REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

ADVANCING ELECTRONIC COMMERCE USING DATA MINING BENEFITS

Mohamed Amro HELAL

Master's Thesis

Supervisor

Asst.Prof. Dr. Ayça Kurnaz TÜRKBEN

Istanbul, 2023

ADVANCING ELECTRONIC COMMERCE USING DATA MINING BENEFITS

Mohamed Amro HELAL

Infromation Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled ADVANCING ELECTRONIC COMMERCE USING DATA MINING BENEFITS prepared by MOHAMED AMRO HELAL and submitted on 28/04/2023 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

____

Asst. Prof. Dr. Ayça Kurnaz TÜRKBEN

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Ayça Kurnaz TÜRKBEN	Department of Computer Engineering, Altınbaş University	_____________
Asst. Prof. Dr. Abdullahi Abdu İBRAHİM	Department of Software Engineering, Altınbaş University	_____________
Asst. Prof. Dr. Zeynep ALTAN	Department of Software Engineering, Beykent University	_____________

I hereby declare that this thesis meets all format and submission requirements of a Master`s Thesis.

Submission date of the thesis to the Graduate Education Institute: ____/____/____

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Mohamed Amro HELAL

Signature

DEDICATION

I would be negligent if I did not thank my parents and instructors for their unwavering love, understanding, inspiration, and encouragement throughout my pursuit of this master's degree. I have a great deal of gratitude for them. My supervisors, parents, and teachers have had the most significant impact on my life, and I will be eternally grateful to them.

ABSTRACT

ADVANCING ELECTRONIC COMMERCE USING DATA MINING BENEFITS

HELAL, Mohamed Amro

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Ayça Kurnaz TÜRKBEN

Date: April / 2023

Pages: 60

Although data mining is now being used in a wide variety of contexts, it has traditionally been put to use in the evaluation of big data sets. Recently, several data mining strategies have been suggested and deployed in the more constrained context of online retail. The e-commerce sector makes use of data mining to examine many kinds of data, such as purchases made by customers, the information logged from websites, and even social media activity. Understanding customer behavior, spotting trends, and fine-tuning advertising initiatives all rely on this data. This paper presents the three primary algorithms used in data mining (D.M.) for online business: association, clustering, and prediction. It highlights several advantages of D.M. to e-commerce businesses, including data pre-treatment, pattern mining for sales, and market pattern analysis, all of which may be accomplished with the help of the three data mining algorithms. Moreover, it also investigates the three data mining algorithms that may aid e-commerce firms with tasks like product planning, sales forecasting, basket analysis, CRM, and market segmentation. This research primarily aims to categorize a product into the four categories of Electronics, Household, Books, and Clothing & Accessories and check the accuracy of this classification.

Keywords: Data Mining, Big Data, E-Commerce, Organization, Data Warehouses.

TABLE OF CONTENTS

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

ANN	:	Artificial Neural Network
ML	:	Machine Learning
NLP	:	Natural Language Processing
OLAP	:	Online Analytical Processing
TFIDF	:	Term Frequency-Inverse Document Frequency

LIST OF SYMBOLS

%	:	Percentage
&	:	And
∑	:	Summation
f	:	Frequency

1. INTRODUCTION

1.1 OVERVIEW

Nowadays, a large volume of organized and unstructured data, referred to as big data, creates the potential for businesses, particularly those that employ internet commerce (e-commerce). Big data describes the massive amounts of information created by modern enterprises and organizations. Big data in e-commerce can be gathered from many different places, including purchases made by customers, data gleaned from the website's analytics, social media activity, and even macroeconomic indicators [1]. E-commerce companies can benefit greatly from big data analysis since it can reveal patterns in consumer buying habits, tastes, and interests. A company can learn what items are selling well, for instance, by looking at client purchase records, which can then inform advertising and stock decisions. Data warehouses, Data mining business intelligence software, and machine learning algorithms are just a few of the various technologies and tools available for handling and analyzing large data in e-commerce. The use of such instruments can aid companies in gaining useful insights and making data-driven decisions that will ultimately benefit their operations and help them expand. Data mining is the practice of applying machine learning, statistics, and database system methods to uncover trends and patterns in massive datasets. Common uses for this tool include learning from and making sense of massive data sets [2].

As a standard practice, data is cleaned and preprocessed before data mining to get rid of any anomalies or outliers. The data is then subjected to a variety of algorithms and analysis methods in order to extract useful insights. By revealing hidden connections and patterns in a company's data, it can aid in the formation of more well-considered judgments.

Data mining is an approach used in the e-commerce industry to generate new concepts or integrated technologies that improve decision-making by combining statistics, databases, and A.I. with other fields. Data mining is seen as a useful tool for promoting an online business. The use of data mining in online business is a hot topic right now [3]. In the context of cloud computing, data mining refers to the process of extracting structured data from semi- or unstructured web data sources. The main idea behind cloud computing from a business perspective is to sell access to shared computing resources on demand [4].

However, the cost of maintaining such services is an issue for virtually all online retailers. Data mining costs vary widely based on various variables, including the shape and data set complexity, the technology used, tool complexity, and the number of human resources and other assets needed to complete the data mining task [5].

As an on-demand service model, cloud computing allows for the rapid deployment of resources., Networking, software, storage analytics, and intelligence are just some of the computer services that can be accessed via the cloud and sold to customers.

Service delivery methods are frequently mentioned when talking about cloud computing. There are three main approaches to delivering cloud computing services: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) [6].

The whole application or program is supplied via the browser or an application programming interface over the internet (API). Consumers only have to worry about adding users to the system when using the service model. Storage capabilities are one of the essential applications of cloud computing. In place of storing data on a user's traditional device or in an on-premises storage system, cloud storage allows users to do so remotely over the internet.

