ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse22942020.pdf
https://doi.org/10.30534/ijatcse/2020/22942020

Telecom Big Data: Social Media Sentiment Analysis

Salam Fraihat[1], Mahmoud AlMomani[2], Malak Fraihat[3], Mohammed Awad[4]
[1]Princess Sumaya University for Technology (PSUT), Jordan, s.fraihat@psut.edu.jo
[2]Princess Sumaya University for Technology (PSUT), Jordan, mr.1momani@gmail.com
[3]Princess Sumaya University for Technology (PSUT), Jordan, fraihatmalak@gmail.com
[4]American University of Ras Al Khaimah, UAE, mohammed.awad@aurak.ac.ae

ABSTRACT

Despite Big Data being one of the trending buzzwords, it is possibly one of the least comprehended terms in business. Big Data revolutionized the way business is organized and managed. Nevertheless, massive data is not as clear as it may appear, especially that originated from social media. Big Data is not only concerned with the amount of data that has changed, but also with the velocity and the structural complexity of the data. Thus, the demand has emerged to look for other techniques and procedures that are structured to deal with this vast amount of data as well as drawing insights out of them. One of the most critical and challenging tasks in social media related data is processing the natural language and defining the implicit and explicit meaning behind unstructured text data. This task refers to understanding and analyzing the text sentiment. Sentiment Analysis (SA) is the process of recognizing the intention or the emotional condition of a client articulation. This paper presents a Business Intelligence (BI) framework using social media data to analyze and define the sentiment meaning of data that comes from digital channel feeds. The implemented Business Intelligence framework consists of four main phases: first, ingesting data from multiple sources: structured and unstructured data. Second, transforming, aggregating, and processing the data through an in-memory processing layer. Third, building the Natural Language Processing (NLP) model to define the meaning behind the customer message or complaint. Finally, the visualization phase, which presents the final output and insights in a single dashboard.

Key words: Big Data, Business Intelligence, Machine Learning, Natural Language Processing, Sentiment Analysis

1. INTRODUCTION

Business Intelligence is an organized way of preparing and using the information to drive business insights. BI is extremely useful in transforming raw data into information that can be used to make decisions [1]. In addition to its innovative method of monitoring and improving companies' and organizations' performance, BI also became a promoting brand for creating and developing businesses [2]. Besides, BI solutions are confronting enormous difficulties because of the dramatic increase of Big Data and Big Data techniques [3]. Thus, utilizing Big Data to enhance BI became a major issue for business [4].

Big Data refers to data sets that cannot be handled using traditional tools, whether in terms of saving, management, or analysis; due to some tool limitations regarding the data characteristics such as size, data types, and speed of flow [5]. Big Data provides companies and organizations with extraordinary insights, hence, enabling them to make better decisions more rapidly and, in return to offer some incentives to their clients [6]. The Big Data collections appear to provide access to forms of learning that were hitherto thought to be inaccessible and insights that were once difficult to be drawn. The value of Big Data into business appears to be irreproachable, and little, if any, would contend that it must be overlooked. Presently, the real issue seems to be how to utilize Big Data set and how to deal with it in an effective, productive manner [6].

The sentiment analysis topic is vast but powerful for extracting useful insights from social media platforms. It is advantageous for companies as it gives a full overview and offers a comprehensive option for a specific subject. Sentiment Analysis indicates the utilization of Natural Language Processing and text examination to automatically identify, extract, and learn about emotional states and subjective information. Sentiment analysis is widely used to analyze reactions and thoughts of clients' items [7]. Sentiment analysis aims to determine the stance of a speaker or a writer concerning the overall contextual polarity or emotional reaction to a document, interaction, or event. An essential task in sentiment analysis is classifying the polarity of a particular content; this could be at a document level, sentence level, or a word level to the positive sentiment, negative sentiment, or neutral sentiment [7]. Sentiment analysis for a language typically relies on manual or semi-automatic constructed lexicons found in dictionaries or corpora. The accessibility to these resources empowers the creation of rule-based sentiment analysis or the construction of training data for

classification tasks [8]. Sentiment analysis can help companies find what their customers think of them, their products, and services. Also, it helps businesses to better understand what their customers think of other competitors, which eventually helps find many hidden patterns such as a customer's intention to churn. Natural Language Processing can directly help companies decide what actions to take to ensure success.

Analyzing social media interactions after launching a specific campaign can indicate the campaign's performance and the trend to follow for any future related actions. Thus, customer experience can be determined via sentiment analysis.

