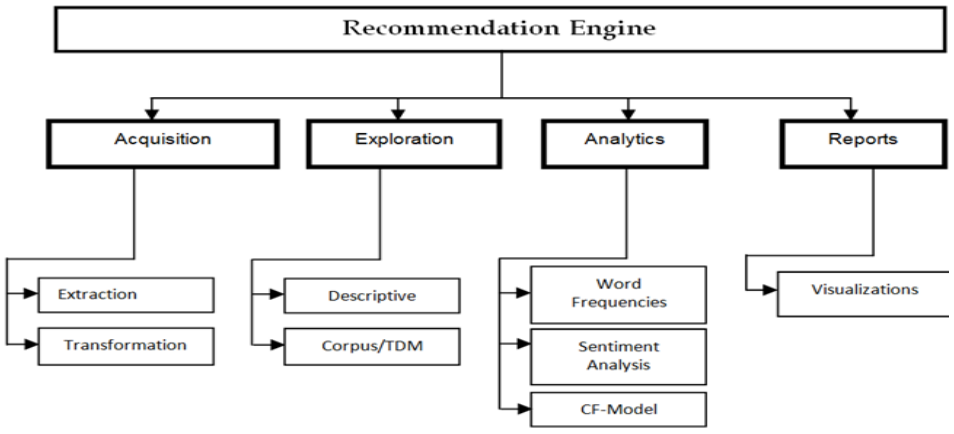

Figure-3. A High-level model of the proposed system.

There is a high cohesion between the subsystems in the high-level model. However, the subsystems are coupled separately to ensure that a modification in a subsystem does not necessarily need to affect the other subsystems. This is to facilitate maintenance and reusability of the analytics codes in any of the subsystems. The acquisition subsystem is decomposed into the extraction and the transformation modules to primarily collect and collate datasets for the system. The exploration subsystem is decomposed into the descriptive and Corpus/TDM modules to serve the data cleaning and preparation activities for the system. The Analytics subsystem is decomposed into word frequencies, sentiment analysis and CF-model modules to work in a holistic manner for the main data analytics operations of the system. The Reports subsystem takes care of producing visualizations like charts and graphs from the analytics subsystem. So, the whole model is coupled tightly to work together as a single system where each subsystem or module contributes a segment for a holistic performance. In practice, the system will basically perform three operations: retrieve datasets, analyze the datasets and recommend site or product.

Figure 4 shows a functional block diagram of the Site Recommendation framework. Reviews from different sites would be retrieved and stored in a Data Lake. These reviews are pulled from the Data Lake for unsupervised sentiment analysis operations. The result is a polarity score of positive and negative. The positive polarities are compared and the site with the highest positive polarity is recommended.

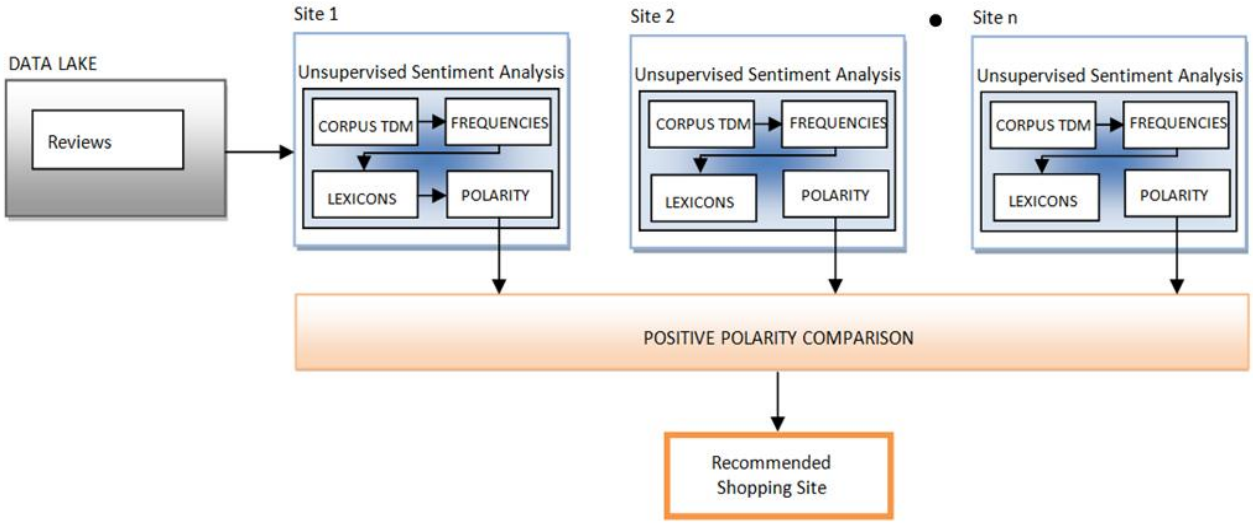

Figure-4. A Functional block diagram of the shopping site recommendation framework.

