International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 03, No. 03, page 118 - 125
Submitted 6/10/2025; Accepted 17/12/2025; Published 18/12/2025

Integrating Structured and Unstructured Data for Enhanced Marketing Intelligence through Text Mining and Business Analytics

Sabreen Hashim Salman
American University of Iraq - Baghdad (AUIB), Airport Road, Baghdad, IRAQ
e-mail: sabreen.salman01@gmail.com

Corresponding Autor: Sabreen Hashim Salman

Abstract
In the digital era, the rapid growth of social media and online platforms has led to an explosion of unstructured textual data that holds significant business value. Traditional marketing strategies, once reliant on structured data such as demographics and purchase history, now benefit from insights derived from text analytics and sentiment analysis. This paper explores the integration of structured and unstructured data to strengthen marketing intelligence and customer segmentation. By utilizing text mining techniques and Natural Language Processing (NLP), unstructured data such as customer reviews and comments can be analyzed to extract sentiments, identify emerging trends, and refine customer relationship strategies. The study proposes an integrated framework that combines data extraction, transformation, and loading (ETL) processes with a data warehouse system for unified analysis. Using clustering algorithms such as K-Means and visualization tools, insights into customer behavior, preferences, and market segmentation are revealed. The paper also discusses the challenges of handling multilingual and context-dependent text, ethical and privacy considerations, and the technical architecture necessary for business intelligence implementation. Findings suggest that effective integration of textual analytics with structured data can lead to more informed decision-making, improved marketing strategies, and stronger customer engagement.

Keywords— Text Mining, Sentiment Analysis, Data Integration, Natural Language Processing (NLP), Business Intelligence

1 Introduction

In recent years, there has been a change in the types of data collected. With the emergence of many social media platforms and online shopping platforms, there has been an increase in the types of unstructured data, especially textual data. The general public is able to express their opinions, thoughts and share reviews on multiple social media platforms and online forums. As for businesses, structured data that was once used for decision making and deciding marketing strategies has now evolved and included semi-structured data and unstructured data. Through negative or positive reviews and comments from users, businesses are able to fine tune their products, services, and strategies. According to Schmidhuber et al. [1], the need to extract valuable information from textual data has led to various techniques of text mining. They state that text mining can be considered an extension of classical data mining techniques intended for unstructured and structured non-textual data. Text analysis can help a business or company find and identify useful information from large volumes of text to support well-informed decisions. Sentiment analysis can be carried out on these textual data (customer opinions) in order to gauge trends or analyze customer interests according to age groups or demographics to enhance customer satisfaction while interacting with the business or service.

1.1 *Data Analytics and Marketing Strategies*

Data analysis is simply defined as the process of collecting, modelling, and transforming data to gain information that will support and influence decision-making. When collecting data, most data is presented as quantitative data; whether discrete or continuous, it can be measured and grouped—such as age, height, weight, years of experience, or average income. Qualitative data, on the other hand, provides depth of understanding and may generate new ideas for research or intervention. This can be collected from focus groups, interviews, and online platforms like web-based forums or business social media platforms.

In the company we have chosen to conduct research on, their current aim is to collect data from their customers to further improve their marketing strategy by clustering customers into segments to identify their most ideal groups. Among the current types of structured data they have collected are year of birth, education level, marital status, income, number of children, date of enrollment, number of days since most recent purchase, and a binary indicator (1 = yes, 0 = no) of whether the customer has raised a complaint in the past two years. All these are structured data and easily visualized for decision-making. What the company could do to gain the upper hand in marketing is to collect unstructured data in the form of text, then analyze it using AI to determine patterns of satisfaction or dissatisfaction. Unstructured data, unlike structured data, cannot be stored in traditional spreadsheets; thus, many companies avoid using it due to the difficulty of large-scale analysis. Customer analysis can nevertheless be improved by incorporating this unstructured data. According to Marr [2], companies can make strongly supported decisions and improve customer relationships if they integrate and analyze data from various sources, such as online reviews and social media mentions, using AI to detect patterns.

