Journal of Computer Science

Original Research Paper

Data Warehouse Design to Support Social Media Analysis in a Big Data Environment

[1]Carlos Roberto Valêncio, [1]Luis Marcello Moraes Silva, [1]William Tenório,
[1]Geraldo Francisco Donegá Zafalon, [2]Angelo Cesar Colombini and [2]Márcio Zamboti Fortes

[1]*Institute of Biosciences, São Paulo State University (Unesp),*
Humanities and Exact Sciences (Ibilce), Campus São José do Rio Preto, São Paulo, Brazil
[2]*Fluminense Federal University (UFF), Niterói, Rio de Janeiro, Brazil*

Article history
Received: 30-09-2019
Revised: 28-11-2019
Accepted: 12-02-2020

Corresponding Author:
Carlos Roberto Valêncio
São Paulo State University
(Unesp), Institute of
Biosciences, Humanities and
Exact Sciences (Ibilce),
Campus São José do Rio Preto,
São Paulo, Brazil
Email: carlos.valencio@unesp.br

Abstract: The volume of generated and stored data from social media has increased in the last decade. Therefore, analyzing and understanding this kind of data can offer relevant information in different contexts and can assist researchers and companies in the decision-making process. However, the data are scattered in a large volume, come from different sources, with different formats and are rapidly created. Such facts make the knowledge extraction difficult, turning it in a complex and high costly process. The scientific contribution of this paper is the development of a social media data integration model based on a data warehouse to reduce the computational costs related to data analysis, as well as support the application of techniques to discover useful knowledge. Differently from the literature, we focus on both social media Facebook and Twitter. Also, we contribute with the proposition of a model for the acquisition, transformation and loading data, which can enable the extraction of useful knowledge in a context where the human capability of understanding is exceeded. The results showed that the proposed data warehouse improves the quality of data mining algorithms compared to related works, while being able to reduce the execution time.

Keywords: Social Media, Data Warehouse, Data Mining, Big Data

Introduction

In the last few years, the amount of data produced in the internet with the advent of web 2.0 technology has increased, especially the data from social media environment (Ghani *et al*., 2018). This had a significant matter in contemporary society due to the ease of sharing and helping communication among people. As a result, this mass of data includes text, pictures, videos and others that are speedily formed and replaced by new data (Wu *et al*., 2014). Therefore, the data from social networks, such as Facebook and Twitter, trends to grow in the next years (Statista, 2019). Due to this, many researchers and enterprises started developing tools for analyzing such data (Storey and Song, 2017). Also, there are efforts to build a business intelligence solution to study the novelty and to extract useful knowledge (Brambilla *et al*., 2017).

The research in social media can offer understanding and support the: Development of new and better products with the opinion of the clients (Ko *et al*., 2017; Jeong *et al*., 2017), assist in the learning processes in high education (Oktavia *et al*., 2017), collaboration in the detection of fake news with computational resources (Shu *et al*., 2017) and others. All those characteristics shows that studying the social media features can help the decision-making process of several areas.

Therefore, the Business Intelligence (BI) is needed because of the nature of the data from the big data universe. Moreover, despite the advantages of this research, there is still the issue of the absence of pattern from the data with the big data context (Wu *et al*., 2014). Thus, traditional technologies used for data storage do not support social media data reasonably, which affects the usage of data. Besides that, the manual analysis of these elements is complex and impracticable because it exceeds the human capability of understanding. These problems lead to a non-effective analysis of the data and high cost of computational resources.

To deal with the problems related to load and the understating of social media data, initially researchers attempted to build Data Warehouse (DW) models for storing the Twitter data (Moalla *et al*., 2017). Rehman *et al*. (2013) attempted to extend the On-line Analytical Processing (OLAP) technique to manage the multidimensional analysis. They used integration methods to bond text integration and opinion mining. A DW to load the semi-structured and the unstructured data from Twitter is showed and a constellation schema is provided.

Similarly, Kraiem *et al*. (2015) described a model to extend the OLAP method to allow the multidimensional analysis. This work also developed a constellation DW approach to improve the OLAP. The proposal was to build a DW to answer more questions and it had focus on the user's activity. In the same context, Moulai and Drias (2018) presented a specific DW called "Information Warehouse" which was made with information fact tables. The paper's goal is to create a generic multidimensional DW with a star schema and validate it with the case of Twitter. In the case, the Data Mining (DM) algorithm Apriori was used to verify the effectiveness of the approach.

On the other hand, Moalla *et al*. (2016) dealt with the conceptual modeling of a DW for Twitter and Facebook to integrate the similar data from both social networks. With this, it was able to make the opinion analysis. Thus, it could attend the drawbacks from state-of-art to make a generic DW. The Extract, Transform and Load (ETL) stage is not done and there is no OLAP or DM application on this work. Besides that, it has problems with some limitations of the information that it loads. Also, there is the issue of carrying redundant data to the DW because of the definition of its schema.

In that way, some issues were identified in the state-of-art and our work tries to answer the questions related:

- How to build a conceptual model DW for social media?
- How to avoid loading redundant data from the DW?
- In which way the DW organization can support DM algorithms?

In that way, despite of the advantages in studying this topic to extract useful knowledge, there are still problems related to the social media data management. Some issues are the absence of pattern in conceptual models and the different goals of the companies and researchers that use such data.

Herewith, the proposal of this paper is to extend the concept of a DW for both Facebook and Twitter to build a normalized constellation schema. Such DW can offer support to the DM application for opinion analysis. Thus, we present our Configurable Load and Acquisition Social Media Environment, called CLASME. This approach deals with the conceptual model of a DW for helping the DM application. In addition, this work differs from the literature in a way that contemplates the ETL stage and the DM application step. Our method attempts to remove redundant data, since it can prejudice the DM performance. To evaluate the work, four DM algorithms had been used to validate the performance of the approach. This paper describes its definitions for building the schema and the opinion analysis, which can be positive, negative or neutral (Balazs and Velásquez, 2016). A discussion is presented with the results and conclusion.