Most e-commerce organizations are enthusiastic about the concept since it reduces the expensive expense of storing large data sets of an organization's operations in remote data centers. The platform also supports free or low-cost B2B and B2C e-commerce commercial apps. Aside from saving money, speed, scalability, and safety, cloud computing is great for online stores since it allows for more efficient operations and more customers to use the same server[7]. The connection between data mining and cloud computing is the cloud, which saves information on the cloud. In contrast, the mining of data used to provide customer-cloud relationships is called a service, with the collection of information depending on legal considerations such as individuality and privacy [5].

1.2 SIGNIFICANCE OF DATA MINING TO BOOST E-COMMERCE

Implementation of data mining with both structured and unstructured data, such as that found in a database and text documents, and social media posts, respectively. The term "e-commerce" is used to describe commercial transactions conducted entirely online. It has changed the dynamic between companies and customers by facilitating the export and purchase of services and goods around the mighty clock. Among the wide varieties of e-commerce are "business to consumer" (B2C) transactions, in which firms sell directly to end users, and "consumer to consumer" (C2C) transactions, in which users sell and buy from one another. In business-to-business (B2B) e-commerce, companies sell their goods and services to other companies, while in consumer-to-business (C2B) e-commerce, individuals sell their goods and services to corporations. The benefits of electronic commerce to both merchants and customers are numerous. It has the potential to outperform the more conventional method of shopping in a store by several measures, including speed, cost, and ease. It paves the way for companies to sell their wares to consumers all around the world. The requirement for safe payment methods and the possibility of fraud are two issues plaguing the e-commerce industry [8].

In certain cases, the end result of data mining will be a prediction model, a mathematical representation of the underlying relationships and patterns in the data. A decision or prediction can be made using this model and subsequent data. A data mining technique may, for instance, be used to construct a prediction model for spotting credit card fraud, which could then be implemented for automatically flagging and reviewing suspect transactions in real-time. Data visualizations, dashboards, and decision support systems are all potential end products of data mining that can help businesses act on the insights and knowledge gained from the process. Data mining's final product allows making the decision to find their customers' purchase history, trendy demand, and geography, thereby facilitating improved strategic decision-making for the company's benefit. As a result, one may spend less on inventory and related costs, allowing the industry to maximize its total gain. Due to the internet's ubiquitous availability, modern businesses rely heavily on online resources and technologies. Therefore, many businesses increasingly engage in e-commerce, calling for the creation of e-commerce applications by a specialist tasked with managing and providing support for the services. When businesses flourish, firm resources might be insufficient to sustain current levels of e-commerce. In this sense, data mining can be used to handle corporate e-commerce facilities and unveil patterns for online clients, ensuring the business's frequent sales and utter productivity [9]. There are several costs connected with data mining, but some of the most significant are:

a. Hardware and Software: Data mining often necessitates the use of high-powered computers and specialized software in order to handle and analyze massive datasets. The data mining project's requirements will determine how much money will be needed.

b. Data Storage: Having enough space to store the information that mining data can produce is very much essential. The utmost necessity to keep data over extended periods of time can result in this being a continuing expense for businesses.

c. Staff Time and Expertise: Data mining may be a time-consuming and complex operation, necessitating a dedicated team with specific knowledge and tools. There are a few options here: either hire/train/outsource someone to do the data mining or do it yourself.

d. Cloud Computing Costs: Data mining performed in the cloud may incur continuous charges for the use of cloud computing resources, which are normally charged on a per-user basis.

The final cost of data mining will be determined by the organization's unique requirements and objectives. Before committing resources to data mining, it's crucial to weigh these expenses against the process's prospective returns. In those associated costs, cloud computing is so much more trivial. Cloud computing thereby revolutionizes company operations by giving companies access to an extensive range of highly flexible and scalable services in a virtual environment hosted in the cloud and accessible over the internet. The advent of cloud computing completely alters the business landscape since a network storage-based service paradigm is provided, new data resource allocation, instantaneous access to data, and mechanisms for processing. Cloud computing allows for the distribution of data mining applications, which enables e-commerce enterprises to consolidate their administrative tasks and storage of data while providing an absolute guarantee of dependability, creativity, and the best security to its consumers, lowering costs and increasing profits [7]

1.3 PROBLEM STATEMENT

a. Primary motivation for this research was to categorize a product into the four categories of Electronics, Household, Books, and Clothing & Accessories and check the accuracy of this classification.

b. In this research, the authors have introduce the three most important algorithms used in data mining (D.M.) for online stores: association, clustering, and prediction.

c. It also investigated how the three data mining algorithms can support e-commerce businesses by providing merchandise techniques, sales prediction, analysis of shopping cart contents, CRM, and niche market breakdown.

1.4 RESEARCH OBJECTIVES

To that end, the following research aims were developed:

a. Use two different models of TF-IDF and Word2Vec to analyze and classify the dataset. TF-IDF and Word2Vec were chosen because both are effective, easy to understand, language-independent, handle synonyms, and handle polysemy. Both are commonly used for tasks like document classification, search engine optimization, and summarisation, but their simplicity belies their potency as tools for information retrieval and text mining.

b. Investigate the classification of e-commerce products after data mining into different categories. The four different categories are Electronics, Household, Books, and Clothing & Accessories.

1.5 RESEARCH QUESTIONS

Based on this, we formulated the following study questions:

Question 1a: Why do big data use in e-commerce?

Question 1b: How is the data gathered to conduct the experiment?