With the vast amount of data being collected, proper data management and storage are essential. Data infrastructure is defined as the setup for storing, maintaining, and organizing data into insightful information. According to Dodds and Wells [3], efficient data infrastructure requires recognizing that it is not only physical assets such as networks and servers that matter; equal importance must be placed on policies that protect data, govern data usage, and ensure secure networking. In the company we studied, data infrastructure is not a primary concern, and only basic equipment is used to store and manage customer information. Given the sensitivity of personal customer data, the ethical dimensions of data management must be systematically considered. According to Edquist et al. [4], if customer or company datasets are breached, sold without consent, or mishandled, the company may face severe penalties, reputation loss, and even legal liability for board members and stakeholders.

Enterprise data refers to information shared across departments and users within an organization, ensuring that employees have accurate and current data needed to perform tasks, with standardized and secure storage. Among the benefits of good enterprise data management is that it enables better business intelligence. According to Malak [5], companies can use business intelligence tools to identify trends and opportunities, enabling well-supported and informed decisions.

1.2 *Literature Review*

The phrase “text mining” is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [6]. It is also considered one of the most useful tools for marketing as it supports Natural Language Processing (NLP). According to Chowdhary [6], it is becoming difficult for humans to discover knowledge and information in natural language text due to the abundant volume produced within a given time. NLP takes unstructured data and formats it into a structured form. Text mining helps a company understand its customers by learning and analyzing the sentiment behind comments and obtaining insights on trends and interests within hours.

2 Problem Statement

As more users across all ages and demographics start to use social media platforms such as TikTok, Instagram, and Twitter, proper techniques and marketing strategies should be used to appeal one’s products and services to the ideal group of customers. According to Romero et al. [7], social media platforms and web platforms enable businesses to gain valuable advantages such as increased customer traffic, improved customer loyalty, enhanced brand awareness, and strategies to attract customers to increase sales and revenues.

2.1 *Context-dependent errors*

Textual data, unlike other forms of data, contains an internal structure. With structured data, there is knowledge that is comprehensible—in other words, information that is easy to understand. According to M. S. Yafooz [8], dealing with large amounts of unstructured or textual data often leads to two common issues in text mining:

insufficient query processing performance and inaccurate information retrieval. On web platforms such as forums, discussion pages, and social media sites, users tend to use simplified language. The textual data available on these platforms may contain grammatical errors, short forms, and misspellings. Online users also often express opinions with sarcasm, masking negative meanings with positive words and phrases. Additionally, polarity in user reviews and comments may be ambiguous, such as using the phrase "not mentionable" to refer to a negative experience.

2.2 *Multilingual Text Mining*

Automatic Language Identification (ALI) is the task of automatically identifying various languages based on the textual content of a document. According to Selamat et al. [9], with the increasing availability and use of social media platforms, online text has shifted from clean, monolingual, and regular forms to short, irregular, and multilingual expressions. Common issues include transliterated text—words from one language written using the script of another—and homophonic confusions such as "accept" and "except," which may be mistyped by users and subsequently misinterpreted during text mining processes.

3 Data Integration

Unstructured data and structured data cannot be easily merged and examined using a relational database [2]. This research aims to bridge the gap between unstructured and structured data by converting unstructured data into column values. Here, we propose an unstructured data integration system that analyzes online reviews and comments using text analytics approaches to extract important information to be used to identify the ideal group of customers and the corresponding marketing strategy.

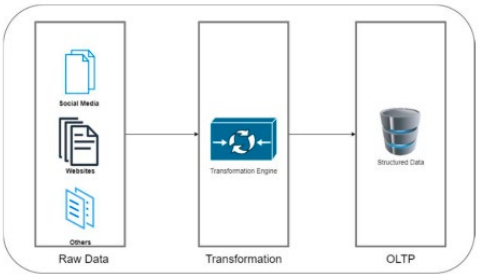

Figure 1. Initial Framework

3.1 *Extraction, Transformation and Loading (ETL)*

ETL processes extracts and reads data from one or more sources of databases, such as the business's social media platforms and web-based platforms used for data collection. Transformation is the process of converting the extracted data from one format to another so that it may be loaded into the data warehouse. The textual data is transformed using rule-based methods and merged with additional information. According to Sharda et al. [10], the three database functions are combined into a single tool that gathers data from several databases and consolidates them into one unified database or warehouse.