This work is organized as follows. Section 2 reviews a background with some definitions from the approach. In section 3 the development of this work is presented. Next, section 4 provides the materials and methods used. Sections 5 and 6 detail the results and discussion obtained with the experiments made. Lastly, section 7 concludes the paper with our scientific contribution and suggests the future work to the research.

Background

For understanding the necessity of making a new conceptual model, it is important to review some features. One feature is the big data universe.

The big data is defined as large data set with no pattern that exceeds the human capacity for understanding. Such property is referenced in computer science nowadays, especially due to its potential in decision-making process and discovering trends and associations (Sivarajah *et al*., 2017). The features from big data are given by its "V's". Those "V's" are split in the first five "V's" that refers to volume, velocity, variety, veracity and value. The others "V's" are variants in the literature, but is important to quote some features: Visualization and variability (Storey and Song, 2017; Gandomi and Haider, 2015). A brief explanation of principals is:

- Volume - is given by its magnitude and its size that can be terabytes, petabytes or exabytes
- Velocity - is given by the rate in which the data is created and by the speed is needed to be analyzed
- Variety - is given by the undefined structure of the data in the database, such as text, images, videos and others
- Veracity - is given by the quality, imprecision and reliability of the data
- Value - is given by identifying useful knowledge to be used in the business context

In this context, a social network holds the features from big data (Ghani *et al*., 2018). The social media is a web-based and mobile-based internet application. It allows the creation, accessing and sharing of data generated by its users (Yoo *et al*., 2018). It also has the feature of being of ubiquitous access (Batrinca and Treleaven, 2015), which allows users to publish and to share publications of any topic. Thus, each publication has some numeric particulars such as likes or favorites, number of comments and number of shares. All these data can come from the Application Programming Interface (API) from both social media, which provides limited public data access.

Merging the similar data from the social media is an approach that could create benefits, since it can combine the multiple source data into one overview among the data. Such approach is called data integration. From this perspective, we must consider the semantic relationships from the data for proceeding with the opinion analysis (Moalla *et al*., 2016). The relationships between the elements from the social media can be:

- Identical - if the feature has the same name and same meaning
- Equivalent – if the feature has different name but the same meaning
- Complementary – if the feature has different name and different meaning

Thereafter, we can identify the relationships presented in the social networks. This is done by comparing the equivalent data. Thus, the analysis of the complementary data allows an overview on the information from the social media.

Herewith, is possible to examine growing and brand impact (Ko *et al*., 2017), user behavior (Yoo *et al*., 2018), opinion analysis (Moalla *et al*., 2016) and others. Studying them, a DW approach is presented and used in this case. Plus, it is possible to extend the DW concept and build a specific solution for a social media problem. One thing that is used in the DW creation is the normalization model, which ensures that the data stored is consistent and not redundant (Inmon, 2002). As will be presented forward, a normalized DW model can help DM techniques to improve the results. Those improvements may be view through features such as: Execution time, memory and CPU usage, quality of the results and others.

Moreover, one of the properties from DW is the ETL stage. The extraction phase consists in obtaining the data from several sources. The transformation phase includes cleaning and standardization of the data, depending on the objective of the application. The load phase consists in mapping and storing the transformed data in the correct part of the DW. This part is a subject-oriented section known as Data Mart.

Once the ETL and the DW are done, the next stage is the data analysis. This phase can be done by DM algorithms or by OLAP techniques. In this scenario, DM methods are used to get consistent results from social media (Moulai and Drias, 2018; Injadat *et al*., 2016; Moro *et al*., 2016; Batrinca and Treleaven, 2015; Wu *et al*., 2014). After this, the final step is interpreting the results for assisting the decision-making process.

The CLASME Approach

This section explains the developed conceptual model and the ETL stage. With this, through our approach it is possible to support DM techniques. This conceptual model consists on the normalized DW for the social media Facebook and Twitter.

Therefore, CLASME enables to select specific public data from Facebook and Twitter. Our approach used both API from social media. After the ETL phase, the specified data are loaded into the DW. So, it assists the analyst in the DM processes, since it integrates both extraction and load functionalities. An overview of our method is presented in Fig. 1.

In the DW, the main focus of the information is the opinion analysis (Moalla *et al*., 2016). Such analysis depends on the necessities of the analyst. As shown in Fig. 1, the analyst sets up the data selection and the ETL stage obtains the data from the both social media.

Later, data is submitted into a transformation and it is integrated in the DW. At last, the stored data can be used for DM, since its organization helps such process.

As showed in Fig. 1, we dealt with the entire process of data analysis since its extraction. In that way, the analyst is fundamental to make the hypothesis about what kind of information he is studying and which DM techniques are most appropriate.

The Conceptual Model

To build the conceptual model, the integration is essential. It consists in finding the semantic relations of the data. This is used to build a single vision of the data, independently from origin.

According to the state-of-art (Moalla *et al*., 2016), the three possible semantics are: Identical, equivalent and complementary. Such data and their semantic relationship are showed in Table 1. Such relationships represent a way to integrate the data in the DW.

Besides that, is important to define the relationships between the data from the posts. In the same way, is possible to find semantic relationships in the social media posts, as shown in Table 2, such as the Number of Positive (NPO) and Negative (NNO) opinions. These relationships are used to integrate the data and through it, we can make the opinion analysis.

If a given semantic relationship is identical or equivalent, it means that it can be stored at the same attribute. On the other hand, if the semantic relationship is complementary, this means that the attribute exists only in the specific social media.

