ISSN 2007-9737

Building a Data Warehouse for Social Media: Review and Comparison

Maha Ben Kraiem*[,1], Jamel Feki[2]

[1] University of Sfax,
Tunisia

[2] University of Jeddah,
Saudi Arabia

maha.benkraiem@gmail.com, jamel.feki@fsegs.rnu.tn

Abstract. The significant advancements in technology over the past few decades have given rise to a relatively straightforward array of Internet applications based on open source software. These applications and services aim to enhance online collaboration for a broad audience, particularly through social networking sites. These platforms have transformed the dynamics of online interaction and information exchange, with millions of users regularly engaging and sharing various digital content. Users express their thoughts and opinions on diverse topics, contributing valuable insights for personal, academic, and commercial purposes. However, the sheer volume and rapid generation of this data present a challenge for decision-makers and the underlying technologies to extract meaningful insights. To leverage the data derived from social networks, researchers have focused on assisting companies in comprehending how to conduct competitive analyses and convert this data into actionable knowledge. This paper offers a comprehensive literature review on data warehouse approaches derived from social networks. We commence by introducing fundamental concepts of data warehousing and social networks, followed by the presentation of three categories of data warehouse approaches, along with an overview of the most notable existing works within each category. Subsequently, we conduct a comparative analysis of these existing works.

Keywords. Data warehouse; social media; opinion analysis; business intelligence; OLAP.

1 Introduction

Over the past two decades, contemporary decision support and information systems have been essential for the efficient operation and expansion of successful global businesses. The cornerstone of decision support in these systems has been the integration of data warehouses and Online Analytical Processing (OLAP).

Widely accepted and employed worldwide, these technologies find application in diverse domains such as manufacturing, telecommunications, e-commerce, healthcare, education, research, and government. Research contributions, coupled with advancements in relevant hardware technology, have matured data warehousing systems, enabling them to manage substantial volumes of data and provide seamless access.

Online Analytical Processing (OLAP) serves as the central element, facilitating multidimensional data analysis, with continuous improvements and extensions made across various domains and datasets. Recent challenges faced by data warehouse technology, including handling multimedia, semi-structured data, text, and streams, have been met with significant and successful efforts.

The initial decade of the 21st century has been characterized by the widespread popularity and utilization of social media in the internet landscape. Billions of users engage with social media platforms for diverse purposes, such as social networking, blogging, information sharing, news discovery, or a combination of these activities.

Since their inception, social media platforms have made users more active in participatory networks, becoming an integral part of daily life by aiding users in connecting with family, keeping up

with friends or colleagues, and contributing to online discussions.

Millions of users regularly interact and share a variety of digital content, expressing their sentiments and opinions on a wide range of topics. Social media platforms have also played a strategic role in the corporate world, establishing links that connect customers to companies. Each of these links provides vast amounts of data, offering companies a substantial competitive advantage.

These links and opinions hold significant value for personal, academic, and commercial applications. However, the sheer volume and speed at which they are generated pose a challenge for decision-makers and the underlying technologies to derive meaningful insights from such data.

The massive data volumes generated by social media are characterized by being semi-structured, unstructured, and dynamic, presenting challenges for companies in terms of utilization, analysis, and storage. Managing and storing such large volumes of data without advanced platforms can be daunting for organizations.

Therefore, many companies opt to leverage the efficient technologies of data warehouses, enabling comprehensive analysis of massive data volumes. This analytical capability proves valuable for conducting competitive analyses and transforming the data into knowledge for decision-makers. The wealth of information generated by social media necessitates analysis by systems that can provide reliable and fast access for processing large amounts of data.

Among these systems, the Online Analytical Processing (OLAP) system stands out, offering interactive online data analysis in an environment capable of handling extensive data volumes. OLAP provides a simple and flexible modeling approach for various types of analyses based on a predefined multidimensional model, including pre-calculated data that accelerates the analytical processing.

With the emergence of social media, decision-makers aim to harness the vast volume of information generated by these platforms to enhance their decision-making processes. Many companies utilize data warehouse technologies to collect both their own data and that of their competitors.

Decision-makers often explore these networks to obtain additional information about companies, leading to better decision-making. In recent years, the explosive growth of social media has resulted in the generation of tremendous volumes of user-related data. This data presents a novel way to gather information in real time, giving rise to the field of *Social Media Analysis* [1].

This area has significant importance for the scientific community, addressing goals such as refining marketing strategies, profiling people's tastes, and targeting advertisements [2]. Exploring these data through an OLAP process could be a strategic opportunity, contributing to a wide variety of analytical needs [3]. Recognizing the importance of social media in the decision-making process, several researchers have focused on studying data warehouse approaches derived from social media.

This paper categorizes these approaches into three major classes: those addressing user behavior analyses, those integrating opinion analysis into data warehouse schema, and those dealing with social business intelligence. The primary objective of this paper is to provide a literature review on existing approaches to data warehouse design from social media.

The subsequent sections of this paper are organized as follows: Section 2 introduces the main concepts of data warehousing and social media. Section 3 presents existing approaches dealing with behavior analysis. Section 4 outlines approaches that integrate opinion analysis into data warehouse schema. Section 5 describes approaches related to social business intelligence. Section 6 provides a comparative study of the presented approaches, highlighting their limitations. Finally, Section 7 concludes the paper, outlining the main perspectives in this work.

2 Background

This section provides an introduction to key concepts and terminology in the realms of data warehousing, Online Analytical Processing (OLAP), and social media. It examines different options for the design and implementation architecture of data warehousing, outlining the

various structural considerations. Furthermore, the section offers a comprehensive overview of selected popular social media platforms, elucidating their growth trajectories and patterns of usage in the current landscape.

2.1 Data Warehouse Concepts

A Data warehouse (DW) is a centrally managed and integrated database containing data from the operational sources in an organization. DW is an integrated repository of data put into a form that can be easily understood, interpreted, and analyzed by the people who need to use it to make decisions.

The most widely cited definition of a DW is from Inmon [4] who states that "a data warehouse is a *subject-oriented*, *integrated*, *non-volatile*, *and time-variant* collection of data in support of management's decisions."

– Subject-oriented: Data is modeled according to the subject area of the respective enterprise, and not according to the application needs of operational systems. A data warehouse does not focus on the ongoing operations; rather it focuses on modeling and analysis of data for decision making.
– Integrated: "A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data".
– Non-volatile: "A data warehouse is kept separate from the operational database and therefore frequent changes in operational database are not reflected in the data warehouse".
– Time-variant: "The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.

The data in a data warehouse are organized according to a multidimensional model. This modeling provides a level of abstraction independently of technical aspects and focusing on decision-making needs [5]. The multidimensional modeling consists in defining the subject to be analyzed as a point in a multidimensional space [6].

In fact, data are organized in such a way to bring out the subject of analysis represented by the concept called *fact*, composed of *measures* corresponding to the additive information of the analyzed activity as well as the *dimensions* of this activity representing analysis axes.

A dimension is composed of attributes expressing the characteristics according to which the measures of the fact are analyzed (i.e., activity). The attributes of a dimension can be organized into hierarchies, from the finer to the most general granularity.

Relying on the fact and dimension concepts, it is possible to build different multidimensional models; the most popular one is the star model. A star model is composed of one central fact surrounded by dimensions, whereas the constellation model consists in defining a set of facts that share common dimensions.

2.2 Social Media Concepts

"Over the last few years, the Web has fundamentally shifted towards user-driven technologies such as blogs, social networks and video-sharing platforms. Collectively these social technologies have enabled a revolution in user-generated content, global community and the publishing of consumer opinion, now uniformly tagged as social media.

This movement is dominating the way we use the Internet and the leading social platforms like Facebook, MySpace, YouTube and now Twitter have moved into the mainstream. These sites are the tip of a redefinition of how the Internet works, with every site now incorporating the features that allow users to publish opinions, connect, build community, or produce and share content" [7].

Social media have become one of the most powerful sources of news updating, online collaboration, networking, viral marketing and entertainment. The terms of social media and social network are used every day. We saw to be more or less the same. In fact, social media include social networking, blogs, forums and platforms.