3.2 Data Warehouse

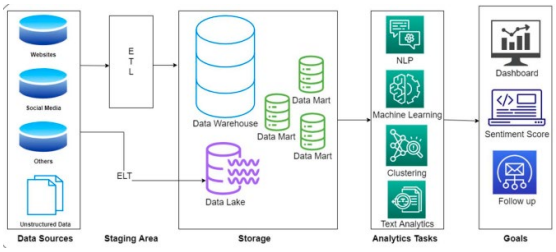

Figure 2. Data Warehouse Infrastructure

When moving data from a database into a data warehouse, it is necessary to extract data from all relevant sources for reviews and comments on the business. These data sources include comments and reviews made about the

business and its products, as well as web analytics data that indicate the business's main strengths and weaknesses. A data warehouse includes a set of business rules that govern how the data will be utilized, including summarization, attribute standardization, and calculation rules, as stated by Sharda et al. [10]. Any problems concerning the quality of data in the source files must be addressed and resolved before the data is loaded into the data warehouse.

3.3 *Results and Discussion*

To come up with a dashboard as shown in Figure which contains insights on customer personality analysis, this research utilizes data sources from Kaggle. The following visualizations are produced with the help of Qviz, a data visualization framework from Jupyter Notebook.

Figure 3. Insights on customer personality analysis

The pie chart in Figure above represents the frequency percentage of marital status. From the pie chart, it can be analyzed that 2/3 of the customers is living with partners at 64.5% while about 1/3 are single at 35.5%. The middle horizontal bar chart shows the average spendings by marital status. Despite the minority, singles spend more money on average compared to customers that have partners. As for the histogram of income distribution of customers, the salaries of customers have a normal distribution with most of the customers earning a salary between 25000 and 85000. Based on the scatterplot chart on the relationship between income vs spendings, we can see that the relationship is linear and customers having higher salaries are spending more. The bottom left pie chart shows the education level percentage distribution of the customers. Based on the pie chart, half of the customers are University students and there are more customers who hold PhD degrees than the customers who participate in Masters.

The middle bottom histogram shows the age distribution of customers, the age of customers is approximately normally distributed, with most of the customers in the age of 40 and 60. On the right of this histogram shows the scatter plot relationship of age vs spendings, we can conclude that there doesn't seem to be any clear relationship between age of customers and their spending habits.

The doughnut chart from top right shows the customers segmentations by age group. More than 50% of the customers are middle-aged adults between the ages of 40 and 60 and the 2nd famous age category would be adults with the age range of 20 and 40. Below the doughnut chart shows the horizontal bar chart of average spendings by age group. The insight we can obtain is that middle-aged adults spend much more than compared to the other age groups.

According to Figure 4, with geospatial analysis a geo map represents the countries where customers log in to the company website. We can see that the customers are mostly from Asian countries. Majority of the customers log in to the website from China with the highest frequency of 10 whereas Greenland and Mongolia are the countries with the lowest frequency of 2.

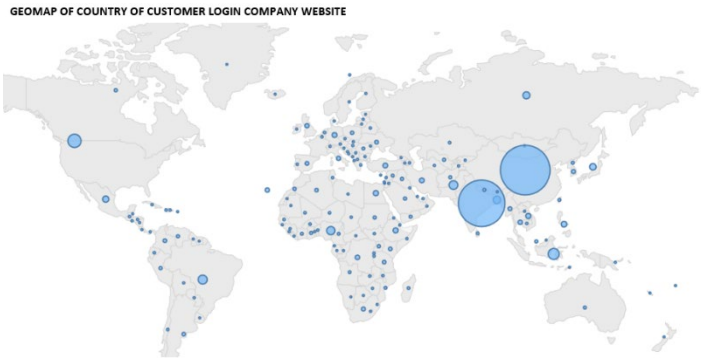

Figure 4. Customers log in to the company website

3.4 *Data Collection Process*

The data collection process starts off with web crawling on e-commerce websites such as Amazon or Taobao which contain feedback/opinions on certain products. The scraper then extracts this data for analysis.

3.5 *Web Crawling*

A crawler/spider is a program that "crawls the web to retrieve web pages, the crawler discriminately collects the data including URLs, meta tags, web pages and store it. Whereas web crawler is an automatic bot that extracts publicly available data from websites. In this paper, web crawling is used to crawl text of consumers' feedback or opinion on products or company websites according to ratings, bestselling products, or website name.

3.6 *Data Cleaning*

Preparing data is an important stage in data analysis process, poorly prepared data can lead to inaccuracy in analysis if not addressed. Data normalization is an approach that could be used to clean text where it converts a word's affixes into its base form. Tokenization is the technique of splitting a text into smaller units known as token whereas normalizing is removing redundant information such as punctuation.