To recommend a product, a smart product recommendation framework would be applied. The functional block diagram in Figure 5 is used to show the flow of operations in the new smart product recommendation framework.

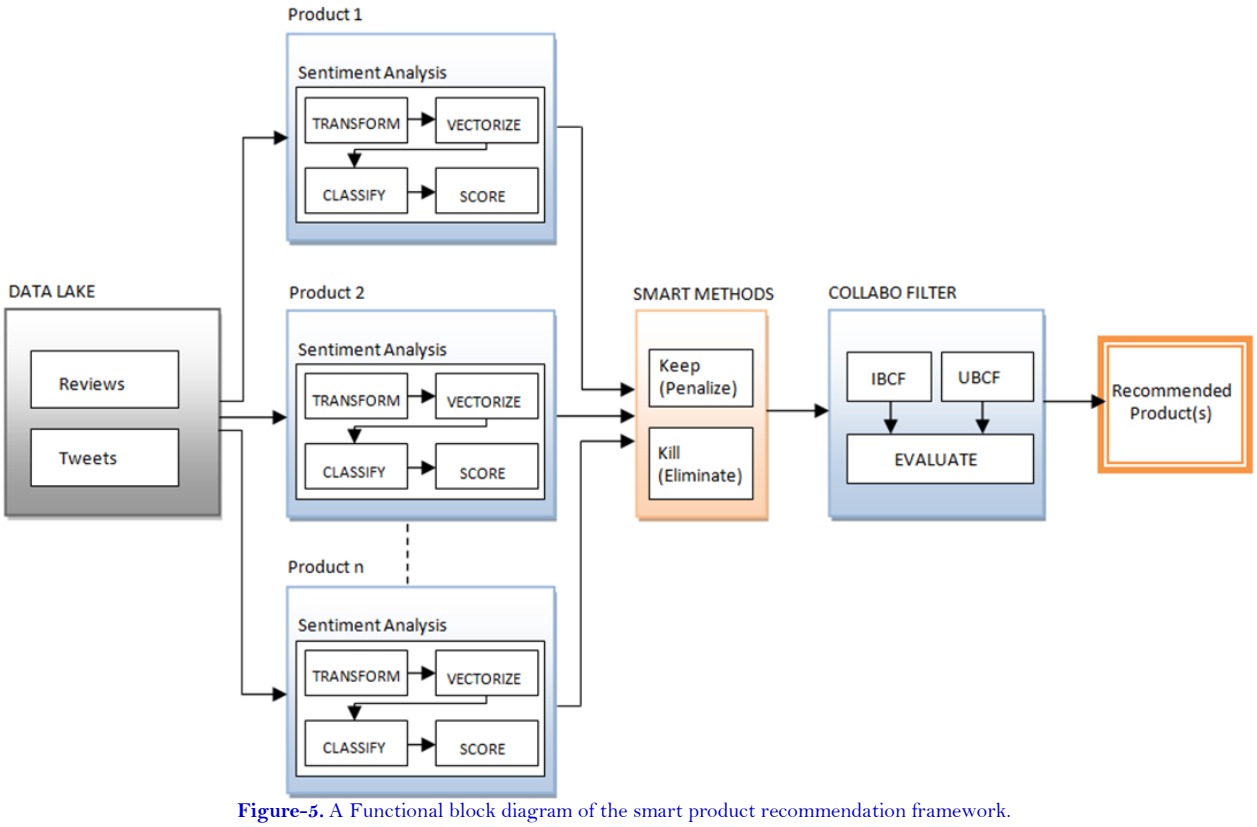

Figure-5. A Functional block diagram of the smart product recommendation framework.

In this framework, reviews for different products from a shopping site or tweets from an e-commerce social media channel for different products would be retrieved and stored in a Data Lake. A supervised sentiment analysis may be performed on the reviews for each product. A smart method is applied to the output to make it serve as an input into the CF system which eventually makes the product recommendation. The two smart methods were described in the modelling phase.