There are several types of social media; each one has features and different purposes. However, many researchers have different classification about types of social media sites ([8, 9, 10]) that emerges as a problem when realizing a study on social media.

In this context, Kaplan et Haenlein [9] classified social media sites into these six types based on social presence, media richness and self-presentation, and self-disclosure. Thus, they have got six forms of social media, which are: (a) collaborative projects (e.g., Wikipedia), (b) blogs (e.g., Wordpress.com, Blogger.com), (c) content communities (e.g., YouTube), (d) social networking sites (e.g., Facebook), (e) virtual game worlds (e.g., World of Warcraft), and (f) virtual communities (e.g., SecondLife)). In 2011, Akar and Topçu [10] add to this classification the microbloggings type such as Twitter.

Shankar and Hollinger [8] classified these new media into three groups: importun (Internet advertising, product placement in video games or advergames et m-commerce), non-importun (Internet advertising, social networking sites, podcasting, buzz or viral marketing) and user-generated (blogs, video site, ratings/recommendations and summary).

Taking into account the different types of social media proposed in the literature, we retain the classification of the most common forms of social media as follows:

– Social Networking sites: “These are sites mainly used for connecting with friends and family. They focus more on person-to-person conversations. Aside from personal conversations, these platforms encourage knowledge sharing. These platforms accommodate the different types of content formats from text to photos, videos, and other creative forms of content. They are considered the center of communication and a jack of all trades. Users are able to create unique interesting content, share their thoughts, and create groups based on similar interests. These sites are user-centered and are built around the social needs of the users and everything that is important to them. Businesses and marketers can fully maximize these platforms because they provide an immense amount of data. Also, they are able to reach the right people through adverts with specific metrics and demographics. They also provide the opportunity to engage with users which helps people connect with your brand on a more personal level. Some of such platforms include Facebook, LinkedIn, and Twitter.”
– Media sharing sites: “they have gained more prominence in recent times. Content like info graphics, illustrations, and images capture the attention of users more. Social media apps like Pinterest, Instagram, and Snapchat are designed to amplify the sharing of images. They say a picture is worth a thousand words, and using this can have lots of positive effects. Video content is one of the most captivating and engaging forms of content. Marketers and businesses have said that they have seen tremendous benefits in using videos. This form of content aids assimilation and understanding, hence why it is largely preferred by users. One major platform that reshaped how people interact with video content is YouTube. With over one billion active users monthly, the platform sometimes serves as a search engine for most users.”
– Discussion forums: “they are very essential because they allow users to ask questions and get answers from different people. These platforms are designed to spark conversations based on shared interests or out of curiosity. Some of such platforms include Quora and Reddit.
– Blogs: they are a great way for businesses and marketers to reach and provide credible information to their target audience. Platforms like Tumblr, Medium, OverBlog, canalblog and blogspot allow users to create a community where people with similar interests can follow them and read all they have to say about certain topics.”

The advent of social media as a novel source of data has significantly introduced new challenges concerning the modeling and handling of data. In the subsequent sections, we will conduct an extensive examination of strategies employed in designing a data warehouse derived from social media.

Our classification encompasses three main categories: (1) data warehousing for behavior analysis, (2) the incorporation of opinion analysis into the data warehouse, and (3) data warehousing tailored for social business intelligence. Section 3 will outline approaches centered around behavior analysis, followed by Section 4, which will elaborate on approaches proposing the integration of opinion analysis into the data warehouse. Lastly, section 5 will scrutinize approaches addressing social business intelligence.

3 Building Data Warehouse for Behavior Analysis

Social media is considered as an environment for human beings so they can express themselves through their interactions. Numerous approaches focused on user's activities on social media in order to help the decision makers to discover new knowledge and to analyze the behavior of people using social media.

The emergence of social media has sparked numerous research initiatives focused on behavior analysis and the extraction of knowledge from the data pertaining to users and their messaging activities. These researches include but are not limited to these works [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. We examine these works which are the most representative of the state of the art that treat data warehouse.

3.1 Approach Proposed by (Bringay et al. 2011)

[11] devised a multidimensional star model tailored for the analysis of a substantial volume of tweets. Their approach incorporated information retrieval methods, introducing a modified metric named "TF-IDF $_{adaptive}$." This metric aimed to identify the most significant words based on hierarchical levels within the cube, specifically focusing on the location dimension. The initial step involved instantiating the multidimensional model of tweets to accommodate new dimensions that vary across different contexts.

Subsequently, the authors introduced the "TF-IDF $_{adaptive}$" measure in the second step, identifying the most relevant words in alignment with the hierarchy level of the cube by incorporating a new dimension termed MotMesh (MeshWords) into their multidimensional model. The final step aimed to determine the context of a tweet.

However, it's important to note that their case study was confined to a specific domain—tracking the progression of diseases, utilizing the MeSH (Medical Subject Headings) thesaurus. The proposed model was specifically crafted for this particular thematic trend.}

Additionally, the absence of schema definition rules for the data warehouse and the lack of adaptation of the proposed data warehouse model to the vast amount of social data present notable limitations in their approach.

3.2 Approach Proposed by (Liu et al. 2012)

A similar approach is put forth in the work of [12], where the authors introduce a text cube for the analysis and modeling of human, social, and cultural behavior (HSCB) from the Twitter stream within a textual database. They developed the Social Cube framework, comprising four core steps.

In the initial step, a framework for the data collection component is proposed, enabling the automatic extraction of relevant data from diverse sources like Twitter. Utilizing Twitter's Application Programming Interface (API), the authors create code for the automated extraction of real-time tweets on a given topic, transforming and loading them into the textual database for subsequent HSCB analysis.

Moving to the second step, the HSCB feature analysis component extracts linguistic features from the text using text analytics tools. These linguistic features serve as fundamental elements for HSCB dimensions, including affect, deception, and a sense of fatalism, as well as any prospective HSCB dimension.

The analysis framework considers the selection of linguistic features with reference to theories and psychological expectations.

The third step involves the design of a star schema to store the linguistic features extracted from various HSCB dimensions. Finally, in the last step, data mining techniques are employed to identify crucial linguistic features and construct predictive models for each HSCB dimension based on the selected features.

The primary objective for decision-makers in this work is to define the architecture of a text cube geared towards organizing social media data in multiple dimensions and hierarchies. However, it's noteworthy that the proposed data warehouse model is not tailored to effectively handle the substantial volume of social data.

3.3 Approach Proposed by (Rehman et al. 2012)

In the same context, [13] introduced a system designed for warehousing streams from Twitter, structured around a five-layer architecture. The layers include: i) The data source layer, utilizing available Twitter APIs, ii) The ETL layer (Extract, Transform, and Load) responsible for extracting data from tweets and processing it into a suitable format for the target database, iii) The Data warehouse layer dedicated to storing data derived from tweets, iv) The Analysis layer specifically designed for OLAP analyses of the tweets, and v) The Presentation layer presenting the results of the analyses.

The objective of this paper is to construct a comprehensive cube for OLAP analysis of tweets. However, it's important to note that the schema definition rules for the data warehouse are overlooked, and the proposed model is not optimized for handling the vast volume of social data. The authors focused on a specific type of social media, and their approach seems tailored to analyze a particular event.

3.4 Approach Proposed by (Cuzzocrea et al. 2015, 2016)

The presented work concentrates on the integration of knowledge mining approaches, specifically FFCA, and OLAP technology for analyzing unstructured data exchanged in social media, facilitating advanced analytics services. With Twitter as the focal point, the authors underscore the significance of implicit information within tweets that goes beyond the explicitly available metadata.

In a subsequent extension of their work, [15] delves into the definition of a multidimensional data model for storing tweet data to facilitate OLAP analysis. The authors commence by outlining the structure of the data cube.

Within their tweet cube model, dimensions are categorized into two types: (i) Semantic Dimension, extracted from the Wikipedia knowledge base, leveraging titles of Wikipedia articles and the Wikipedia category graph, and (ii) Metadata Dimension, encompassing information about the tweet derived from its metadata, such as timestamp, user, hashtag, location, etc.