3.7 *Data Mining*

Text clustering is the implementation of data mining in cluster analysis to text-based documents. It uses machine learning and natural language processing to group documents from a large collection that are similar in characteristics into cluster to determine their similarities. The k-means clustering algorithm divides n documents into k - clusters in the context of text data based on the distance between points and cluster centers. The four basic steps of K-Means are as follows:

1. Determining the centers
2. Assigning points to clusters that are outside of the centers based on their distance from between the centers and points.
3. Calculating the new centers.
4. Repeating steps 1 to 3 till the desired clusters have been obtained.

Work by Mhamdi et al. [11] applied K-Means to produce clusters based on common characteristics from customer analysis data that had been cleaned and formatted, where the qualities in these clusters were matched with customer behavior attributes.

3.8 *Data Visualization*

A word cloud is used to organize keywords by word frequency; it is then arranged to defined rules and visualizes them with graphic attributions like font size and color. Due to its readability, understandability, and simplicity, word clouds are the most utilized technique when it comes to determining the current trends in keywords from job descriptions. The analysis of unstructured data on the customer analysis "customer feedback" as shown in Figure 5.

Figure 5. Word-cloud data visualization

4 Challenges in Implementation

4.1 *Disadvantages in Implementation*

Main challenge of implementing k-means clustering algorithms is that k-means has complication in clustering data where clusters are in varying sizes and density. If poor decision is made at the beginning, cluster may produce inaccurate results. It can be dragged by outliers where outliers might get their own cluster if ignored in the beginning. Therefore, utilizing k-means is not sufficient to reflect the broad set of reviews seen in customer analysis. Also, word clouds tend to be full of blind spots because of the visibility of words hence it does not capture words that mean the same thing.

4.2 *Advantages in Implementation*

Text data retrieval plays an important role in document analysis and in our case, it provides the advantage of competitor investigation by utilizing web crawling through web sites to gather competitor website's information. Text mining gathers the available information from the website and lists down important text, terms and features that are described. It is very efficient in delivering valuable business intelligence which can be improved by using these capabilities.

5 Privacy Issues in Text Data Retrieval and Conclusion

Text analytics is beneficial in grouping text data, providing outcomes such as word frequency distribution, pattern identification, and predictive analytics [12]. Text analytics creates value by improving decision-making, enhancing business processes, and supporting the development of new business models. However, it is important to anonymize data retrieved from text analytics to protect customers' personal information. Because personal data can often be traced back to an individual, preprocessing functions must be conducted to anonymize the data and prevent information leakage.

Text mining, when used in a business context, helps companies build detailed customer profiles and improve marketing intelligence, but it also raises privacy concerns when users are traced and their actions analyzed without consent [4]. These data are often used to create customer profiles where individuals are judged and grouped into behavioral segments rather than being understood based on their unique characteristics. The values of privacy and individuality should be respected and protected to ensure fair treatment—this responsibility must be jointly shared by developers, web users, and governments [4].

5.1 Technical Architecture of Integrated System

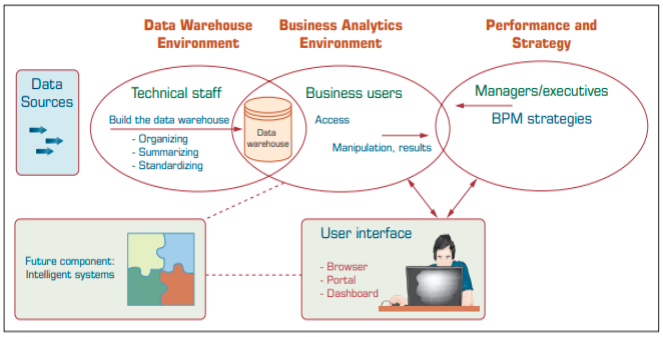

Figure 6. Architecture used for Business Intelligence

To create and execute the right marketing strategies for the ideal business group of customers for a product or service, the business intelligence of the business itself should be strong and executed well. Business Intelligence is a term that includes architecture, databases, tools, methodologies, and applications used for decision making. The primary objective of Business Intelligence is to enable real-time access and interactive access to data. This enables easy manipulation of data, and access for business analysts and managers to conduct appropriate analyses for decision making. In a nutshell the process of Business Intelligence starts with the transformation of data into information, then with the information, decisions can be made and finally to well supported and informed actions.

