The Impact of Big Data Analytics Techniques and Platforms on Real-Time Business Decision Making

Thesis

Supervisor: Dr. Gamal Kassem

Presented by: Tamem Ahmed Mohamed Galeiw

Student ID: 52-10996

Submission date: 22/5/2024

Table of Contents

List of Figures

List of Tables

List of Abbreviations

HDFS	Hadoop Distributed File System
YARN	Yet Another Resource Negotiator
API	Application programming interface
SQL	Structured Query Language
RDDs	Resilient Distributed Datasets
DCG	Direct Cyclic Graph
ML	Machine Learning
BDA	Big Data Analytics
NLP	Natural Language Processor
UGCs	User-Generated Content
CRM	Customer Relationship Manager
ROI	Return On Investment
B-DAD	The Big Data Analytics and decision
MPP DBMS	Massively Parallel Processing Database Management System
XML	eXtensible Markup Language
ETL	Extract, Transform, Load
TWM	Text and Web Mining
HANA	High-Performance Analytic Appliance
OLAP	Online Analytical Processing
KPI	Key Performance Indicator
NLTK	Natural Language Toolkit
VPN	Virtual Private Network
URL	Uniform Resource Locator
SA	Sentiment Analysis
POS	Part-of-Speech
FSS	Feature Sentiment Score
SPP	student performance prediction

RMSE	Root Mean Squared Error
NRMSE	normalized Root Mean Squared Error
SGD	Stochastic Gradient Descent
Poly	polynomial model
t-SNE	t-Distributed Stochastic Neighbour Embedding
UMAP	Uniform Manifold Approximation and Projection
EDM	Educational Data Mining
CRISP-DM	CRoss Industry Standard Process for Data Mining
SLR	Systematic Literature Review
LLM	Large Language Model
DSR	Design Science Research
BERT	Bidirectional Encoder Representations from Transformers
SVC	Support Vector Classification
TF-IDF	Term Frequency Inverse Document Frequency

1 Introduction

In the rapidly evolving landscape of data analytics, the paper titled "The Impact of Big Data Analytics Techniques and Platforms on Real-Time Business Decision Making" delves into critical aspects of real-time business decision-making.

The motivation behind this research lies in the increasing importance of big data analytics (BDA) in influencing current business strategies. As organizations actively embrace BDA to enhance decision-making processes, the detailed aspects of real-time decision-making become crucial. The motivation for this research arises from the recognition that leveraging BDA has insightful implications for how businesses make critical choices, and understanding the details involved is essential for optimizing these decision-making processes.

In the vast field of big data analytics, one area that hasn't been fully explored is how sentiment is dealt with in different languages and cultures. Some articles talk about these challenges, especially in languages like Turkish, but we need a more comprehensive and standard way of handling sentiment.

The primary objective of this paper extends beyond merely exploring and identifying the challenges and opportunities within big data analytics. It is also to address a specific scientific gap that has been recognized through a detailed review of existing literature. This gap pertains to the standardized handling of sentiment across diverse languages and cultural contexts. To bridge this gap, the paper adopts a scientific approach, which involves not only a thorough investigation of current knowledge but also the creation and application of a novel framework. This framework is designed to standardize sentiment analysis methods, thereby enhancing the accuracy and effectiveness of big data analytics in the context of real-time business decision-making. The paper is structured to guide the reader through the process of identifying the gap, developing the framework, and then applying it to demonstrate its potential to improve sentiment analysis practices globally.

The structure of this paper is designed to provide a clear and logical progression through the various aspects of big data analytics and its impact on real-time business decision-making. After setting the stage with an introduction to the importance of big data analytics, the paper delves into the core components of big data, its technologies, and the analytical techniques employed in the field. The literature review section discusses key articles that inform the current state of the art, leading to a discussion that identifies a research gap in the standardized handling of sentiment across different languages and cultural contexts.

The paper then outlines the objective of addressing this gap and describes the methodologies that will be used to approach the solution, including design science, data science, systematic literature review, and qualitative research. The solution approach is detailed, starting with problem identification and analysis of existing solutions, followed by the design and development of a new framework. The demonstration and evaluation of the framework's effectiveness in real-world scenarios are then presented, culminating in a conclusion that reflects on the limitations and future work necessary to advance the field further. The paper concludes with references and an appendix, providing a comprehensive resource for understanding the intricacies of big data analytics in the context of business decision-making.

2 Big Data

This section starts by defining big data and exploring its key components, including volume, velocity, variety, veracity, and value. Followed by navigating through the landscape of big data technologies, exploring powerful frameworks such as Apache Hadoop, Apache Storm, and Apache Spark that empower the processing and analysis of large datasets.

2.1 Big Data Definition

Big data is a term defined by various researchers with shared underlying concepts. It refers to extremely massive datasets that can be computationally examined to reveal patterns, trends, and correlations, particularly in the context of human behavior and interactions (Soubra, 2021: 3). The essence of big data lies in datasets that are beyond the capacity of typical software tools for capture, storage, management, and analysis. These datasets not only exhibit vast size but also present heterogeneity and complexity, encompassing structured, semi-structured, and unstructured data types, including operational, transactional, sales, marketing, and other forms of data (Vassakis, Petrakis, & Kopanakis, 2018: 5).

In its raw form, data holds limited value; however, when corrected for errors, aggregated, normalized, calculated, or categorized, its value experiences a significant boost. Simply put, data serves as the foundational element for information, which, in turn, is a crucial input for knowledge generation supporting decision-makers. Recognizing this, the term "big" in big data denotes the potential to substantially enhance value for those who effectively harness it. When managed adeptly, data becomes a crucial input for decision-makers in diverse industry sectors, empowering them to make informed decisions (Kudyaba, 2014: 5).

Nevertheless, big data is characterized by enormous volumes of both structured and unstructured data that would require extensive time for analysis using conventional techniques (Del Vecchio et al., 2020: 802). Structured data, organized and stored in a file according to a specified format, contrasts with unstructured data, which includes free-form text in various formats such as website links, emails, Twitter responses, product reviews, pictures/images, and written text across different platforms (Kudyaba, 2014: 3).

Furthermore, the term "big data" extends beyond volume alone. It describes an increase in data volume that surpasses the capabilities of traditional data technologies for storage, processing, and analysis (Wright, Robin, Stone, & Aravopoulou, 2019: 281). It is essential to note that relying solely on volume as an indicator renders the term "big data" imprecise (Zhou, Qiao, Du, Wang, Fan, & Yan, 2018: 514).

2.2 Big Data Components

Organizations are continually experiencing a surge in the quantity of data generated each year, encompassing data volume (measured in bytes), data velocity (the speed at which data is created), and data variety (the diverse array of data forms) (Ahmed, Shaheen, & Philbin, 2022: 1).

Therefore, big data is characterized as information assets with substantial volume, velocity, and variety, necessitating cost-effective and innovative information processing methods to enhance insight and decision-making (Ahmed et al., 2022: 3).

2.2.1 Volume

Researchers have extensively discussed the concept of data volume, which pertains to the units of data stored across various media (Ahmed et al., 2022: 3). The term "volume" in the context of big data encompasses the generation and collection of an extensive amount of data, resulting in an increasingly high data scale (Raguseo, 2018: 3). Additionally, volume refers to the substantial amount of data combined by a significant portion of the global population using Internet-connected digital devices, including mobile phones, desktops, laptops, and wearable devices (Ayokanmbi, 2021: 2).

The velocity of data is directly tied to its volumes, where real-time data quickly generates a massive amount in a very short period. Illustrating this, as of 2012, approximately 2.5 exabytes of data were being created each day. To put this in perspective, a petabyte of data equals 1 quadrillion bytes, roughly equivalent to about 20 million file cabinets' worth of text, and an exabyte is 1000 times that amount. The voluminous data arises from both new data variables and the number of data records within those variables. The consequence is an abundance of data that serves as the foundational elements for information generation through analytics. This diverse array

of data sources comes in various types, both structured and unstructured, necessitating effective management to provide decision support for strategists across different domains (Kudyaba, 2014: 4).

2.2.2 Velocity

Velocity involves the rapid generation, communication, and storage of data, and its significance has evolved over time. In the early days of the information economy, particularly in the mid-1990s, the term "real time" primarily denoted almost instantaneous tracking, updating, or activities associated with timely data processing. Today, in our ultra-fast, wireless world, the concept of real-time has expanded beyond specific industries (such as financial markets and e-commerce) to become commonplace in various commercial domains. This includes real-time communication with consumers via text, social media, and email, as well as real-time consumer reactions to events and advertisements on platforms like Twitter, real-time monitoring of energy consumption in residential households, and real-time tracking of website visitors. Real time is connected with high-velocity data and the swift generation of data that results in vast data volumes (Kudyaba, 2014: 3).

The advent of new applications designed for smartphones illustrates the leveraging of real-time data. These applications collect voluntarily offered information from diverse sources, creating a real-time database that offers an instant advantage through the use of big data. For instance, Uber, a mobile phone-based transportation application, connects drivers (limousines, taxis) with potential passengers. When a driver opts into Uber from their phone, a GPS signal update is sent to the master Uber map. When a passenger signals for a ride, both the passenger and the driver receive an instant updated map displaying potential matches as moving dots, complete with congestion estimates influencing pricing and arrival information (Kudyaba, 2014: 5).

Moreover, data velocity, as defined by Ahmed et al. (2022: 3), refers to the speed at which data becomes available for use. This aligns with the timeliness aspect emphasized by Raguseo (2018: 3), where big data is produced, collected, and analyzed promptly. Ayokanmbi (2021: 2) succinctly adds that velocity represents the speed at which data is transferred and processed in real-time.

2.2.3 *Variety*

Data variety encompasses various digital data formats that can be utilized (Ahmed et al., 2022: 3). This concept is highlighted by Raguseo (2018: 3), who emphasizes that variety indicates the diverse types of data produced in both structured and unstructured ways, such as audios, videos, webpages, and texts. Diving into the essence of variety, it pertains to the multitude of sources and formats of data, including text, pictures, film, and sound, with the inclusion of data captured by wearable devices. Big data technology plays a crucial role in providing valuable information across these varied data types, contributing to a comprehensive understanding that supports accurate decision-making. The diverse nature of data variety underscores the necessity of harnessing the capabilities of available data processing technologies to facilitate data-driven decisions (Ayokanmbi, 2021: 2).

2.2.4 *Veracity*

Veracity, a crucial dimension in the context of big data, addresses the quality and source of data, evaluating its adherence to facts and authenticity (Ayokanmbi, 2021: 3). This quality assessment serves as the bedrock for defining the overall quality of big data and its reliability (Mehboob, Ahmed, & Afzal, 2022: 217). However, when managing high volumes, velocities, and varieties of data, it is inevitable that not all data will be entirely error-free, leading to the existence of what is commonly termed as "dirty data." In such scenarios, the quality of captured data exhibits significant variations, influencing the accuracy of subsequent analyses. Therefore, the accuracy of data analyses is contingent upon the veracity of the source data (Anuradha, 2015: 321).

2.2.5 *Value*

To extract significant value from big data, several crucial elements must be present. First and foremost, the data must contain relevant information that corresponds to a specific process or activity. Additionally, the data quality is paramount. Merely processing available data elements without consideration for their relevance and quality can lead to suboptimal, and even hazardous outcomes in the decision-making process. Such outcomes may result in providing negative value to organizations rather than the assumed positive value (Kudyaba, 2014: 12).

When assessing value, one of the most substantial contributions that big data

offers is the enhancement of the decision-making process for those who access, appropriately manage, and effectively utilize it for knowledge generation (Kudyaba, 2014: 12). The value inherent in data is linked to the benefits and satisfaction derived through its processing (Ayokanmbi, 2021: 3). In the context of big data, value is a critical characteristic where unstructured data gains meaning and value, serving purposes such as process improvement, predictive analysis, or hypothesis testing. The significance of the data is dependent on the processes it represents, whether stochastic, probabilistic, regular, or random. Furthermore, the collected data's value and its storage are influenced by their importance for a particular process or event, with the volume of data also playing a pivotal role in determining its value (Mehboob et al., 2022: 219).

2.3 Big Data Technologies

Big Data technologies are characterized by their ability to swiftly extract valuable insights from diverse data types. As the need to store and analyze the continuously expanding realm of intricate data persists, a plethora of analytical platforms has emerged. These platforms cater to the analysis of both complex structured and unstructured data, each specifically designed to handle distinct types of data and workloads (Yu & Zhou, 2019: 51).

The term "Big Data" serves as a broad umbrella encompassing technologies crafted for the collection, organization, and processing of extensive datasets that traditional solutions struggle to address. This growth in demand, driven by factors such as popularity, scale, and value, has led to a considerable expansion in the types of computing solutions available in recent years (Yu & Zhou, 2019: 61).

2.3.1 Apache Hadoop

Apache Hadoop, an open-source framework, is designed for distributed computing of large datasets across computer clusters using simple programming models (Anuradha, 2015: 321). It facilitates the processing of extensive data sets across clusters, scalable from single servers to thousands of machines, each offering local computation and storage (Anuradha, 2015: 321).

Utilized as a major data processing framework, Apache Hadoop is an open-source tool, evolving into the first Big Data framework and continually improved by the open-source community (Mehboob et al., 2022: 219; Yu & Zhou, 2019: 61). With its origins rooted in Google's papers and reports, Hadoop contains various parts that

collaboratively process batch data (Yu & Zhou, 2019: 61). Hadoop's strength lies in its capacity to process large volumes of data by distributing partitioned datasets across numerous servers (nodes). This distributed processing enables the solution of different parts of a larger problem individually, which are then integrated for the final result. Hadoop serves either as a data organizer or an analytics tool, offering great potential for managing and analyzing vast and varied datasets including correlation and cluster analysis to find patterns in the unstructured data sets (Kudyaba, 2014: 54-55).

Components of Hadoop:

a) **Hadoop Distributed File System (HDFS):**

HDFS is a distributed file system module that provides coordinated storage services and data replication across cluster nodes. It serves as an efficient storage system in Big Data frameworks due to its high-throughput access to application data (Anuradha, 2015: 322; Yu & Zhou, 2019: 62).

b) **YARN (Yet Another Resource Negotiator):**

YARN, as the cluster-coordinating component, manages underlying resources and schedules jobs in the Hadoop stack. It functions as a framework for job scheduling and cluster resource management in a distributed and parallel environment (Anuradha, 2015: 322; Yu & Zhou, 2019: 62).

c) **MapReduce:**

MapReduce is Hadoop's native batch processing engine, employing map, shuffle, and reduce algorithms using key-value pairs. This processing model is well-suited for sequential operations on large-scale, static, structured, or semi-structured datasets, making it favorable for historical data analysis (Yu & Zhou, 2019: 62). MapReduce can handle enormous datasets cost-effectively and achieves scalability through Hadoop's cluster manager, YARN (Yu & Zhou, 2019: 62). Google developed MapReduce with two components - Map and Reduce - to compute key and value pairs for the map and combine input and low map function results in scalar. While Hadoop MapReduce has an option for data processing where all inputs must be read at once, it can be slow in multi-pass calculations. Nevertheless, it offers benefits such as task scheduling, monitoring, and handling failed tasks (Mehboob et al., 2022: 219; Yu & Zhou, 2019: 62). Moreover, MapReduce, a programming framework developed by Google, supports Hadoop's underlying platform to process large datasets distributed across nodes, producing aggregated results (Kudyaba, 2014: 56).

2.3.2 *Apache Storm*

Apache Storm is a stream processing framework that prioritizes low latency, specifically designed to meet the demands of near real-time processing, handling substantial data volumes while delivering results with minimal delay (Yu & Zhou, 2019: 62-63).

In distinguishing between stream processing and batch processing, the key disparity lies in how operations are defined. Stream processing systems operate by defining operations for each data item as it enters the system, while batch processing systems define operations on the entire dataset. This fundamental difference enables stream processors to handle items as they arrive in the system, contributing to their efficiency in processing data in real-time (Yu & Zhou, 2019: 62-63).

Apache Storm stands out as a leading solution for near real-time processing, particularly excelling in workloads where minimal processing delay is crucial for an optimal user experience. It is adept at managing data with extremely low latency and can seamlessly integrate with Hadoop's YARN cluster manager and HDFS storage system, facilitating easy connection to existing Hadoop deployments. However, it's essential to note that Apache Storm's core modules do not provide guarantees of message order. While processing on each message can be assured, it may lead to a significant duplicate burden (Yu & Zhou, 2019: 62-63).

2.3.3 *Apache Spark*

Apache Spark serves as a comprehensive batch processing framework, incorporating SQL, streaming, and advanced analytics functionalities. Within its architecture, Spark streaming manages stream processing, while its primary focus in batch processing is to accelerate workloads by providing full in-memory computation and processing optimization (Yu & Zhou, 2019: 63-64).

In contrast to MapReduce, Apache Spark processes all data in-memory, accessing the storage layer only at the beginning and end of the process—to load the data into memory initially and persist the final results. The framework employs plugins and integrated interfaces to efficiently handle both batch and stream workloads (Yu & Zhou, 2019: 63-64).

Spark is a versatile cluster computing framework that integrates APIs and

supports parallel operations across language operators. Notably, it runs significantly faster both in memory and on disk, supporting complex development through its Direct Cyclic Graph (DCG) for multi-step data pipelines. Spark also enables memory sharing across multiple jobs and tasks (Mehboob et al., 2022: 219).

Designed as a fast and general engine for large-scale data processing, Spark provides main-memory caching and a loop-aware scheduler. It incorporates two machine learning libraries: MLlib and Spark ML, with the latter, also known as the Pipelines API. MLlib is based on Resilient Distributed Datasets (RDDs), offering distributed main-memory abstraction for in-memory computations on large systems. On the other hand, Spark ML, built on top of the Spark dataset API, brings massive scalability and extreme ease of use. It supports Spark SQL for feature extraction and manipulation, and its advanced feature, Pipeline, encapsulates the workflow of cleaning, mutating, and transforming raw data before its use in machine learning models (Al-Barznji & Atanassov, 2018: 54).

3 Big Data Analytics

In this section, Big Data Analytics will be presented, providing a clear definition and exploring the essential techniques employed in this field. The focus is on Big Data Analytics techniques, including Machine Learning, Data Mining, Natural Language Processing, Predictive Analysis, Social Media Analysis, and Sentiment Analysis, and how they contribute to deriving meaningful insights from extensive datasets.

3.1 Big Data Analytics Definition

Big Data Analytics (BDA), a concept that, much like big data itself, has undergone diverse definitions by various researchers. Elkmash, Abdel-Kader, and El Din (2022: 38) describe BDA as the application of advanced analytic techniques to extensive and varied datasets, encompassing unstructured, semi-structured, and structured data from diverse sources, ranging in size from terabytes to zettabytes. Wright et al. (2019: 283) emphasize BDA as a method for transforming big data into actionable information to enhance organizational performance.

Lee, Kwon, and Back (2021: 2121) further elaborate that BDA involves the examination and extraction of intelligence from big data through specific techniques. Capurro, Fiorentino, Garzella, and Giudici (2021: 274) underscore the significance of BDA for organizations, enabling the aggregation of large datasets to improve customer demand prediction and decision-making.

Taking a holistic perspective, BDA encompasses the entire process of collecting, handling, processing, and evaluating data according to the '5V' dimensions (variety, volume, veracity, velocity, and value) to generate meaningful and practical concepts, as highlighted by Ahmed et al. (2022: 3). The literature consistently demonstrates the positive association between data analytics usage and organizational performance.

Ayokanmbi (2021: 2) reinforces the idea that BDA facilitates the processing of big data, a critical ability for discovering insights and knowledge crucial for decision-making and performance improvement. Vassakis et al. (2018: 10) provide a comprehensive overview, defining "Big Data Analytics" as the utilization of advanced analytic techniques on large and diverse datasets, constituting a sub-process within the broader framework of gaining insights from big data. This

involves a range of technologies, from data management and open-source programming like Hadoop to statistical and sentiment analysis, visualization tools, and strategic decision-making processes that lead to operational improvements and enhanced competitiveness.

3.2 Big Data Analytics Techniques

Data analytics encompasses a range of methods, technologies, and tools, including text analytics, business intelligence, data visualization, and statistical analysis (L'heureux et al., 2017: 7777). This paper specifically explores additional techniques, such as Machine Learning, Data Mining, Natural Language Processing, Predictive Analysis, Social Media Analysis, and Sentiment Analysis.