Impressions, reactions, likes, comments, tweets, retweets, loves, and clicks are metrics in which a company can analyze its customers' interactions and implicitly listen to their voice. Many businesses and industrial domains have adopted social media sentiment analysis. This paper describes the implementation of a complete Business Intelligence solution beginning with the data extraction process till the final process of presenting the insights on a dashboard. The data in this paper is gathered from Telecom customers' complaints over various platforms. We analyzed these complaints using Natural Language Processing and sentiment analysis. The implemented BI solution consists of several phases. First, the three main phases: Extract, Load, and Transform. Then, the processing, analyzing, and modeling phases. Finally, the visualization phase, which presents the final output and other insights on a dashboard.

People, who use digital channels such as the Internet and social media to consume content, engage with brands and carry transactions are known as digital customers [9]. In this paper, we focus on three main social media channels: Facebook, Twitter, and the company's mobile application. The BI solution consists of two major components, namely the sentiment analysis of the complaints and customized reports to monitor the company's customer service response.

This paper presents a Business Intelligence solution to analyze, handle, and follow the digital customer complaints messages, which cannot be handled using the traditional databases and data warehouse solutions. Hence, Big Data storage and process (Apache Spark) and NLP techniques are applied on gathered data to define the customer sentiments and to display different measures analysis on a dashboard. This BI solution's motive is the need to hear from customers to ensure customers' satisfaction by analyzing their experience and eventually having a predictive customer behavior model. Thus, the company guarantees a lower churn rate and a higher retention rate.

The rest of this paper is organized as follows: section 2 presents the literature review; section 3 highlights the system requirements; section 4 illustrates the solution's machine learning and NLP phase; section 5 contains the experiment results; section 6 describes the proposed dashboard, and Section 7 concludes the paper.

2. RELATED WORK

Martin et al. discussed in [1] the performance of a BI project that uses data mining techniques. They used Real Genetic Algorithm to predict the quantitative business performance. On the other side, they used Ant Miner algorithm to predict the qualitative business performance. Each user receives a customized report that contains the right information; these reports are built using Fuzzy Multi-Criteria Decision Support System (FMCDS). Kusuma, Rivai, and Wang proposed a BI infrastructure of medical records to facilitate the detection of possible dangerous diseases [10]. CECÍLIA reviewed the adoption of BI in companies, specifically in the retail chain highlighting the high importance of the requirement engineering stage and how it can cause significant problems in case of a misunderstanding [11]. Finally, the author presented, from manager's perspective, how the decision making improved with help of BI [11]. Barone et al. in [12] conducted a BI modeling research that is a hospital use case (Toronto hospital) to test and examine whether the implementation of BI is supported by models founded on business views. Fortunately, they found that BI projects enhance the communication of project teams and stockholders and that the business modeling is a vital activity to correctly capture the BI requirements.

In their paper, Atoum and Al-Jarallah shed some light on the benefits of implementing big data analytics to improve the health care industry [13]. According to their research, the use of big data can improve the operational efficiency in addition to other benefits [13]. Business Intelligence solutions have been implemented in several other domains such as real estate [14], mobile money systems [15], and stock market [16].

3. REQUIREMENTS AND ANALYSIS

This section describes the model requirements and provides analysis on it.

3.1 Business Objective

The business objective is to aggregate customer experience data from social media with customer profile data, and call details data to better handle and follow customers' messages and complaints. This aggregation will enable the company to directly hear from customers and accordingly evaluate the company's action process in an effective manner. The output of this solution is an online dashboard, where users can draw insights out of the company's business data. The insights will allow assessing the company's situation accurately. Thus, making informed decisions to improve the customer's journey. Users of the mentioned BI dashboard can view data and other related insights regarding the customer's interactions on the official company's social media channels, specifically, on Facebook, Twitter, and the company's mobile application.

3.1 Data System

The solution uses Data Lake as a tool to handle Big Data with all its characteristics: volume, complex, and speed. In the proposed solution we select the Apache Hadoop platform to store all available data and aggregate them in a multidimensional analysis to provide a centralized data that is accurate. The data side consists of four main phases:

A. Source of Records:

Sources of the data are mainly the Call Detail Records (CDRs), which provide the details of phones and other telecommunication transactions. Dimelo which feeds the data from social media websites. In this BI solution it will feed data related to both Facebook and Twitter, and the last source of data is the Customer Relationship Management (CRM), that is an oracle based system used in the telecom industry to manage, track, and analyze the interactions with customers. Figure1 below shows the entity relationship diagram of the used data. Four entities are presented: subscriber Info, Dimelo, CDR, and Digital_User.