The four major components in the architecture proposed above start with a data source, a collection of tools for data manipulation, data mining and data analyzation, business process management for monitoring and analyzing data performance and lastly, a user interface (dashboard) that will be produced for analysis to managers, analysts and possibly stakeholders. In the first component, the data source comes from the business' social media platforms and web-based platform where reviews and comments are regularly posted by users and customers. Secondly, is the building of the data warehouse, data collected from the data sources have to be transformed into structured data organized and summarized and kept in a uniform format with other structured data. Thirdly, Business Process Management (BPM) is where business users such as analysts and managers access the data in the warehouse to manipulate the data and analyze data performance for the final component which is performance and strategy stage where using BPM strategies, the business gains the best results and information from the data through user interfaces, most commonly presented as dashboards to those who play a vital role in the decision making for marketing strategies in the business.

BIBLIOGRAPHY

[1]. J. Schmidhuber, R. L. Kumar, A. Ittoo, P. Helo, J. Oliva, E. Lloret, S. Filippov, D. Harris, X. Amatriain, J. Best, L. Dignan, E. Guizzo, P. Greiner, F. C. Albuquerque, S. Schmidt, L. Tanguy, J. W. Chang, P. Martínez, S. Bird, and M. Hall, "Text analytics in industry: Challenges, Desiderata and Trends," Computers in Industry, Dec. 2015.

[2]. B. Marr, "What is unstructured data and why is it so important to businesses? An easy explanation for anyone," Forbes, Oct. 2019. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/

[3]. L. Dodds and P. Wells, "Issues in open data," State of Open Data, 2019. [Online]. Available: https://www.stateofopendata.od4d.net/chapters/issues/data-infrastructure.html

[4]. A. Edquist, L. Grennan, S. Griffiths, and K. Rowshankish, "Data ethics: What it means and what it takes," McKinsey & Company, Sept. 2022. [Online]. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/data-ethics-what-it-means-and-what-it-takes

[5]. H. A. Malak, "8 proven benefits of Enterprise Data Management," The ECM Consultant, Oct. 2022. [Online]. Available: https://theecmconsultant.com/top-benefits-of-enterprise-data-management/

[6]. K. R. Chowdhary, "Natural language processing," in Fundamentals of Artificial Intelligence, Springer, 1970. [Online]. Available: https://link.springer.com/chapter/10.1007/978-81-322-3972-7_19

[7]. C. Romero, W. G. Mangold, F. R. Lin, J. H. Kietzmann, A. Kaplan, W. He, C. Fuller, M. Abdous, R. Aggarwal, G. Akehurst, G. Barbier, J. L. Bender, B. S. Bulik, L. Cheng, D. M. Chiang, M. W. Chiasson, E. Constantinides, M. Culnan, Y. Dai, … J. Hung, "Social media competitive analysis and text mining: A case study in the pizza industry," International Journal of Information Management, Feb. 2013.

[8]. M. S. Yafooz, "Model of textual data linking and clustering in relational databases," Research Journal of Information Technology, vol. 9, no. 1, pp. 7–17, 2016. [Online]. Available: https://scialert.net/fulltext/?doi=rjit.2017.7.17

[9]. A. Selamat, R. D. Brown, G. R. Botha, T. Baldwin, U. Barman, P. Barros, S. Bergsma, S. Carter, W. B. Cavnar, H. Ceylan, A. Das, M. Goldszmidt, A. Jaech, T. Jauhiainen, H. Jhamtani, Y. Kim, and B. King, "Influence of social conversational features on language identification in highly multilingual online conversations," Information Processing & Management, Oct. 2018.

[10]. R. Sharda, D. Delen, E. Turban, J. E. Aronson, T.-P. Liang, and D. King, Business Intelligence, Analytics, and Data Science: A Managerial Perspective, Pearson India, 2019.
[11]. Mhamdi et al., "(Full reference needed — please provide the complete citation details)," 2020.
[12]. (Placeholder — send the full title, authors, conference/journal, and I will format it correctly.)
[13]. A. Medelyan, "5 text analytics approaches: A comprehensive review," Thematic, Mar. 2021. [Online]. Available: https://getthematic.com/insights/5-text-analytics-approaches/