Furthermore, the authors introduce a measure that exploits a wikification service, representing a sentence with a set of Wikipedia concepts. Subsequently, a summarization algorithm is proposed to select the most representative tweets for each cube based on the OLAP dimensional fact model.

A case study is presented, addressing microblog summarization through the utilization of timed fuzzy lattice generated from the execution of time-aware FFCA on the unstructured content of tweets.

3.5 Approach Proposed by (Yangui et al. 2017)

The researchers in [16, 17] present a four-stage methodology for defining a data warehouse schema from social networks. Initially, they design the initial data warehouse schema using a classical approach based on existing methods and tailored to structured and heterogeneous sources. In the second stage, they articulate a set of transformation rules facilitating the conversion of data warehouse schema concepts to specific concepts within a NoSQL database.

This stage comprises three key steps: defining features, generating clusters, and determining multidimensional concepts. Subsequently, social network profiling data is clustered based on user requirements, enabling the dynamic discovery of multidimensional concepts. To achieve this, the

SHICARO (Semi-supervised Hierarchical Clustering based on ranking features using Ontology) approach is employed.

Finally, the discovered multidimensional concepts enrich the NoSQL data warehouse schema to ensure schema evolution. However, it's important to note that in this work, the schema definition rules for the data warehouse are overlooked, and the proposed model is not adaptable to the substantial volume of social data.

3.6 Approach Proposed by (Moulai and Drias 2018)

In their work, Moulai and Drias [18] introduced a specific type of data warehouse named "Information Warehouse," primarily designed with information fact tables. The authors put forth a generic information warehouse architecture intended for the storage and analysis of various information sources, including scientific papers, press articles, and social media.

The outlined infrastructure is then applied to the case of Twitter, where a multidimensional information model is defined. The collected information flow is subjected to analysis using the A-priori algorithm, a data mining technique, to uncover association rules indicative of the topics discussed in the Twitter collection. The results obtained are promising, confirming the potential of the proposed paradigm.

However, despite the robust foundation of their approach for identifying a multidimensional structure suitable for social media data, the authors did not emphasize the semantics of the analyzed text in favor of a specific application domain. Additionally, they did not address the challenges related to the volume and velocity of social media data.

3.7 Approach Proposed by (Jenhani et al. 2019)

In their work, the authors in [19] introduce a large-scale system designed for structured information extraction from streaming and voluminous social media text, with the aim of easily integrating this information into a data warehouse.

They implement a novel approach within a large-scale architecture comprising Storm and Hadoop for extracting events from streaming social media text. Leveraging Storm's real-time processing capabilities; they collect tweets from the Twitter Streaming API and employ clustering techniques for data filtering.

To facilitate this process, the authors propose a snowflake schema for modeling event data. This schema enables both independent analysis of social media events and their integration with the existing enterprise data warehouse.

Additionally, the authors utilize the power of Hadoop for batch processing of large volumes of data, focusing on structured information extraction, specifically entities and events.

This entire process is considered a data preparation stage preceding the well-known ETL (Extract, Transform, Load) process. Once events are extracted, they can be loaded into the Social Media Data Warehouse (SMDW) using any ETL tool, and standard OLAP, data mining, and BI tools can be employed for further analysis.

For the data warehouse design, the authors propose a customized conceptual model specifically tailored for event data type modeling. This multidimensional design allows for the separate analysis of social media events and their integration with the existing enterprise data warehouse (EDW) data, enabling more accurate analysis. The connection between the SMDW and the EDW is facilitated through the addition of an intermediate bridge table.

3.8 Approach Proposed by (Ben kraiem et al. 2020)

In their work, [20] applied data warehousing technology to facilitate a comprehensive analysis of massive data volumes generated by the Twitter social network. They introduced a multidimensional model dedicated to online analytical processing (OLAP) of data exchanged through tweets.

The model comprises a set of facts and dimensions constructed from the structure of tweets, designed to be generic and not limited to predefined analytical requirements, thus offering broad analytical potential and capacity to address ad-hoc needs. Special considerations were given to the specifics of tweet data, including links between tweets and tweet responses. To

accommodate this, the authors extended the concept of a fact by proposing a new type named “reflexive fact”, allowing connections between instances of the fact table and one or several instances of the same table.

Various options for enriching the multidimensional model were suggested, such as adding new elements like measures and hierarchies. To validate their proposals, the authors developed a software prototype called TweetOLAP, demonstrating through extensive experimentation how the resulting data warehouse can be used for various analytical tasks.

Additionally, a solution based on five OLAP operators was proposed to support analyses considering the specificities of the proposed multidimensional model called “Tweet Constellation”.

In a related work [21], facing large volumes of data with a significant amount of missing data, the researchers proposed extended versions for conventional OLAP operators, namely Null-Drilldown, Null-Rollup, and Null-Select. These extended operators process OLAP queries on datasets with missing data, providing options for handling missing data in analysis results.

The researchers introduced the options of All, $All_{NullLast}$, or Flexible, with $All_{NullLast}$ reorganizing the multidimensional table by moving non-significant rows to the bottom and Flexible displaying percentages of non-null data. Furthermore, to exploit the reflexive relationship on fact instances, two specific OLAP operators, FDrilldown and FRollup, were proposed.

These operators facilitate intuitive navigation between different levels within the fact, catering to decision-making applications and enabling diverse analyses by showcasing how information propagates through each tweet.

4 Building Data Warehouse for Opinion Analysis

Sentiment analysis, also known as opinion mining, involves the computational study of opinions, sentiments, and emotions expressed in text. The integration of opinion data has become a prominent topic in various research communities, notably within Data Warehousing and Decisional Support.

The aim of this section is to thoroughly examine the existing approaches for integrating sentiment analysis into the schema of data warehouses. Specifically, we scrutinize the research conducted by the most notable authors [22, 23, 24, 25, 26, 27, 28, 29, 30] which are the most representative works.

4.1 Approach Proposed by (Moya et al. 2011)

[22] introduced a multidimensional data model designed to integrate sentiment data extracted from Web 2.0 customer opinion forums into the corporate data warehouse.

This model comprises two primary components. The first part focuses on corporate information, sourcing data from internal documents and company databases through traditional Extract, Transform, and Load (ETL) processes. Corporate facts in this section typically encompass standard Business Intelligence (BI) measures such as sales and profits.

The second part is dedicated to sentiments within the data warehouse, containing information derived from user reviews on products obtained from opinion forums. This part operates at two levels of granularity: the overall sentiment regarding the product and specific sentiments about the product's features, as mentioned by users in their opinion posts.

The initial step in the proposed approach involves gathering customer opinions from the web to identify products that have been subject to customer feedback. Subsequently, the authors identified potential features influenced by opinion words. They compiled a list of opinion words through the intersection of adjectives from two lists, manually verifying and supplementing it with adverbs and verbs of context-independent polarity.

The second step involves classifying potential features based on their importance, determined by both functional relevance and feature frequency [31]. Finally, the authors suggested calculating characteristics of synonym groups using the Jaccard distance function, which considers both the lexicon and overlapping word synonyms.

4.2 Approach Proposed by (Costa et al. 2012)

[23] introduced an architecture that emphasizes the integration of social networks and sentiment analysis with user decision-making processes. The primary focus of this work is on extracting data from Twitter and applying sentiment analysis to generate a data warehouse.

The proposed software architecture, named Online Social Networks Business Intelligence (OSNBIA), is structured as follows: (1) Social Networks Crawling: Utilizing application programming interfaces (API) provided by Social Network Sites, the authors retrieve tweets containing the text ''lenovo ThinkPad". (2) Data Cleansing: In this phase, inconsistencies in the data are corrected before moving on to the next step.

Aspects such as completeness, consistency, validity, conformity, accuracy, and integrity are addressed. For missing data, the constant 'NOT AVAILABLE' is used to fill attributes when certain data attributes are inaccessible or lack content. (3) Analysis using Mining Algorithms: With the cleaned data, [23] employed link mining and opinion mining algorithms to identify the sentiments expressed in 58,906 tweets.