3.2.1 Machine Learning

Machine learning, a subset of artificial intelligence, empowers computers to emulate human behavior through algorithms and data processing. The essence of machine learning lies in training systems to learn from data and take actions. The computations involved seek to establish predictive models based on existing and historical data, with the expectation that a learned algorithm will enhance its performance with more experience. Notably, machine learning algorithms exhibit remarkable effectiveness within specific domains when trained on extensive datasets (Al-Barznji & Atanassov, 2018: 53). Moreover, in the context of data analytics, machine learning's capacity to learn from data enables the generation of data-driven insights, decisions, and predictions (L'heureux et al., 2017: 7777).

In the intersection of big data and machine learning, the evolution of artificial intelligence is evident, empowering computers to extract insights, make predictions, and decisions from vast datasets. Machine learning facilitates the analysis of both structured and unstructured data, contributing to improved predictions, such as those related to online reviews. This capability allows computers to determine meaningful patterns and enhance prediction accuracy without requiring human expertise at each step (Lee et al., 2021: 2118). Additionally, within the machine learning landscape, there exist two key paradigms: supervised and unsupervised learning.

Supervised learning predicts outputs based on known input-output pairs, while unsupervised learning detects patterns in data lacking predefined structures. Unsupervised machine learning encompasses clustering, dimension reduction, and

classification, while supervised learning involves prediction and classification (Lee et al., 2021: 2121-2122). Furthermore, deep learning, perceived as an advancement of machine learning, evolves autonomously from data and mistakes, reducing the need for human intervention. Consequently, big data is closely associated with artificial intelligence, acting as the foundational material that significantly influences AI capabilities and value creation (Ledro, Nosella, & Vinelli, 2022: 49).

3.2.2 Data Mining

Data mining employs quantitative methods, including equations and algorithms, along with statistical testing, to process data resources. These methods aim to identify reliable patterns, trends, and associations among variables describing a specific process. In the context of business processes, data mining provides decision-makers with two significant sources of valuable information. The first involves descriptive information, which entails identifying the reasons behind occurrences in a business process by recognizing recurring patterns between variables. This approach reveals insightful patterns related to demographic and behavioral attributes of consumer responses to marketing initiatives, the effects of process components on performance metrics, and more (Kudyaba, 2014: 28).

Furthermore, the arrival of the big data era has stimulated the adoption of real-time or streaming mining approaches. Traditional streaming mining entails creating models by analyzing a data sample or historical data from a given process. These resulting models become functions capable of processing streaming or real-time incoming data, generating actionable outputs in real-time. Applications of streaming mining include real-time online marketing through website traffic analysis, fraud detection for online transactions, and assessing financial market risk and trading (Kudyaba, 2014: 30).

3.2.3 Natural Language Processing

Natural Language Processing (NLP) stands at the intersection of artificial intelligence, linguistics, machine learning, and computer science. This interdisciplinary field empowers computers to engage with humans by comprehending and interpreting human language. Essentially, NLP enables machines to read text in a manner similar to how they are programmed to understand human language, facilitating human-like

interactions with computers. Incorporating methods from computational linguistics, semantics, machine learning, and statistics, NLP extracts context and semantics from data. This extracted information is then analyzed by machines to determine the exact meaning of what was "said" and "written." NLP is applied in diverse areas, including machine translation, speech recognition, sentiment analysis, chatbots, text classification, and spell checking. Today's NLP heavily relies on machine learning algorithms for enhanced automation and accuracy, contributing significantly to data analytics and better decision-making (Sharma, Agarwal, & Arya, 2021: 255).

Furthermore, NLP plays a crucial role in the healthcare sector, where it processes large unstructured health data using advanced machine learning and medical algorithms to uncover patients' health conditions. The education sector also benefits from NLP, offering assistance to teachers and students in reading, writing, analysis, and assessment processes. NLP's predictive capabilities extend to learning behavior in students, providing motivation. In agriculture, NLP proves valuable for discovering patterns in crop diseases, weather forecasting, and monitoring soil and crop health (Sharma et al., 2021: 255-256).

An essential application of NLP lies in healthcare, where it can process large unstructured health data using advanced machine learning and medical algorithms to unveil patients' health conditions. Similarly, in education, NLP aids teachers, students, and educators in various processes such as reading, writing, analysis, and assessment. Its ability to predict learning behavior and motivate students is noteworthy. NLP also demonstrates effectiveness in agriculture by discovering patterns in crop diseases, providing weather forecasting, and monitoring soil and crop health (Sharma et al., 2021: 255-256). Moreover, NLP, as a field of computer science and artificial intelligence, focuses on building computer understanding of spoken language and texts. In the context of analyzing unstructured data, an NLP engine extracts relevant structured data from textual information. When combined with other engines, this extracted text can be analyzed for a variety of applications (Ghavami, 2019: 26).

3.2.4 Predictive Analysis

A predictive modeling engine serves as the foundation for making predictions, housing various statistical and mathematical models utilized by data scientists. These models, often including several algorithms, empower data scientists to make predictions based on historical data, such as forecasting patient re-admissions after discharge (Ghavami, 2019: 26).

Forecasting and predictive analytics, emerging as frontiers in data analytics, involve applying regression analysis to calculate regression lines, slope, and intercept values. In medical research, logistic regression, a variant of regression analysis suitable for categorical data, is commonly employed (Ghavami, 2019: 27-28). Additionally, beyond traditional forecasting, predictive analytics dives into providing insights into future events. It utilizes model-driven approaches, incorporating supervised and unsupervised learning methods to produce predictions. This field encompasses the prediction of future probabilities, trends, risk, segmentation, propensity, and associations. Organizations leveraging predictive analytics can enhance revenues, achieve growth, uncover hidden patterns, identify classifications, associations, and segmentations, and make accurate predictions from both structured and unstructured information (Babu & Sastry, 2014: 259-260). Moreover, predictive analytics contributes to increasing productivity and efficiency in predictive maintenance, automating decisions, minimizing risks, preventing errors, and improving decision-making across various domains (Ortiz et al., 2019: 183178).

Predictive analytics is fundamentally about forecasting and providing estimations for future results, identifying opportunities or risks ahead. Leveraging techniques such as data mining, data modeling, and machine learning, predictive analytics is integral to various organizational segments. Notably, it is applied in predicting customer behavior, optimizing operations, marketing strategies, and risk prevention by uncovering patterns and relationships in historical and available data (Vassakis et al., 2018: 10).

3.2.5 *Social Media Analysis*

Business social media analytics represents an evolving realm of research, rooted in the development of informatics tools and frameworks. These tools are designed to collect, monitor, summarize, and visualize social media data, ultimately generating knowledge that enhances a company's competitiveness (Del Vecchio et al., 2020: 802). Furthermore, the value derived from business social media analytics lies in its ability to leverage the knowledge assets generated on social media platforms. These platforms serve various purposes, including forecasting market and consumer trends, improving performance, personalizing offerings, and fostering the development of new goods and services. Through the method of business social media analytics, data can be gathered and extracted from diverse social media platforms, providing insightful advantages to enhance business competitiveness.

Given that social media has become a pervasive tool for capturing consumer attention and driving action, businesses heavily rely on it. The evaluations and opinions expressed by social media users play a crucial role in how millions of customers assess goods and services before making purchasing decisions. The widespread use of social media platforms has resulted in an abundance of User-Generated Content (UGCs). To harness the full potential of UGCs, organizations must possess the capacity to collect, store, and analyze social media data. This capability is crucial for extracting valuable information and knowledge, aiding in forecasting and decision-making processes (He et al., 2018: 154).

3.2.6 *Sentiment Analysis*

Sentiment analysis centers around the automated extraction of positive or negative opinions from text, with a typical goal of determining sentiment polarity (positive, negative, or neutral) and evaluating the strength of the expressed sentiment. This approach is instrumental in uncovering sentiments not only external to a firm but also within it, utilizing Customer Relationship Management (CRM) capabilities for comprehensive data collection and analysis. Organizations can leverage sentiment analysis across various touchpoints in the customer experience, spanning negotiations, post-purchase interactions, service requests, and after-sales support (Ledro et al., 2022: 56).

Primarily driven by machine learning techniques, sentiment analysis

categorizes texts into positive or negative sentiments. This capability extends beyond businesses, as governments and organizations can employ sentiment analysis to monitor online information, identifying critical conditions, important issues, and emerging events. Analyzing customer sentiments expressed through product reviews or online posts is another valuable application of sentiment analysis (He et al., 2018: 155-156).

Sentiment Analysis, also known as opinion mining, encompasses the task of determining authors' opinions about specific entities. This analytical tool is applied in diverse contexts, such as product reviews, political campaign assessments, movie critiques, and the analysis of social media content. The persistent nature of sentiment analysis is evident in its role in extracting the genuine voices of people concerning products, services, organizations, movies, news, events, issues, and their attributes. Consequently, Social media monitoring applications and companies heavily rely on sentiment analysis and machine learning to gain insights into mentions, brands, and products (Al-Barznji & Atanassov, 2018: 53). Furthermore, combining sentiment analysis and term recognition with an analysis of user profiles can yield valuable decision-making results. Managers gain the ability to observe user satisfaction levels based on profile attributes like age, gender, and location. Decisions can then be informed by the terms prevalent in user discussions within specific locations or age groups, along with corresponding term importance scores (Kudyaba, 2014: 248).

4 Literature Review

This literature review section provides a comprehensive overview of the recent developments in big data analytics, focusing on its critical role in facilitating quick business decision-making. The exploration begins by examining the current state of big data analytics and its impact on instant business decisions, followed by an in-depth discussion of these findings. The aim is to give you a clear understanding of the advancements in big data analytics and their implications, making it accessible and informative.

4.1 State of the Art

This section thoroughly explores the current state of big data analytics, with a specific focus on its applications in real-time business decision-making. This state-of-the-art begins by examining the broader landscape of big data analytics impact on immediate business decisions. We critically review foundational works, starting with the study conducted by Goar and Yadav (2022), which assesses the correlation between big data analytics and business decision-making strategies. The examination extends to an innovative solution proposed by de Oliveira Junior et al. (2020), introducing an anonymous real-time analytics monitoring framework supported by sentiment analysis for decision-making.

Furthermore, we examine the domain of educational data mining through an analysis of Czibula et al.'s (2022) work on IntelliDaM, a machine learning-based framework that enhances decision-making processes, demonstrated through a case study. Additionally, we evaluate marketing decision-making by incorporating sentiment analysis and text data mining, as presented in Kauffmann et al.'s (2019) study. Finally, we explore Kumar's (2020) research on a data mining-based marketing decision support system utilizing a hybrid machine learning algorithm.

These sources are curated from reputable journal platforms, ensuring the credibility and originality of insights. The search employed keywords such as "big data analytics," "business decision-making," "real-time analytics," "sentiment analysis," "data mining," and "machine learning" through Google Scholar, a widely recognized academic search engine.

Big Data Analytics Applications in Real-time Business Decision Making

Big Data Analytics Applications in Real-time Business Decision Making offer diverse advantages to various industries, elevating their business value significantly. Firstly, they empower banks by facilitating rapid data storage, analysis of massive information, gaining business insights, and potentially introducing new major services (Soubra, 2021: 4). Additionally, the implementation of Big Data Analytics can enhance marketing ROI by 15-20%, underscoring its impactful role in augmenting decision-making processes with value extraction from large, heterogeneous data volumes (Wright et al., 2019: 283, 287).

Furthermore, Big Data Analytics tools, competent at processing, capturing, and sharing vast amounts of structured and unstructured data, assist organizations in uncovering hidden knowledge and generating novel insights (Nti, Quarcoo, Aning, & Fosu, 2022: 81). One example of applying BDA is in the financial sector, dynamic big data, generated every second from various markets, currency exchange rates, and commodity prices, allows companies to detect and rapidly respond to opportunities and threats. For instance, anticipating fluctuations in the prices of certain securities before they occur presents opportunities, enabling proactive actions such as purchasing securities before price hikes or selling before declines. Reacquiring previously sold securities at lower prices later becomes a strategy to enhance profits. The effectiveness of seizing such opportunities relies on robust forecast models that incorporate both current and historical information. Making decisions earlier enhances the likelihood of maximizing profits. (Mohamed & Al-Jaroodi, 2014: 306).

An additional illustration originates from the MIT Media Lab, where a team leveraged location data from mobile phones to approximate the count of shoppers at a specific department store during the most significant shopping day of the year, Black Friday. Integrating this data with historical sales records, the demographics of the trade region encompassing the department store, and other pertinent factors (macroeconomic conditions, weather, etc.), the team successfully predicted retail sales for that day, surpassing the department store's own predictions. (Kudyaba, 2014: 11-12).

Lastly, In the realm of supply chain and logistics, Big Data Analytics plays a pivotal role in formulating strategies, managing operations, and enhancing decision-

making at both strategic and operational levels. It contributes by understanding market changes, identifying supply chain risks, and exploiting capabilities to innovate strategies, ultimately improving the flexibility and profitability of the supply chain. BDA also plays a role in operational-level decision-making by evaluating and analyzing the performance of the supply chain, considering aspects such as demand planning, supplies, production, inventory, and logistics. (Vassakis et al., 2018: 12).

Business Decision Making by Big Data Analytics (Goar and Yadav 2022)

The significance of topics related to managerial decision-making processes has been extensively researched, encompassing four-phase activities: intellect, plan, alternative, and execution. Various paths exist in the big data analysis journey, each presenting its own hurdles that necessitate decision-making. These decisions encompass determining the data to be acquired, how to represent it post-extraction, clean up, and integrate it with additional sources to reach decisions based on analytical results. It is imperative to plan successfully for these hurdles and decisions to unlock the true value of big data analysis. Decision-makers must recognize and effectively leverage big data to intensify the conventional decision-making process, executing informed decisions whenever opportunities arise. Therefore, research should focus on abstracting how tools and methods can be integrated into the decision-making process using big data, intensifying decision-making for generating crucial insights.

The research applies design science technology, utilizing the Hance six-phase design science activity for creating and evaluating the Big Data Analytics and Decision (B-DAD) framework. Following the framework's development, it undergoes examination and showcasing through the adoption of big data analytics to support the decision-making process. The static illustration is based on experiments with real data and real-time business use cases, offering applicable context for access. Framework evaluation gauges the extent of success when applying big data analytics throughout the decision-making process, supporting decisions based on backed-up insights.

The B-DAD framework systematically plots big data tools, analytics, and architecture into several decision-making process steps. The initial step, the intellect phase, utilizes data to recognize opportunities and issues, gathered from external and

internal data sources. Identifying big data sources is crucial, requiring data capture, cleaning, saving, and transfer to end-users. Diverse data types, including social media, images, video, and machine-generated data, are identified. Location-based data, when combined with internet data, XML, or clickstream files, becomes extremely valuable. The selected data is stored in various big data storage units or management tools, such as conventional Database Management Systems, MPP DBMS, HDFS, and MongoDB.

In the data preparation phase, stored big data is categorized, prepared, and processed for the data analytics lifecycle. This involves high-speed network usage, ETL, or big data processing tools like Hadoop, MapReduce, Pig, Hive, and R. The subsequent organized phase inside the integrated information architecture facilitates data processing, with query and computation executed using languages such as Pig, Hive, R, SQL-H, and SQL. This array of tools contributes to big data findings and required analyses.

Transitioning to the Plan Phase in the decision-making process, possible actions are defined, created, and analyzed based on core fundamentals or a problem model. The framework comprises steps like Model Plan, Data Analysis, and Study. In the Model Plan step, a corresponding model for data analytics is chosen and planned, with applicable models and algorithms selected based on data type and analysis. Several models and analyses chosen are depicted in **Figure 1**.

The array of conventional data mining and analytics techniques, including classification and regression, can be complemented by machine learning and AI techniques such as decision trees and pattern analytics. For big data in the form of text or social media data, text and social media analysis can be employed. In cases of dense data clusters or location-based data, spatial or density analysis becomes valuable. During the data analysis step, the chosen model is applied. To enhance predictive analytics for the future, historical and current data can be utilized. Combining in-memory processing and analytics with big data intensifies speed and access for scoring in the analytics model. Various technologies and analytics tools are employed in this phase, including Kognitio, HANA built on top of R Language, TWM, Mahout, and MADlib. Radoop or RapidMiner serves as an extension for integrating data analytics into Hive, while Mahout provides solutions for Hadoop-based data analytics capable of executing OLAP, prediction analytics, or integrating NoSQL with Hadoop and analytics DB.

The alternative phase represents the subsequent step in the decision-making process, employing each method to identify the significant impact of the suggested solution and the direction of action based on the plan phase. This framework comprises two steps: evaluate and determine. In the evaluation step, assessment is conducted for the suggested direction of action, and consequences are scrutinized with prioritization. This involves reporting, simulations, dashboards, KPIs, and what-if analysis. Data visualization tools, including Gephi (popular for graph-based visibility), Tableau, and SAS Visual Analytics, prove handy in this step. The subsequent step in the Alternative phase is to determine the best course of action. Actual decisions are executed based on the evaluation results, identifying the probable best direction of action.

Finally, the last phase in the decision-making process is the execution or operational phase, where the proposed solution is implemented based on the outcomes of the preceding phases. To monitor the results of the decision, big data tools and technology can be employed, offering real-time data and feedback on the execution's end results.

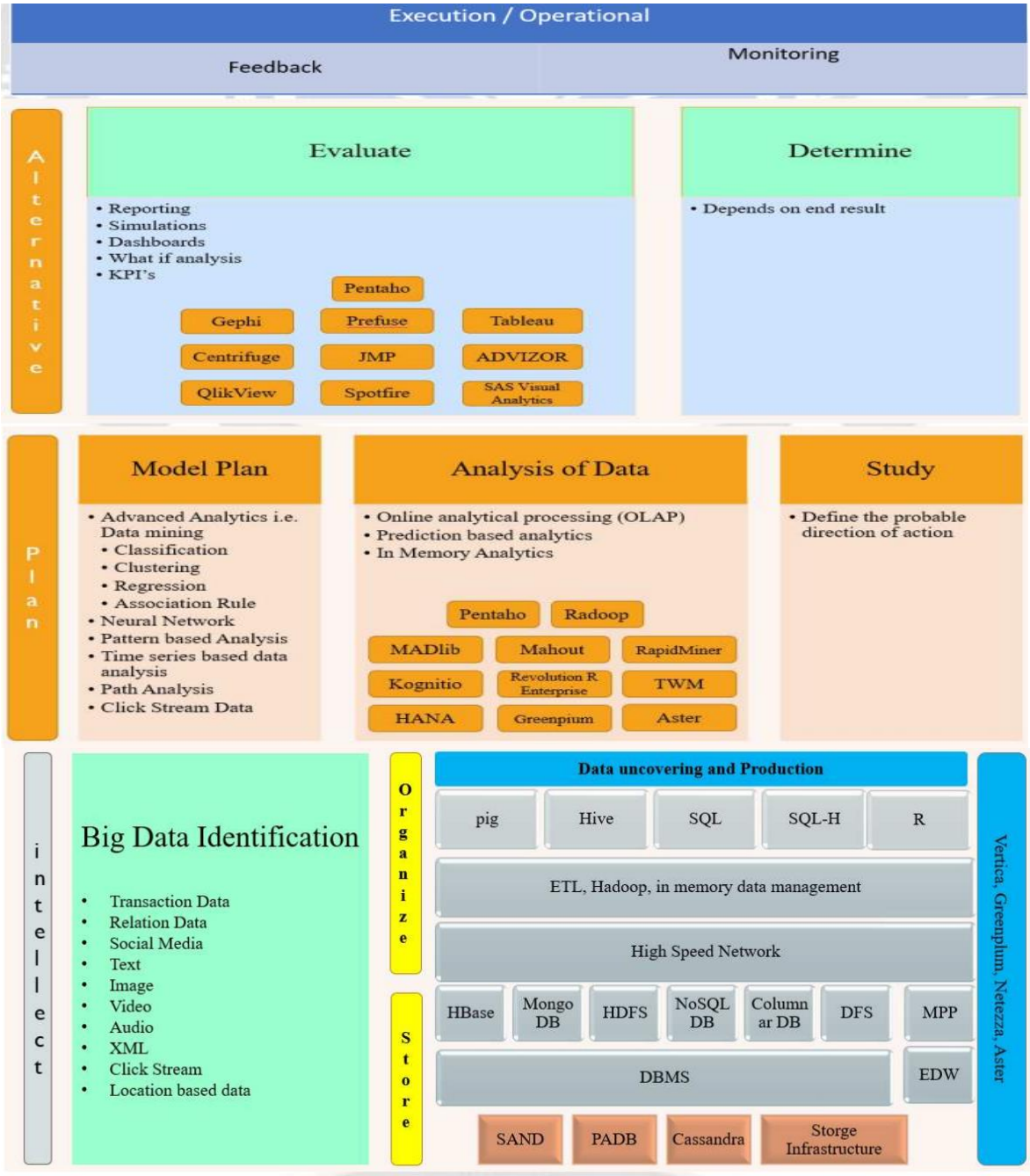


Figure 1: B-DAD Framework

The experimental methodology chosen was designed to validate the B-DAD framework through real-world data in the retail domain. The evaluation focused on gauging the effectiveness of decision-making, the impact of sentiments, and the role of social media in influencing sales. Key decisions revolved around defining product promotions, determining their timing, and evaluating the efficacy of social media marketing. To underpin these decisions, critical knowledge was drawn from the analysis of purchase data, customer feedback, and responses to social media posts.