Question 2: What are the main algorithms in data mining?

Question 3: How the data mining can help e-commerce?

Question 4: How does cloud storage help data mining?

1.6 ROADMAP

The rest of this dissertation will follow the following format to accomplish the goals mentioned above:

Chapter 1: This chapter describes the background, in brief, research aim, research objectives, and academic and industrial implications in the rationale and dissertation research questions.

Chapter 2: An overview of the comprehensive literature review of the properties of data mining, data clouding and preprocessing market and application, and markets segmentation, rigorous analysis of data using two robust models. It also shows the existing study regarding product knowledge and customer relationship.

Chapter 3: This section describes the whole process of this study from beginning to end. This chapter outlines the techniques utilized to examine if the methodological operationalization of these for the goals mentioned above was viable. To achieve the aims of this study, the accurate tools for method analysis that are now accessible are also briefly presented. The quantitative research method was described here briefly.

Chapter 4: Deals with the analysis and results to show how the three data mining algorithms can help e-commerce businesses by providing merchandise techniques, sales prediction, analysis of shopping cart contents, CRM, and niche market breakdown.

Chapter 5: This section describes the influence of sales engineer product knowledge and customer relationship on the development of a company. Chapter 4 is discussed, and regression correlations are constructed in terms of the many underlying components.

2. LITERATURE REVIEW

2.1 INTRODUCTION

Several data mining procedures are discussed, and the present incarnation of the art in electronic commerce is presented, all of which are made possible by the availability of massive amounts of data and the rapid pace at which it is being mined. A review of the pertinent literature is first carried out in this chapter which represents data clouding, clustering, data prediction, and analysis of data using two prediction models. Special attention is paid to different techniques of data mining and then analysing data to categorize those into four categories. Finally, this chapter reviews the most significant literature on two models that are TF-IDF and Word2Vec to analyze and classify the dataset.

2.2 E-COMMERCE HISTORY

In the past decade, not only businesspeople but also academics have shown interest in e-commerce and its benefits and inner workings as the contemporary event of the global monetary condition [10] . The importance of examining and enhancing efficiency, and augmenting the tools of e-commerce, necessitates the development of novel solutions to scientific challenges in this area [11].Traditional definitions of e-business and e-commerce center on the transaction of products and services via the internet. In its broadest sense, E-commerce refers to online transactions for purchasing and selling goods and services. This includes the availability of funds as well as the transfer of data, which is why it is often called "two-way transportation [12]." Since the development of the World Wide Web, e-commerce, also known as electronic commerce, has reaped several benefits in product exchange. According to studies, the following factors account for the booming e-commerce industry: Because it is conducted entirely online, e-commerce has the potential to reach far more people than conventional retail stores do [13]. The nature of business is evolving due to electronic commerce. It enables more effective operations, new marketing techniques, a wider variety of products, and better customer management. The widespread adoption of cutting-edge data mining methods is a crucial factor in this trend [14]. As a cost-cutting measure, online businesses have centralized their data in one place, increased their focus on customizing offerings to each individual buyer, and made it easier for customers to engage with their data at a cheap per-transaction price [15]. Therefore, not only is e-commerce a kind of creative entrepreneurship, it also serves as a laboratory where ideas for enhancing both can be tried out and eventually implemented. It has been and will continue to be an area where creativity is actively implemented and disseminated.

2.3 DATA MINING

In order to uncover pertinent patterns and connections, data miners comb through vast amounts of data stored in repositories. Abstractions, aggregations, data characteristics, and summarization are crucial tools for data production [16]. In addition to facilitating online transactions, websites often fulfill other functions. Customers can find information about goods and services on a website. For instance, IBM anticipated that providing support information to consumers via the web would result in a cost of $2 billion in savings made in 2000 [17]. Several requirements must be met for data mining to be successful. However, these requirements are rarely seen in practical implementations. With proper design, some customers are happy with e-commerce, while others can be satisfied [18]. In this way, the significance of any observed patterns may be guaranteed statistically, and an overfitting scenario is diminished by having several records match each pattern. Even for small sites, it is possible to acquire clickstream data comprising information on page views quite quickly. Assuming a 2% conversion value and eight pages each session, a website that sells five goods on average every hour will generate almost 1.4 million page views per month. Yahoo! delivers more than 1 billion page hits per day, indicating that the log files alone use nearly 10 G.B. of storage per hour. Given the volume of data, valid sampling techniques like sampling, done by consumers (using cookie-based analytics) or, less ideally, by training, seem crucial for a preliminary analysis. Data mining has transformed the e-commerce industry during the last decade. There is more than one type of data that can be mined with this technique. Data mining may be used for any sort of data source; however, the techniques and tactics used to mine different types of data may differ. Distinct types of data provide different issues. The mining of data is utilized in various kinds of databases, including flat files, data warehouses, and object-oriented records.

Tables in a relational database hold information about entities, such as the values of their characteristics or the states of their relationships with one another. In tables, the rows stand in for tuples, and the columns for their respective features. A tuple is a unit of data in a relational database table, and it can represent an individual item or a relationship between objects [16]. SQL is a functional prevalent query language for relational databases, and it allows you to alter and obtain data from tables. Data mining techniques that use relational databases are more flexible than data mining algorithms that use flat files. SQL may help with data selection, transformation, and consolidation in data mining [19].