4.1. Advantages of the New System

The new system promises an effective business technique for the e-commerce sector especially in the efforts to reduce the prevailing information overload challenges. The significant of the recommender engine is on its ability in information filtering. It filters the high volume of items (products/services) available on the e-commerce website and presents to the user the most relevant information for the user's consumption. To justify the rationale for the new system, the following strengths and advantages were identified:

i) Data analytics becomes a mainstay of the e-commerce system thereby offering informed and insightful decision making for shopping sites recommendation. In the same manner, product recommendation becomes more precise with low error. This is demonstrated in the System Testing section.

ii) The new system is able to recommend suitable shopping sites for a chosen product thereby reducing the search time spent in looking for the product in other sites.

iii) The new system is able to extract reviews from different sites and analyse them for critical management decisions.

iv) The system can extract timely business intelligence about how products or services are perceived by customers.

v) The customer feedbacks are put to active use in the e-commerce system. The reviews were not only assisting prospective customers to make decision to buy a product as it is in the existing system but is now being used to make recommendations to the customer on new products. So, an improved customer interaction and feedback is being achieved.

vi) Customer's sentiments/emotions are duly analyzed to offer a personalized and targeted recommendation/marketing with a promise for improving the customer satisfaction rate.

4.2. Application Areas of the New System

The primary application area of the system is in the e-commerce domain. It can be applied to the shopping reviews page as well as the social network channels in the following ways, to:

i) Measure the impact of new products.

ii) Analyze the reactions or feedbacks of customers to company's news/events.

iii) Evaluate user experience in the delivery of customer services.

iv) Discover business trends and gauge customers' ideologies or bias to a brand or product.

v) Analyze consumers' product preferences and develop effective marketing and production strategies.

vi) Monitor the impact of products and services and collect appropriate social intelligence on them by government control agencies in order to protect the consumer community.

5. CONCLUSION

This work has been able to articulate Big Data frameworks by hybridizing the techniques of sentiment analysis and collaborative filtering to enrich the e-commerce business objectives. It has proposed an analytics system that is able to make smart recommendation for products and shopping sites using customers' opinions thereby contributing to solving the challenge of information overload. The system will be relevant to product manufacturers through its ability to provide automated ways of receiving consumer feedback about their products from different outlets in order to improve on the product in subsequent versions. If implemented with an API using the application of the Kill method, the recommender will be able to block off products with negative reviews from being recommended to new customers. The business implication of the system is that only good and relevant product(s) would be recommended to new customers. The system will discourage the sales and marketing of bad products hence it will compel manufacturers to address the concerns raised by the customer reviews. It will discourage product sellers from continual sales of products with negative reviews. This will bring sanity into the e-commerce world; it will help create an e-commerce system with good and trusted products that actually meet the needs of the customers. Our future work will show the implementation of the proposed frameworks in practice, the model developed from them and their various testing/evaluation results.

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
Acknowledgement: All authors contributed equally to the conception and design of the study.

REFERENCES

Alenezi, S., & Mesbah, S. (2015). Big data spatial analytics in social networks using Hadoop. *International Journal of Computer Applications*, *128*(14), 21-26. Available at: https://doi.org/10.5120/ijca2015906745.

Batool, R., Khattak, A. M., Maqbool, J., & Sungyoung, L. (2013). *Precise tweet classification and sentiment analysis.* Paper presented at the IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS).

Cheng, O., & Lau, R. (2015). Big Data stream analytics for near real-time sentiment analysis. *Journal of Computer and Communications*, *3*(1), 189-195. Available at: 10.4236/jcc.2015.35024.

Denecke, K. (2008). *Using sentiwordnet for multilingual sentiment analysis*. Paper presented at the In 2008 IEEE 24th International Conference on Data Engineering Workshop, IEEE.

Esfandiari, K., Honarvar, A., & Aghamirzadeh, S. (2016). Improvement of recommender systems considering big data of users' comments on chosen items. *Journal of Fundamental and Applied Sciences*, *8*(2), 882-891. Available at: https://doi.org/10.4314/jfas.8vi2s.141.

Flesch, B. (2014). Design, development and evaluation of a big data analytics dashboard, [M.Sc Thesis]. Retrieved from: http://studenttheses.cbs.dk/bitstream/handle/10417/4945/benjamin_flesch.pdf?sequence=1.