New attributes resulting from this analysis are added to the tables, creating an analyzed data repository that is inserted into the data warehouse. (4) Data Warehouse Insertion: The researchers then inserted these files into a data warehouse to analyze the sentiments expressed in tweets relative to sales performance.

Following the generation of the data warehouse, QlikView was utilized to develop a Business Intelligence (BI) analysis application, providing greater flexibility for data analysis. However, it is important to note that a drawback of this approach is the absence of a data warehouse schema.

4.3 Approach Proposed by (Rehman et al. 2013)

[24] extends their previous work [13], aiming to enhance Online Analytical Processing (OLAP) for multidimensional analysis of data from social networks. The extension involves integrating text mining methods, opinions, and knowledge discovery techniques with a data warehousing system. The researchers initiate the process by identifying the facts, measures, hierarchies, and dimensions of the Twitter data warehouse.

This proposed data warehouse adopts the aggregation-centric multidimensional data model, facilitating drill-down and roll-up operations. Subsequently, the researchers enrich and extend the social media dataset to provide new analytical aspects for business analysts. Text and opinion mining algorithms, along with sentiment analysis, are applied to support both exploratory and predictive analysis of social media data.

Two APIs, namely AlchemyAPI for sentiment analysis and OpenCalais for topic extraction and concept tagging, are utilized to ensure uniformity of results. Using decision tree classifications, the authors mine the dataset for features classifying tweets into multiple popularity classes, considering hashtags, sentiment, and user popularity as input features for the model.

In the final stages, the researchers modify concepts related to slowly changing dimensions as presented in [6]. They update the name and screen name attributes by replacing existing data with new ones, focusing on changes in dimensions and hierarchies within the data warehouse. The ETL process (extraction, transformation, and loading) is repeated each time new data are uploaded into the data warehouse.

The resulting data warehouse was used in an attempt to perform analyses during the 2012 European Football Championship final played between Spain and Italy on July 1, 2012.

4.4 Approach Proposed by (Walha et al. 2016)

[25] introduces the integration of social opinion data in multidimensional design, combining sentiment analysis techniques and Extract, Transform, Load (ETL) design to present a novel approach for social ETL design. The researchers define a lexicon opinion analysis approach that extracts sentiment polarity from informal text expressed on the Twitter social network.

They propose a new algorithm, POLSentiment, based on lexical resources to extract opinion words and emoticons from tweets and then determine their positive or negative polarity. The process involves the following steps: (1) *Creating an*

Opinion Dictionary: The researchers construct an opinion dictionary based on the AFINN word list, specifically designed for microblogs and considered a standard for opinion analysis.

This dictionary includes 2477 English words and phrases. They further enrich the dictionary with a list of positive and negative opinion words, sentiment words for English, and emoticons. (2) *Automatic Lexicon-Based Method*: The researchers propose an automatic lexicon-based method to determine tweet polarity based on the opinion lexicon used in the tweet, which includes emoticons and opinion words. (3) *Tweet Preprocessing*: This step involves cleaning the tweet by removing diacritics, useless characters, URLs, repetitive characters, etc. (4) *Tweet Tokenization*: The text is segmented into words, phrases, and symbols called tokens. (5) *Detecting Tweet Polarity*: This process determines whether a piece of writing is positive, negative, or neutral.

The authors extract the opinion lexicon used in the tweet, including opinion words, their modifiers, and emoticons. The lexicon is determined from previously defined opinion and emoticon dictionaries. The POLSentiment algorithm is then used to calculate tweet polarity and perform opinion analysis. (6) *Loading Step*: In the final step, opinion analysis subject and axes are defined in a Data Warehouse Bus (DWB) star schema, which includes dimensions, measures, facts, attributes, and parameters.

4.5 Approach Proposed by (Ahsene Djaballah et al. 2019)

In 2019, Ahsene Djaballah et al. [26] introduced an approach for analyzing terrorism-related activities in social networks through the utilization of data warehousing and OLAP analysis.

The architecture of their proposed approach comprises five layers: (1) the data source layer, which is represented by available APIs for searching social network data and their metadata, including external sources such as location data; (2) the ETL layer, responsible for extracting data from heterogeneous sources, performing necessary treatments and cleanings using text mining techniques, and loading the processed data into the data warehouse; (3) the Data Warehouse layer, characterized by a star multidimensional model designed for analyzing business processes; (4) the analysis layer, incorporating an OLAP server that translates users' queries into requests on the data warehouse and provides results to decision support tools; and (5) the presentation layer, consisting of reporting tools for different visualizations of the analysis layer.

The researchers implemented their approach on the Twitter social network, employing the FEEL dictionary (a French Expanded Emotion Lexicon) for sentiment analysis to determine positive scores, specifically tweets inciting terrorism. However, a drawback of this approach is its reliance on a single type of social media, and the proposed model may not be adaptable to the vast amount of social data.

4.6 Approach Proposed by (Valêncio et al. 2020)

[27] introduced a normalized data warehouse schema designed for modeling social media data originating from two distinct platforms, Facebook and Twitter. This normalized data warehouse model aims to eliminate redundant data storage by focusing on quantitative attributes from publications on social media, thereby enhancing the efficiency of data mining algorithms and reducing execution time.

The authors presented the Configurable Load and Acquisition Social Media Environment (CLASME), a tool facilitating data preparation from Facebook and Twitter to uncover valuable knowledge and support analysts in decision-making. The ETL (Extract, Transform, Load) phase involves obtaining data from various sources, followed by data cleaning and standardization during the transformation phase, depending on the application's objectives.

The load phase maps and stores the transformed data into the appropriate section of the data warehouse, known as Data Mart. Subsequently, qualitative data are categorized as positive, negative, or neutral, and opinions about the posts are determined before loading quantitative data.

Once the ETL and data warehousing processes are complete, data mart algorithms are applied during the data analysis phase to validate the classification model. Finally, the results obtained

are interpreted to facilitate the decision-making process.

4.7 Approach Proposed by (Gutiérrez-Batista et al. 2021)

[28] aimed to enhance decision-making based on a substantial volume of text extracted from social media, spanning various topics and timeframes, by introducing a fuzzy sentiment analysis dimension. The authors employed OLAP for text storage and extraction within their multidimensional model.

The fuzzy dimension's hierarchy levels encompassed five tiers, ranging from the most general to the most specific. Texts underwent clustering, sentiment scores were assigned, and Fuzzy Logic was applied for processing. Tools like Text Blob and VADER, known for their efficacy with social texts, were utilized as unsupervised tools to ensure a more generic applicability.

A comparative study conducted on real tweets and movie reviews, employing six machine learning algorithms, demonstrated the high performance and accuracy of this method. The proposed approach involves four primary processes. Initially, a Fuzzy Sentiment Dimension is created from texts extracted from social networks to facilitate multidimensional sentiment analysis.

Subsequently, an automatic process of document clustering is established, considering the sentiments expressed in the texts. Sentiment evaluations are automatically assigned to each document using *linguistic labels*, and a hierarchical structure is constructed to enable sentiment analysis at different granularity levels.

An adaptive process is then developed for the automatic selection of linguistic labels and the definition of membership functions for these labels. Finally, storage and query extensions are defined to support the Fuzzy Sentiment Dimension.

The first extension allows the definition of structures such as cubes, dimensions, hierarchies, and levels to facilitate fuzzy multidimensional analysis of users' opinions.

The second extension enables querying the Fuzzy Sentiment Dimension by defining operations in the multidimensional model, such as roll-up and drill-down. The notable advantage of this method lies in its fully automated and unsupervised nature.

4.8 Approach Proposed by (Moalla et al. 2017, 2022)

In 2017, Moalla et al. [29] propose a new method of opinion analysis based on machine learning that determines the polarity of users ‘comments shared on different social media. The latter will be integrated in the ETL (Extract, Transform and Load) process to analyze the users' opinions.

The proposed method is based on the n-grams technique to construct a semi-automatic dictionary for positive and negative keywords that is used in the learning phase to establish the prediction model. In addition, they propose a new features vector specific for social media for classifying the comments as positive, negative or neutral.