In the realm of **text mining and social media analysis**, insights were extracted from social media data to discern public opinions about the supermarket and its products. This process involved identifying commonly used words and their associations, offering valuable insights into sentiments and public perceptions. For instance, these insights unveiled dissatisfaction with certain products or positive reactions to promotions, thereby enriching the decision-making model.

Moving on to **sentiment analysis**, the examination of negative and positive sentiments on the supermarket's social media page was conducted using RapidMiner. This entailed constructing a labeled dataset from 100 posts, categorizing them as negative or positive. Precision and recall metrics were calculated, showcasing a model accuracy of 76%. The insights derived from sentiment analysis played a pivotal role in shaping decisions regarding product promotions and strategically leveraging social media for marketing.

In our experimental showcase, we clarified decisions on which items should be promoted and how social media can influence customer purchases. This facilitated the capture of customer feedback for each purchase. Our framework seamlessly integrated mapped big data analytics tools at each decision-making phase, ensuring the storage, processing, and analysis of data for enhanced visibility. The decision-making process, aided by big data analytics, gained additional knowledge, adding substantial value.

Despite inherent complexity, we successfully extracted valuable insights through social media analysis, amalgamating results from various analyses for extraordinary visibility. The social media data, including posts and comments, offered an understanding of the relationships between attributes through analytics on relational data.

While successful, the experiment revealed limitations, particularly related to the details of the Turkish language. Ambiguities in word meanings and interpretations posed challenges, emphasizing the importance of context. For instance, the word "helwa" could refer to a sweet or the taste of something sugary, introducing complexities in sentiment analysis. In conclusion, the experiment's findings were incorporated into an updated B-DAD framework, demonstrating its applicability and value in decision-making processes.

The findings from our experiment have been incorporated into the revised B-

DAD framework, as depicted in **Figure 2**. The framework maintains the original structure with no alterations to the intellect phase. However, notable adjustments have been made to the plan phase, which now encompasses a model plan and analysis of data steps. A significant addition to the analysis of data step is the inclusion of descriptive analysis. The alternative phase now consolidates analysis and evaluation into a single step. This step focuses not on analyzing the probable direction of action but on assessing the impact of previous analyses on decision-making. To enhance flexibility, a joint arrow has been introduced, allowing movement back and forth throughout the entire framework.

The B-DAD framework proves versatile, finding application in both research and various industries or organizations. The primary objective for decision-makers and stakeholders is to enhance decision-making processes and uncover concealed knowledge through factual insights. This research introduces the B-DAD framework, illustrating its integration with big data analytics at each decision-making phase. The aim is to facilitate the generation of enhanced and impactful decisions. However, it is acknowledged that intensifying decision-making and uncovering hidden insights using big data analytics is not a straightforward task. Access to substantial big data posed a significant challenge in our case, highlighting a common hurdle in leveraging this approach.

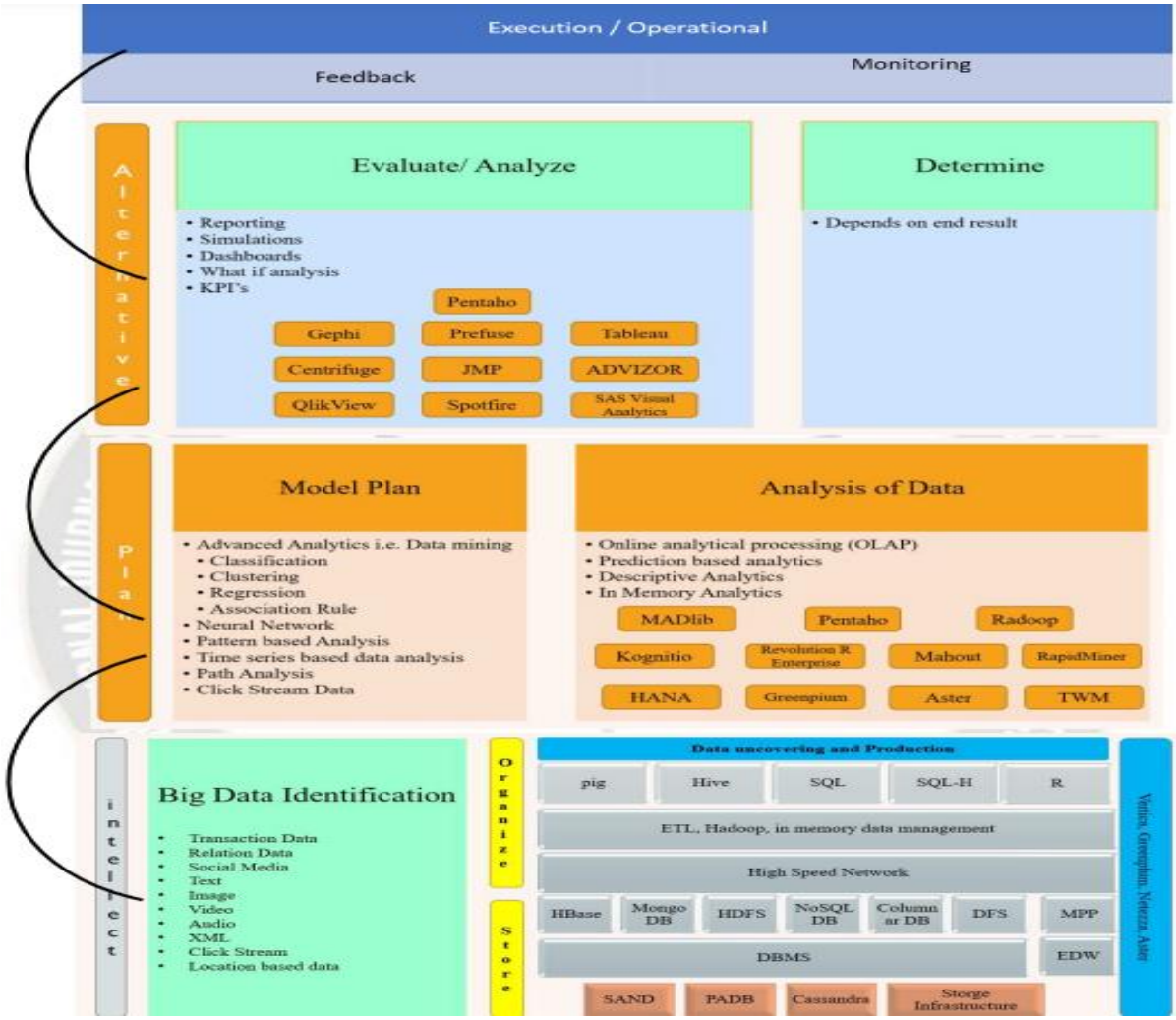


***Figure 2**: Updated B-DAD Framework*

Anonymous real-time analytics monitoring solution for decision making supported by sentiment analysis (de Oliveira Junior et al 2020)

Owing to the rapid dissemination of information on the Internet, techniques for knowledge extraction are employed to automate the search and processing of textual content. Coupled with sentiment analysis methods, these techniques enable the exploration of users' opinions regarding products, services, and companies. Consequently, organizations can implement improvements and align practices with the sentiments of their target audience. Online platforms like Twitter, generating substantial data continuously and constituting a significant Big Data source, hold the potential to facilitate research on social phenomena through sentiment analysis. This, in turn, aids in the quest for innovative solutions to extract valuable knowledge from vast datasets.

Given that Twitter embodies social networking features conducive to mining, it serves as the preferred system for detecting users' opinions. Tweets indicating the sentiments of their authors, user statistics, logical user metrics, hashtags, retweets, mentions, likes, and user mapping through graphs are extracted and analyzed based on the analyst's areas of interest. The classification process employs an algorithm utilizing a lexical approach, allowing real-time categorization of Twitter users' opinions into positive, negative, and neutral sentiments.

In this study, our focus is specifically on examining the data analysis practices employed by professionals working with social media data to advise decision-makers. Consequently, researchers dealing with social media data encounter various methodological and technical challenges, raising questions about the appropriate conduct of online social data research, encompassing considerations such as validity, ethics, and reproducibility.

As per researchers, the pervasive use of the Internet results in the continuous generation of vast amounts of data by individuals every second. The challenge at hand involves navigating how to handle this immense volume of information and exploring how organizations can harness these data. Notably, a significant portion of this knowledge resides within textual content, emphasizing the importance of real-time data analysis.

In response to this need, our work introduces OctopusViz, a comprehensive framework featuring a suite of applications designed to monitor and collect a substantial volume of tweets in real time, ensuring anonymity and online accessibility. OctopusViz further automates the processing, searching, viewing, and categorization of message sentiments into three distinct categories: positive, negative, and neutral. The framework operates by capturing tweets aligned with the analyst's specified interests and subsequently employs an algorithm implementing a lexical classification approach to categorize sentiment. Ultimately, OctopusViz presents the results in graphical form, facilitating real-time analyses and comparisons of metrics and sentiments across diverse Twitter users and topics as shown in **Figure 3**. This capability aids in informed decision-making processes across various environments and scenarios, including commercial, police, military, etc.

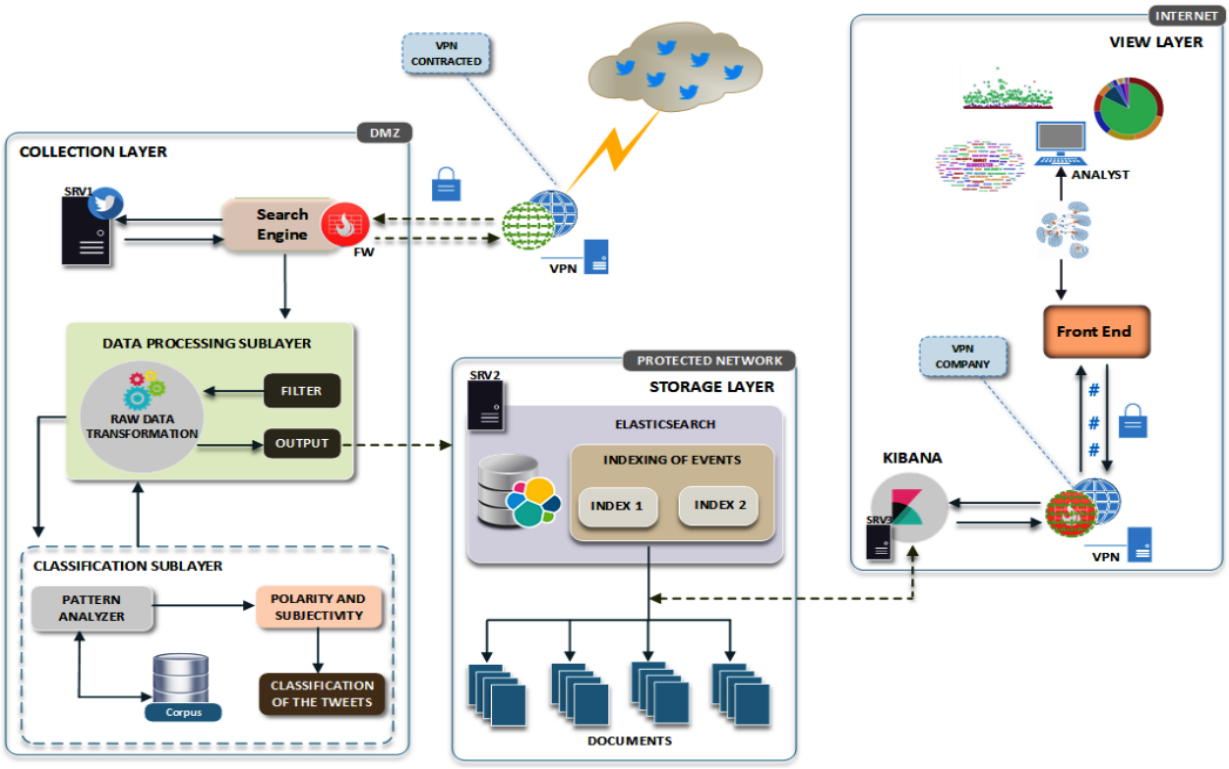


Figure 3: *OctopusViz architecture.*

The development of the proposed architecture took place in five phases, with Phase 1 dealing with the data collection layer; Phase 2 with the data processing sublayer; Phase 3 with the classification sublayer; Phase 4 with the distributed storage layer; and Phase 5 with the real-time tweets' visualization aspects. The details of each phase are explained below.

Phase 1: Data Collection Layer

The primary objective of this phase is to facilitate the real-time collection of data pertinent to the study from the Twitter platform. The Tweepy library, implemented in Python, is employed for authentication and data gathering purposes. This library interfaces with the Twitter API and facilitates user authentication through application keys and tokens. In the architectural setup, the search engine is configured to authenticate via the Virtual Private Network (VPN) established as part of the project, thereby ensuring both privacy and confidentiality of the collected data.

Phase 2: Data Processing Sublayer

Within the processing sublayer, the transformation of raw data into relevant information takes place. Given the absence of a standardized writing format on social

networks, a series of procedures is implemented to enhance the refinement of information. This transformative process is executed using the TextBlob and NLTK libraries within a Python script, enabling both data transformation and centralization. The TextBlob API, known for its capabilities in Natural Language Processing (NLP), offers functionalities such as sentiment analysis, classification utilizing algorithms like naive Bayes and decision tree, tokenization, translation, and spelling correction. This robust combination of tools ensures a comprehensive processing of the data to derive meaningful insights.

Prior to initiating any processing, the system performs language detection, translation, and automatic correction to ensure uniformity in the English language. The language detection is facilitated by the Google Translate API, employing methods such as **get_languages(), detect_language(), and translate().** Corrections are implemented using the **correct()** method from the TextBlob library. This dynamic translation and correction process, supported by the extensive language coverage of more than 100 languages and numerous language pairs, enhances the adaptability and inclusivity of the proposed system. **Table 1** outlines the functions involving translation and correction methods within the environment.

Table 1: *Function for translation and correction of tweets.*

Example of Data Input in Portuguese	O Brasil jogou muito bem contra a Costa Rica
Data Preprocessing (Translation and Correction)	*tweet = TextBlob("O Brasil jogou muito bem contra a Costa Rica")* *if tweet.detect_language() != 'en':* *translate_to_english = TextBlob(str(tweet.translate(to='en')))* *correct_tweet = translate_to_english.correct()* *print (correct_tweet)* *else:* *tweet.correct()* *print (tweet.correct())*
Data Output	*Brazil played very well against Costa Rich*

As part of the pre-processing activity, the system employs the **Stop Words and Special Characters technique** to eliminate words with minimal analytical value, including articles, prepositions, punctuation, conjunctions, and pronouns. The NLTK library's corpus stop words and methods such as **stopwords.words()** and **string.punctuation** facilitate this removal process. Additionally, the removal of URLs is incorporated, considering their lack of relevance to the sentiment analysis requirements in our work. **Table 2** illustrates the special characters, punctuation, and select stop words removed during pre-processing, while **Table 3** provides an example of the

function employed to eliminate stop words and special characters.

Table 2: *Corpus words stop words and special characters*

Methods	Method Description	Data Output
stopWords = set(stopwords.words('english')) *print(stopWords)*	Corpus words stop words	*['i', 'me', 'my', 'we', 'our', 'ours', 'his', 'y', 'your', 'it']*
string.punctuation	Scores and special characters	*'!"#$%&'()*+,-./:;<=>?@[]^_'{|}'*

Table 3: *Function for cleaning tweets*

Example of Data Input	Brazil is an excellent soccer team :) !!!
Data Preprocessing (Stop Words and Special Characters)	*tweet = TextBlob("Brazil is an excellent soccer team :) !!!")* *translation_correction(tweet)* *stopwords_english = stopwords.words('english')* *words = tweet.words* *words_clean = []* *for word in words:* *if word not in stop words_english:* *if word not in string.punctuation:* *words_clean.append(word)* *print (words_clean)*
Data Output	*['Brazil', 'excellent', 'soccer', 'team']*

Tokenization, a crucial pre-processing step, involves breaking down texts into individual words, phrases, or symbols. In our work, the method **textblob.tokenizers.WordTokenizer()** from the TextBlob library is employed to achieve this, facilitating subsequent analysis and tasks within the Classification Sublayer. **Table 4** provides an example of the function utilized for tokenization..

Table 4: *Function for tokenization of tweets.*

Example of Data Input	Brazil played very well against Costa Rica
Data Preprocessing (Tokenization)	*tweet = TextBlob("Brazil played very well against Costa Rica")* *translation_correction(tweet)* *tweet_clean_stop words(tweet)* *print (tweet.words)*
Data Output	*['Brazil', 'played', 'very', 'well', 'against', 'Costa', 'Rica']*

Phase 3: Classification Sublayer

The primary goal of this layer is to conduct sentiment analysis on tweets, identifying behaviors that provide insights into public opinion. The TextBlob library is configured for processing textual data, specifically for the classification of tweets.

Sentiment Analysis

Within the **textblob.sentiments()** module, two sentiment analysis algorithms are available:

1. PatternAnalyzer: Based on the Patterns library.
2. NaiveBayesAnalyzer: An NLTK classifier trained on a corpus of movie reviews.

For this work, the Pattern Analyzer algorithm and a lexical corpus were employed. Following data transformation, the information is directed to the sentiment analyzer. The Pattern Analyzer algorithm references the lexical corpus and classifies tweets based on polarity, subjectivity, and intensity.

- Polarity Score: Ranges from -1.0 to 1.0, where (0.01, 1 = positive), (−0.01, −1 = negative), and (0.0 = neutral).
- Subjectivity: Operates within the interval of (0.0, 1.0), with 0.0 being very objective and 1.0 being very subjective.

The **_text.py** class is responsible for calculating sentiment. **Table 5** illustrates the function used for classifying tweets based on polarity and subjectivity.

Table 5: *Function for tweets' classification.*

Example of Data Input	Brazil is an excellent soccer team :) !!!
Data Classification (Polarity and Subjectivity)	*tweet = TextBlob("Brazil is an excellent soccer team :) !!!")* *translation_correction(tweet)* *tweet_clean_stop words(tweet)* *tokenization(tweet)* *if tweet.sentiment.polarity > 0:* *print (tweet.sentiment)* *print ('Polarity: Positive')* *elif tweet.sentiment.polarity == 0:* *print (tweet.sentiment)* *print ('Polarity: Neutral')* *else:* *print (tweet.sentiment)* *print ('Polarity: Negative')*
Data Output	*Sentiment(polarity = 0.98828125, subjectivity = 1.0)* *Polarity: Positive*

Phase 4: Distributed Storage Layer

This layer focuses on indexing and searching large volumes of data. The process involves a separate host on the internal network utilizing the Elasticsearch tool. Elasticsearch is responsible for storing the complete structure of real-time monitoring

data from tweets with distributed storage. This approach enhances the understanding and interpretation of behaviors collected from Twitter.

Phase 5: Visualization Layer

The purpose of viewing tweets and retweets in this environment is to facilitate analysts in interpreting the data, enabling them to anticipate information and propose efficient measures from a data interpretation perspective. Real-time monitoring allows the observation of various elements such as tweets, retweets, mentions, hashtags, entity relationships through graphs, the number of likes, georeferenced information, and user sentiments on a specific topic. Kibana is employed in this process. Kibana provides a feature-rich interface, enabling advanced analytical queries, visualization, and interaction with data stored in Elasticsearch indexes.

After conducting tests, the framework demonstrated its capability to capture large amounts of tweets in real time. As a distinctive feature, the environment facilitates sentiment analysis, information extraction, user metrics and statistics, hashtag tracking, tweets and retweets analysis, and social bots' identification through outlier analysis. The quantitative data can be configured based on the specific needs and interests of those tasked with analyzing high-volume and high-speed data.

*IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining (*Czibula et al 2022)

The application of data mining techniques to extract meaningful insights from diverse data types has garnered significant interest, particularly for enhancing decision-making processes across various domains. Machine learning (ML) emerges as a pivotal tool, offering a plethora of models and techniques capable of unveiling concealed patterns in data from diverse practical domains. Examples include bioinformatics, meteorology, software engineering, medicine, computer vision, and educational data mining.

In this article, we introduce IntelliDaM, a versatile machine learning-based framework designed to elevate the efficacy of data mining tasks and, consequently, augment decision-making processes. IntelliDaM incorporates distinct components tailored for feature analysis, unsupervised learning-based data mining, and supervised

learning-based data mining, facilitating the extraction of hidden knowledge from data. The key components of IntelliDaM encompass: (1) a feature analysis and selection component; (2) an unsupervised learning-based data analysis component, offering data visualizations and performance evaluation metrics; and (3) a supervised learning-based data mining component that provides performance assessments for predictive models. **Figure 4** provides an overview of the IntelliDaM framework.

This framework proves instrumental in unraveling valuable insights, making it applicable across various domains. The feature analysis component enables a focused examination of relevant data attributes, while the unsupervised learning component aids in visualizing data patterns and evaluating performance metrics. On the other hand, the supervised learning component facilitates the creation and evaluation of predictive models, adding a layer of precision to decision-making processes. Through the seamless integration of these components, IntelliDaM stands as a robust tool for knowledge discovery and enhanced decision-making.

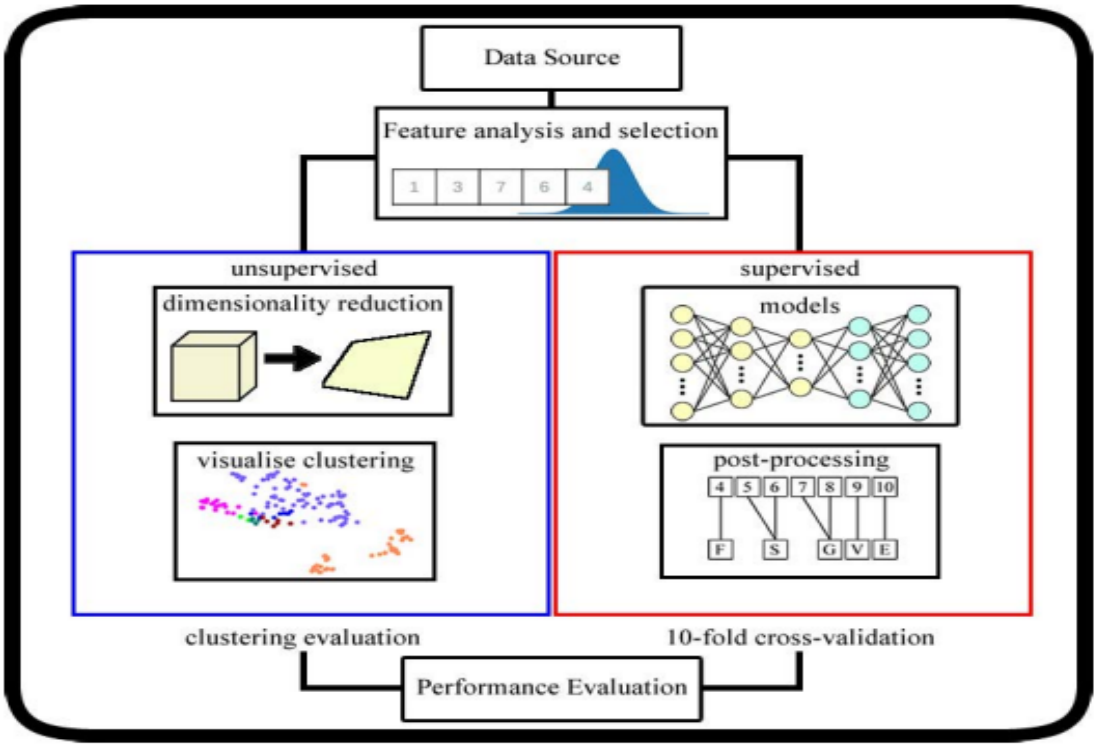


Figure 4: *Overview of IntelliDaM framework*

1. FEATURE ANALYSIS AND SELECTION

As in any machine learning task, independent of the type of learning (supervised or unsupervised), the relevance of the features used for characterizing the input instances is crucial for obtaining high performance. Theoretically, for a machine learning task, we would need features that are independent, and additionally, for supervised learning tasks, features highly correlated with the target output (in this case in any machine learning task, regardless of the learning type (supervised or unsupervised), the significance of the features used to characterize input instances is paramount for achieving high performance. For a machine learning task, it is ideal to have independent features and, especially for supervised learning tasks, features highly correlated with the target output. The feature analysis component of IntelliDaM encompasses three main functionalities.

Firstly, the statistical-based feature analysis employs Pearson and Spearman rank correlation coefficients to study the correlation between features and the target output. Pearson measures the linear relationship between two features, while Spearman rank correlation describes the strength and direction of the monotonic relationship between two variables. Next, the strength of association of two variables by using one of these two correlation coefficients can be interpreted as shown in **Table 6**.

Table 6: *Strength of association of two variables by using correlation coefficients*

Range of the absolute value of the correlation coefficient	Strength of association
[0, 0.2)	Very weak
[0.2, 0.4)	Weak
[0.4, 0.6)	Moderate
[0.6, 0.8)	Strong
[0.8, 1]	Very strong

Secondly, the feature selection process involves using the ReliefF algorithm, an extension of the Relief algorithm. ReliefF supports multi-class classification and is robust in the face of interactions between features, making it suitable for our student performance prediction (SPP) task.

Thirdly, the feature sets' quality analysis introduces a metric, QF(A, St), to measure the predictive performance or relevance of a feature/attribute set A for differentiating the target output among input instances. This metric aids in

understanding how well certain features characterize the academic performance of students.

2. UNSUPERVISED LEARNING-BASED ANALYSIS

Moving on to the unsupervised learning-based analysis, IntelliDaM employs t-Distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. Both methods, rooted in neighbor graphs, offer insights into local and global data structures. To enhance visualizations, a k-means labeling step is included, compacting data into groups with similar characteristics.

For performance evaluation of clustering models, external metrics such as homogeneity, completeness, and V-measure are used. These metrics, relying on ground truth labeling, provide insights into the effectiveness of clustering.

3. SUPERVISED LEARNING-BASED ANALYSIS

In the supervised learning-based analysis, IntelliDaM incorporates three regression models: Tweedie regressor, Stochastic Gradient Descent (SGD), and a polynomial model (Poly). Each model implements a different approach, contributing to the robustness of the framework.

Performance evaluation for regression models involves two measures: Root Mean Squared Error (RMSE) and normalized Root Mean Squared Error (NRMSE). Additionally, the SPP regression task is transformed into a multi-class classification problem, enabling the use of accuracy, precision, recall, and F-measure metrics.

Through these comprehensive analyses and evaluations, IntelliDaM proves to be a versatile framework capable of enhancing decision-making processes by extracting valuable knowledge from various data types.

Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining (Kauffmann et al 2019)

The continuous success of companies significantly dependent on their ability to effectively satisfy customer needs. Essentially, the pursuit of customer satisfaction aims at establishing brand value, a pivotal factor for a company's enduring success. Consequently, considerable financial investments are made by many companies in marketing research to glean insights into consumer preferences and demands. From this wealth of information, a critical aspect is comprehending consumer perspectives on the products they purchase, vital for formulating relevant branding and positioning strategies. With widespread internet access, an abundance of data is generated, offering a promising avenue to determine consumer opinions about purchased and experienced products. Organizations seek to leverage this data, transforming it into actionable information for informed decision-making—a process facilitated by the analysis of extensive data, commonly referred to as "big data." Consumer comments on online forums emerge as a valuable source for unearthing consumer insights, with user-generated content (UGC) standing out as a promising alternative for identifying potential customer needs.

Despite prior research revealing the utility of sentiment scores and the examination of positive and negative product features in decision-making, a gap exists in studies that integrate product price, quantitative star scores from users, sentiment scores from sentiment analysis (SA) tools in global reviews, and sentiment scores for each specific extracted feature. This study focuses on Sentiment Analysis (SA) techniques and the application of Natural Language Processing (NLP) tools in shaping marketing decision-making strategies. Initially, customer preferences, as indicated by star scores from users in UGC and sentiment scores, were analyzed. Subsequently, the review was divided into positive, neutral, and negative sections, unraveling distinct sentiments expressed by customers. Lastly, the identification of primary product features evoking positive, neutral, and negative sentiments among clients was undertaken. The research encompassed three significant analyses: (1) a comprehensive sentiment analysis at the review level to measure overall product likability, (2) an in-depth examination of distinct phrases at the sentence level to distinguish buyer sentiments, and (3) the extraction of positive, neutral, and negative product features at

the aspect level. The central contribution of this research lies in presenting a well-structured and effective architecture, segmented into stages, to navigate this landscape of consumer feedback and decision-making.

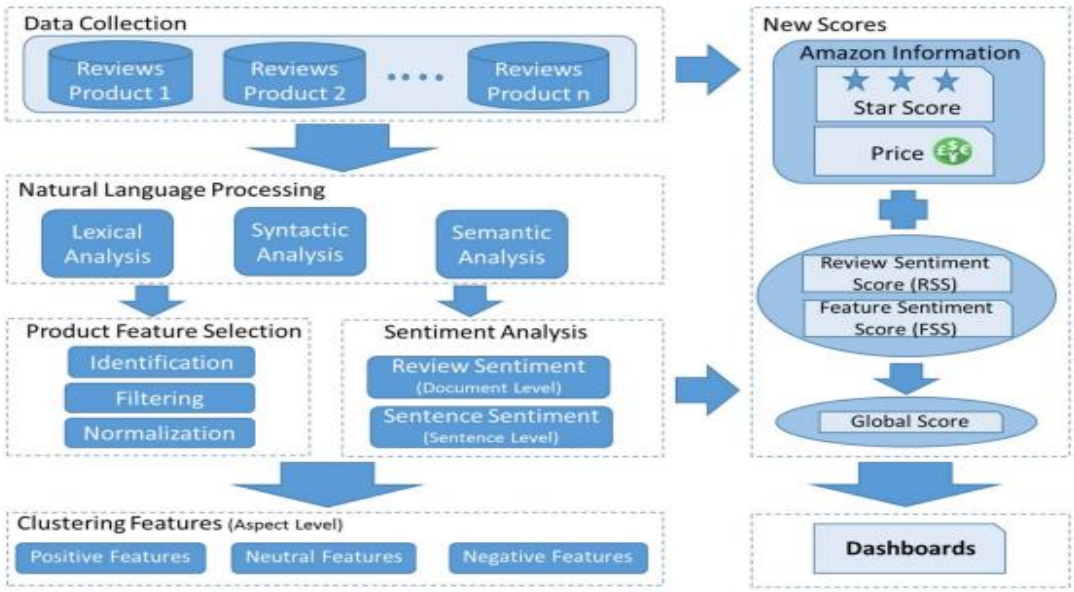


***Figure 5**: The proposed architecture using sentiment analysis (SA) and text data mining to identify the main positive/negative product features*

As depicted in **figure 5**, our process unfolds in seven distinct stages: (1) data collection, (2) review preprocessing using Natural Language Processing (NLP) techniques, (3) product feature selection, (4) sentiment analysis, (5) clustering features, (6) new scores, and (7) dashboards. The architecture allows for a distinct analysis of reviews at various levels: (1) At stage 4, Sentiment Analysis (SA) provides a global score for the entire review (document level), evaluating overall product attractiveness. (2) Simultaneously, in stage 4, scores for each sentence are computed (sentence level), unveiling specific sentiments expressed by buyers regarding the product. (3) Progressing to stage 5, clustering features, scores for each feature are derived (aspect level), categorizing them as positive, neutral, or negative. The subsequent sections elaborate on each stage.

1. Data Collection Stage:

We accumulated product reviews along with pertinent information such as price, brand, and categories. This data serves as a foundation for our analysis, aiming to unveil insights that aid managers and users in informed decision-making. Reviews typically

include an explicit star score and unstructured textual comments. The star score is global, encapsulating the overall product experience, while the textual comments encompass diverse opinions on different aspects of the product.

2. Review Preprocessing Using NLP Techniques:

NLP preprocessing is employed on textual reviews, encompassing lexical (analyzing individual words), syntactic (study of sentence structure and grammar), and semantic analyses (meaning of a sentence or text). This process enriches words with Part-Of-Speech (POS) tags (lexical information) and semantic information of the different words. Moreover, Product features are selected based on product descriptions, using NLP tools to extract lexical, syntactic, and semantic information. This information aids in ranking positive, neutral, and negative elements for showcasing in dashboards.

3. Product Feature Selection Stage:

A domain ontology, specifically tailored for cell phones/mobile phones, is employed for detecting primary product features. This ontology serves as a reference list, and product descriptions are mined to identify the most frequently used nouns. Alternative techniques explore statistical patterns in the text, unveiling frequently mentioned words and phrases.

4. Sentiment Analysis Stage:

Two scores are calculated in this stage: a global sentiment score for each review and a specific sentiment score for each main feature of the product. The Afinn affective lexicon is utilized for sentiment analysis, assigning sentiment scores based on the textual comments of product reviews. The sentiment scores act as crucial criteria for determining the best products within a category or brand.

5 Clustering Features Stage

To evaluate the sentiment polarity of a product feature, the sentiment score of each phrase in which the feature appears is evaluated, and the average of these scores is calculated using Equation (1). This is done for each product feature. The features are then classified as having a positive, neutral, or negative score.

$$sentiment\_score(feature) = \frac{\sum_{p \in P} sentiment\_score(p)}{|P|},$$

where $P = \{p | p \; is \; a \; phrase \; of \; review \; \wedge$ feature in $p\}$.

6 New Score Stage

In this crucial stage, our process computes two innovative scores, each playing a pivotal role in evaluating the comprehensive performance of a product. The Feature-based Score (FBS) is ascertained by averaging the sentiment scores associated with all features of the product, as articulated by Equation (2):

$$feature\_based\_score(product) = \frac{\sum_{f \in F} sentiment\_score(f)}{|F|},$$

where $F = \{f | f \ is \ a \ feature \ of \ product\}$.

Subsequently, a novel score known as the feature sentiment score (FSS) is introduced for a product. This FSS is then combined with the product price, star score, and review sentiment score (RSS) to compute the global score for the product, as shown in Equation (3). Each variable in this computation is assigned a specific weighting, aligning with consumer priorities. Significantly, the product price holds the utmost importance for the consumer, followed by the sentiment score, RSS, and FSS, each assigned weightings greater than that of the star score.

$$\begin{aligned}
&\text{GlobalScore(product)} = \\
&\text{NormalizedPrice(product)} * 0.3 + \\
&\text{NormalizedStarScore(product)} * 0.2 + \\
&\text{NormalizedSentimentScore(product)} * 0.25 \\
&\text{NormalizedFeatureSentimentScore(product)} * 0.25 \\
&\text{NormalizedPrice(product)} = \frac{(MaxPrice - Price + 1)}{MaxPrice} \\
&\text{NormalizedScore(product)} = \frac{Score - MinScore}{(MaxScore - MinScore)}.
\end{aligned}$$

7 Dashboards Stage

In the final stage, extracted data are presented through various dashboards, such as word clouds. These dashboards offer valuable insights into positive/negative features of a product or the overall ranking of top products using the global score. They serve as a powerful tool for companies to track the evolution of their products based on in-depth consumer reviews.

In conclusion, our proposed methodology was applied to a corpus of reviews focusing on cell phone products sourced from Amazon. Our analysis revealed that while star scores provided a general overview of a user's perspective on the product, reviews

allowed users to specifically highlight features they either appreciated or disliked. It's important to note that customer preferences are shaped by distinct product features and the significance attached to these attributes. Therefore, delving into a consumer's opinions regarding these specific features holds paramount importance for managers. The perception consumers hold of a brand profoundly influences the brand's value concerning these attributes. This approach aids managers in gaining a deeper understanding of their firm's product positioning. By elucidating consumer opinions on the main features of a product, our methodology provides insights into the market structure and brand positioning. This, in turn, assists marketing managers in making informed decisions and navigating their decision-making processes more effectively.

Data mining based marketing decision support system using hybrid machine learning algorithm (Kumar 2020)

With the ongoing advancements in technology, computer-based business decision-making has become a common practice, even among small-scale organizations, aiming to gain insights into the organization's future. Decision support systems have emerged to alleviate the human workload involved in analyzing organizational data, offering prompt decisions based on available information to enhance organizational quality in a short timeframe. Recognizing that data quality plays a pivotal role in defining an organization's quality, timely decisions are crucial for steering growth in the right direction. Decision support systems excel in selecting and analyzing data to identify trends, facilitating the formulation of strategies and solutions. Given the centrality of data in the decision-making process, these support systems present information graphically or textually, leveraging expert artificial intelligence.

In this research, a novel data mining-based decision support system is proposed, employing a hybrid approach with decision trees and artificial neural networks to estimate marketing strategies for organizations. It is noted that many decision support systems rely on conventional statistical models, with limited exploration of machine learning models, presenting an opportunity for improved efficiency. Addressing this gap, the proposed hybrid model is designed to offer enhanced decision support for the marketing domain. Applied to a manufacturing company, the hybrid model aids in decision-making regarding marketing ideas, addressing existing issues within the company's marketing section. The company's choice of marketing strategies

significantly impacts its image and stakeholder relations, emphasizing the need for a decision support system to develop production and marketing strategies, ultimately boosting company profits.

The process flow of the proposed hybrid model, as illustrated in Figure 6, consists of three phases: problem setting, model development, and action development. This structured approach ensures a systematic and comprehensive decision-making process tailored to the unique challenges and opportunities present in the manufacturing company's marketing landscape.

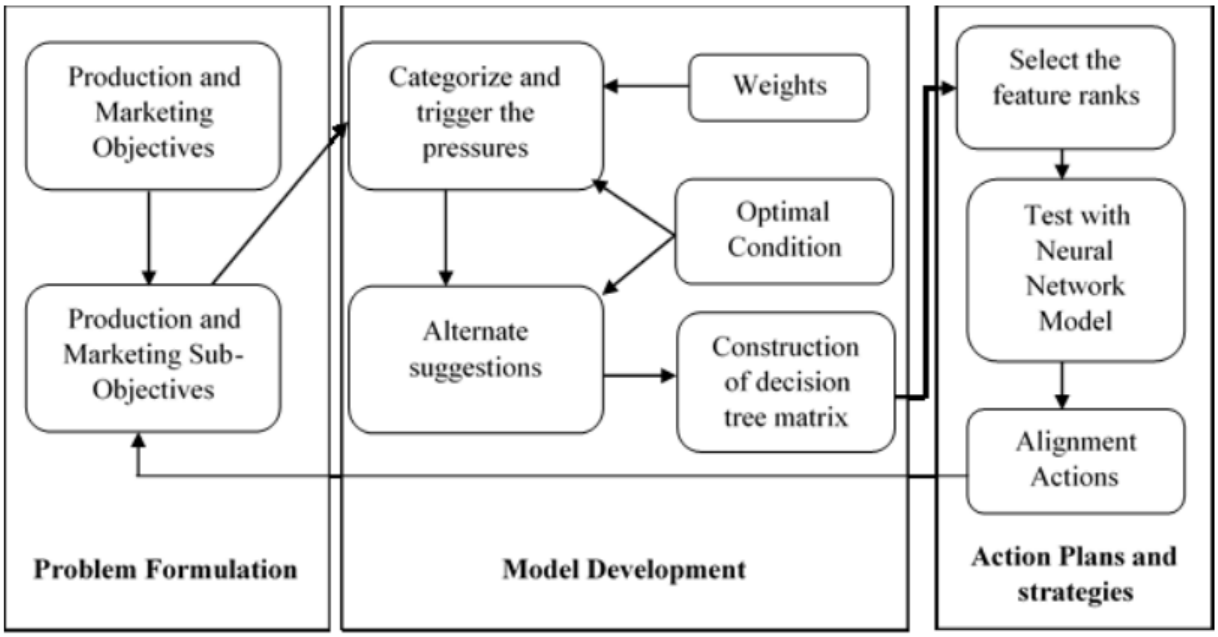


Figure 6: *Proposed Hybrid Model*

The initial **problem formulation** phase is used to define the scope and identifying the unavoidable issues in the industry. The route-cause analysis and its relationship is essential for decision support system, in order to identify the root a framework is setup for production and marketing as objectives. In this level, the essential production and marketing objectives are further subdivided in order to reach the framework from top level to bottom level employees. Based on hierarchical order, these objective functions are created.