Several different foundational data mining techniques are used to develop. Typically, people will try one of these methods:[20, 21]:

a. Rules of the Association

An important and widely-studied data mining technique, "association rule mining," was first described by [22]. The program's ultimate goal is to "mine" databases used for transactions and other types of data storage that have "interesting correlations, frequent patterns, linkages, or casual structures." The main purpose of this tactic is to identify interesting connections and relationships between information stored in databases or pools. Based on the definition, we know that massive data stores are the focus of data mining. Because of the enormous amount of information, it can be difficult to keep up with the necessary changes to the found rules as the data evolves over time. There have been studies [23] conducted on preserving the results of research into previously unknown association rules. In order to manage the addition of fresh transaction data, the FUP(Fast Update) algorithm was presented by [23]. Finding the huge item sets considering a given database, let's call it D U db, here D and db are collections of historical data and new updates to the database, respectively., is a challenge for incremental updates. Sample-based methods for updating association rules have also been proposed [24]. The algorithm Difference Estimations for Large Item sets (DELI) takes samples from the data to calculate a maximum acceptable difference between the previous and current association rules. A third method identifies the incremental database's large item sets. If the negative border of the huge item sets grows larger than the primary database's, then the primary database is scanned.[25] For this reason, it is possible to locate all huge collections with just a single scan of the original database. Association norms are widely employed in a variety of fields. The implication of a typical association rule is of type A.B., where A is a set of items and B is a set of items where every atom has the same condition [26].

b. Clustering

By forming many categories out of similarly structured elements, it describes a wide variety of different types of systems, or it refers to the organization of data in classes. Furthermore, clustering class labels are unidentifiable; thus, the clustering algorithm must determine which classes are appropriate. Unsupervised classification is a term used to describe clustering. The rationale for this was that categorization could not be reduced to a set of categories. Clustering is a method for grouping disparate items into cohesive categories, whether the items are concrete or conceptual [26]. Clustering techniques have close ties to many other fields of study. Mathematical and scientific studies [27] have traditionally employed clustering. provide a foundational introduction to the pattern recognition paradigm. Speech and character recognition are two common examples of this technology in action. Clustering methods from the field of machine learning were used in the context of technology for dividing up images and analyzing them [28]. [29] worked with statistical methods of pattern recognition. The process of clustering can be thought of as an issue of density estimation. A multivariate statistical estimation has long been used to study this phenomenon [30]. Vector quantization, a data-compression method used in picture processing [31], makes extensive use of clustering. With the "divide and conquer" concept of cluster analysis, a complex system can be broken down into more manageable chunks for easier planning and execution.

c. Prediction

There has been a lot of focus on the forecast because of the importance of accurate prediction in the business world. It is possible to classify forecasts into two broad categories: The first is forecasting missing data values, and the latter is that after a classification figure is built on a training stage, the object's classifier can be anticipated from the attribute values of the object. The forecast of missing numerical quantities is more commonly referred to as prediction [26]. Logistic regression and multiple discriminant analysis and analysis are two examples of the classic statistical methods that are used in data mining for predictive modelling. In addition to the conventional ways, artificial intelligence has also contributed to expanding data mining techniques. Neural networks and decision trees are the two most prominent models here. Because academics in the accounting field are already familiar with classical statistics and can find them in dictionaries devoted to the subject, we will not be covering them here. Instead, we'll go through neural networks and decision trees in the next chapters. These two data mining methods outperform more standard statistical approaches (like non-linear regression) at modeling complicated non-linear and interaction relationships [32].

2.4 COMMON DATA MINING TOOLS

2.4.1 Weka

Weka is the first of the three animals. You'll need the correct tool for the dataset you're mining if you want precise data mining results. Weka, nonetheless, makes it possible to put learning algorithms into practice. The program has several benefits as it has all of the usual techniques of data mining, for example, data preprocessing, clustering, association, classification, attribute selection, and regression. There are Java and non-Java editions, a visualization app, and plenty of options for personalization [33]. Weka is an extremely popular package and one that finds extensive application in the classroom. In tandem with Ian H. Witten and Eibe Frank's book Data Mining: Practical Machine Learning Tools and Techniques, there is a matching software suite. For their work on the Weka system and associated book, Weka has been honored with the ACM SIGKDD Service Award for 2005. According to Gregory Piatetsky-Shapiro, who covered the story for KDnuggets news on June 28th, 2005, "Weka is a watershed system in the development of machine learning and data mining since it is the only toolkit to achieve such extensive adoption and longevity." (11 years have passed since Weka's initial version has been made available).

2.4.2 NLTK

The majority of its applications are in the realm of language processing, and it features several different processing tools, as well as machine learning, data mining, sentiment analysis, data scraping, and other language processing tasks. To make use of the NLTK library, a user must first download and install the full package on his computer. It's written in Python, so a user may develop applications on top of it and play about with it to his heart's content. All three of the tools described above are open source. The fact that NLTK is developed in Python means that it benefits from the language's clear syntax, powerful string-handling capabilities, and intuitive design. As an object-oriented language, Python programming allows for an interactive interpreter, data encapsulation, and a rich library. The treebank tokenizer is used by NLTK, and the POS tagger makes use of the tags used by the Penn Treebank database, often built using a Maximum Entropy prototype on the PENN treebank corpus. A Maximum Entropy model is used in the ACE database with the help of the chunking and NER modules [34]. Methods like tokenization, morphological analysis, POS tagging, stemming, chunking, and named entity recognition (NER) are just a few of the natural language processing (NLP) tasks that the NLTK test Python modules can handle. There are a number of corpus tests included in NLTK, including the CMU Pronunciation Dictionary, Brown Corpus, and the CoNLL-2000 Chunking Corpus, P.P. Attachment Penn Treebank, the SIL box of shoe corpus design, Corpus, NIST IEER Corpus, and.