The evaluation results performed on the both publicly data sets Stanford Twitter Sentiment (STS) and Sanders dataset showed a high accuracy level. In 2022, [30] presented an extension of their previous work. They propose a new approach for building a data warehouse from social media for opinion analysis.

The proposal consists of four phases: Data extraction and cleansing, Transformation, loading, and analysis. They have presented the different stages of data extraction and cleansing. These steps are intended to model data marts for each social media. In the transformation phase, the authors have detailed the mapping and merging step to obtain a generic data warehouse schema.

In addition, they have summarized the opinion analysis step. After that they presented the implementation of a data warehouse under a database NoSQL- oriented documents. Finally, in the analysis and reporting steps, they performed some queries on our data warehouse.

5 Building Data Warehouse for Social Business Intelligence

Business intelligence (BI) is the set of devices, the querying tools and the data analysis used to drive a company and help it in the decision making. Today, BI decision-making processes are informed by social media pattern. Social networks are an essential aspect of the information infrastructure. These social media sites have attained an unparalleled degree of penetration for users,

customers, and enterprises to provide the professional environment with a valuable information source.

In this context, a novel area known under the name of Social Business Intelligence (SBI) appeared which refers to the discipline that aims at combining corporate data with user generated content (UGC) to let decision makers analyze and improve their business based on the trends and moods perceived from the environment [33].

The purpose of this section is to study in depth the proposed existing approaches dealing with behavior analysis. More precisely, we examine these works of [32, 34, 35, 36, 37, 38, 40].

5.1 Approach Proposed by (Gallinucci et al. 2013, 2015)

In 2013, Gallinucci et al. [32] focused in their study on the social business intelligence, which enables to combine corporate data with the user-generated content (UGC) to help the decision makers to improve their company and aggregate subjects at different levels.

Therefore, they proposed to model topic hierarchies in ROLAP systems called meta-stars. This approach is based on the combination of the traditional dimension tables and the navigation tables to deal with the dynamics of the subject area. Gallinucci et al. [32] introduced the architecture for SBI (social business intelligence) that integrates both the corporate data and the sentiment data of the Web users (UGC).

In the implementation of this architecture, the researchers manually defined the topics and roll-up relationships. After that, they presented a cube to analyze the sentiments expressed by the Web users. Furthermore, they defined a set of relationships between the topics in the hierarchy roll-up. In fact, they proposed to model topic hierarchies on ROLAP platforms combined with classical dimension tables and with recursive navigation tables.

Then, they extended the obtained result by using meta-modeling called meta-stars. Therefore, the authors tested some OLAP queries to evaluate the performance of meta-stars against star schema. In fact, they noted that meta-stars are better in terms of space efficiency and query expressiveness and lower in terms of time.

In [33] improved their work by extending their meta-stars model. Firstly, they took non-covering and non-strict hierarchies. Secondly, they dealt with the techniques of slowly changing topics and levels. Thirdly, they supported the semantics queries on topic hierarchies.

Finally, they evaluated the meta-star approach proposed on wider set of tests. Nevertheless, these works are limited to the use of one type of social media. Additionally, the proposed model is not adaptable to the huge amount of social data.

5.2 Approach Proposed by (Francia et al. 2014)

The authors of [34] propose an iterative methodology for designing and maintaining SBI applications that reorganizes the activities and tasks normally carried out by practitioners. They proposed architecture for the SBI process where the information resulting from clip analysis is stored into a data mart in the form of multidimensional cubes to be accessed through OLAP techniques.

This architecture is composed of six features: (1) An ODS (Operational Data Store) that stores all the relevant data about clips, their topics, their authors, and their source channels; to this end, a relational database is coupled with a document-oriented database that can efficiently store and search the text of the clips and with a triple store to represent the topic ontology.

(2) A data mart that stores clip and topic information in the form of a set of multidimensional cubes to be used for decision making. (3) A crawling component that runs a set of keyword based queries to retrieve the clips (and the related meta-data) that lies within the subject area.

(4) An ETL (Extraction, Transformation, and Loading) component that turns the semi-structured output of the crawler into a structured form and loads it into the ODS, and then periodically extracts data about clips and topics from the ODS to load them into the data mart. (5) A semantic enrichment component that works on the ODS to extract the semantic information hidden in the clips, such as the topic(s) related to the clip, the syntactic and semantic relationships between words, and the sentiment related to a whole sentence or to each single topic it contains.

(6) An OLAP front-end to enable interactive and flexible analysis sessions of the multidimensional cubes. The disadvantage of this approach is the lack of data warehouse schema.

5.3 Approach Proposed by (Kurnia et al. 2018)

In the same context, Kurnia et al. [35] developed a business intelligence dashboard to observe the performance of each Topic or channel of news posted on Facebook and Twitter. For this reason, a data warehouse model and software for the business intelligence system are designed and implemented. The architecture of the proposed system is composed of four stages.

Firstly, data collection is extracted from Facebook and Twitter through the API available on each platform. Secondly, content analysis used text classification techniques like Naive Bayes, Decision Tree and SVM to attribute a category or class to the data retrieved based on the characteristics of the document.

But before going into the classification algorithm processing, the data will go through the preprocessing stage such as case folding, tokenizing, filtering, and stemming to eliminate data noises. Thirdly, the Data warehousing design method used is [6] method which there are 4 stages that must be passed in the design of data warehouse that is select the business process, declare the grain, identify the dimensions, and identify the facts.

Therefore, a star schema model is proposed to show the number of comments, tweets, likes, etc., for each topic. Fourthly, the design of business intelligence with the Carlo Vercellis method, where in this method there are four main phases, namely analysis, planning, planning, implementation and control. Nevertheless, the proposed model is not adaptable to the huge amount of social data.

5.4 Approach Proposed by (Girsang et al. 2020)

Similarly, [36] presented a Business Intelligence application dashboard using a data warehouse to provide a solution for powerful, effective, and limitless news sources to the journalistic community. The proposed approach is divided into four main phases. The first phase is the data collection which consists of collecting data from selected Social Media Platforms.

Data retrieval will be carried out periodically by crawlers who have been created with the Social Media API Token input. Thus, the results of data retrieval will be saved on the database as raw data. The second phase is content analysis s including retrieving data for content analysis from each data on each Social Media platform. Then, the data is processed for text classification using SVM into 10 news categories.

The results of text processing will be stored in the database as analysis data. The third phase is the data warehouse process. This is done by defining the transformation and loading procedures of the ETL process.

Finally, the last phase is the client side. A business intelligence dashboard is created when the data is stored which can help journalism in the analysis of news information. The disadvantage of this approach is the use of one type of social media. Additionally, the proposed model is not adaptable to the huge amount of social data.

5.5 Approach Proposed by (Mouyassir et al. 2021)

The authors of [37] presented an analysis of the applicability of social media in BI that helps businesses to get a global and well-defined perception of consumer's sentiments and emotions. This process aims to analyze data according to several important components. The first step in this process is data collection.

The researchers use Apache Flume as data collection tool. In this second step, they deal with the data cleaning, including several errors that require filtering and sorting, discarding irrelevant data, meaningless data, eliminating redundant data. The third stage includes evaluating the quality of the social media data after it has been filtered, and using text classification.

Mouyassir et al. used a set of text classification algorithms like SVM, Decision Tree. The fourth step is store data. At this phase, the data will be processed in a data warehouse in a distributed and non-volatile method. Then, the authors use the NoSQL language to collect all the data that has been deposited in a distributed manner.

Finally, reports are created as part of this pre-process to help the end-users understand the results. These end-users would be able to get a better understanding of consumer behavior, allowing them to interpret the data and making it understandable.

Reporting is about transforming data into information, while analysis is the process of transforming information into knowledge. Nevertheless, in this work the schema definition rules of data warehouse are ignored. Besides, the authors have not specified a type of social media. Therefore, the proposed approach can be used only for ensuring a special treatment.

5.6 Approach Proposed by (Aramburu et al. 2021)

[38] define the special requirements of the analysis cubes of a Social Business Intelligence (SoBI) project. They present a new data processing method for SoBI projects whose main contribution is a phase of data exploration and profiling that serves to build a quality data collection with respect to the analysis objectives of the project. The authors propose a new data processing methodology that consists of three main phases: Collection Construction, Data Preparation and Data Exploitation (see Figure 1).