Model development is the second phase which includes formal models. These models are developed based on the sub-objectives, so that the present manufacturing and marketing aspects could be rearranged. Considering the normalizing, constituting and decision tree weight functions which is given as a matrix function where the column

of the matrix is used to represent the alternate solutions and row of a matrix is used to represent the triggered pressures.

Artificial neural network is used to refine the results of decision tree model. **Figure 7** depicts the neural network model used in the proposed decision support system for classifying the decision tree hierarchical results.

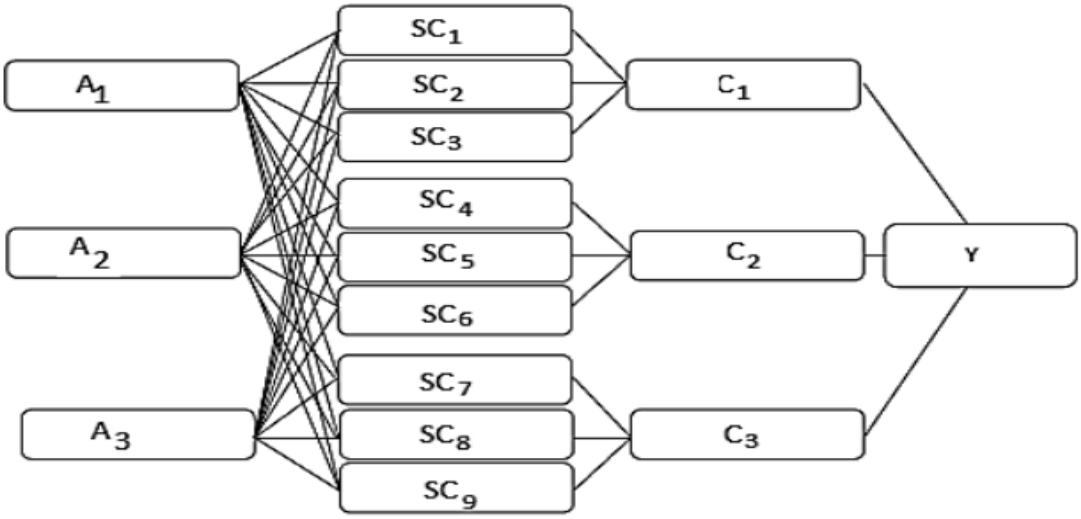


Figure 7: *Neural Network Model*

The process initiates with inputs from the decision tree system at three levels, which are then meticulously processed into nine different intermediate levels based on the predefined sub-objectives within the proposed decision support system. The interconnected nature of this process signifies the system's engagement across all organizational levels, fostering interaction from top-level management to bottom-level employees. The cumulation and classification of these sub-objectives ultimately contribute to the formulation of a comprehensive strategic plan for both manufacturing and marketing processes.

In the final stages, responses are analyzed using the decision tree matrix and undergo training through an artificial neural network to enhance the accuracy of classification strategies. The experimental results are observed and subsequently compared with conventional models such as the hidden Markov Model and support vector-based decision support systems. The findings indicate that the proposed model consistently achieves superior classification results, demonstrating its applicability and effectiveness across various domains.

4.2 Discussion

This section provides a comprehensive examination of key research articles in the domain of big data analytics and decision-making, offering insights into their objectives, areas of focus, original contributions, methodologies, frameworks/architectures, and associated limitations.

1. Objective:

In examining the objectives of the analyzed articles, each publication takes a distinct approach. The work by Goar and Yadav (2022), titled "Business Decision Making by Big Data Analytics," aims to develop and evaluate the B-DAD framework, with a focus on its application in real-world retail scenarios. This involves a thorough examination of decision-making enhancements, particularly concerning product promotions, timing, and the impact of social media on sales. In contrast, de Oliveira Junior et al. (2020)'s contribution, "Anonymous Real-Time Analytics Monitoring Solution for Decision Making Supported by Sentiment Analysis," centers around proposing a framework for real-time analytics monitoring on Twitter. The emphasis is on leveraging sentiment analysis to capture and interpret large volumes of tweets for decision-making insights.

Moving on to Czibula et al. (2022)'s work, "IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes," shifts the focus to the educational domain. The primary objective is to introduce and evaluate the IntelliDaM framework, a machine learning-based approach in the context of Educational Data Mining (EDM), aiming to enhance the performance of data mining tasks, particularly in the realm of education.

Shifting attention to Kauffmann et al. (2019)'s contribution, "Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining," explores the domain of marketing decision-making. The objective is to manage marketing decisions through sentiment analysis and text data mining, evaluating the main features of products using these techniques.

Lastly, Kumar (2020)'s research, "Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm," takes a broader approach by proposing and analyzing a hybrid decision support system for marketing. The focus is on improving manufacturing and marketing strategies by addressing issues in existing decision support models.

2. Area of Focus:

As we explore the areas of focus, each article brings a unique perspective. "Business Decision Making by Big Data Analytics" centers on the integration of big data analytics into managerial decision-making processes within the retail sector. "Anonymous Real-Time Analytics Monitoring Solution for Decision Making Supported by Sentiment Analysis" focuses on real-time analytics monitoring, sentiment analysis, and the consequential impact on decision-making processes, particularly in the context of social media platforms like Twitter.

In "IntelliDaM," the spotlight shifts to the machine learning framework's application in the educational domain, specifically targeting knowledge discovery and decision-making enhancement. "Managing Marketing Decision-Making with Sentiment Analysis" narrows its focus to marketing decisions, employing sentiment analysis and text data mining to evaluate consumer sentiments towards main product features. In contrast, "Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm" proposes a hybrid decision support system for marketing, utilizing a combination of decision trees and artificial neural networks.

3. Originality:

Regarding original contributions, Goar and Yadav (2022) introduce the B-DAD framework, showcasing novelty in seamlessly integrating big data analytics into the decision-making fabric of retail. This innovative approach holds promise for enhancing decision-making processes in business. De Oliveira Junior et al. (2020)'s work introduces OctopusViz, a comprehensive framework tailored for real-time analytics monitoring, particularly supported by sentiment analysis on Twitter. The originality lies in its ability to capture and interpret large volumes of real-time tweets for decision-making insights.

In Czibula et al. (2022)'s "IntelliDaM," the original contribution lies in the introduction of a versatile machine learning-based framework designed for educational data mining. Kauffmann et al. (2019)'s article brings innovation through the combination of sentiment analysis, natural language processing techniques, and text data mining in managing marketing decision-making. Kumar (2020)'s research introduces a novel data mining-based decision support system, employing a hybrid machine learning algorithm to enhance organizational performance through efficient data mining approaches.

4. Methodology:

Methodologies adopted by the articles diverge significantly. "Business Decision Making by Big Data Analytics" employs a design science technology methodology, unfolding in six distinct phases. These phases systematically integrate big data tools, analytics, and architecture into decision-making processes. De Oliveira Junior et al. (2020)'s work proposes a five-phase methodology tailored for real-time Twitter monitoring. This includes configuring search engines, transforming and centralizing data, employing sentiment analysis algorithms, and providing tools for analyst interpretation.

In Czibula et al. (2022)'s "IntelliDaM," the methodology involves a comprehensive machine learning-based approach encompassing diverse components. Kauffmann et al. (2019)'s article adopts sentiment analysis and text data mining to evaluate consumer sentiments towards main product features, emphasizing the analysis of consumer opinions from online forums and user-generated content. Kumar (2020)'s research proposes a hybrid model integrating decision tree and artificial neural network techniques for marketing decision support, spanning three phases: problem setting, model development, and action development.

5. Framework/Architecture:

When it comes to frameworks or architectures, Goar and Yadav (2022) introduce the B-DAD framework, designed for seamless integration of big data analytics into managerial decision-making processes. This architecture offers a structured approach, particularly focused on retail contexts. De Oliveira Junior et al. (2020) presents OctopusViz, a comprehensive framework personalized for real-time analytics monitoring, particularly supported by sentiment analysis on Twitter. Czibula et al. (2022)'s "IntelliDaM" is introduced as a versatile machine learning-based framework designed for educational data mining. Kauffmann et al. (2019)'s article proposes a well-structured architecture for managing marketing decision-making using sentiment analysis, natural language processing techniques, and text data mining. Kumar (2020)'s research proposes a novel data mining-based decision support system, employing a hybrid machine learning algorithm.

6. Limitations:

As for the associated limitations, each article grapples with unique challenges. "Business Decision Making by Big Data Analytics" acknowledges the experimental limitations associated with the unique characteristics of the Turkish language, making sentiment analysis in Turkish challenging due to its intricate nature. "Anonymous Real-Time Analytics Monitoring Solution for Decision Making Supported by Sentiment Analysis" recognizes the potential limitation in the exclusive reliance on a lexicon-based sentiment analysis approach, particularly the Pattern Analyzer algorithm. The study anticipates that findings and applicability might not seamlessly extend to other platforms or contexts beyond the scope of Twitter.

In Czibula et al. (2022)'s "IntelliDaM," limitations become apparent in the context specificity of the case study conducted at a particular educational institution. The observed strong correlation between the unsupervised and supervised learning components may not universally hold across diverse scenarios and domains, indicating potential limitations in generalizability. Kauffmann et al. (2019)'s study acknowledges limitations stemming from the reliance on the Afinn lexicon for sentiment analysis, with the acknowledgment that alternative sentiment analysis tools could yield different results.

In Kumar (2020)'s research, the effectiveness of the proposed hybrid model is contingent on employee responses, potentially introducing subjectivity or bias into the decision-making process. Furthermore, the study's specific focus on internal and external pressures and target criteria may limit its applicability to decision-making contexts beyond those explicitly addressed in the research. The constraints of the model could become apparent when applied to diverse organizational contexts.

Table 7: *Results of the Analysis*

Article	Objective	Area of Focus	Originality	Methodology	Framework/ Architecture	Limitations
Article 1	Develop and evaluate B-DAD framework in retail domain.	Integration of Big Data Analytics into Managerial Decision-Making Processes, Retail Analytics, Social Media Analysis for Decision Support.	High, introduces B-DAD framework in retail decision-making.	Design science technology methodology, six-phase process, integration of big data tools, analytics, and architecture.	B-DAD framework for integrating big data analytics into managerial decision-making in retail.	Challenges in Turkish sentiment analysis, context-dependent nature of language.
Article 2	Propose real-time analytics monitoring on Twitter using sentiment analysis.	Real-Time Analytics Monitoring, Sentiment Analysis for Decision Support, Social Media Analytics Impact on Decision-Making.	Innovative OctopusViz framework with anonymity emphasis.	Five-phase methodology, VPN for collector anonymity.	OctopusViz framework for real-time analytics monitoring supported by sentiment analysis.	Reliance on lexicon-based sentiment analysis, VPN usage challenges.
Article 3	Introduce and evaluate IntelliDaM framework in educational data mining.	Machine Learning Framework for Decision-Making, Educational	Versatile IntelliDaM framework for knowledge discovery	Comprehensive machine learning-based methodology, feature	IntelliDaM framework for educational data mining.	Case study specificity, potential correlation challenges between

		Data Mining, Performance Enhancement in Decision-Making.	and decision-making in education.	analysis, unsupervised, and supervised learning.		unsupervised and supervised learning.
Article 4	Manage marketing decision-making through sentiment analysis and text data mining.	Marketing Decision-Making, Sentiment Analysis in Marketing, Text Data Mining for Product Feature Evaluation.	Innovative combination of sentiment analysis, text data mining, and clustering for product feature evaluation.	Proposed architecture using sentiment analysis and text data mining, seven-stage process.	Architecture for managing marketing decision-making using sentiment analysis and text data mining.	Reliance on Afinn lexicon, manual derivation of product features introduces subjectivity.
Article 5	Propose and analyze hybrid decision support system for marketing.	Data Mining in Marketing, Decision Support Systems for Marketing, Hybrid Machine Learning Algorithm for Decision Making in Marketing.	Innovative hybrid model using decision tree and artificial neural networks.	Three-phase methodology, engagement of employees across organizational levels.	Hybrid model integrating decision trees and artificial neural networks.	Potential subjectivity or bias in employee responses, specific focus on internal and external pressures.

4.3 Research Gap

In the analysis of the state-of-the-art in big data analytics techniques and platforms for real-time business decision-making, a notable gap emerges, primarily in the realm of sentiment analysis. While existing literature showcases advancements in analytics tools, scalability, and decision-making frameworks, a crucial aspect that remains under-addressed is the need for a comprehensive and standardized approach to handling sentiment analysis across diverse languages and cultural contexts.

In decision-making fields like marketing and social media analytics, effective cross-cultural communication is vital. Current sentiment analysis tools often struggle when applied to languages and cultures beyond their primary focus, leading to inaccuracies. A standardized approach accommodating cultural and linguistic diversity can significantly enhance the reliability and effectiveness of sentiment analysis.

Considering this, the related scientific question emerges: "How can sentiment analysis tools be adaptively standardized to ensure accurate interpretation of sentiment across diverse languages and cultural contexts, thereby enhancing their reliability and effectiveness in global decision-making processes?"

5 Objective

The primary objective of this thesis is to analyze and explore the challenges and opportunities associated with big data analytics techniques and platforms. Specifically, the research aims to investigate the limitations highlighted in existing literature, focusing on the standardized handling of sentiment across diverse languages and cultural contexts in the realm of big data analytics.

The paper's gap will be addressed by adopting design science, specifically by creating a standardized approach to sentiment analysis across diverse languages and cultural contexts within big data analytics. This approach will contribute to enhancing the reliability and effectiveness of sentiment analysis in global decision-making processes

Design science will be employed due to its emphasis on solving problems in practice by proposing innovative solutions and evaluating their effects. It fits well with the objectives of this research, as it allows for the development of a concrete, testable solution that addresses the identified gap in the field of big data analytics and sentiment analysis.

6 Methodologies

This chapter discusses the methodologies that can be used for the solution approach which are: design science, data science, systematic literature review, and qualitative research.

6.1 Design Science

For design science, the methodology can be broken down into several steps, including problem identification, solution objectives, design, demonstration, evaluation, and communication.

1. Problem Identification:

Clearly define the research challenge and explain why a solution is valuable. It could be helpful to conceptually break down the problem so that the solution can adequately represent the complexity of the problem, as the problem definition will be utilized to create an effective artifactual solution. Two goals are achieved by defending the worth of a solution: it encourages the researcher and the research's audience to pursue the answer and accept the findings, and it facilitates comprehension of the logic underlying the researcher's comprehension of the issue. Knowledge of the problem's current condition and the significance of its solution are among the resources needed for this task. (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007: 56).

2. Objective & Analysis:

Determine a solution's goals based on the description of the problem. The goals can be qualitative, such as when a new artifact is projected to assist answers to issues not previously addressed, or quantitative, such as terms in which a desirable solution would be preferable than present ones. The problem statement should be used to reasonably derive the objectives. Knowledge of the current state of issues, their remedies, and their efficacy, if any, are among the resources needed for this (Peffers et al, 2007: 56).

3. Design & Development:

The artifactual solution should be created. These artifacts could be models, representations, constructs, or methods, each specified generically. This task involves focusing on the architecture and desired function of the artifact before actually building it. One of the resources needed to go from objectives to design and development is theory knowledge that can be applied to a problem (Peffers et al, 2007: 57).

4. Demonstration:

Provide evidence of the artifact's ability to effectively address the issue. This can involve applying it to a case study, simulation, experiment, proof, or other suitable task. Effective understanding of how to utilize the artifact to address the problem is one of the resources needed for the demonstration (Peffers et al, 2007: 57).

5. Evaluation:

Evaluate and quantify the extent to which the artifact facilitates a problem-solving approach. In this exercise, the goals of a solution are contrasted with the real-world outcomes that are observed when the artifact is used in the demonstration. Understanding relevant metrics and analytic methods is necessary. Evaluation criteria could include things like comparing the functionality of the artifact with the solution objectives, objective quantitative performance measures like budgets or items produced satisfaction surveys, client feedback, or simulations, depending on the nature of the problem venue and the artifact. The researchers have the option to continue with communication and save additional enhancement for later projects, or they can go back and try to improve the effectiveness of the artifact. According to Peffers et al. (2007: 57), the capability of a particular iteration may depend on the characteristics of the study venue.

6. Communication:

When applicable, communicate to researchers and other relevant audiences—such as practicing professionals—the nature of the problem and its significance, the artifact, its uniqueness and utility, the careful planning of its design, and its effectiveness. Just as the nominal structure of an empirical research process (problem definition, literature review, hypothesis development, data collection, analysis, results,

discussion, and conclusion) is a common structure for empirical research papers, researchers may use the structure of this process to structure their work in scholarly research publications. Understanding the disciplinary culture is necessary for effective communication (Peffers et al, 2007: 58).

6.2 Data Science

Data mining can benefit from a standardized process to bridge the gap between business challenges and technical solutions. This process should guide data transformation, technique selection, and evaluation of results, while also capturing lessons learned.

The CRISP-DM (CRoss Industry Standard Process for Data Mining) project offers a solution by outlining a data mining project framework. This framework is applicable across industries and technologies, providing a structured approach regardless of the specifics. It consists of six phases that will be explained: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. (Wirth & Hipp, 2000: 30).

1. Business Understanding:

This first stage focuses on understanding the project's requirements and goals from a business standpoint. Based on this understanding, a data mining problem definition and a rough project schedule are then created (Wirth & Hipp, 2000: 33).

2. Data Understanding:

The initial data collection is the first step in the data understanding phase, which then includes activities to familiarize with the data, identify issues with its quality, gain preliminary understanding of the data, or identify intriguing subsets to generate theories about hidden information. Business understanding and data understanding are closely related concepts. A basic comprehension of the available data is necessary for the creation of the data mining challenge and the project plan (Wirth & Hipp, 2000: 33).

3. Data Preparation:

Data preparation is the process of creating the final dataset (the data that will be used to enter into the modelling tools) from the first raw data. The steps involved in data preparation will probably be taken more than once and not necessarily in that order. Table, record, and attribute selection, data cleansing, the creation of new attributes, and

data transformation for modelling tools are among the tasks (Wirth & Hipp, 2000: 33-34).

4. Model Building:

Various modelling techniques are chosen and used during this phase, and their parameters are calibrated to ideal values. Usually, the same data mining problem type can be solved using a number of different strategies. Specific data formats are required by some processes. Modeling and Data Preparation are closely related. While modelling, one frequently becomes aware of data issues or has ideas for creating additional data (Wirth & Hipp, 2000: 34).

5. Testing and Evaluation:

Before moving forward with the model's final deployment, it's crucial to perform a more thorough evaluation of the model and assess the procedures taken to build the model in order to make sure it correctly achieves the business objectives. Identifying any significant business issues that have not been appropriately taken into account is one of the main goals. A choice on the application of the data mining results should be made at the conclusion of this stage (Wirth & Hipp, 2000: 34).

6. Deployment:

In most cases, the project doesn't finish with the creation of the model. The acquired knowledge typically has to be arranged and presented so that the client may make use of it. The deployment step can range in complexity from establishing a repeatable data mining process to something as simple as producing a report, depending on the objectives. Across several situations, the user will perform the deployment stages rather than the data analyst. In any case, it's critical to be aware of the steps that must be taken in advance in order to apply the developed models effectively (Wirth & Hipp, 2000: 35).

6.3 Systematic Literature Review

This section outlines the key steps involved in conducting a Systematic Literature Review (SLR). An SLR aims to provide a reliable and unbiased method for researchers to gain a comprehensive understanding of a specific research topic (Van Dinter et al., 2021). Torres-Carrión et al. (2018: 1371) break down the SLR process into three main phases: planning, conducting the review, and reporting the review.

Phase A: Planning

The planning phase lays the groundwork for a successful SLR. It consists of three sub-phases:

1. Defining the Research Problem: A strong understanding of the research problem forms the foundation for any scientific process, including an SLR (Torres-Carrión et al., 2018: 1372). The researcher's approach to the problem and their clarity regarding its nature significantly impact all subsequent phases of the SLR (Torres-Carrión et al., 2018: 1372).

2. Formulating Research Questions: Researchers emphasize the importance of research questions as a key component of the problem statement. These initial questions formulated in this sub-phase guide the entire SLR process. They capture the researcher's core interests and existing knowledge within the specific field of study (Torres-Carrión et al., 2018: 1372).