2.4.3 Spider Miner

Built-in Java, this data mining software eliminates the need for human input. Spider miner's tool capabilities include comprehensive analytics using template-based design. It is a very versatile and user-friendly application that can display, forecast, data preprocessing, deploy statistical modeling, and progress report activities, as well as data mining. Learning techniques, algorithms, and models from WEKA, as well as an R script, are included in the tool, making it more effective. All three of the tools described above are free and open source. Sometimes, instead of referring to a cryptocurrency mining program, the phrase "spider miner" is used to describe a computer program that does data mining. A spider miner, in this sense, is a program made to scour the web for information, and it can do so by accessing websites, social media sites, and databases. Market research, consumer trend analysis, and brand mention monitoring are just some of the many applications of this information. Spider miners, like bitcoin miners, can be a burden on a computer's resources and can be tricky to identify and remove [33].

2.5 DATA MINING IN E-COMMERCE

Data mining in e-commerce is an important part of repositioning an e-commerce firm by providing the necessary information about the business. Due to the sheer volume of clickstream and transactional data, e-commerce has emerged as a significant field for data mining [35]. Due to their recent embrace of e-commerce, most companies now store vast volumes of information in internal databases. Mining this data to improve making choices or enable intelligence in business is the only way to get the most out of it. Before data can be exploited or converted into knowledge in e-commerce data mining, it must go through three essential processes. Data mining procedures in electronic commerce are outlined in Figure 2.1.

The initial and least difficult step of every data mining project is preparing the raw data. Data cleansing is the first step in the data mining process in which irrelevant information is removed. As a result, the procedure will improve the whole productivity in data mining while boosting output data accuracy and reducing the time required for real mining. At least 80% of all processing time is spent preprocessing the procedures for data selection, cleaning, and reformatting if the organization does not already have a target data warehouse. Still, if it does not, the process will take at least 80% of the selection, cleaning, and transformation of data [36]. One of the most significant challenges in the well-known knowledge Discovery from the Data process is the one that data pretreatment for Data Mining (D.M.) concentrates on. Information gathering is typically an organized process [37] . There is a high probability that the data contain flaws such as discrepancies, mistakes, out-of-range numbers, impossibly unlikely data pairings, missing data, or—most importantly—data that is insufficient to start a D.M. process [26]. Additionally, the need for more complex processes to evaluate it is necessitated by the ever-increasing volume of data used in today's business applications, science, industry, and academia. Preprocessing the data allows us to turn the impossible into the doable by tailoring it to meet the specific needs of each D.M. algorithm's input parameters [38] . When working with a lot of data, the time spent in the phase of preparation can be prohibitive. Reduction of information, which try to degrade the data logging by deleting unimportant and disturbing aspects using potential selection, instance selection, or methods of discretization, is also included in data preprocessing. A complete set of data that can be taken into consideration as valid, again helpful as well, while subsequent D.M. algorithms the anticipated outcome of a reliable connection of data preparation operations [39].

The second phase is pattern mining, which means the ways or approaches adopted to generate a set of guidelines or a prototype from an enormous data collection. The term also goes by data mining techniques or algorithms. The most common e-commerce patterns are those based on prediction, grouping, and association principles. The latest phase, analysis of pattern, is used to check and shed more light on the discovered model to show us the way for the data mining result to be used. The statistics and principles of the pattern utilized were heavily emphasized in the study, which was done by watching them after several users had visited them [22]. Applications in speech recognition and temporal database matching methods like time travel disruption have both made significant strides in recent years. For temporal data, there are essentially two distinct sorts of similarity inquiries that have developed so far: There are two types of matching in this scenario: entire matching, where the repeating unit and the patterns in the dataset are of equal number, and subsequence matching, where the sequence may be smaller than the patterns in the dataset as well as a result may occur at any arbitrary location. [40]. Some key distinctions between the various methods proposed in the literature are as follows. First, we choose a certain metric of similarity [41]. Second, we can distinguish between a period and changed evaluations. [23]. Third, the flexibility of the method in accommodating subsequences of varying lengths, as well as in scaling and translating them, is a key distinguishing feature [42]. Finally, several methods have been investigated to lessen the need for comparisons and search space in mining [43].

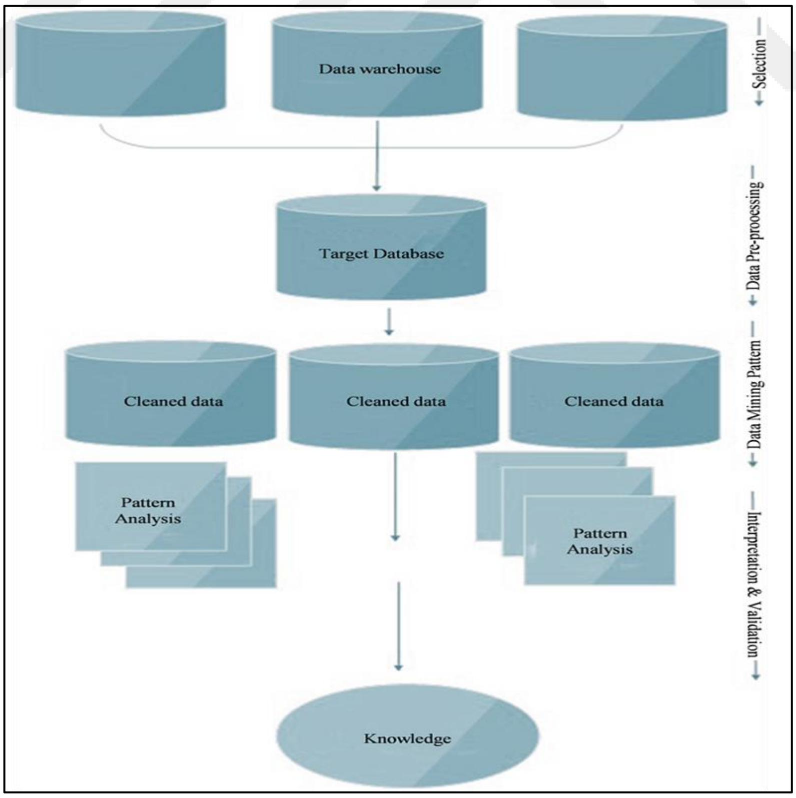


Figure 2.1: E-Commerce Data Mining Procedure [44]