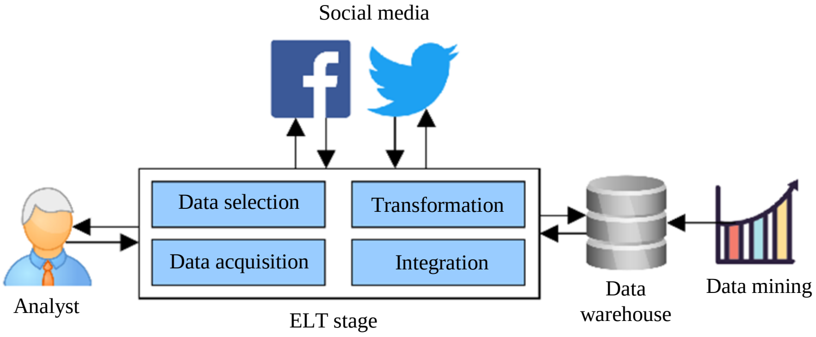

Fig. 1: Overview of the approach

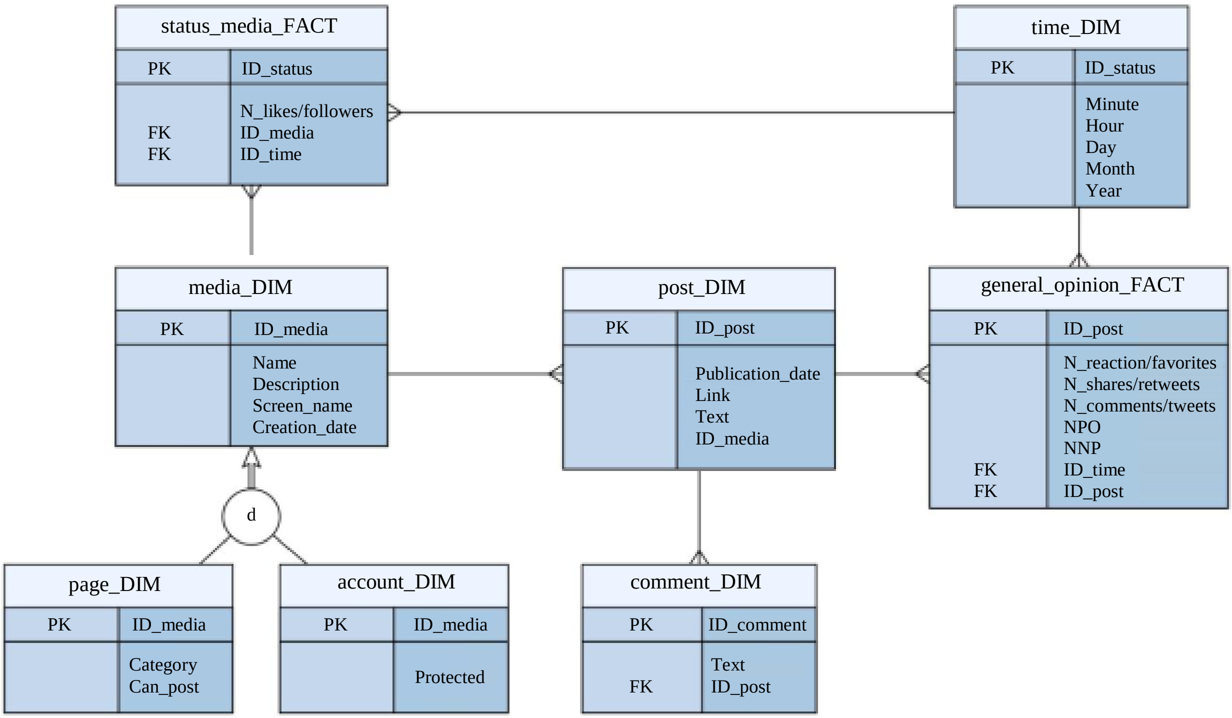

Fig. 2: Developed data warehouse schema

Table 1: Media semantic

Facebook	Twitter	Relation
ID	ID	Identical
Name	Name	Identical
Creation date	Creation date	Identical
Description	Description	Identical
Number of likes	Number of followers	Equivalent
User name	Screen name	Equivalent
Category	-	Complementary
Can post	-	Complementary
-	Protected	Complementary

To study these relationships, the developed DW was built to provide assistance to a temporal analysis of the opinion analysis. The concept of normalization was used to guarantee consistency and absence of duplication.

Table 2: Post semantic

Facebook	Twitter	Relation
ID	ID	Identical
Publication date	Publication date	Identical
Link	Link	Identical
Message	Text	Equivalent
Number of reactions	Number of favorites	Equivalent
Number of comments	Number of tweets	Equivalent
Number of shares	Number of retweets	Equivalent
NPO	NPO	Identical
NNO	NNO	Identical

To assist reading, the developed schema is presented in Fig. 2. According to the literature (Rehman *et al*., 2013) the tables with the "DIM" label are described with the qualitative attributes. Also, the tables with the

"FACT" label are described with the quantitative attributes known as measures (Rehman *et al*., 2013).

We started building our model defining the "media_DIM" entity. This entity is an abstraction obtained through generalization from social media Twitter and Facebook and the "d" represents that both pages and accounts are disjoint. We used it to load the common attributes from both social media, while the specific data are inserted on the "page_DIM" or "account_DIM" entity. The attributes with identical and equivalent relationships presented on Table 1 offer the set of attributes from the "media_DIM" entity. The complementary relationship gives the attributes of "page_DIM" and "account_DIM" entities.