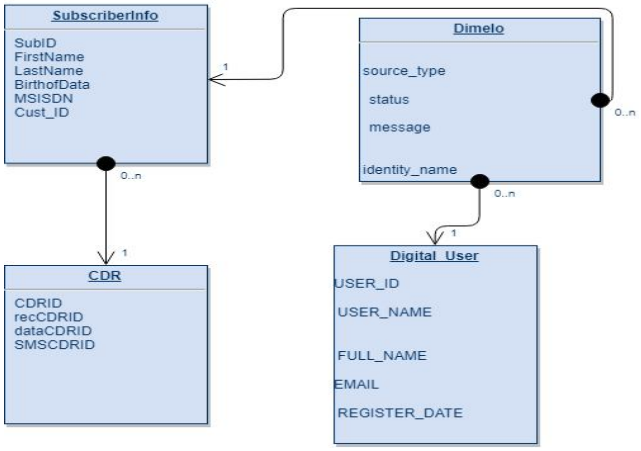

Figure 1: Entity Relationship Diagram

B. Data Lake Phase:

Data Lake was implemented as a storage layer using commodity hardware and open source software: Apache Hadoop Data Lake as proposed solution. Figure 2 shows data lake architecture and the apache Hadoop parts in details: HDFS is used to store the unstructured and semi-structured data and apache kudu to store the structured data. ELT, which is Extract, Load, Transform, is the process used to extract data from data sources and ingest into the Hadoop platform.

C. Data Ingestion Phase:

(1) Data ingested from relational database. In our case: Apache Sqoop; as it is designed to transfer data from oracle based system to Hadoop, (2) ingest social media data in real time mode (proposed Apache Kafka, and Apache Flume, (3) ingest system files, and logs in near real time mode, proposed Apache Flume. Flume is generally used for large data file batches and log data files. CDR's are very large files with constant change, hence, making flume the most suitable service to collect and aggregate this data.

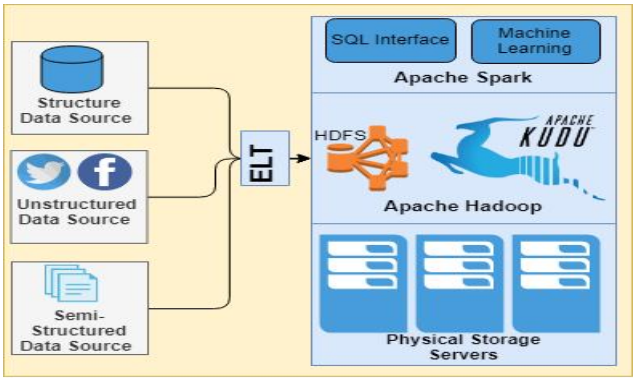

Figure 2: Data Lake Architecture

D. Data Transformation Phase:

Transformation layer using Apache Spark is implemented and all transformation is applied. Additionally, the preparation and cleaning processes take place in this layer. Furthermore, the machine learning model is built, and NLP techniques are used in this layer as well.

E. Data Visualization Phase:

Business Intelligence dashboard is built using Power BI tool.

A summary of all phases is shown in Figure 3. The first process for BI use case is ingesting data from all sources and integrating them together. The second process is storing data ingested from first process. The third process is regarding data transformation and aggregation to a useful format and performing in-memory analysis to achieve fast processing. The last process is the modeling and visualization the collected data.

Three software are used for the data extraction and ingestion part of this BI solution, namely: Scoop, Flume and Kafka. They are used to extract data from respective sources and ingest them in the Hadoop storage system. Apache Scoop is built to extract data from relational database servers such as MySQL or Oracle and transfer them to the Hadoop. Apache Flume is used to load bulk of streaming batch logs data files from various data producers into Hadoop. Apache Kafka is used for real time data. Figure 4 illustrates the interaction between the mentioned software.

Apache Spark is an open-source engine designed for fast large complex interactive queries since it is based on an in- memory distributed computing analytics approach. In other words, data is queried from the Random Access Memory (RAM) instead of the hard disk. Consequently, no I/O operations are performed. Apache Spark packs a number of libraries for machine learning modeling. Spark MLlib is a machine learning library with a framework to build ML pipeline.