All of this, though, has to do with how iterative the entire process is, as well as how you interpret visual data at each sub-step. Therefore, the process of data mining consists of five stages below:

Selection of Data: In this step, we will establish the parameters of the data mining project, including the data type, the goals, and the tools. Appropriate input attributes and task-related output data are selected. The selection of data in data mining is the procedure of picking out a subset of data from a larger dataset for further study. It may be required to pick data for a number of reasons, including the following: The dataset may be too big to process as its whole [45]. Compared to the distribution reconstruction approach proposed by [Agrawal and Srikant 2000], it is more efficient in terms of the amount of information lost throughout the process. Despite having access to a massive amount of data, this technique, which is on the Expectation Maximization (E.M.) technique, approaches, with increasing probability, the original distribution from the changed data. In addition, it was noted that the E.M. technique was indistinguishable from the Bayesian reconstruction presented in [46], save for the interval approximation used in the latter [47].

Transformation of Data: This phase is all about eliminating irrelevant information and, sorting data into functional categories and, altering the format of data so that it can be used in another context, normalizing the necessary data, finally, deciding on data-driven planning to complete missed data. In data mining, the term "data transformation" refers to the process by which raw data is converted into a more manageable form prior to analysis. Numerous activities, such as: are possible within the realm of data transformation. Aggregation means bringing together information from several places or organizations. By introducing a way to guarantee just partial disclosure while still letting a miner delve deeply into the material, [48] did just that. Using this method, one constructs localized data-driven decision-making, and training data is generated by randomly exchanging values between entries inside a parent node of the tree. Labeling the category: considered a personal attribute; therefore, only that will be swapped. This method addresses statistical accuracy versus safety as a trade-off, where safer methods include less precise results. When developing a classification tree learned from training examples, Agrawal and Srikant [49] investigated a scenario in which the individual records values were disrupted in addition to probability-based random numbers. The appearance and distribution of the generated data records differ considerably from those of the original milestones.

It presented reliable estimates of the original data range, necessitating a new method of reconstruction. It is impossible to recreate original values for individual data records. The authors acknowledge that there is some inevitable loss of data throughout the distribution reconstruction process, but they maintain that this is often acceptable in practice.

Data mining step in and of itself: Data mining methods are developed after the miner has converted data and mined it in order to find patterns of interest using any of the available techniques. To protect the confidentiality of individual purchases, Evfimievski et al. [Evfimievski et al. The objective here is to replace some components of each transaction with others that weren't there before. It appears to have achieved a fair amount of privacy protection by eliminating some accurate information and adding some misleading information. If you use simple uniform randomization, this method can help you recover association rules that are less frequent than they were before. While privacy is maintained on average, While the chance of information leakage is low, it is nevertheless a serious concern due to uniform randomization for a subset of all transactions.

Result interpretation and validation: The resilience of the data mining application test is checked for a better comprehension of the data and its synthesized knowledge, as well as its validity span. The information collected can also be assessed by comparing it to previous application domain expertise. Using a data distortion strategy, Boolean association rules have recently been studied [50]. Once again, the goal is to distort the data so that it becomes impossible in order to piece together the monetary details of a certain deal. Still, the rules discovered using the skewed data remain intact. The work's flexibility in defining privacy is an intriguing aspect. Successfully learning a '0' from the skewed data could be seen as less of a privacy risk than correctly guessing a '1'. Probabilistic distortion of user data is the foundation of this approach, which also includes an analytical formula and a privacy score. While this approach protects the user's anonymity and ensures reliable mining results, it can be prohibitively expensive in terms of both time and storage to mine the distorted database instead of the original.

Integration of newly acquired information: Knowledge discovery results are presented to the responsible party so that any conflicts with previously extracted knowledge may be resolved and the recently found pattern can be put to use [51].

2.6 PROS OF DATA MINING IN ONLINE BUSINESSES

"Data mining" in the context of "online commerce" means areas where data mining could be applied to benefit an e-commerce business. As it is known, while purchasing virtually, individuals typically miss essential data that companies can store in their systems. These numbers indicate data, either unstructured or structured, that can be mined to give the company a strategic advantage. Some of how firms might benefit from data mining in the e-commerce sphere are as follows:

2.6.1 In-Depth Analysis of Client Profiles

This is also known as "customer-centric" or "user-centric" design in the e-commerce context. E-commerce enterprises can more effectively organize their activities and operations by drawing on business insights gained from customer data. By segmenting visitors into those with a higher propensity to make a purchase, businesses can save money on advertising. Businesses can learn about customers' online shopping intentions by analyzing their clickstream data and seeing if customers are seeking a specific product or just browsing. These aid businesses in planning and improving their infrastructure [20].