Following, there is a relationship one-to-many between "media_DIM" and "post_DIM" entities. The "post_DIM" table is described with the following attributes: ID, publication date, link and text. Publication date refers to the creation of the post; the link refers to the URL in the social media and text is the message presented in the post. Additionally, there is a relationship in between "comment_DIM" and "post_DIM" entities with cardinality one-to-many. Such table was defined with the descriptive attribute text of the comment.

After that, we defined a fact table called "general_opinion_FACT". Such table is described with the following measures: Number of reactions/favorites, number of shares/retweets, number of comments/tweets, NPO and NNO. Such measures were defined through the identical and equivalent relationship illustrated in Table 2. It represents the state of an opinion through two dimensions.

Therefore, it is defined another fact table which is "status_media_FACT". Such entity is characterized with the measure number of likes/followers. With it, we can analyze such measure through the dimensions "time_DIM" and "media_DIM".

Lastly, the "time_DIM" table is described with time related attributes, such as: minute, hour, day, month and year. This dimension is used for describing the insertion time of a given "status_media_FACT" or "general_opinion_FACT" instances.

As presented in Fig. 2, there are two fact tables described in our constellation schema. Thereby, it is possible to make *ad hoc* queries in "general_opinion_FACT" through two dimensions, which are "post_DIM" and "time_ DIM". Also, it is possible to analyze posts from just one specific social network, differently from the state-of-art that usually focus only one social media. Lastly, we can manage studies involving the popularity of the social media over time due to "status_media_FACT".

The remaining relationships that are associated with the qualitative data needed for further studies and those data are defined by the analyst. Such conceptual model implies that is possible to retrieve all data origin and rebuild the post in same way it was inserted.

For studying the popularity of a page or account, the "status_media_FACT" entity is used. Such entity enables the analysis of trends in the accounts of Twitter and the pages of Facebook, differently from the state-of-art. This can support the analyst to study only the specific growing a page or account, for instance.

Another contrast from the paper of Moalla *et al.* (2016) is that same posts from the page and account with the same name but in different social media are stored independently. In that study, the authors loaded only one post published by the same page and account, while our method considered loading both posts to enable different analyses over them.

Additionally, comments need to be stored just once and further studies can be made with them. The "time_DIM" entity is used to facilitate the insertion, because all attributes from it do not have the necessity to be in both fact tables.

On the other side, the measures from "general_opinion_FACT" deal directly with the opinion analysis. Here, DM algorithms can be applied to extract useful information and predicting trends. Further, it enables the study of which posts are the most positive, negative and neutral.

The ETL Stage

As shown in Fig. 1, our work dealt with the ETL stage to prepare the data initially. The loading method from CLASME is made first defining the pages and accounts to study. This is done with the decision maker.

Herewith, the selected public data from pages and accounts are consulted by the environment. After that, the qualitative data from pages and accounts are loaded only once. Following, the attributes of a non-existing post in the DW are loaded on the DW once. The timestamp of this load are registered as well.

The attributes from the comments are classified and the opinion about the post is given. The message from the comment is standardized to lower case and URLs and irrelevant characters are removed. Later, they are stored as well. The software executes this function only when is trigged by the user.

Once the qualitative data are loaded, there is no need to store such data again. Only the quantitative data are needed to load, since its evolution tells the opinion of a given post over time. Further, to store the history of a publication, the data in the DW needs to have a new load. This means that a new query must be made in API to load the possible changes in the post's measures. A given publication is checked up to three days.

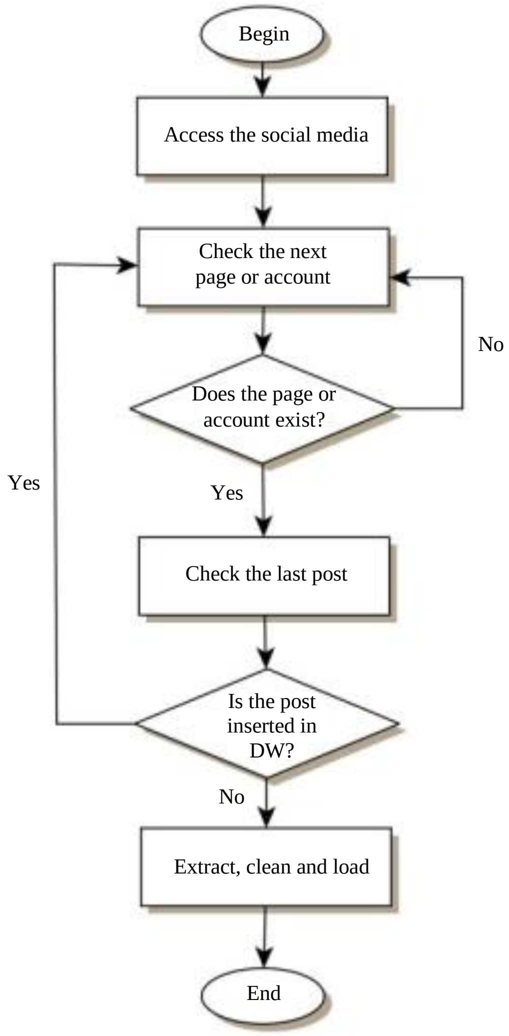
Begin
Access the social media
Check the next page or account
Does the page or account exist?
No
Yes
Yes
Check the last post
Is the post inserted in DW?
No
Extract, clean and load
End

Fig. 3: Loading method for new publications

The method used for new publications works as it follows. Our approach accesses the accounts and pages through API. For each page or account in the set defined by the analyst, it is checked if the media instance exists. If it does, the last post of this media is verified and if the publication is not inserted, it is obtained, clean and loaded. This includes its comments as well. On the other hand, if the media do not exist or the last post is already in the DW, our approach verifies the next page or account. This process is illustrated in Fig. 3.