The first phase is the construction of a collection of tweets through an exploratory process executed by the user and directed by the quality of the recovered data. When a quality data collection is ready, in the data preparation phase, the facts of the analytical cubes are extracted from the posts and then exploited in the last phase of the process.

During the Collection Construction phase, the user executes some data exploratory and profiling tasks to assess and improve data coverage and data quality until obtaining a quality collection that meets the project's analysis objectives. More specifically, this phase consists of two complementary and iterative tasks: Evaluating the subject coverage of the collection with respect to its topics and users and analyzing and improving the quality of the collection by filtering the posts of low quality or out of the scope.

In the Collections Construction and Data Preparation phases, the processing of tweets to extract the measures and values that serve both to clean the collection and to feed the analysis cubes, can be made in different ways. Some values are directly available in the tweets metadata, such as post-date and number of followers of the user. Other values can be calculated with a simple processing like counting tweets over a period. Evaluating the grammatical richness of a post is executed by a process that calculates some textual measures [39].

Finally, in the Data Exploitation phase, the analysis cubes constructed by processing the tweets collection can be stored into the corporate Data Warehouse for future uses. OLAP applications, or any other Business Intelligence or Data Mining tools, can be applied to analyze and extract new insights from these cubes. However, in this work the schema definition rules of data warehouse are ignored. Additionally, the proposed model is not adaptable to the huge amount of social data.

5.7 Approach Proposed by (Lanza-Cruz et al. 2023)

[40] propose a methodology for author profiling (AP) in Twitter based on social business intelligence roles. The method allows the unsupervised construction of a labeled dataset that serves as input to different text classification tasks. They automatically build a training dataset from unlabeled user descriptions by making use of the multidimensional user profile knowledge model provided by the analysts.

The proposed methodology relies on semantic knowledge encapsulated in ontologies provided by analysts at the commencement of a Social Business Intelligence (SBI) project. From these ontologies, basic linguistic information is extracted to identify potential unlabeled user profiles. Consequently, the generated training data are directly associated with the concepts represented in the knowledge multidimensional model, such as users' roles.

This integration allows [40] to verify the consistency and identify conflicts within the training data. This approach contributes to the existing body of knowledge by offering an integrated perspective on ontologies and predictive models for Audience Profiling (AP) in social networks.

Table 1. A comparative study of data warehouse design approaches for behavior analysis

Approach	Modeling level		ETL process		Social media	Analysis process	Objectives	Drawbacks
	Conceptual	Logical	Tools used	Language				
Bringay et al. 2011	Star schema	Not mentioned	Postgre SQL and Pentaho Mondrian	Not mentioned	Twitter	Classic OLAP operators	Leveraging a multidimensional star model for the examination of tweets and proposing relevant metrics conducive to knowledge exploration.	Utilizing a partial modeling approach with a predefined data warehouse schema.
Liu et al. 2012	Star schema	Not mentioned	Not mentioned	Not mentioned	Twitter	Classic OLAP operators	Text cube architecture is presented for analyzing human social and cultural behavior with the capacity to develop prediction models and perform analyses.	Insufficient theoretical framework for opinion analysis.
Rehman et al. 2012	X-DFM	Not mentioned	BaseX and Microsoft SQL Server	Not mentioned	Twitter	Classic OLAP operators	The authors propose an exhaustive cube for OLAP analysis of tweets. The suggested model may be utilized to complete any tasks based on the mining or aggregation of data.	Absence of defined rules for schema.
Cuuzzocrea et al. 2015, 2016	DFM	Not mentioned	Not mentioned	Not mentioned	Twitter	Classic OLAP operators	OLAP technology used with knowledge mining methods (i.e., FFCA) to analyze multidimensional tweet streams of unstructured social media data.	The vast volume of social data cannot be accommodated by the suggested approach.
Yangui et al.. 2017	X-DFM	NOSQL Data base	Not mentioned	Not mentioned	Twitter		The suggested methodology makes use of the established design methods' maturity, the NOSQL Data Base's scalability, and the capacity to dynamically identify multidimensional concepts using clustering algorithms.	The schema definition rules of data warehouse are ignored.
Moulai and Drias. 2018	Star schema	Not mentioned	Not mentioned	Not mentioned	Twitter		A specific DW called “Information Warehouse” which focused on semantic extraction and modeling. It is the structure which stores data having meaning and significance such as text, image, video, etc.	Absence of the semantic of analyzed text to the profit of a specific application domain. Absence of volume and velocity problems of social media data.
Jenhani et al. 2019	Snowflacke Schema	Not mentioned	Hadoop	Not mentioned	Twitter		An approach in a large-scale architecture based on distributed storage and parallel processing for event extraction from streaming social media data.	the design process is not present in detail by presenting the rules that lead to the data warehouse schema.
Ben kraiem et al. 2020	Extended conceptual constellation schema	ROLAP	JAVA and ORACLE 10	Not mentioned	Twitter	Extended OLAP operators	The conceptual model takes into account the specificity of tweet and tweet response and missing data. Proposal of new OLAP operators to deal with the specificities of the proposed model.	The proposed model is not adaptable to the huge amount of social data.

Furthermore, the method is adaptable to dynamic scenarios where semantic knowledge requires updating to accommodate new roles or dismiss others. In such instances, the method constructs a new training dataset and develops new predictive models based on the updated knowledge.

Another noteworthy aspect of this approach is that the utilization of user profiles, rather than their posts or metrics, is deemed sufficient for characterizing their business roles. Previous methods, relying on posts and metrics, yielded poor results due to the redundant, often shared, and heterogeneous nature of social network content.

The multidimensional AP approach addresses the demand for analysis based on dynamic dimensions inherent in social media. It aids information systems in characterizing the audience of popular topics and news.

However, a drawback of this approach is the absence of a data warehouse schema. Additionally, the authors concentrated on a specific type of social media.

Table 2. A comparative study of data warehouse design approaches for opinion analysis

Approach	Modeling level		ETL process		Social media	Analysis process	Objectives	Drawbacks
	Conceptual	Logical	Tools used	Language				
Moya et al. 2011	Constellation schema	-	SQL server Business Intelligence Studio	Not mentioned	web	Classic OLAP operators	The display of a sentiment-integrated multidimensional data model. Sentiment data extraction yields a semantically rich data collection that supports sophisticated queries.	Lack of DW schema.
Costa et al. (2012)		-	Data Manager and ORACLE	Not mentioned	Twitter	Classic OLAP operators	The establishment of a business intelligence tool that combines social network and sentiment analysis with the decision-making processes of the user.	Lack of schema definition rules.
Rehman et al. 2013et al. 2012	X-DFM	-	BaseX and Microsoft SQL Server	Not mentioned	Twitter	Classic OLAP operators	The integration of opinion mining methods and knowledge discovery techniques into the data warehousing system, in order to perform multidimensional social media analysis.	Insufficient theoretical framework for opinion analysis.
Walha et al. 2016	Star schem	-	Not mentioned	Not mentioned	Facebook	Not mentioned	An ETL design technique that incorporates user opinions as given on the well-known social network Facebook. The list of each element of the ETL process is defined.	The massive volume of social data cannot be accommodated by the suggested.
Ahsene Djaballah et al. 2019	Star schema	Not mentioned	Postgre RDBMS	-	Twitter	Classic OLAP operators	Proposition of a data warehouse using data mining technique to analyze a related to terrorism in the social network twitter.	Lack of schema definition rules.
Valêncio et al. 2020	Constellation schema	Not mentioned	PostgreSQL9.5 JAVA	-	Facebook Twitter		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 supports the application of techniques to discover useful knowledge.	Lack of a theoretical approach for opinion analysis.
Gutiérrez-Batista	Star schema	-		-	Twitter Movie reviews	Extended OLAP operators	Creating a Fuzzy Sentiment Dimension from texts extracted from social networks that facilitates the multidimensional sentiment analysis in social networks. Establishing an automatic process of documents clustering, taking into account the sentiments expressed in the.	The proposed model is not adaptable to the huge amount of social data.
Moalla et al. 2017, 2022	X-DFM	Presented	Microsoft SQL Server and XML (2017) MongoDB	-	Facebook Twitter Youtube		Proposition of a technique for social media opinion analysis that uses machine learning to determine the degree of polarity in user comments.	Insufficient theoretical framework for opinion analysis.