PowerBI is a data visualization tool, powered by Microsoft, used for business analytics. PowerBI brings data to static or interactive dashboards and reports to get more of the business data.

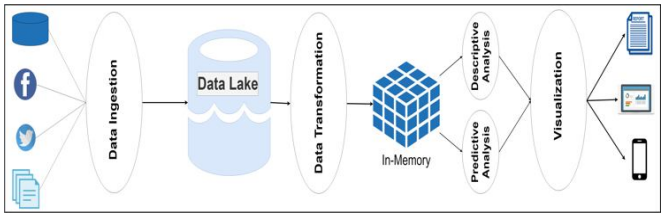

Figure 3: Information and Data Architecture

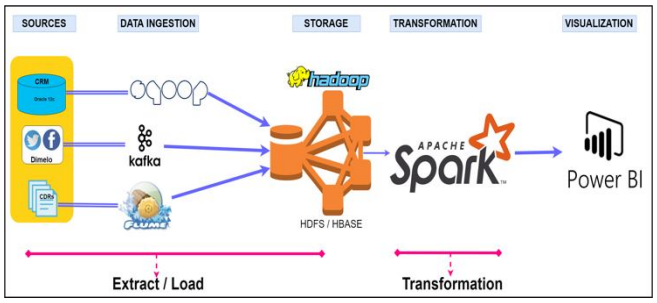

Figure 4: Technical and Product Architecture

4. NLP AND MACHINE LEARNING MODEL

The main objective of this solution is to extract sentiment from messages and complaints of digital customers. As illustrated in Figure 5, every NLP and machine learning system can be divided into three main stages: pre-processing stage. In this stage the dataset is pre-processed by a set of operations to ensure it is in a form that is ready for the next two stages. The second stage is feature extraction stage, in which a set of features are defined and extracted from the dataset. Those features will be used in the last stage that is Training a Machine Learning Model stage. In this stage, the features extracted from dataset are used as input to the machine-learning algorithm to train its parameter. The architecture of the proposed solution is shown in Figure 5.

4.1 Data Preprocessing

Preprocessing operations are needed to process the dataset and reduce the variations and unnecessary components in the dataset. The importance of this stage comes from its effect on machine learning and the learning process. As the dataset has a clear and unambiguous structure, this will help and increase the accuracy of the learning process of the machine learning model. In the solution a set of preprocessing tasks was used as explained below:

A. Normalization:

normalizing the dataset means removing unnecessary symbols and letters from the dataset. In this BI project, pyArabic Python library was used for this task.

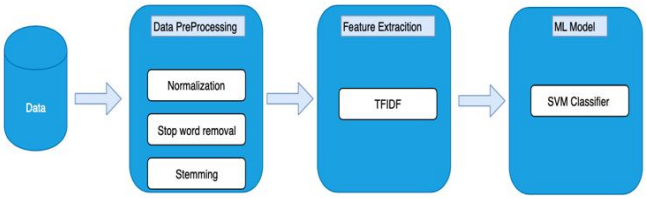

Figure 5: NLP Modelling Pipeline

B. Stop word removal:

stop words are usually referring to the most common words in a language. These words should be generally removed from a text dataset. Since they are the majority of the words and will reduce the presence of unique and important words. In this BI project, Python NLTK predefined list of Arabic stop words was used.

C. Stemming:

stem a word is to reduce it to its root. This will decrease the variations between words and reduce them to a set of unified word. In this BI project, Python NLTK ISRI Arabic stemmer is used.

4.2 Feature Extraction

For the feature extraction, Term Frequency-Inverse Term Frequency (TF-IDF) [17] is used as a feature inside the solution. The TF-IDF is a numerical measurement that is expected to reflect how vital a term is to a document in a collection or corpus. The measure TF-IDF is the most known in the literature for the calculation of the frequency representation of documents. The term's weight increases in a proportional manner in accordance to the number of occurrences of the term in the document. Additionally, the term's weight varies according to its frequency in the corpus. The TF-IDF is a proved weighting method, where effectiveness has been demonstrated in several works.

The weight of a term t_i in a document d_j is calculated as follows:

$$TF - IDF\ (ti, dj)\ =\ TF\ (ti, dj)\ x\ IDF\ (ti) \qquad (1)$$

The $TF\ (ti, dj)$ is calculated by:

$$TF(t_i, d_j) = \frac{f_{i,j}}{\sum_k f_{k,j}} \qquad (2)$$

Where, $f_{i,j}$ is the number of times where the term t_i appears in the document d_j and the denominator is the total number of terms in the document d_j.