Similarly, our approach has other method to deal with already loaded publications. Such method operates accessing the same data from the API. If the account or page exists, then it is verified if its publication is inserted in the DW. If the post exists in the DW, then our approach makes a new insertion containing the current quantitative data about that post. If that post does not exist in the DW, then we verify the next media. This method is presented in Fig. 4.

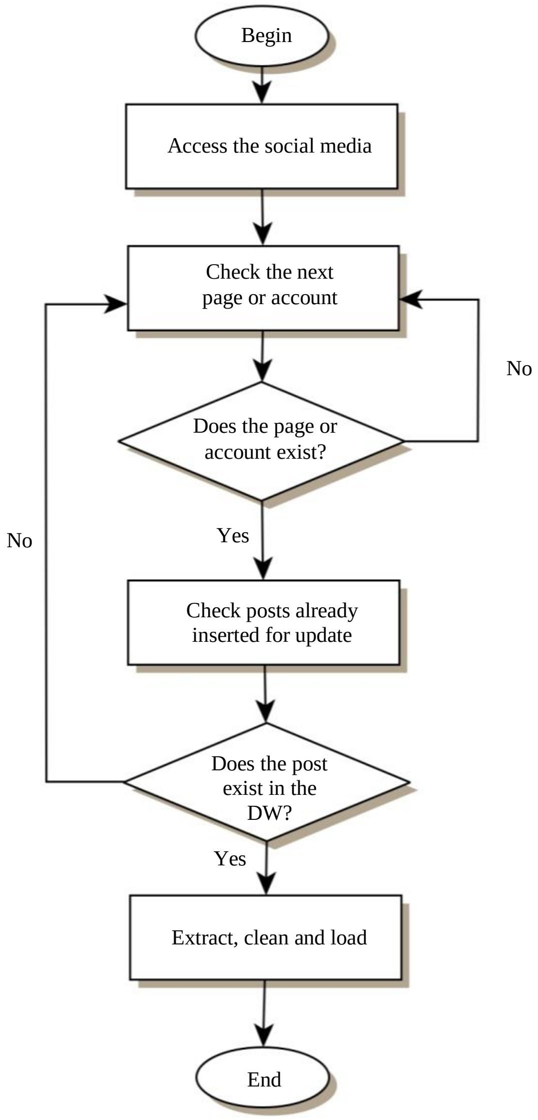
Begin
Access the social media
Check the next page or account
Does the page or account exist?
No
No
Yes
Check posts already inserted for update
Does the post exist in the DW?
Yes
Extract, clean and load
End

Fig. 4: Loading method for old publications

Materials and Methods

This section specifies the used hardware and software features, as well as the database and the experimentation method applied.

The DW model was implemented over PostgreSQL9.5. The manipulation of the API and the ETL stage were developed using the JAVA programming language. The experiments were conducted in a computer with the following hardware and software specifications: Intel Core i5 7200, 8 GB RAM DDR4, Windows 10, Java 8, WEKA 3.8.

The database used in the experiments was obtained by the developed ETL tool, according with the CLASME approach. Those data were collected in a period of ten days through manual execution. One hundred pages from Facebook and one hundred accounts from Twitter were selected. They represent a set of topics containing newsletters, product announcements and political content.

Each page and accounts were checked for new posts once a day and each publication was verified two days after its load on the DW, both by manual trigger. Due to limitations from API, only the most recent publication from each page or account was loaded. However, the already inserted posts were searchable and the quantitative attributes from it were loaded as well to study the changes. Public data available through API was used to compose the database. These data represent part of the publications and comments usually found in the social media. Here, the goal is to load the DW according to the ETL and specifications previously made. We were not intended to analyze data to obtain value and veracity. Those features need a business analyst to verify the data later on.

To validate the developed environment, the DW model from Moalla *et al.* (2016) was implemented as well. The comparison between our approach and the DW proposed by Moalla *et al.* (2016) was conducted to analyze the capability of each approach to enable useful knowledge discovery through the application of DM algorithms. Due to the fact that our model was normalized; the loaded data through our approach should present itself consistent and without redundancies.

The DM algorithms were applied in each DW through the software WEKA. These algorithms are well known methods to retrieve relevant and useful knowledge from data, according to the state-of-art (Batrinca and Treleaven, 2015; Balazs and Velásquez, 2016; Ghani *et al*., 2018).

The experiment 1 was conducted considering the Naïve Bayes (NB) algorithm. It was used to classify the opinion from both DWs using the post's origin as a label. The classifier is a supervised method that, after trained, can infer which social media a given data comes from. To validate the results, we considered the execution time in seconds, accuracy (Equation 1), precision (Equation 2), recall (Equation 3) and relative absolute error (Equation 4), such as done in similar works from scientific literature (Balazs and Velásquez, 2016). In the cited equation, *TP*, *FP*, *TN* and *FN* represents the number of true positive results, the number of false positive results, the number of true negative results and the number of false negative, respectively. Also, *P* represents a predicted value and *R* represents a real value.

In the experiment 2, we used linear regression. Differently from the first experiment, this algorithm attempts to generate a numeric function based on the input data. To estimate the value of an attribute, the other attributes were considered as input data to the algorithm. Thus, the approach aims to discover predictive patterns. To validate the results, we considered the execution time in seconds, the relative absolute error (Equation 4) as well as the CC representing the correlation coefficient (Equation 5) (Hayes and Montoya, 2017). In the cited equation, *X* and *Y* represent the independent variable and the dependent variable, respectively. Also, $\bar{X}$ and $\bar{Y}$ represent the average value of *X* and *Y*, respectively.

The experiment 3 was conducted considering the K-means algorithm. It was used to cluster the data using the post's origin as a label. Herewith, the *K* was set as two, because there are only two possible social media. The Euclidian distance was used. The values were analyzed through execution time in seconds, error rate in the classification and number of interactions (Hossny *et al*. 2018).