Gupta, P., Kumar, P., & Gopal, G. (2015). Sentiment analysis on Hadoop with Hadoop streaming. *International Journal of Computer Applications*, *121*(11), 0975 – 8887.

Howe, M. (2009). Pandora's music recommender. Retrieved from: https://courses.cs.washington.edu/courses/csep521/07wi/prj/michael.pdf.

Kabiljo, M., & Ilic, A. (2015). Recommending items to more than a billion people. Retrieved from: https://engineering.fb.com/2015/06/02/core-data/recommending-items-to-more-than-a-billion-people/.

Kim, D., Park, C., Oh, J., & Yu, H. (2017). Deep hybrid recommender systems via exploiting document context and statistics of items. *Information Sciences*, *417*, 72-87. Available at: https://doi.org/10.1016/j.ins.2017.06.026.

Kismet, K. (2017). Netflix: Recommendations worth a million. Retrieved from: https://rpubs.com/kismetk/Netflix-recommendation.

Kumar, S. P., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A. (2014). *An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites*. Paper presented at the 5th International Conference of the Next Generation Information Technology Summit (Confluence).

Larose, D. T., & Larose, C. (2015). *Data mining and predictive analytics* (2nd ed.). New Jersey: John Wiley & Sons, Inc., Hokoken.

Linden, G., Smith, B., & York, J. (2003). *Amazon.com recommendations item-to-item collaborative filtering*. Paper presented at the IEEE Computer Society.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and Emotions*. UK: Cambridge University Press.

Luo, Y., Miao, F., & Xiaoxia, Z. (2011). *The design and implementation of feature-grading recommendation system for e-commerce*. Paper presented at the IEEE International Conference on Information and Automation (ICIA).

Mali, D., Abhyankar, M., Bhavarthi, P., Gaidhar, K., & Bangare, M. (2016). Sentiment analysis of product reviews for E-commerce recommendation. *International Journal of Management and Applied Science*, *2*(1), 127-130.

Murugavalli, S., Bagirathan, U., Saiprassanth, R., & Arvindkumar, S. (2017). Feedback analysis using sentiment analysis for e-commerce. *International Journal of Latest Engineering Research and Applications (IJLERA)*, *2*(3), 89 – 90.

Nair, A. S., & Sreelakshmi, K. (2017). Movie recommendation system using sentiment analysis. *International Journal For Trends In Engineering & Technology*, *24*(1), 28-30.

Ogbuju, E., Ejiofor, V., Ihinkalu, O., & Ajulo, E. B. (2017). Sentiment analysis for rules-driven instant messaging. *Confluence Journal of Pure and Applied Sciences (CJPAS)*, *1*(1), 241-155.

Piatetsky, G. (2014). CRISPDM: still the top methodology for analytics, data mining, or data science projects [Web log post]. Retrieved from: http://www.kdnuggets.com/.

Priyadharsini, R. L., & Felciah, M. L. (2017). Recommendation system in e-commerce using sentiment analysis. *International Journal of Engineering Trends and Technology (IJETT)*, *49*(7), 445-450.

Ray, P., & Chakrabarti, A. (2017). *Twitter sentiment analysis for product review using lexicon method*. Paper presented at the Paper presented at the International Conference on Data Management, Analytics and Innovation (ICDMAI) 2017. IEEE.

Salvi, K., Pawar, V., Kadu, P., & Dike, O. D. (2016). Shopping site recommendation using sentiment analysis. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(4), 6990-6993. Available at: 10.15680/IJIRCCE.2016.0404103.

Singh, V. K., Piryani, R., Uddin, A. P., & Marisha, W. (2013). *Sentiment analysis of textual reviews, evaluating Machine Learning, unsupervised and sentiwordnet approaches.* Paper presented at the Proceeding in IEEE 2013 5th International Conference on Knowledge and Smart Technology (KST).

Song, H., Fan, Y., Liu, X., & Tao, D. (2011). *Extracting product features from online reviews for sentimental analysis.* Paper presented at the Proceedings of the 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT).

Taylor, J. (2016). Decision modeling and notation standard [PowerPoint slides]. Retrieved from: http://www.decisionmanagementsolutions.com/solutions/big-data-decisions/.

Wang, J., & Tang, Q. (2015). Recommender systems and their security concerns.Retrieved from: https://eprint.iacr.org/2015/1108.pdf