The inverse document frequency (IDF) measures the importance of the term in the corpus. The objective is to give more weight to terms that appear in some documents. This

weight is calculated by considering the logarithm of "inverse" relative document frequency that containing the descriptor in the corpus:

$$IDF(t_i) = log_2 \frac{|D|}{|DF(t_i)|} \quad (3)$$

where, |D| is the number of documents in the corpus, and |DF (t_j)| the number of documents that contain the term t_k.

4.3 Machine Learning Modeling

After preparing the dataset and extracting the features from it, the training of a machine learning model comes in. Support Vector Machine (SVM) algorithms are used as the machine learning model for the learning process [18].

5. EXPERIMENT RESULTS AND DISCUSSION

This section presents the experimental results and discussion of models in addition to a comparative analysis for the models used to predict Sentiment Analysis.

5.1 Dataset

The used dataset to train the model is a Twitter Dataset for Arabic Sentiment Analysis [7], which consists of two attributes for 2000 tweets and a class for each tweet. Table 1 below shows a sample from dataset.

Table 2 shows the attribute information of positive and negative classes in the twitter dataset.

10-fold cross-validation technique was used to evaluate the models, then mean of the 10-folds for the evaluation's measures was calculated.

5.2 Performance Measures

In this paper, precision, recall, f-measure, and accuracy are used as the evaluation metrics.

Where: TP true positives, TN true negatives, FP false positives, and FN false negatives.
Accuracy: the percentage of correctly predicted examples to all the examples [19].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision: measures how many of the record classified as positive are actually positive [19].

$$precision = \frac{TP}{TP + FP} \quad (5)$$

Recall: measures how many of the total positive records were classified correctly as positive [19].

$$recall = \frac{TP}{TP + FN} \quad (6)$$

After applying the SVM model on the Twitter dataset, the results were significantly good with 88% precision as shown in Table 4.

Table 1: Sample of Twitter dataset

Tweet	Class
الوضع سيء، ما قدرت أصبر! الجو كثير حامي والمكان كثير ضيق والناس فوق بعضها البعض، الله بعين.	0
الحل الوحيد هو القرب من الله وذكره (الا بذكر الله تطمئن القلوب (قال الله تعالى	1

Table 2: Attribute information of each class of the Twitter dataset

	Positive Samples	Negative Samples
Total tweets	(1000) 50%	(1000) 50%
Total words	7189	9769
Avg. words in each tweet	7.19	9.97
Avg. characters in each tweet	40.04	59.02

Table 3: Tweets affiliation between classifier prediction and reality

	Positive Tweet	Negative Tweet
Prediction in Positive class	TP	FN
Prediction in Negative classes	FP	TN

Table 4: Machine learning SVM modeling results

Precision	Recall	F-measure	Accuracy
0.88	0.77	0.81	0.82

6. BI DASHBOARD

The results are represented in a dashboard. The dashboard has two main interfaces as shown in Figure 6. (1) The overview, which contains five graphs about the customer base information: gender distribution graph, age distribution graph, type of subscription, source of the customers data (weather it is from Facebook, twitter or the company's mobile application), and the geographical distribution of the customers graph. (2) The second part of the dashboard's interface displays all of the customer's complaints with their negative/positive score that is found through the previously explained prediction step. Furthermore, the dashboard has a slider to filter data per some features such as channel type and subscriber type. Additionally, the dashboard has the ability to view a specific time period as needed.

7. CONCLUSION AND FUTURE WORK

Applying BI techniques and analytics with Big Data enables enterprises to get unprecedented knowledge and insights. Therefore, enhancing customer service and revenue. The main factor in the success of this process is the evolution of technology and its ability to deal with Big Data. Without the assistance of this technology, these data would have remained hidden along with its value.

Big Data analytics has reshaped the business world and will keep doing so for the foreseeable future, particularly as Big Data technology becomes more advanced.

We believe our solution contributes to increasing customer satisfaction. The customer is the capital of almost any enterprise. Thus, customer preservation is a substantial goal. In this context, sentiment analysis facilitated figuring out customers' opinions, feelings, and needs, which eventually helped to improve customer experience.