The experiment 4 was conducted considering the Apriori algorithm. This algorithm is used for frequent item set mining and association rule learning over relational databases (Agrawal *et al*., 1993). We intended to analyze the capability of our approach to enable the discovery of more useful association rules when compared to the approach proposed by Moalla *et al.* (2016). To define an association rule, let's considerer $I = \{i_1, i_2,\ldots,i_n\}$ a set of *n* binary attributes called *items* and $T = \{t_1, t_2,\ldots, t_n\}$ a set of transactions called *database*. Each transaction t_i is called *ItemSet* and $t_i \subseteq I$. So, a rule is defined as an implication of the form $X \to Y$ where $X, Y \subseteq I$ and $X \cap Y = \varnothing$ and an association rule is a rule that comply with the user-specified minimum support and confidence values. The support *supp*() of the rule is the percentage of the transactions in database that contain $X \cup Y$, i.e., the frequency of the rule in the database. The confidence *conf*() of the rule is the percentage of the transactions containing *X* which also contain *Y*, i.e., the reliability of the rule (Valêncio *et al*., 2018).

Considering these definition, we have analyzed the execution time in seconds, number of rules, greater confidence, greater lift (Equation 6) and greater conviction (Equation 7), which are metrics commonly used to analyze algorithms that mine association rules. The origin attribute, which defines from which social media the data comes from, was not taken into consideration at the first experiment and then was used. The support and confidence values were defined as 0.1 and 0.5, respectively, such as done in similar works involving the Apriori algorithm.

The experiments were executed five times to ensure consistency of the results and the reported results represent the average value between these five executions.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$Precision = \frac{(TP)}{(TP + FP)} \tag{2}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{3}$$

$$RAE = \frac{\sqrt{\sum_{i=1}^{N}(P_i - R_i)^2}}{\sqrt{\sum_{i=1}^{N}(R_i)^2}} \tag{4}$$

$$CC = \frac{\sum_{i=1}^{N}(X_i - \bar{X}_i)(Y_i - \bar{Y}_i)}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X}_i)^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2}} \tag{5}$$

$$lift(X \to Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)} \tag{6}$$

$$conv(X \to Y) = \frac{1 - supp(X)}{1 - conf(X => Y)} \tag{7}$$

Results

This section presents the obtained results. In Table 3 to 7, DW A represent the proposed data warehouse and DW B represent the data warehouse proposed by Moalla *et al*. (2016).After loaded the data into the developed DW, a total of 5.452 tuples were presented in the main entity “general_opinion_FACT”. After loaded the data into the DW proposed by Moalla *et al*. (2016), a total of 388.042 tuples were presented. Although the quantity of tuples was different, it represented the same information. Comparisons about these values are showed in the next section.

The capability of classifying an opinion through its origin was verified in the first experiment, comparing both DW strategies. The results are presented in Table 3.

On the experiment 2, the objective was to create a descriptive numerical function to predict the attributes of an opinion. The execution time was analyzed over the five attributes from the opinion. The results related with the performance from each attribute are presented in Table 4. Table 5 shows the quality metrics for each function described by each the five attributes and compares such results between the DWs.

Then, the experiment 3 was conducted to compare the capability of both DWs to cluster the data through the origin of the opinion. Table 6 illustrates the results of this technique. To calculate the error, the origin of the opinion was given as an input.

In the final experiment, we executed the Apriori algorithm. Our goal was to find and analyze the discovered association rules and evaluate its quality. Table 7 presents the number of rules discovered and the features related. The greater metrics were selected to emphasize the quality of the discovered rules. In the next section, we discuss these results and the implications of the findings.

Table 3: Naïve Bayes results

Metric	DW A	DWB
Execution time (s)	0.8000	3.4900
Accuracy (%)	98.4000	97.2000
Precision (%)	98.4000	99.3000
Recall (%)	98.4000	97.2000
RAE (%)	3.4982	143.7436

Table 4: Linear regression performance results

	Execution time (s)	
Attribute	DW A	DWB
Reacts/favorites	0.01	0.66
Shares/retweets	0.02	0.66
Comments/tweets	0.01	0.82
Positive comments	0.01	2.47
Negative comments	0.01	1.35

Table 5: Linear regression quality results

Attribute	CC	RAE (%)
Reacts/favorites (DW A)	0.7030	71.8075
Reacts/favorites (DW B)	0.1980	92.5140
Shares/retweets (DW A)	0.8826	40.2845
Shares/retweets (DW B)	0.4914	87.1933
Comments/tweets (DW A)	0.9001	31.1845
Comments/tweets (DW B)	0.6233	90.5237
Positive comments (DW A)	0.7904	40.2159
Positive comments (DW B)	0.3894	89.5878
Negative comments (DW A)	0.7478	46.7156
Negative comments (DW B)	0.3543	90.8951

Table 6: K-means results

Metric	DW A	DWB
Execution time (s)	0.0300	2.2800
Error (%)	28.4483	47.7142
Number of interactions	4.0000	24.0000

Table 7: Apriori results

	Unlabeled		Labeled	
Metric	DW A	DWB	DW A	DW B
Execution time (s)	2.01	106.7	2.50	187.96
Number of rules	50.00	0	180.00	5.00
Greater confidence	1.00	-	1.00	0.99
Greater lift	2.92	-	2.92	1.00
Greater conviction	1033.30	-	1077.04	1.00

Discussion

This section presents an analysis of the obtained results. All experiments conducted over our approach has finished the execution in less than three seconds. On the other hand, the DW designed by Moalla *et al*. (2016) reached times greater than 100 sec. In terms of quality, the four experiments confirmed that the designed DW brought improvement.