2.6.1.1 Service customization

What we mean by "personalization" is the process of making particular adjustments to a user's experience based on their preferences and past actions. Information mining with a personalized twist is the practice of adjusting how a user is served data following his or her unique characteristics. To do this, computers and machine learning approaches can examine user data and create predictions about what the user will find interesting or useful. Personalization can be used in a variety of contexts to enhance the user experience and boost participation. Using a customer's purchase and browsing history as examples, online stores can make product recommendations. Search engines and social media platforms can both benefit from personalization by displaying more relevant results for a user's search or feed. In the grand scheme of things, personalization can be a potent tool for enhancing the efficiency of data mining operations and delivering a more satisfying and satisfying user experience. Research on data mining for personalization has mostly concentrated within the context of content-based recommendation systems and associated areas like collaborative filtering. In the data mining sector, recommender systems have received a lot of attention. Collaborative filtering, community data mining, and content-based systems are the three main categories. These methods often take the form of a user profile and grow and learn from either direct or indirect user input. In evaluating data sources generated by a team of many persons as while doing their regular activities, analysis of social data may be a source of great benefit for information for businesses [19].

2.6.1.2 Basket analysis

Data miners might use a method called "basket analysis," sometimes known as "market basket analysis," the most frequently bought items together. Case in point, stores can utilize this data to understand their customers' shopping preferences better and use that knowledge to place products in their stores better.

Typically, a retailer's sales data is analyzed to conduct a basket analysis. Details about the time and place of each purchase are recorded here, along with the things bought.

Before performing a basket analysis, the data must undergo preprocessing to eliminate noise and duplicates. After then, we look for trends in the items that appear together in several transactions. Methods like frequent item set mining and association rule mining can help with this.

Using the information gleaned through basket analysis, businesses may make educated choices about things like which goods to cross-promote, which ones to stock next to one another on shop shelves, and which ones to highlight in highly specific email campaigns. Every shopper's basket has anything to say, and the term "market basket analysis" (MBA) refers to a common type of statistical study used in retail analytics, a method of business analytics that helps stores learn more about their client. Making the most of the analysis of market baskets can be done in a few different ways:

a. Product affinities recognition: monitoring and using not-so-obvious product affinities is the actual problem in retail. Customers who buy Barbie dolls at Walmart prefer those three chocolate bars. Using sophisticated market basket analytics, one can unearth previously unknown connections like this, leading to more effective advertising campaigns.

b. Promotion of up and cross-selling; these display items are bought simultaneously, enticing printer purchasers to purchase premium paper and add-ons.

c. The importance of planning and product pairings are used to enhance stock management by concentrating on items that are likely to be purchased together, creating attractive bundle deals and user-friendly store layouts.

d. Shopper profile: employing data mining to analyze shopping carts over time can reveal demographic information about your clients, including their age, income, occupation, interests, and preferred brands [19].

2.6.1.3 Sales forecasting

The term "sales forecasting" refers to the method used by businesses to predict their sales in the future. It's a crucial aspect of every company's financial strategy because it allows for better budgeting, stockpiling, and resource distribution. Among the many tools available for predicting future sales are:

Analyzing qualitative data: These techniques rely on the judgment and subjective analysis of experts in order to predict market trends, customer mood, and the competitive landscape, all of which have the potential to affect sales.

To produce predictions, quantitative approaches use data and statistical analysis. Time series analysis, regression analysis, and exponential smoothing are all methods that are frequently used.

To foretell sales, causal approaches look to uncontrollable factors like the weather and marketing budget.

Businesses collect information on historical sales and any external factors that may affect future sales in order to generate a sales forecast. After collecting this information, it is examined with one or more forecasting methods to project potential revenue.

An accurate sales accuracy of a prediction is proportional to the thoroughness of the data and the reliability of the chosen method of prediction, so keep that in mind when making projections. Therefore, it is normal practice for companies to construct many prediction scenarios, including best case, most likely, and worst case, to deal with the inherent uncertainty in any given forecast. Predicting how long it will take for a satisfied client to make a purchase and whether or not that consumer will make a repeat purchase are two critical aspects of sales forecasting. This type of investigation can benefit planned obsolescence strategies and related product recommendations. In sales forecasting, cash flow may be forecasted in three ways: pessimistic, optimistic, and realistic. If sales are lower than anticipated, this helps ensure you have a cushion of cash to ride out the storm [19].

2.6.1.4 Merchandise planning

Merchandise planning involves estimating future demand for a company's goods and figuring out how to allocate resources best to meet that need. With its support, stores and online marketplaces can stock their shelves with the correct items at the right times in sufficient amounts.

Merchandise preparation requires a number of important factors, such as:

Predicting the future demand for a company's products is a key part of sales forecasting. Inventory planning, also known as "assortment planning," is deciding what goods to stock and in what amounts. Calculating how much stock to keep on hand and when to reorder is part of inventory management. Allocation is the process of determining where and in what quantities to individual stock products. When it comes to replenishment, it's important to plan ahead so that you know how to resupply your shelves and shelves when they run out. Seasonal or annual planning cycles are common for merchandise, with adjustments made as necessary to account for fluctuations in demand and other factors. Having the correct products on hand to meet client demand while eliminating excess inventory and associated costs is essential to managing a retail or e-commerce firm. Merchandise planning is beneficial for both online and physical stores. With the help of merchandising planning, online retailers can figure out how much stock to keep on hand and where to keep it, while offline retailers wanting to grow through shop openings can estimate how much product they'll need by studying the floor plan of an existing store. These questions can be answered with certainty if you take the right approach to product planning:

a. Pricing: Database mining will aid in selecting optimal prices for goods and services through the process of revealing customer sensitivities.

b. Choosing items; by analyzing client data, online retailers may learn which items are most in demand as well as knowledge of competitors' products.

c. Stock balancing: by mining the retail database, it is possible to establish the correct and particular quantity of stock required, in other words, just right, especially during busy sales periods and the course of an entire business year.