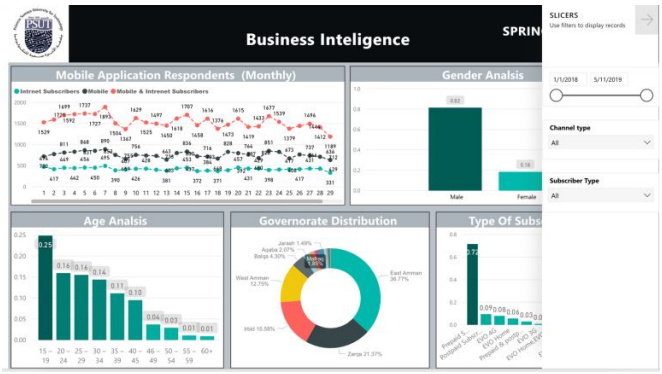

Figure 6: BI Dashboard Overview Interface

REFERENCES

1. A. Martin, T. M. Lakshmi, and V. P. Venkatesan. **A Business Intelligence framework for business performance using data mining techniques**, *2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET)*, 2012.
2. E. Turban and L. Volonino, ***Information technology for management: improving strategic and operational performance.*** Hoboken, NJ: John Wiley & Sons, 2011.
3. S. Fan, R. Y. Lau, and J. L. Zhao. **Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix,** *Big Data Research*, vol. 2, no. 1, pp. 28–32, 2015.
4. Z. Sun, H. Zou, and K. Strang. **Big Data Analytics as a Service for Business Intelligence**, *Open and Big Data Management and Innovation Lecture Notes in Computer Science*, pp. 200–211, 2015.
5. M. Sewell. **Ensemble learning,** RN, vol. 11, no. 02, 2008.
6. C. Kimble and G. Milolidakis. **Big Data and Business Intelligence: Debunking the Myths**, *Global Business and Organizational Excellence*, vol. 35, no. 1, pp. 23–34, 2015.
7. Abdulla, N.A., Ahmed, N.A., Shehab, M.A. and Al-Ayyoub, M., 2013, December. Arabic sentiment analysis: Lexicon-based and corpus-based. In 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT) (pp. 1-6). IEEE.
8. S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth. **Multilingual sentiment analysis: from formal to informal and scarce resource languages,** *Artificial Intelligence Review*, vol. 48, no. 4, pp. 499–527, 2016.
9. "Digital Customer - Gartner IT Glossary." [Online]. Available: https://www.gartner.com/it-glossary/digital-customer/. [Accessed: 20-May-2020].
10. Dewo Wishnu Setya Kusuma, Muhammad Akbar Rivai, and Gunawan Wang. **Business Intelligence Infrastructure of Medical Record Data History System to help Doctor in differencing rare and dangerous disease in patient**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, Vol. 9 No. 1, pp. 664-672, 2020. https://doi.org/10.30534/ijatcse/2020/93912020
11. Olexová, C. E. C. Í. L. I. A, Business Intelligence adoption: a case study in the retail chain," 2014 *WSEAS transactions on business and economics, 11(1), 95-106.*
12. D. Barone, T. Topaloglou, and J. Mylopoulos. **Business Intelligence Modeling in Action: A Hospital Case Study**, *Advanced Information Systems Engineering Lecture Notes in Computer Science*, pp. 502–517, 2012.
13. Ibrahim A. Atoum and Nasser A. Al-Jarallah. **Big Data Analytics for Value-Based Care: Challenges and Opportunities**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, Vol. 8 No. 6, pp. 3012-3016, 2019. https://doi.org/10.30534/ijatcse/2019/55862019
14. B. Abutahoun, M. Alasasfeh, and S. Fraihat, **A framework of business intelligence solution for real estates analysis**, *Proc. Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems*, pp. 1-9, 2019.
15. L. AlWreikat, A. AlShawa, D. Al-Rimawi et al., **Business intelligence and data analytics system for mobile money**, *Proc. Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems*, pp. 1-6, 2019.
16. B. AlArmouty, and S. Fraihat, **Data Analytics and Business Intelligence Framework for Stock Market Trading**, *Proc. 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pp. 1-6, 2019.
17. A. Rajaraman and J. D. Ullman. **Mining Data Streams,** *Mining of Massive Datasets*, pp. 108–138. https://doi.org/10.1017/CBO9781139058452.005
18. "1.4. Support Vector Machines," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html. [Accessed: 25-May-2020].
19. Tharwat, A., 2018. Classification assessment methods. Applied Computing and Informatics.