An important property is the problem of absence of pattern in modeling the DW. Those DW are highly variable depending on the context. The approach to build them depends on a specific problem. In addition, most of the literature works deals on building a model only for one social media. The model proposed by Moalla *et al*. (2016) had some issues related to the data redundancy because it was build following the star schema. Still, it was taking into account only the business rules. With this approach, DM algorithms had their performance and quality prejudice, as shown in this paper. Also, the number of tuples obtained made the execution time be greater in all experiments relative to the approach of Moalla *et al*. (2016), as represented by the results.

According to the algorithm used in experiment 1, due to the replication problem in the model presented by Moalla *et al*. (2016), the data organized in the developed model was better. In this experiment, the results showed a higher accuracy and higher recall. The precision, yet, is better in the DW presented by Moalla *et al*. (2016), it means it was better in classifying the true positives, but it doesn't take into account the true negatives. Also, RAE illustrates the advantage of our model. This experiment showed that even with more input data, the data consistency is important to secure the quality.

Following, the experiment 2 presents better results in all the features. Besides the smaller execution time, the CC and the RAE were better in all cases as well. This means that all the predictive functions obtained have a better description of the data in DW. If the four quantitative measures were given, is possible to calculate the estimated value with the functions of our DW with less chance of missing. As showed in this experiment, a greater data input with redundancy prejudice the creating of the functions, so avoiding duplications is needed.

In the experiment 3, the clustering technic had a better result as well. The performance had better results as well as the quality. The results show that the algorithm was better in partitioning the organized data from the developed model. Besides the error value being smaller, the computational cost related to the number of iterations was six times smaller. In that way, the duplications interfered in the results showing that such feature does not bring any benefit.

In the last experiment, the biggest difference in terms of execution time is presented. In the first case, without the labels, the DW developed by Moalla *et al*. (2016) could not find any association rule, while our DW discovered fifty. In the second case, even more rules were found in our model because of the labels. The other features presented in Table 7 shows that quality of the rules was better. Moreover, after analyzing the association rules in both cases, the rules obtained by our model could describe some ordinary features such as: *number of comments* = 0 → *NPO* = 0. On the other hand, the association rules discovered through the model from Moalla *et al*. (2016) presented fake facts such as: *NPO* = 0 → *origin* = *Facebook*, since that fact does not occur in the obtained database. So, the redundancy and the lack of consistency can affect negatively the knowledge discovery related to Apriori algorithm.

Thereby, the quality metrics reflected on how good results were. Better results in comparison with the state-of-art means a better data model for the used DM. As represented by the results, the proposed model improved the quality and the performance of the algorithms. This indicates that this model helped areas in DM that are classification, clustering and association rules. This was possible by developing a normalized DW. The implications of this model are that normalizing the DW has brought benefits in terms of quality, consistency and processing. This means also that further analysis based on the DM algorithms applied should be less costly in computational terms. So, the proposed fact table dealt with the possibility to provide useful data to DM. Only with few transformations needed for each algorithm, it was possible to execute and evaluate the knowledge extraction.

With this, we can consider that the DW organization is one important feature in DM application. Moreover, the computational resources involved in the four experiments cases were better used with our approach. In a more convoluted case, the data entry would be bigger. This is what justifies the importance of studying solutions for the social media data to make it less costly. Finally, the decision maker can spend less time and computational expenses using our DW organization to have support in the decision-making process.

Lastly, through the experiments we can conclude that the decision between the use of a specific DM technique need to be guided by an analysis of the context of the loaded data. For example, when considering large datasets where data can be grouped by labels, Naïve Bayes is recommended since it is able to perform faster than other classifiers. Also, in cases where the loaded data does not have labels, K-means is useful to handle exploratory analyses.

Conclusion and Future Works

This paper presented a normalized data warehouse schema for modeling social media data from two different social media platforms: Facebook and Twitter.

The normalized DW avoids redundant data to be stored, which can efficiently reduce the execution time of data mining algorithms. The ETL stage was described and four DM algorithms were applied to validate the model. Our experiments showed that our model can efficiently assist the analyst in the decision-making process, while being able to execute faster than the related works.

In addition, the DW has focus on opinion analysis, which means that we do not concern about the content of the post or different analysis. Our major study considered the quantitative attributes from the publications and the classification of the comments into positive, negative and neutral.

As future work, we recommend the use of a NoSQL database to treat the scarcity and the excess of attributes. It is also interesting to adapt the DW to include content of others social network, despite the fact that related works commonly deal with only one.

Acknowledgement

The authors acknowledge Brazilian National Council for Scientific and Technological Development (CNPq) and Institute for the UNESP Development (FUNDUNESP) for financial support.

Author’s Contribution

Carlos Roberto Valêncio: Participated in all the projects decisions that defined the scientific contributions of this work; definition and analysis of the data, writing and reviewing the manuscript, head of the research group in which this research was developed.

Luis Marcello Moraes Silva: Participated in all the projects decisions that defined the scientific contributions of this work, definition and analysis of the data, writing and reviewing the manuscript, responsible for coding activity resting and creation of all byproducts related to the work.

William Tenório, Geraldo Francisco Donegá Zafalon, Angelo Cesar Colombini and Márcio Zamboti Fortes: Participated in all the projects decisions that defined the scientific contributions of this work, definition and analysis of the data, writing and reviewing the manuscript.

Ethics

This paper is original and contains unpublished material. The authors confirm that are no conflict of interest involved.