3. Conceptual Framework (Mentefacto Conceptual): An optional tool that can be helpful during the planning phase is the conceptual framework, also known as mentefacto conceptual, designed by researchers. This tool allows researchers to visually represent the key concepts involved in their research question (Torres-Carrión et al., 2018: 1372).

Phase B: Developing a Review Protocol

The development of a review protocol is the second phase of the SLR process. This phase consists of two sub-stages:

1. Definition of Inclusion and Exclusion Criteria: These criteria determine which studies are relevant (included) and irrelevant (excluded) for your review (Torres-Carrión et al., 2018: 1373). Establishing clear inclusion and exclusion criteria during the planning of the search process is crucial. These criteria consider both general and specific factors, along with any additional parameters that may be relevant to your research question (Torres-Carrión et al., 2018: 1373).

2. Preparing a Data Extraction Form: Researchers need a system for organizing the information they extract from the studies they select (Torres-Carrión et al., 2018). This sub-stage focuses on preparing a data extraction form, which may involve using spreadsheets, bibliographic management software, or other tools to specify and configure how the results will be organized (Torres-Carrión et al., 2018: 1373).

Phase C: Conducting the Review

Once the protocol is established, researchers can move on to the review phase. This phase is broken down into five sub-stages:

1. Identification of Research: Building on the protocol developed earlier, this sub-stage focuses on creating search strategies using relevant databases and keywords (Torres-Carrión et al., 2018: 1374). It's important to consider potential publication bias, document retrieval methods, and to document the search process for transparency purposes (Torres-Carrión et al., 2018: 1374).

2. Selection of Primary Studies: This sub-stage involves obtaining and reviewing the full text of the studies identified through the search strategy (Torres-Carrión et al., 2018: 1375). Pre-defined selection criteria are applied to ensure the studies directly address the research question. Final decisions regarding inclusion or exclusion are only made after a full-text review of each study (Torres-Carrión et al., 2018: 1375).

3. Study Quality Assessment: In addition to relevance, the quality of the selected studies is also assessed during this sub-stage. This evaluation considers factors such as the study's methodology, the quality of the bibliographic sources used, the relevance and academic prestige of the authors, and the impact factor of the journal in which the study was published (Torres-Carrión et al., 2018: 1375).

4. Data Extraction and Monitoring: Here, researchers design forms to systematically capture key information from each included study (Torres-Carrión et al., 2018: 1375). These forms may include details like the review name, date of data extraction, title, authors, journal, publication details, and space for additional notes (Torres-Carrión et al., 2018: 1375)

5. Data Synthesis and Monitoring: The final sub-stage of the review phase focuses on data synthesis and monitoring, which ultimately defines the quality of the systematic review (Torres-Carrión et al., 2018: 1376). The synthesis process itself can be either descriptive (non-quantitative) or involve a combination of descriptive and quantitative analysis, known as a meta-analysis. Researchers further elaborated on the specific attributes that meta-analysis should possess, depending on whether it's qualitative or quantitative in nature (Torres-Carrión et al., 2018: 1376).

D. Reporting the review

Communication of the systematic review's findings to the scientific community is crucial. This allows for peer review, which strengthens the review's validity (Torres-Carrión et al., 2018:1376). Systematic reviews hold significant value in organizing

research results within a specific scientific field over an extended period. They are frequently integrated as a chapter within PhD theses and can also be presented at conferences or published in specialized journals relevant to the research area (Torres-Carrión et al., 2018:1376).

6.4 Qualitative Research

The qualitative research method relies on non-numerical data like interviews, documents, and observations to understand and explain social phenomena. In the field of Information Technology and Communication (ICT), research has shifted its focus from purely technological aspects to the managerial and organizational impacts of technology. This shift has led to a growing interest in applying qualitative research methods. These methods, originally developed in the social sciences, allow researchers to delve into social and cultural aspects of technology use (Relacion, 2018).

Qualitative research involves collecting data that reflects people's experiences. This can include personal stories, self-reflection, interviews, observations, interactions, and visual materials that hold significance in people's lives (Relacion, 2018). Here are some key qualitative research methods used specifically in ICT research:

1. Action Research: This method focuses on investigating and implementing changes within a specific context. Researchers become directly involved in the situation and learn from the change process itself (Relacion, 2018).

2. Case Study: This method involves a deep dive into a single case, such as a company or project, to understand a phenomenon or contribute to existing theories (Relacion, 2018).

3. Ethnography: This method involves immersing oneself in a specific setting, like a community using a particular technology, to observe and understand their behavior and experiences (Relacion, 2018).

4. Grounded Theory: This method emphasizes analyzing data, often gathered through interviews, to develop new theories or models about a phenomenon (Relacion, 2018).

5. Content Analysis: This method involves systematically examining the content of documents, recordings, or other communication materials to identify patterns or themes (Relacion, 2018).

7 Solution Approach

This chapter presents the proposed solution to the identified problem: the lack of standardization in sentiment analysis tools across diverse languages and cultural contexts. The solution aims to enhance the reliability and effectiveness of these tools in global decision-making processes. The Design Science Research (DSR) methodology is used to develop and evaluate the proposed solution.

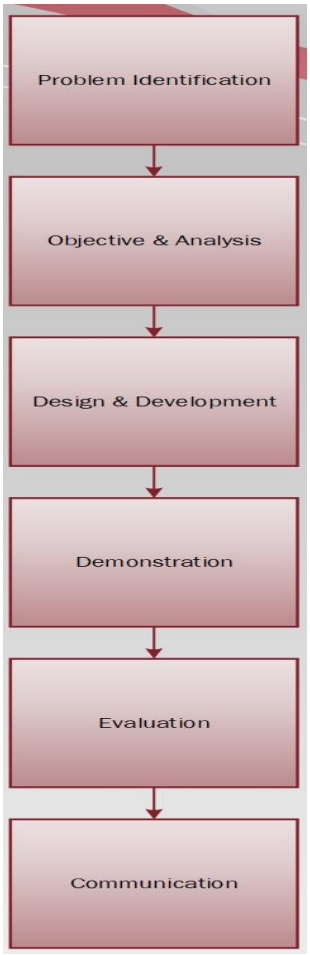


Figure 8: *DSR Process*

7.1 Problem Identification

The problem identified in the thesis revolves around the need to enhance real-time business decision-making through the utilization of big data analytics techniques and platforms. Specifically, the challenge lies in understanding and analyzing how sentiment is handled across different languages and cultures within the context of big data analytics.

7.2 Objective & Analysis

This section aims to determine the goals of the proposed solution based on the problem description. The goals can be both qualitative and quantitative, and they are derived from the problem statement. This section also includes an analysis existing solutions.

7.2.1 Objective

The primary objective of this research is to develop a standardized approach for sentiment analysis that can accurately interpret sentiment across diverse languages and cultural contexts, thus enhancing the reliability and effectiveness of these tools in business decision-making processes. This objective is both qualitative, as it involves the development of a new artifact (the standardized approach), and quantitative, as the proposed solution is expected to outperform existing ones in terms of accuracy and reliability.

7.2.2 Analysis of Existing Solutions

Sentiment analysis is a critical tool in today's data-driven decision-making processes. However, as businesses become increasingly global, the need for sentiment analysis tools that can accurately interpret sentiment across diverse languages and cultural contexts has become apparent. Current sentiment analysis tools often struggle when applied to languages and cultures beyond their primary focus, leading to inaccuracies in interpretation. Some existing solutions attempt to address this gap by using machine translation or bilingual sentiment lexicons, but these approaches often fall short when dealing with the nuances of different languages and cultural contexts.

7.2.2.1 A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports (Miah et al 2024)

As This study addresses the challenge of performing sentiment analysis on foreign languages. The authors identified a significant gap in sentiment analysis for non-English languages due to the lack of sufficient labeled data. To address this gap, they proposed an ensemble model of transformers and a large language model (LLM) that performs sentiment analysis of foreign languages by translating them into a base

language, English. They used four languages - Arabic, Chinese, French, and Italian - and translated them using two neural machine translation models: LibreTranslate and Google Translate. The translated sentences were then analyzed for sentiment using an ensemble of pre-trained sentiment analysis models: Twitter-Roberta-Base-Sentiment-Latest, bert-base-multilingual-uncased-sentiment, and GPT-3, which is an LLM from OpenAI. The experimental results showed that the accuracy of sentiment analysis on translated sentences was over 86% using the proposed model. This indicates that sentiment analysis in foreign languages is possible through translation to English, and the proposed ensemble model works better than the independent pre-trained models and LLM. This study demonstrates improved accuracy and reliability in sentiment analysis compared with individual pre-trained models or LLMs alone.

7.2.2.2 Ensemble Language Models for Multilingual Sentiment Analysis (Hasan, 2024)

This article addresses the gap of lack of standardization in sentiment analysis tools across diverse languages and cultural contexts. The author identifies a significant research gap in understanding human sentiment based on the content shared on social media. This gap is particularly prominent for low-resource languages like Arabic, which have seen less research due to resource limitations.

To address this gap, the author proposed an approach that involved exploring sentiment analysis on tweet texts from SemEval-17 and the Arabic Sentiment Tweet dataset. The author chose to utilize four pretrained language models, namely AraBERTv2, RoBERTa, multilingual BERT, and XLM-RoBERTa. These models have been trained on large amounts of data and are capable of understanding the nuances of language and sentiment. Each of these models underwent an individual fine-tuning process with both English and Arabic datasets. Fine-tuning is a process where a pretrained model is further trained on a new dataset with a similar task. This process helps the model to adapt to the specific characteristics of the new dataset. In this case, it optimized them for the nuanced linguistic characteristics of each language. After the fine-tuning process, the author proposed an ensemble approach. The four models were combined into a language-independent ensemble model. Ensemble models combine the predictions of multiple models to make a final prediction, often leading to improved performance. In this case, the ensemble model was designed to provide a more holistic

and robust sentiment analysis solution that transcends linguistic boundaries. This ensemble model was then trained on the combined dataset. This is expected to provide a sentiment analysis tool that is not only effective across diverse languages but also robust against the cultural contexts embedded in these languages.

Through this detailed approach, the author was able to address the identified gap in sentiment analysis across diverse languages and cultural contexts. However, it's important to note that further research is needed to validate and improve upon these findings. This approach provides a valuable reference for addressing the lack of standardization in sentiment analysis tools across diverse languages and cultural contexts.

The results of this research were promising. The findings showed that monolingual models exhibited superior performance and ensemble models outperformed the baseline. In particular, the majority voting ensemble outperformed the English language. However, the author acknowledges that further research is needed and plans a comprehensive evaluation against State-of-the-Art Deep Learning Models.

7.3 Design & Development

The design phase will focus on creating a framework or methodology that can effectively handle sentiment analysis in various languages and cultural contexts within the domain of big data analytics as shown in **figure 9**.

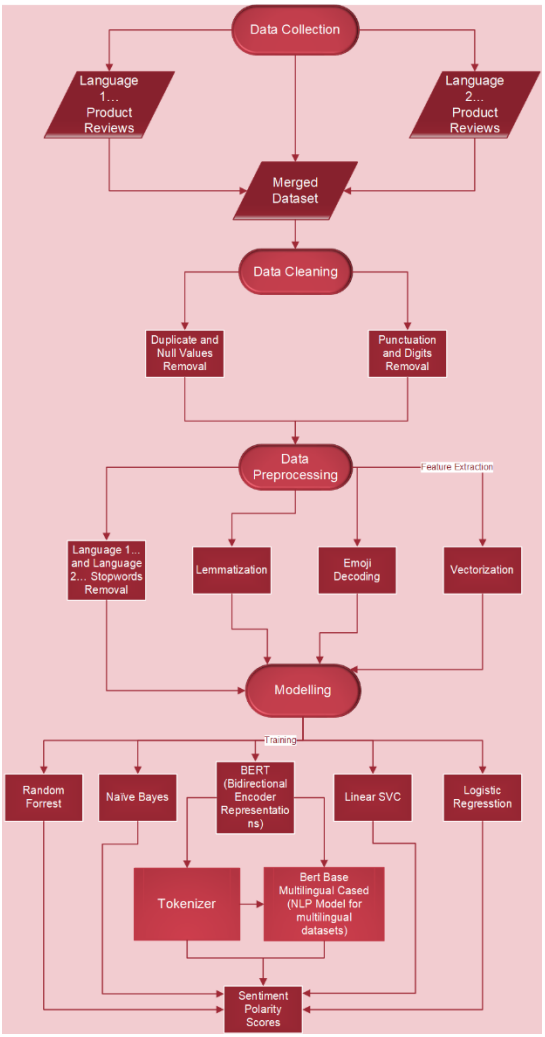


Figure 9: *Multilingual Sentiment Analysis Framework*

The framework's methodology is grounded in best practices for natural language processing and machine learning. It emphasizes the importance of understanding the cultural context within which sentiments are expressed, especially when dealing with languages that have rich and complex linguistic structures like Arabic and Turkish.

The proposed framework integrates a series of methodical steps, each designed to process and analyze sentiment data effectively. The primary components of the framework include:

Data Collection Module: This module is responsible for aggregating a set of reviews in both Arabic and Turkish languages from Kaggle, ensuring a rich dataset that encapsulates a wide range of sentiments and cultural expressions.

Data Preprocessing Module: This module is a critical step in the framework, involving several sub-processes to clean and standardize the data. After initial cleaning and handling of duplicates and null values, the text undergoes language-specific preprocessing. This includes: Stopwords Removal, Lemmatization, and Emoji Handling.

Feature Extraction Module: Utilizing the TfidfVectorizer, this module transforms the preprocessed text into a numerical representation that captures the importance of words within the dataset. The feature extraction process is fine-tuned for each language to ensure that linguistic subtleties are adequately represented.

Model Training Module: A selection of machine learning models, including Random Forest, Multinomial Naive Bayes, Linear SVC, Logistic Regression, and a BERT-based model, are trained on the extracted features. Each model is evaluated for its ability to accurately predict sentiment, with separate instances trained for Arabic and Turkish texts to account for language-specific characteristics.

Evaluation Module: The performance of each model is rigorously assessed using a suite of metrics such as accuracy, precision, recall, and F1 score. The evaluation process is supported by visual tools like confusion matrices and classification reports, providing a detailed analysis of the models' predictive capabilities.

Testing & Prediction Module: New, unseen texts are processed through the framework to predict their sentiment. The **predict_sentiment** function encapsulates the end-to-end process, from text preprocessing to sentiment scoring, demonstrating the framework's applicability in real-world scenarios as shown in **figure 10**.

```python
def predict_sentiment(text, models, vectorizer, id2label):
    # Preprocess the text with returnCleanText method that cleans and process and the text
    clean_text = returnCleanText(text)
    # Check the language of the text and use the appropriate vectorizer
    if is_arabic(clean_text):
      # Vectorize the text
      text_vector = arabic_vectorizer.transform([clean_text]).toarray()
       # Make predictions with each model
      predictions = {}
      for model_name, model in models_ar.items():
          predicted_label_num = model.predict(text_vector)[0]
          predicted_label = id2label[predicted_label_num]
          predictions[model_name] = predicted_label
    elif is_turkish(clean_text):
      text_vector = turkish_vectorizer.transform([clean_text]).toarray()
      predictions = {}
      for model_name, model in models_tr.items():
          predicted_label_num = model.predict(text_vector)[0]
          predicted_label = id2label[predicted_label_num]
          predictions[model_name] = predicted_label
    return predictions
```

Figure 10: *predict_sentiment Function*

7.4 Demonstration

The demonstration phase is where the theoretical framework is put into practice, showcasing its effectiveness in handling sentiment analysis across different languages and cultures within the realm of big data analytics.

The initial phase involves understanding the objectives and requirements from a business perspective. Recognizing the need for a sentiment analysis framework that can handle multiple languages and cultural nuances within big data analytics is crucial for the **real-time business decision making**.

Data Collection

The data collection phase involves collecting initial data and becoming familiar with it. The Arabic dataset, sourced from Kaggle, included a substantial number of product reviews, totaling 40,046, from a variety of companies such as Talabat, Swvl, Telecom_Egypt, and others. These reviews provided a rich source of customer feedback across different services and products.

In contrast, the Turkish dataset was smaller, with 313 reviews, but it was still valuable for understanding customer sentiment in the Turkish context. After combining the Arabic and Turkish reviews into one dataset, visualizations were created to analyze

the sentiments expressed in these reviews.

Three pie charts were developed: one representing the sentiment distribution of the entire dataset, another breaking down the sentiments of the Arabic reviews, and a third for the Turkish reviews. Additionally, a bar plot was created to provide a comparative view of the sentiments across the entire dataset. These visualizations, illustrated in **figures 11 through 14**, helped in gaining a clearer understanding of the overall sentiment trends within the reviews.

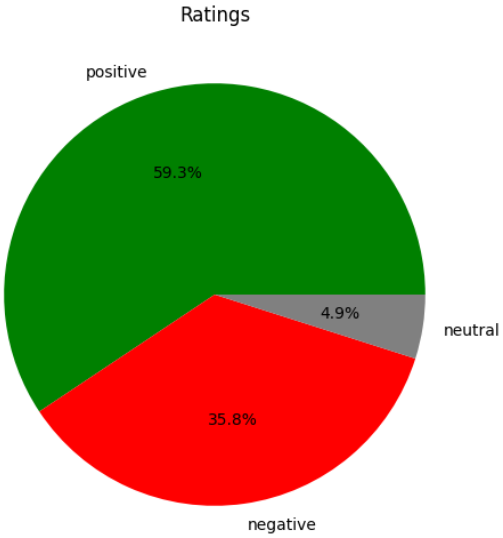


Figure 11: *Merged Reviews' Sentiment*

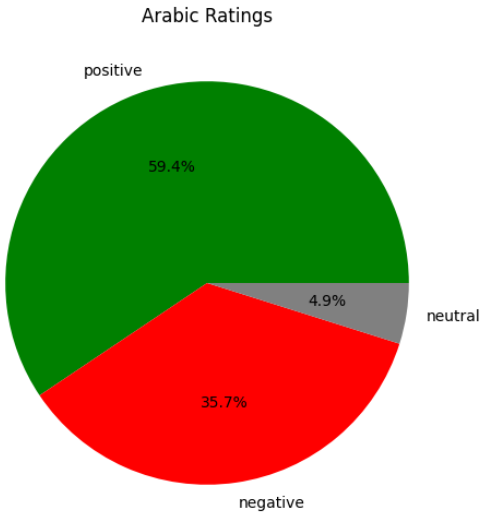


Figure 12: *Arabic Reviews' Sentiment Percentage*

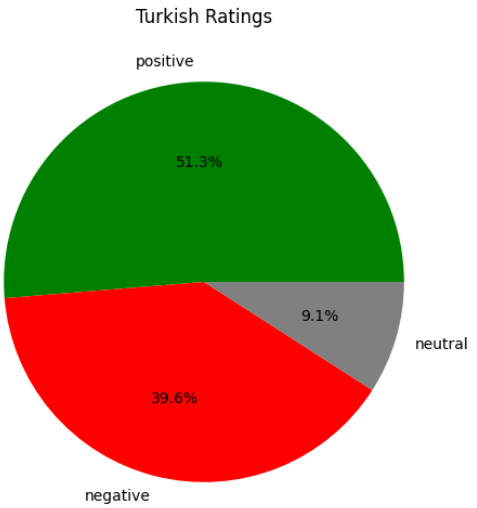


Figure 13: *Turkish Reviews' Sentiment Percentage*

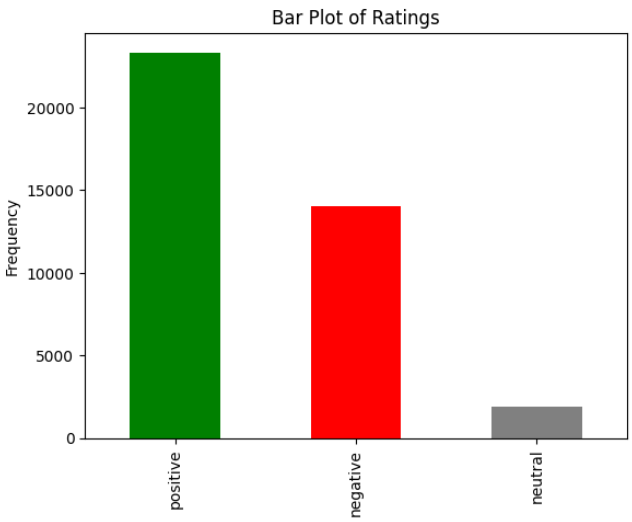


Figure 14: *Merged Reviews' Sentiment*

Data Preprocessing

Data preparation is about constructing the final dataset from the initial raw data. It is an essential phase in the sentiment analysis framework, ensuring that the input data is of the highest quality before it is fed into the machine learning models. This phase encompasses a series of fundamental steps:

Digit Removal: Digits are removed from the text as they generally do not contribute to sentiment analysis. This step helps to focus the analysis on textual content that is more likely to carry sentiment.