6 A Comparative Study of the Existing Approaches

Social media, as a burgeoning data source, has introduced novel challenges in data analysis and manipulation. The research landscape has witnessed a surge in studies focusing on the analysis of data extracted from social media, leading to the emergence of new analytical domains.

However, a limited number of studies have delved into the realm of multidimensional data modeling of data warehouses derived from social media.

Table 1 provides a comparison of behavior analysis approaches, while Table 2 delineates a comparative overview of methods integrating sentiment analysis into data warehouse structures.

Table 3 presents a comparative analysis of approaches integrating social business

Table 3. A comparative study of data warehouse design approaches for social business intelligence

Approach	Modeling level		ETL process		Social media	Analysis process	Objectives	Drawbacks
	Conceptual	Logical	Tools used	Language				
Gallinucci et al. 2013, 2015	Meta-star schema	-	Talend	Not mentioned	Web	-	Model suggestion for ROLAP platforms that considers topic hierarchies The ability to enable OLAP queries with increasing expressiveness and complexity, starting with queries that solely use static levels and progressing to queries that take semantics into account.	Topic and roll-up relationships are defined manually.
Francia et al. 2014	Meta-star schema	-	MongoDB	Not mentioned	Twitter	-	The proposal of an interactive methodology for designing and maintaining Social Business Intelligence (SBI) applications.	Lack of schema definition rules.
Kurnia et al. 2018	Star schema	-	CodeIgniter framework	-	Facebook + Twitter	-	Development of a business intelligence dashboard to evaluate the performance of each topic posted on Facebook and Twitter.	Partial approach at the modeling level (fixed DW schema).
Girsang et al. 2020	Star schema	-	Pentaho	Python script	Twitter	-	The presentation of a Business Intelligence application dashboard employing a data warehouse to assist and provide the journalistic community with a solution for powerful, effective, and limitless news sources.	The proposed model is not adaptable to the huge amount of social data.
Mouyassir et al. 2021	Meta-star schema	-	Apache Flume	NoSQL	Not mentioned		An analysis of the applicability of social media in BI that helps businesses to get a global and well-defined perception of consumer's sentiments and emotions.	Lack of schema definition rules.
Aramburu et al. 2021	Star schema	-	-	-	Twitter		Building a quality data collection in the data preparation phase of Social Business Intelligence projects. A methodology that considers collection construction as an iterative exploration process in which the user analyses the current collection from the point of view of the analysis objectives and discovers clues about how to improve it.	Lack of DW Schema.
Lanza-Cruz et al. 2023	Constellation schema	-	-	-	Twitter		The application of author profiling (AP) in order to characterize both the contents generators and the audience that is interacting with these contents.	The proposed model is not adaptable to the huge amount 0.

intelligence into data warehouse schemas. Several criteria are employed for comparison:

– Modeling level: Indicates the level of modeling, encompassing conceptual (star model, constellation model, snowflake model, DFM, x-DFM, etc.) and logical models (ROLAP, MOLAP, HOLAP).

– ETL process (extract–transform–load) : Encompasses the tools and modeling languages employed in the ETL process.

– Social media: Identifies the specific social media platforms used as data sources.
– Analysis process: Reveals whether the approach utilizes classic OLAP operators (drill down, rollup) or defines specific operators.
– Objective: Provides a succinct overview of the general idea behind each approach.
– Drawbacks: Describes the limitations and shortcomings of each approach.

Upon examining the modeling level, it is apparent that most approaches explicitly address conceptual modeling, with [20] being an exception as it presents both logical and conceptual modeling.

Regarding the social media criterion, the majority of approaches focus on a single social media platform, such as Twitter, Facebook, or the web. Only a few studies tackle the challenge of modeling data warehouse schemas from multiple social media platforms [29, 33, 35, 27, 28].

In terms of the ETL process criterion, which is crucial in data warehouse construction, only [25, 36, 37]) explicitly address this stage. Notably, other researchers do not provide a detailed definition of the various functions of the ETL process, and diverse tools are employed in this process.

Concerning the analysis process criterion, most approaches leverage classical OLAP operators, with a notable lack of specific operator definitions, except for the approach proposed by [20], which introduces new OLAP operators enhancing existing solutions and dealing with missing values and reflexive relationships on fact instances.

Each of the discussed approaches has its strengths but also exhibits certain weaknesses. Notably, existing design methods predominantly focus on Twitter as a data source, neglecting other social media platforms. The design process lacks detailed presentation, particularly in defining rules guiding data warehouse schema creation.

Despite the use of commercial tools for opinion analysis in the reviewed approaches, there is a noticeable absence of a theoretical framework for opinion analysis. Moreover, only a few works have concentrated on creating data warehouses under Hadoop, MapReduce, and NoSQL databases to handle the voluminous and massive data generated from social media.

To address these limitations, we propose a novel approach leveraging data warehousing technology to comprehensively analyze massive data volumes from various social media platforms. This approach aims to define a method for opinion analysis within a decisional system and utilize NoSQL databases to efficiently handle large amounts of data and enhance the analysis process.

7 Conclusion

In this work, we reviewed the research on social media data warehouse architecture strategies. To be more explicit, we discussed the fundamental ideas behind data warehouses and social media, and we classified social data warehouse design methods into three groups, namely behavior analysis, integration of sentiment analysis and social business intelligence in data warehouse schema.

Subsequently, based on the criteria we had established, we offered a comparison of the existing approaches. These criteria include modeling level, ETL process, the used social media, the objective and the drawbacks of each approach. Although the contributions that have been given are strong, they have significant flaws.

They may be summed up as the absence of schema defining standards, the usage of just one particular social networking platform, and the exclusive reliance on relational databases for storage. As future work, we intend to apply the data warehousing technology to enable comprehensive analysis of massive data volumes generated by the most popular social media. We aim to propose a multidimensional model dedicated to the on-line analytical processing (OLAP) of the data exchanged through social media.

We will ensure that this model is generic, that is, not limited to a set of pre-determined analytical requirements, which gives it a broad analytical potential and capacity to respond to ad-hoc needs. Besides, we will also take into account the specificities of such data. It would be interesting to define an approach enabling OLAP to keep up with volatile data using the concepts of slowly changing

dimensions to enable analysis of both the recent state of data and any of its previous states. Also, it would be interesting to define new OLAP operators that take into consideration the specificities of data extracted from social media.

These operators will allow facilitating the interpretation of the results of the multidimensional analyses on the tweets and their metadata. We also expect to exploit the "Text Mining" techniques in order to extract knowledge from data and strengthen more semantics.