References

Agrawal, R., T. Imieliński and A. Swami, 1993. Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22: 207-216. DOI: 10.1145/170035.170072

Balazs, J.A. and J.D. Velásquez, 2016. Opinion mining and information fusion: A survey. Inform. Fus., 27: 95-110. DOI: 10.1016/j.inffus.2015.06.002

Batrinca, B. and P.C. Treleaven, 2015. Social media analytics: A survey of techniques, tools and platforms. Ai Society, 30: 89-116. DOI: 10.1007/s00146-014-0549-4

Brambilla, M., S. Ceri, E.D. Valle, R. Volonterio and F.X.A. Salazar, 2017. Extracting emerging knowledge from social media. Proceedings of the 26th International Conference on World Wide Web, Apr. 03-07, ACM, Switzerland, pp: 795-804. DOI: 10.1145/3038912.3052697

Gandomi, A. and M. Haider, 2015. Beyond the hype: Big data concepts, methods and analytics. Int. J. Inform. Manage., 35: 137-144. DOI: 10.1016/j.ijinfomgt.2014.10.007

Ghani, N.A., S. Hamid, I.A.T. Hashem and E. Ahmed, 2018. Social media big data analytics: A survey. Comput. Human Behav., 101: 417-428. DOI: 10.1016/j.chb.2018.08.039

Hayes, A.F. and A.K. Montoya, 2017. A tutorial on testing, visualizing and probing an interaction involving a multicategorical variable in linear regression analysis. Commun. Meth. Measures, 11: 1-30. DOI: 10.1080/19312458.2016.1271116

Hossny, A.H., T. Moschuo, G. Osborne, L. Mitchell and N. Lothian, 2018. Enhancing keyword correlation for event detection in social networks using SVD and k-means: Twitter case study. Soc. Netw. Anal. Min., 8: 49-49. DOI: 10.1007/s13278-018-0519-9

Injadat, M.N., F.S., Fadi and A.B. Nassif, 2016. Data mining techniques in social media: A survey. Neurocomputing, 214: 654-670. DOI: 10.1016/j.neucom.2016.06.045

Inmon, W.H., 2002. Building the Data Warehouse. 3rd Edn., John Wiley and Sons, USA, ISBN-10: 0-471-08130-2, pp: 473.

Jeong, B., J. Yoon and J. Lee, 2017. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. Int. J. Inform. Manage., 48: 280-290. DOI: 10.1016/j.ijinfomgt.2017.09.009

Ko, N., B. Jeong, S. Choi and J. Yoon, 2017. Identifying product opportunities using social media mining: Application of topic modeling and chance discovery theory. IEEE Access, 6: 1680-1693. DOI: 10.1109/ACCESS.2017.2780046

Kraiem, M.B., J. Feki, K. Khrouf, F. Ravat and O. Teste, 2015. Modeling and OLAPing social media: The case of Twitter. Soc. Netw. Anal. Min., 5: 47-47. DOI: 10.1007/s13278-015-0286-9

Moalla, I., A. Nabli, L. Bouzguenda and M. Hammami, 2016. Data warehouse design from social media for opinion analysis: The case of Facebook and Twitter. Proceedings of the IEEE/ACS 13th International Conference of Computer Systems and Applications, Nov. 29-Dec. 02, IEEE Xplore Press, Morocco, pp: 1-8. DOI: 10.1109/AICCSA.2016.7945627

Moalla, I., A. Nabli, L. Bouzguenda and M. Hammami, 2017. Data warehouse design approaches from social media: Review and comparison. Soc. Netw. Anal. Min., 7: 5-5. DOI: 10.1007/s13278-017-0423-8

Moro, S., P. Rita and B. Vala, 2016. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. J. Bus. Res., 69: 3341-3351. DOI: 10.1016/j.jbusres.2016.02.010

Moulai, H. and H. Drias, 2018. From data warehouse to information warehouse: Application to social media. Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications, May 02-05, ACM, Morocco, pp: 24-24. DOI: 10.1145/3230905.3230914

Oktavia, T.U. Fauziyah, S. H. Karin and A. F. Siregar. 2017. The influence of social media to support learning process in higher education institution: A survey perspective. Proceedings of the International Conference on ICT for Smart Society, Sept. 18-19, IEEE Xplore Press, Tangerang, Indonesia, pp: 1-5. DOI: 10.1109/ICTSS.2017.8288866

Rehman, N.U., A. Weiler and M.H. Scholl. 2013. OLAPing social media: The case of Twitter. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Aug. 25-28, IEEE Xplore Press, Canada, pp: 1139-1146. DOI: 10.1145/2492517.2500273

Shu, K., A. Sliva, S. Wang, J. Tang and H. Liu, 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorat. Newsletter, 19: 22-36. DOI: 10.1145/3137597.3137600

Sivarajah, U., M.M. Kamal, Z. Irani and V. Weerakkody. 2017. Critical analysis of big data challenges and analytical methods. J. Bus. Res., 70: 263-286. DOI: 10.1016/j.jbusres.2016.08.001

Statista, 2019. Most famous social network sites worldwide as of July 2019, ranked by number of active users (in millions).

Storey, V.C. and I. Song, 2017. Big data technologies and management: What conceptual modeling can do. Data Knowl. Eng., 108: 50-67. DOI: 10.1016/j.datak.2017.01.001

Valêncio, C.R., G.H. Morais, M.Z. Fortes, A.C. Colombini and L.A. Neves *et al*., 2018. A user-driven association rule mining based on templates for multi-relational data. J. Comput. Sci., 14: 1475-1487. DOI: 10.3844/jcssp.2018.1475.1487

Wu, X., X. Zhu, G. Wu and W. Ding, 2014. Data mining with big data. IEEE Trans. Knowl. Data Eng., 26: 97-107. DOI: 10.1109/TKDE.2013.109

Yoo, S., J. Song and O. Jeong, 2018. Social media contents based sentiment analysis and prediction system. Expert Syst. Applic., 105: 102-111. DOI: 10.1016/j.eswa.2018.03.055