Punctuation Removal: Punctuation marks are stripped away, except for those that serve as sentiment indicators, such as exclamation marks. This standardizes the text and avoids unnecessary complexity in the feature space.

Null Value Treatment: Null values in the dataset are addressed to prevent any errors during the modeling process. Rows with null values, especially in the review description column, are removed to maintain data integrity.

Duplicate Removal: Duplicate entries in the dataset are identified and eliminated. This ensures that the models are trained on unique data points, which helps to prevent bias and overfitting.

Stopwords Removal: Common words that do not carry significant meaning, known as stopwords, are removed from the text. This includes language-specific stopwords for both Arabic and Turkish, streamlining the dataset to focus on words that are more likely to influence sentiment.

Emoji Transformation: Emojis are transformed into corresponding text descriptions using custom functions shown in **figure 15**. This innovative step captures the sentiment expressed through emojis, which is particularly prevalent in user-generated content. The first function **emoticons_to_emoji** transforms emoticons into emojis and the next function **checkemojie** checks if there is any emojis are in the input text and append it to a list and returns the list. The last function **emojiTextTransform** calls **emoticons_to_emoji** to transform any emoticons to emojis then calls **emoticons_to_emoji** to transform the text and make it ready for the modelling.

```python
from langdetect import detect, LangDetectException
import emoji

def emoticons_to_emoji(text):
    for emoticon, emoji in emotions_to_emoji.items():
        text = text.replace(emoticon, emoji)
    return text

def checkemojie(text, lang):
    emojistext=[]
    for char in text:
        if any(emoji.distinct_emoji_list(char)):
            if lang == 'ar':
                if char in arabic_emojis.keys():
                    emojistext.append(arabic_emojis[emoji.distinct_emoji_list(char)[0]])
            else:
                if char in turkish_emojis.keys():
                    emojistext.append(turkish_emojis[emoji.distinct_emoji_list(char)[0]])
    return " ".join(emojistext)

def emojiTextTransform(text):
    text = emoticons_to_emoji(text)  # Transform emoticons to emojis
    cleantext = re.sub(r'[^\w\s]', '', text)
    try:
        lang = detect(cleantext)  # Detect the language of the text
    except LangDetectException:
        lang = 'unknown'
    return cleantext + " " + checkemojie(text, lang)
```

Figure 15: *Emoji Transformation Functions*

Lemmatization: Words are lemmatized to their base form, consolidating different inflected forms of a word. This process is crucial for languages like Arabic and Turkish, which have rich morphological variations.

Feature Extraction with TF-IDF Vectorizer: The TF-IDFVectorizer is employed to convert the preprocessed text into a structured numerical format. This feature extraction technique evaluates the importance of words within the dataset and across the corpus, resulting in a feature set that is primed for the machine learning models.

Since the number of rows of Arabic language is much higher than the Turkish language, I vectorized the Turkish and Arabic reviews separately. I extracted all words of the Turkish language and the most 10,000 frequent words of the Arabic reviews to ensure that the models have a balanced and representative set of features to learn from. as shown in **figure 16**.

```python
# Vectorize Turkish texts
turkish_vectorizer = TfidfVectorizer()
turkish_features = turkish_vectorizer.fit_transform(turkish_texts)

# Vectorize Arabic texts
arabic_vectorizer = TfidfVectorizer(max_features=10000)
arabic_features = arabic_vectorizer.fit_transform(arabic_texts)
```

Figure 16: *TF-IDF Vectorizer*

Modelling

In the modelling phase, various modeling techniques are selected and applied, adjusting their parameters for optimal performance. Models like Random Forest, Naive Bayes, Linear SVC, Logistic Regression, and BERT are trained, each tailored to the linguistic characteristics of Arabic and Turkish.

The modeling module of the sentiment analysis framework is a critical component where various machine learning algorithms are employed to understand and predict sentiments from text data. It is where the sentiment analysis framework's core analytical processes occur. This module leverages the features extracted through the TF-IDF Vectorizer to train models that can determine the underlying sentiment in product reviews written in Arabic and Turkish. Here's a detailed explanation of the modeling module:

Training and Test Sets: The feature sets for Arabic and Turkish are split into training and test sets, providing a basis for model evaluation and validation. The **train_test_split** function ensures that the models are tested on unseen data to assess their generalization capabilities.

Separate Model Training: For both the Arabic and Turkish datasets, separate models are trained to account for the unique linguistic characteristics of each language. This approach ensures that the models are specialized and optimized for their respective languages. The models trained include:

Random Forest: An ensemble method that builds multiple decision trees to improve prediction accuracy.

Naive Bayes: A probabilistic classifier that's efficient for high-dimensional text data.

Linear SVC: A support vector machine with a linear kernel, suitable for large feature spaces.

Logistic Regression: A model that estimates probabilities and is widely used for classification tasks.

BERT: Unlike the other models, the BERT model is trained on a combined dataset of 3,000 Arabic reviews and all Turkish reviews. This deep learning model benefits from a larger, more diverse training set, which helps it capture the context and nuances of both languages more effectively.

Model Evaluation: Each model's performance is evaluated using metrics such as accuracy, precision, recall, and F1 score. The evaluation process includes confusion matrices and classification reports, offering insights into the predictive strengths and weaknesses of each model that will be shown in the next subchapter which is Evaluation.

Practical Application: The trained models are applied to new, unseen text data to predict sentiment. The **predict_sentiment** function, that is shown in **figure 10**, demonstrates the framework's end-to-end prediction capabilities, from preprocessing to sentiment scoring.

7.5 Evaluation

The evaluation phase will involve assessing the success and impact of the developed framework in enhancing real-time business decision-making through standardized sentiment analysis across diverse linguistic and cultural backgrounds. It evaluates the performance of sentiment analysis models developed for both Arabic and Turkish languages. It presents a comprehensive analysis of each model's accuracy, precision, recall, and F1 score, along with insights drawn from confusion matrices.

7.5.1 Model Performance

Random Forest: Achieved testing accuracy, precision, recall, and F1 score of 81.89% for Arabic and 80.65% for Turkish, showing strong performance in identifying negative and positive sentiments. However, it struggled with neutral sentiments, often confusing them with Positive.

Naive Bayes: Exhibited testing accuracy, precision, recall, and F1 score of 82.37% for Arabic and 82.26% for Turkish, with similar challenges in classifying neutral sentiments. The model's precision and recall for neutral sentiments were notably low, indicating an area for improvement.

Linear SVC: Recorded testing accuracy, precision, recall, and F1 score of 81.96% for Arabic and 83.87% for Turkish, with high precision and recall for negative and positive classes but lower scores for neutral sentiments.

Logistic Regression: Attained testing accuracy, precision, recall, and F1 score of 83.01% for Arabic and 79.03% for Turkish, showing a good balance of precision and recall for negative and positive sentiments but a significant drop in recall for neutral sentiments.

BERT: The deep learning model showed an improvement over epochs, achieving a validation accuracy of 78.70%. However, the increasing validation loss suggests potential overfitting, which could be mitigated with further tuning.

7.5.2 Confusion Matrices Analysis

The confusion matrices provided deeper insights into each model's classification capabilities. Firstly, Models generally performed well in classifying negative and positive sentiments for both languages. Next, Neutral sentiments posed a challenge across all models, with many neutral instances being misclassified as negative or positive. Finally, the confusion matrices highlighted the need for more balanced datasets or improved modeling techniques to enhance neutral sentiment classification.

7.5.3 Testing and Prediction Analysis

The evaluation of the sentiment analysis models was conducted using new, unseen data to test their predictive capabilities. The models were assessed based on their ability to accurately classify the sentiment of text samples in both Arabic and Turkish.

Arabic Model Predictions

The Arabic text samples were classified by the models as follows:

- **Negative Sentiments:** All models consistently predicted negative sentiments for text samples that expressed dissatisfaction or negative experiences.
- **Positive Sentiments:** Similarly, all models correctly identified positive sentiments in text samples that conveyed satisfaction or positive

feedback.

- **Neutral Sentiments:** The models faced challenges in predicting neutral sentiments, often misclassifying them as positive.

Turkish Model Predictions

The Turkish text samples yielded the following results:

- **Negative Sentiments:** The models, except for Linear SVC, struggled to predict negative sentiments accurately, often classifying them as positive.
- **Positive Sentiments:** All models successfully predicted positive sentiments in text samples that depicted favorable opinions or experiences.
- **Neutral Sentiments:** As with the Arabic models, the Turkish models were unable to accurately predict neutral sentiments, tending to classify them as positive.

BERT Model Predictions

The BERT model was evaluated separately, showing an improvement in validation accuracy over training epochs. However, the increasing validation loss indicated a need for further tuning to prevent overfitting and improve generalization.

BERT model prediction was more accurate than the other four models, especially in predicting the sentiment of Turkish reviews. It predicted the positive and negative sentiments correctly in both languages. Unfortunately, it was unable to accurately predict neutral sentiments, they were always classified as positive.

8 Conclusion

In my paper, I've explored the dynamic field of sentiment analysis within big data analytics, focusing on the nuances of language and culture. I developed a unique framework to standardize sentiment analysis across different languages, aiming to improve global real-time business decision-making. While I've made significant developments, challenges remain in fully capturing the different ways people express themselves across cultures.

The pursuit of a standardized approach to sentiment analysis in big data analytics is an ongoing endeavor. The findings of this paper lay the groundwork for future research and practical applications that can leverage the power of sentiment analysis to drive global business success and real-time decision making. As the world becomes increasingly connected, the ability to understand and analyze sentiment across linguistic and cultural barriers will become ever more critical.

8.1 *Limitations*

Although the research has been enlightening, there have been challenges along the way. One notable limitation is the models I've developed, while effective for Arabic and Turkish, may require further refinement to be as effective for other languages. Furthermore, the challenge of accurately predicting neutral sentiments points to a potential imbalance in the training data, which may have skewed the models' learning towards more expressive positive and negative sentiments. Moreover, the use of advanced models like BERT requires significant computational resources, which may not be readily available to all organizations.

8.2 *Future Work*

Looking ahead, there are several paths I plan to pursue to enhance the framework I've developed. I aim to refine NLP techniques to better capture cultural subtleties and expand the datasets to include a broader spectrum of sentiments. Optimizing the models to prevent overfitting and improve their generalization across various datasets is another critical step. Finally, I intend to explore ways to make these advanced sentiment analysis tools more accessible, ensuring that businesses of all sizes can harness the power of sentiment analysis for their global operations.

References

Ahmed, R., Shaheen, S., & Philbin, S. P. (2022). The role of big data analytics and decision-making in achieving project success. Journal of Engineering and Technology Management, 65, 101697.

Al-Barznji, K., & Atanassov, A. (2018, May). Big data sentiment analysis using machine learning algorithms. In Proceedings of 26th International Symposium" Control of Energy, Industrial and Ecological Systems, Bankia, Bulgaria, 53-58.

Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and Hadoop technology. Procedia computer science, 48, 319-324.

Ayokanmbi, F. M. (2021). The impact of big data analytics on decision-making. International Journal of Management IT and Engineering, 11(4), 1-5.

Babu, M. P., & Sastry, S. H. (2014, June). Big data and predictive analytics in ERP systems for automating decision making process. In 2014 IEEE 5th international conference on software engineering and service science (pp. 259-262). IEEE.

Capurro, R., Fiorentino, R., Garzella, S., & Giudici, A. (2021). Big data analytics in innovation processes: which forms of dynamic capabilities should be developed and how to embrace digitization?. European Journal of Innovation Management, 25(6), 273-294.

Czibula, G., Ciubotariu, G., Maier, M. I., & Lisei, H. (2022). IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining. IEEE Access, 10, 80651-80666.

De Oliveira Júnior, G. A., de Oliveira Albuquerque, R., Borges de Andrade, C. A., de Sousa Jr, R. T., Sandoval Orozco, A. L., & García Villalba, L. J. (2020). Anonymous real-time analytics monitoring solution for decision making supported by sentiment analysis. Sensors, 20(16), 4557.

Del Vecchio, P., Mele, G., Passiante, G., Vrontis, D., & Fanuli, C. (2020). Detecting customers knowledge from social media big data: toward an integrated

methodological framework based on netnography and business analytics. Journal of Knowledge Management, 24(4), 799-821.

Elkmash, M. R. M., Abdel-Kader, M. G., & El Din, B. B. (2021). An experimental investigation of the impact of using big data analytics on customers' performance measurement. *Accounting Research Journal*, 35(1), 37-54.

Ghavami, P. (2019). Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing. Walter de Gruyter GmbH & Co KG.

Goar, V. K., & Yadav, N. S. (2022). Business decision making by big data analytics. International Journal on Recent and Innovation Trends in Computing and Communication, 10(5), 22-35.

Hasan, M. A. (2024). Ensemble Language Models for Multilingual Sentiment Analysis. arXiv preprint arXiv:2403.06060.

He, W., Zhang, W., Tian, X., Tao, R., & Akula, V. (2018). Identifying customer knowledge on social media through data analytics. Journal of Enterprise Information Management, 32(1), 152-169.

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2019). Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. Sustainability, 11(15), 4235.

Kudyba, S. (2014). Big data, mining, and analytics: components of strategic decision making. CRC Press.

Kumar, D. T. S. (2020). Data mining based marketing decision support system using hybrid machine learning algorithm. Journal of Artificial Intelligence and Capsule Networks, 2(3), 185-193.

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. Ieee Access, 5, 7776-7797.

Ledro, C., Nosella, A., & Vinelli, A. (2022). Artificial intelligence in customer relationship management: literature review and future research directions.

Journal of Business & Industrial Marketing, 37(13), 48-63.

Lee, M., Kwon, W., & Back, K. J. (2021). Artificial intelligence for hospitality big data analytics: developing a prediction model of restaurant review helpfulness for customer decision-making. International Journal of Contemporary Hospitality Management, 33(6), 2117-2136.

Mehboob, T., Ahmed, I. A., & Afzal, A. (2022). Big Data Issues, Challenges and Techniques: A Survey. Pakistan Journal of Engineering and Technology, 5(2), 216-220.

Miah, M. S. U., Kabir, M. M., BinSarwar, T., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports, 14, 9603.

Mohamed, N., & Al-Jaroodi, J. (2014, July). Real-time big data analytics: Applications and challenges. In 2014 international conference on high performance computing & simulation (HPCS) (pp. 305-310). IEEE.

Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics, 5(2), 81-97.

Ortiz, G., Caravaca, J. A., García-de-Prado, A., & Boubeta-Puig, J. (2019). Real-time context-aware microservice architecture for predictive analytics and smart decision-making. IEEE Access, 7, 183177-183194.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. International Journal of Information Management, 38(1), 187-195.

Relacion, P. J. (2018). Qualitative Research Methods: Definition of Qualitative Research.

Sharma, R., Agarwal, P., & Arya, A. (2022). Natural language processing and big data: a strapping combination. In New Trends and Applications in Internet of Things (IoT) and Big Data Analytics (pp. 255-271). Cham: Springer International Publishing.

Soubra, L. (2021). BIG DATA, CUSTOMER CENTRICITY AND SUSTAINABILITY IN THE BANKING INDUSTRY. *BAU Journal-Creative Sustainable Development*, *3*(1), 11.

Torres-Carrión, P. V., González-González, C. S., Aciar, S., & Rodríguez-Morales, G. (2018, April). Methodology for systematic literature review applied to engineering and education. In 2018 IEEE Global engineering education conference (EDUCON) (pp. 1364-1373). IEEE.

Van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. Information and Software Technology, 136, 106589.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: Applications, prospects and challenges. Mobile big data: A roadmap from models to technologies, 3-20.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).

Wright, L. T., Robin, R., Stone, M., & Aravopoulou, D. E. (2019). Adoption of big data technology for innovation in B2B marketing. Journal of Business-to-Business Marketing, 26(3-4), 281-293.

Yu, J. H., & Zhou, Z. M. (2019). Components and development in Big Data system: A survey. Journal of Electronic Science and Technology, 17(1), 51-72.

Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., & Yan, X. (2018). Measuring customer agility from online reviews using big data text analytics. *Journal of Management Information Systems*, *35*(2), 510-539.

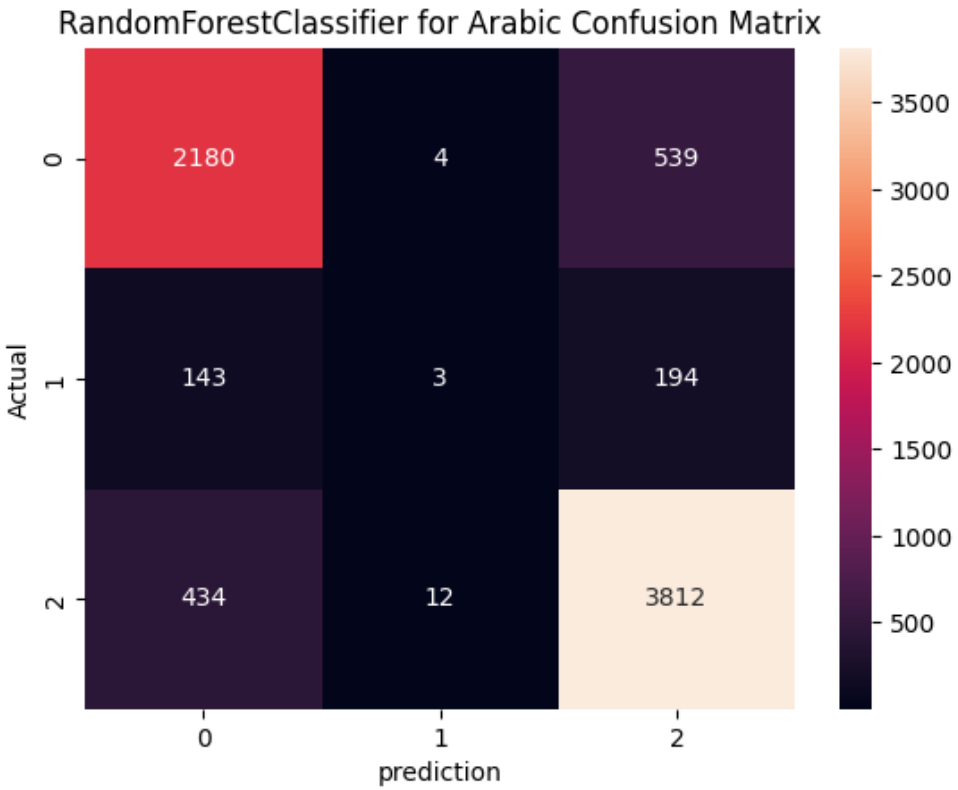
Appendix

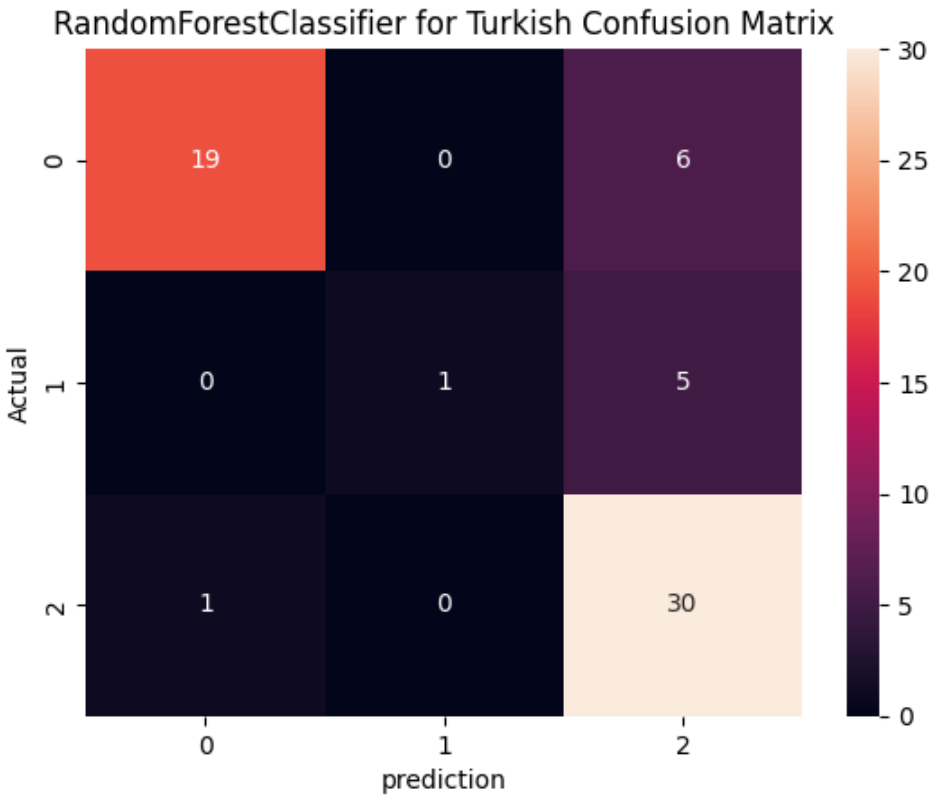
Datasets:

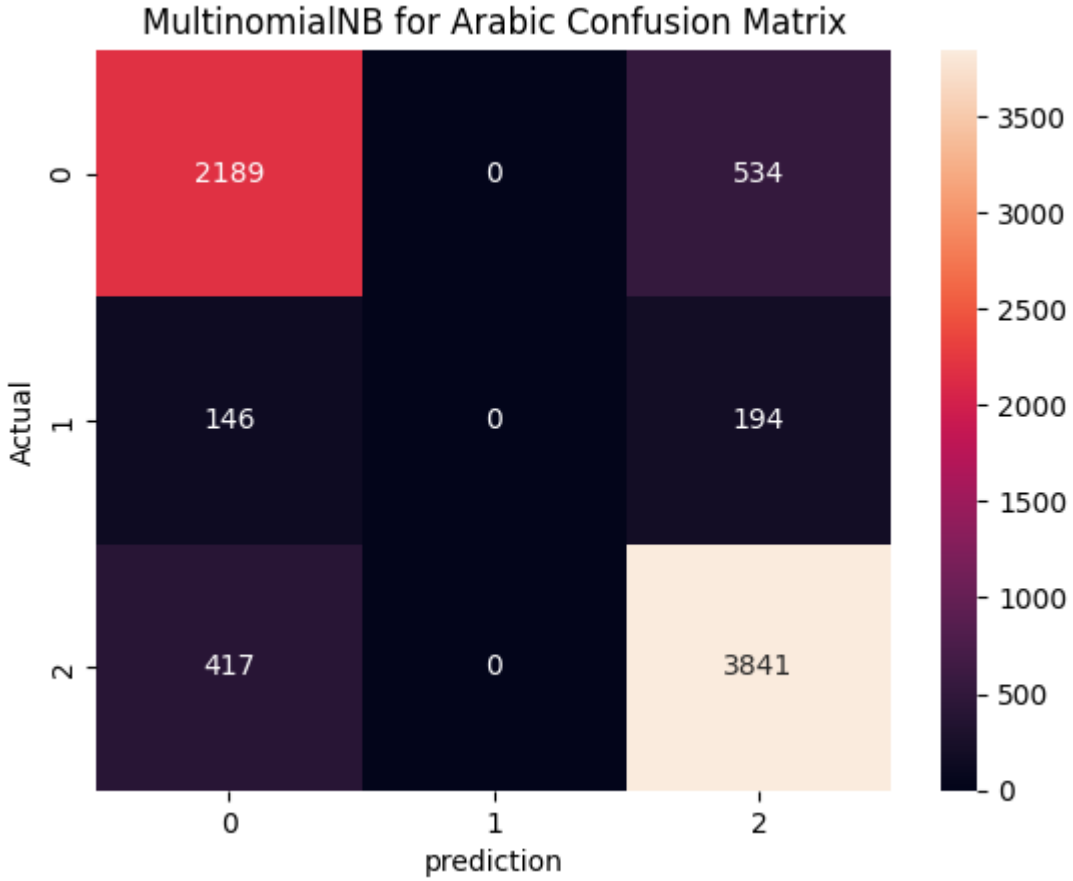
https://www.kaggle.com/datasets/fahdseddik/arabic-company-reviews

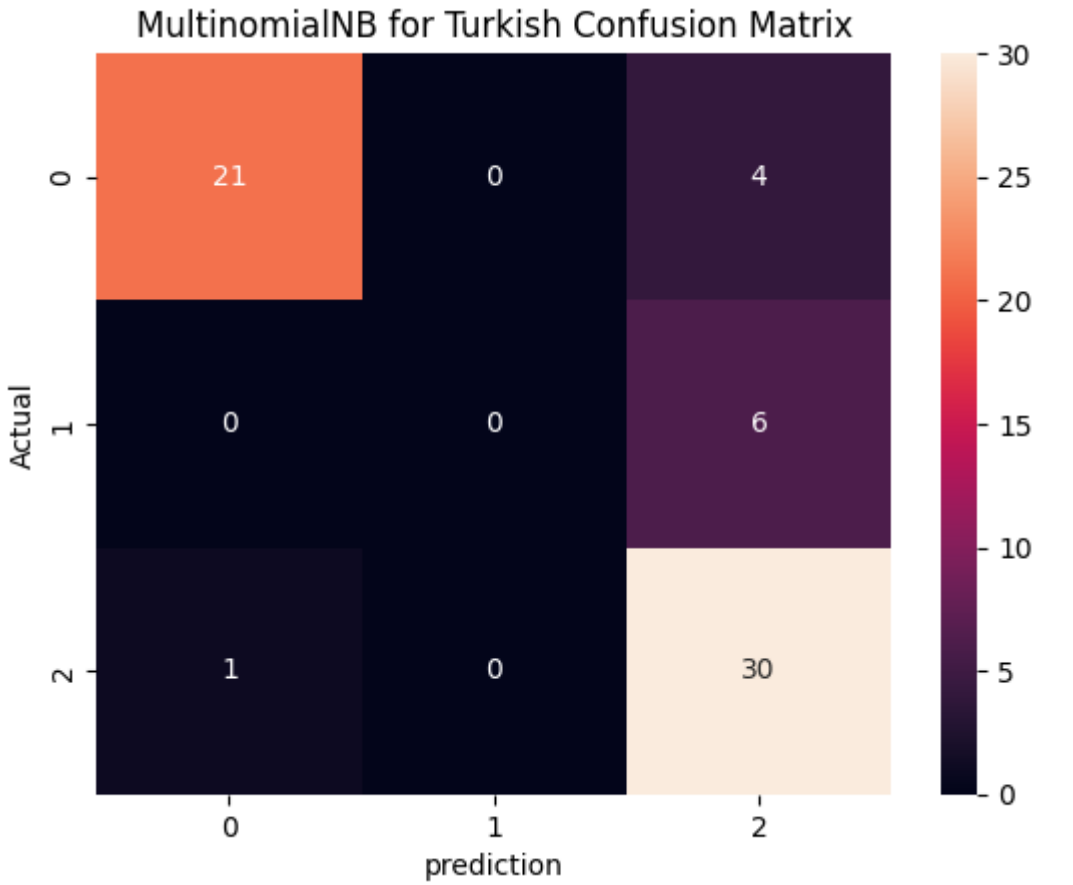
https://www.kaggle.com/datasets/selahattincanler/turkish-product-comment-sentiment-analysis

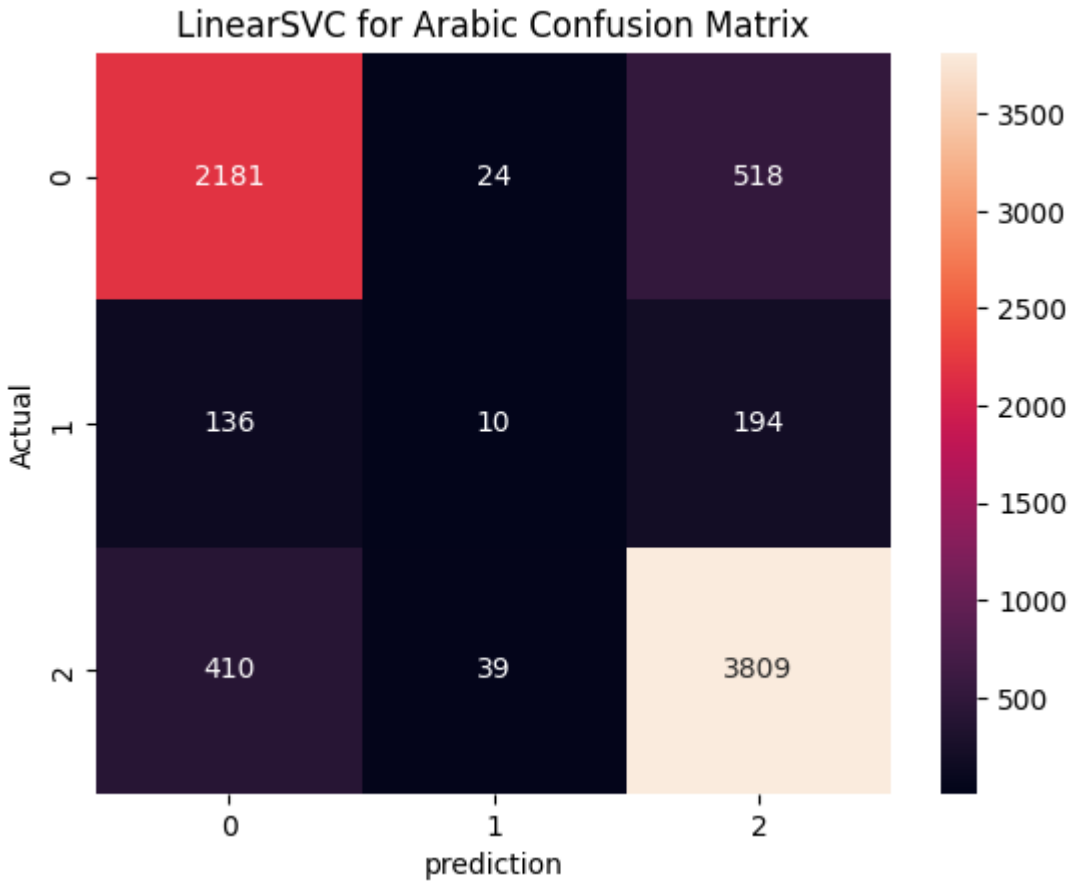
Confusion Matrices:

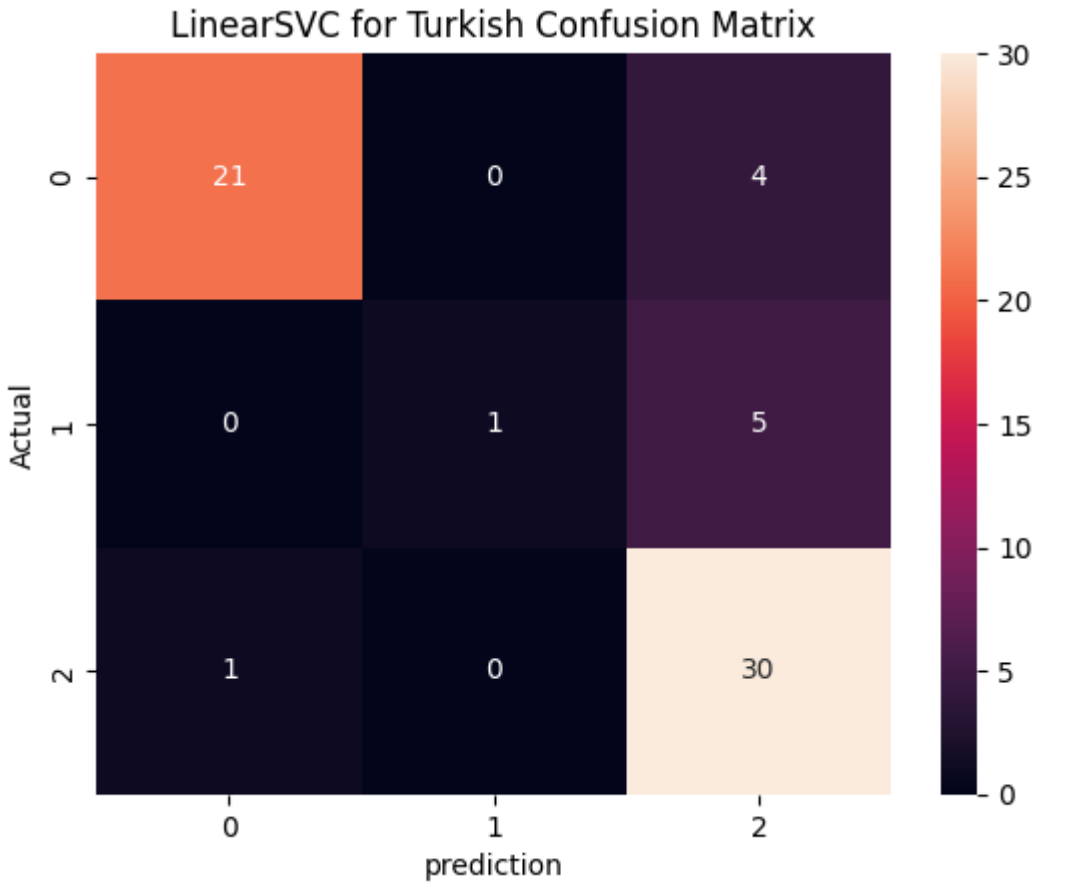


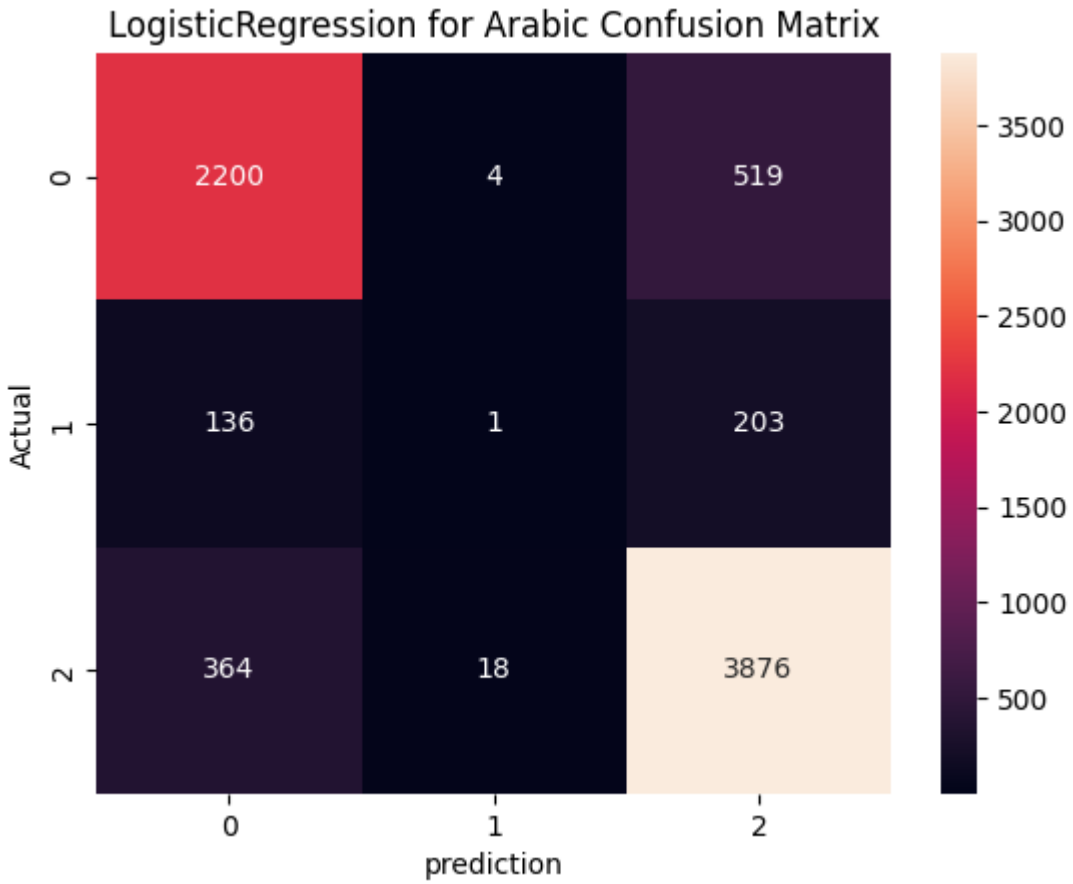


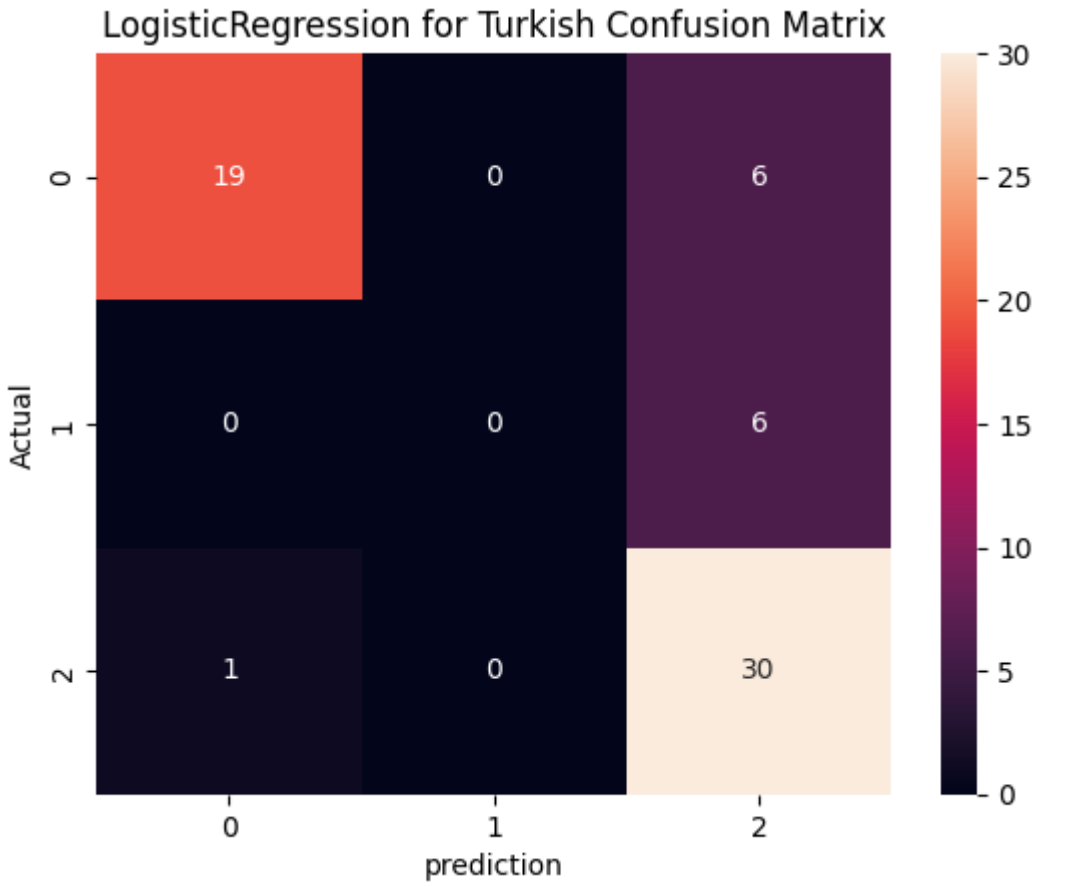

MultinomialNB for Arabic Confusion Matrix
2189
0
534
146
0
194
417
0
3841
Actual
prediction
0
1
2
3500
3000
2500
2000
1500
1000
500
0


MultinomialNB for Turkish Confusion Matrix
21
0
4
0
0
6
1
0
30
Actual
prediction
0
1
2
30
25
20
15
10
5
0


LinearSVC for Arabic Confusion Matrix
2181
24
518
136
10
194
410
39
3809
Actual
prediction
0
1
2
3500
3000
2500
2000
1500
1000
500


LinearSVC for Turkish Confusion Matrix
21
0
4
0
1
5
1
0
30
Actual
prediction
0
1
2
30
25
20
15
10
5
0


LogisticRegression for Arabic Confusion Matrix
2200
4
519
136
1
203
364
18
3876
Actual
prediction
0
1
2
3500
3000
2500
2000
1500
1000
500


LogisticRegression for Turkish Confusion Matrix
19
0
6
0
0
6
1
0
30
Actual
prediction
0
1
2
30
25
20
15
10
5
0