8 Declaration

- **Declaration of interests:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- The authors declare the following financial interests/personal relationships which may be considered as potential competing interests
- **Financial interests:** The authors have no relevant financial or non-financial interests to disclose.
- The authors have no conflicts of interest to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

References

1. **Kaplan, A. M., Haenlein, M. (2010).** Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, Vol. 53, No. 1, pp. 59–68. DOI: 10.1016/j.bushor.2009.09.003.
2. **Cuzzocrea, A., De Maio, C., Fenza, G., Loia, V., Parente, M. (2015).** Towards OLAP analysis of multidimensional tweet streams. DOLAP '15: Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP, pp. 69–73. DOI: 10.1145/2811222.2811233.
3. **Chaudhuri, S., Dayal, U. (1997).** Data warehousing and OLAP for decision support. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, pp. 507–508. DOI: 10.1145/253260.253373.
4. **Rizzi, S. (2007).** Conceptual modeling solutions for the data warehouse. Data Darehouses and OLAP. Concepts, Architectures and Solutions, IGI Global, pp. 1– 26.
5. **Kimball, R. (1996).** The data warehouse toolkit: practical techniques for building dimensional data warehouses. John Wiley & Sons, Inc.
6. **Smith, T. (2009),** Conference notes - The social media revolution. International Journal of Market Research, Vol. 51, No. 4, pp. 559–561. DOI: 10.2501/S1470785309200773.
7. **Shankar, V., Hollinger, M. (2007).** Online and mobile advertising: current scenario, emerging trends, and future directions. Marketing Science Institute, Vol. 31, No. 3, pp. 206–207.
8. **Akar, E., Topçu, B. (2011).** An examination of the factors influencing consumers' attitudes toward social media marketing. Journal of Internet Commerce, Vol. 10, No. 1, pp. 35–67. DOI: 15332861.2011.558456.
9. **Bringay, S., Béchet, N., Bouillot, F., Poncelet, P., Roche, M., Teisseire, M. (2011).** Towards an on-line analysis of tweets processing. Database and Expert Systems Applications: 22nd International Conference, DEXA´11, Springer Berlin Heidelberg, pp. 154–161. DOI: 10.1007/978-3-642-23091-2_15.
10. **Liu, X., Tang, K., Hancock, J., Han, J., Song, M., Xu, R., Manikonda, V., Pokorny, B. (2012).** SocialCube: A text cube framework for analyzing social media data. 2012 International Conference on Social Informatics, pp. 252–259. DOI: 10.1109/SocialInformatics.2012.87.

11. **Rehman, N. U., Mansmann, S., Weiler, A., Scholl, M. H. (2012).** Building a data warehouse for twitter stream exploration. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE, pp. 1341–1348.
12. **Cuzzocrea, A., De Maio, C., Fenza, G., Loia, V., Parente, M. (2016).** OLAP analysis of multidimensional tweet streams for supporting advanced analytics. Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 992–999. DOI: 10.1145/285 1613.2851662.
13. **Yangui, R., Nabli, A., Gargouri, F. (2016).** Automatic transformation of data warehouse schema to NoSQL data base: comparative study. Procedia Computer Science, Vol. 96, pp. 255–264. DOI: 10.1016/j.procs.2016. 08.138.
14. **Yangui, R., Nabli, A., Gargouri, F. (2017).** DW4SN: A Tool for dynamic data warehouse building from social network. Research in Computing Science, Vol. 134, pp. 191–205.
15. **Moulai, H., Drias, H. (2018).** From data warehouse to information warehouse: application to social media. Proceedings of the international conference on learning and optimization algorithms: Theory and applications. pp. 1–6. DOI: 10.1145/3230905. 3230914.
16. **Ferdaous, J., Gouider, M. S. (2022).** Large-scale system for social media data warehousing: the case of twitter-related drug abuse events integration. International Journal of Data Warehousing and Mining (IJDWM), Vol. 18, No. 1, pp. 1–18. DOI: 10.4018/IJDWM. 290890.
17. **Kraiem, M. B., Feki, J., Khrouf, K., Ravat, F., Teste, O. (2015).** Modeling and OLAPing social media: the case of twitter. Social Network Analysis and Mining, Vol. 5, No. 47, pp. 1–15. DOI: 10.1007/s13278-015-0286-9.
18. **Kraiem, M. B., Alqarni, M., Feki, J., Ravat, F. (2020).** OLAP operators for social network analysis. Cluster Computing, Vol. 23, pp. 2347–2374. DOI: 10.1007/s10586-019-03006- z.
19. **Moya, L. G., Kudama, S., Cabo, M. J. A., Llavori, R. B. (2011).** Integrating web feed opinions into a corporate data warehouse. Proceedings of the 2nd International Workshop on Business intelligencE and the WEB, pp. 20–27. DOI: 10.1145/1966883. 1966891.
20. **Costa, P. R., Souza, F. F., Times, V. C., Benevenuto, F. (2012).** Towards integrating online social networks and business intelligence. Proceedings of the international conferences web based communities and social media, pp. 21–32.
21. **Rehman, N. U., Weiler, A., Scholl, M. H. (2013).** OLAPing social media: The case of Twitter. Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining, pp. 1139–1146. DOI: 10.1145/2492517.2500273.
22. **Walha, A., Ghozzi, F., Gargouri, F. (2015).** ETL design toward social network opinion analysis. Computer and information science 2015, Springer International Publishing, pp 235–249. DOI: 10.1007/978-3-319-23467-0_16.
23. **Djaballah, K. H., Boukhalfa, K., Bouassid, O. (2019).** Datawarehouse-based approach for the analysis of terrorism-related activities in social networks. Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).
24. **Valêncio, C. R., Silva, L. M. M., Tenório, W., Zafalon, G. F. D., Colombini, A. C., Fortes, M. Z. (2020).** Data warehouse design to support social media analysis in a big data environment. Journal of Computer Science, pp. 126–136. DOI: 10.3844/jcssp.2020. 126.136.
25. **Gutiérrez-Batista, K., Vila, M. A., Martin-Bautista, M. J. (2021).** Building a fuzzy sentiment dimension for multidimensional analysis in social networks. Applied Soft Computing, Vol. 108. DOI: 10.1016/j.asoc. 2021.107390.
26. **Moalla, I., Nabli, A., Hammami, M. (2017).** Integration of a multidimensional schema from different social media to analyze customers' opinions. 2017 11th International Conference on Research Challenges in Information

Science (RCIS), IEEE, pp. 391–400. DOI: 10.1109/RCIS.2017.7956564.

27. **Moalla, I., Nabli, A., Hammami, M. (2022).** Data warehouse building to support opinion analysis in social media. Social Network Analysis and Mining, Vol. 12, No. 1, pp. 123. DOI: 10.1007/s13278-022-00960-2.
28. **Zhang, L., Liu, B., Lim, S. H., O'Brien-Strain, E. (2010).** Extracting and ranking product features in opinion documents. Coling 2010: posters, pp. 1462–1470.
29. **Gallinucci, E., Golfarelli, M., Rizzi, S. (2013).** Meta-stars: multidimensional modeling for social business intelligence. Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP. ACM, pp 11–18. DOI: 10.1145/2513190.2513195.
30. **Gallinucci, E., Golfarelli, M., Rizzi, S. (2015).** Advanced topic modeling for social business intelligence. Information Systems, Vol. 53, pp. 87–106. DOI: 10.1016/j.is.2015.04.005.
31. **Francia, M., Golfarelli, M., Rizzi, S. (2014).** A methodology for social BI. Proceedings of the 18th International Database Engineering & Applications Symposium, pp. 207–216. DOI: 10.1145/2628194.2628250.
32. **Parama-Fadli, K. S. (2018).** Business intelligence model to analyze social media information. 3rd International Conference on Computer Science and Computational Intelligence, Vol. 135, pp. 5–14. DOI: 10.1016/j.procs.2018.08.144.
33. **Girsang, A. S., Isa, S. M., Ginzel, M. E. C. (2020).** Implementation of a journalist business intelligence in social media monitoring system. Advances in Science, Technology and Engineering Systems Journal, Vol. 5, pp. 1517–1528. DOI: 10.25046/aj0506182.
34. **Mouyassir, K., Hanine, M., Ouahmane, H. (2021).** Business intelligence model to analyze social media through big data analytics. SHS Web of Conferences Vol. 119, pp. 07006 DOI: 10.1051/shsconf/202111907006
35. **Aramburu, M. J., Llavori, R. B., Lanza-Cruz, I. (2021).** Quality management in social business intelligence projects. ICEIS, Vol. 1, pp. 320–327. DOI: 10.5220/0010495703200327.
36. **Gupta, A., Kumaraguru, P., Castillo, C., Meier, P. (2014).** Tweetcred: real-time credibility assessment of content on twitter. Proceedings of the 6th International Conference on Social Informatics, pp. 228–243. DOI: 10.1007/978-3-319-13734-6_16.
37. **Lanza-Cruz, I., Berlanga, R., Aramburu, M. J. (2023).** Multidimensional author profiling for social business intelligence. Information Systems Frontiers, pp. 1–21. DOI: 10.1007/s10796-023-10370-0.

Article received on 04/08/2023; accepted on 20/04/2023.
**Corresponding author is Maha Ben Kraiem.*