2.6.1.5 Market segmentation

Segmenting a market involves separating potential buyers into subsets defined by shared preferences and buying habits. In order to better target certain groups of customers with tailored products and services, businesses often engage in market segmentation. There are numerous approaches to market segmentation. Demographic segmentation refers to the process of dividing a market into smaller subsets based on demographic factors, including age, gender, income, education level, and geographic location. The term "psychographic segmentation" refers to the practice of splitting the market into distinct subsets depending on factors such as consumers' values, interests, and personality characteristics. The term "behavioral segmentation" refers to the practice of splitting the market into subsets based on consumer actions like purchasing patterns, usage patterns, and brand loyalty. Geographical segmentation divides a market into smaller niche markets (region, country, or city). The term "firmographic segmentation" refers to a method of market division based on the characteristics of the firms being targeted, such as their industry, company size, and geographic location. As a result of market segmentation, firms can learn more about their customers and cater their products, promotions, and support to each segment's unique characteristics and preferences. It's a vital part of any marketing plan that aims to get the word out and get people interested in a product or service. Data mining is a powerful tool since it can be used for customer segmentation. Customer data like revenue, age, ethnicity, and employment can be utilized to segment audiences for targeted email marketing and search engine optimization strategies. Market segmentation can also aid in the identification of a company's rivals. Even though the retail company already knows that the frequent responses point to the same customer dollars as the existing company, this data may help them see that there are other companies with which they may compete [25]. Segmenting a retail firm's database will increase conversion rates since the corporation can target an attractive subset of consumers with its advertising. In addition, this helps the retail company learn about its competitors at every stage of the process, which helps in making products that will appeal to the target demographic [25].

2.7 E-COMMERCE DATA MINING PROBLEMS

Despite the high implementation, capabilities, and benefits of using data mining in electronic commerce, there are some challenges and disadvantages as follows:

2.7.1 Spider Analysis

It is generally understood that data mining's fundamental purpose is to discover actionable insights hidden within large datasets. Data for online stores primarily comes from web pages. Therefore, online retailers need to comprehend how search engines work so that they may monitor the rate of change, the nature of that change, and the timing of its appearance. When a search engine needs to find new content, it dispatches "spiders" or computer programs. Bots or crawlers are other names for these spiders.

It's a piece of software used by search engines to seek and retrieve web pages. A search engine follows a link from one website to another and then asks for a copy of the target page to be downloaded to their server. This is the data used by search engines' ranking algorithms, and it's what users see when they search. Here, the challenge is that search engines need a replica of the page to index it properly. The search engine's data is sent into the algorithm, and the e-commerce website must be legible and visible. Data mining algorithms must be adequately trained to produce reliable findings, and tools must-have features that allow them to automatically eliminate unnecessary data that will be turned into information [18].

2.7.2 Transformation of Data

In this case, data mining techniques have difficulty overcoming the data transformation challenges. Right now, there are only two places to get the data needed for transformation: a fully operational infrastructure for the database system. Moreover, there are a few manipulations, including adding additional columns, grouping data, and totaling it all up.

The first procedure requires only periodic modifications; examples include when there is a modification to the site, and the resulting data set makes data mining more challenging [18].

2.7.3 Data Mining Algorithms' Ability to Scale

In light of Yahoo's massive data volume and over 1.2 billion daily page views, scalability is an important issue.

a. A significant amount of data may be collected from the website in a short amount of time, and the data gathering algorithm can manage or analyze it as much as is needed, primarily because of the apparent scalability nonlinearly.

b. It can be challenging for viewers to grasp the intended meaning of the generated simulations because of their seeming complexity [18].

2.7.4 Assist Business Users in Understanding Data Mining Models

Business users, such as merchandisers, web designers, and marketers, ought to be able to comprehend data mining's findings before acting on them. More model types must be developed and defined as a method for delivering these models to business users. Where do we go from here in terms of developing and providing regression models? (Even simple linear regression might be challenging for non-statisticians to understand). How do we visualize nearest-neighbor models, for instance? How can thousands of rules from association rule algorithms be presented to customers without overwhelming them?[18]

2.7.5 Support Slowly Changing Dimensions

The model's assumptions about visitors shift as couples have more children, earn more money, and see their kids grow up faster. Product qualities change, including introducing new features, product design and packaging modifications, and enhanced or diminished quality. "Slowly Changing Dimensions" is a term used to describe specific attributes that alter over time. The challenge here is keeping up with the changes while simultaneously advocating for the one the research highlights [4].

2.7.6 Business Users Access to Data Transformation and Model Building

The capacity to deliver definitive answers to queries posed by particular business users' technical understanding of analytical tools and familiarity with data transformations are prerequisites. Many commercial report designers and online analytical processing (OLAP) solutions are difficult for business users to comprehend. Two favored options in this instance are (i) providing templates for the expected queries (e.g. OLAP cubes and suggested transformations for mining) and (ii) offering expert advice through consulting or service business. The primary objective of the above activity is to devise a method to aid business users in assessing data accurately and efficiently [4].

The following is noted based on the literature review offered in this chapter:

Most of the experimental investigations were conducted on e-commerce and advancing e-commerce by using various sorts of applications related to online marketing. Researchers have also reported that e-commerce is getting more and more focused because of its recent advancement as people want to make their lives comfortable. It is important to understand under what circumstances the progress of e-commerce is happening. To contribute to this understanding, data mining has come into light as it’s an emerging idea for the advancement of e-commerce using its advantage. Numerous experimental investigations have been carried out to find out the advantage of data mining. Few researchers examined data mining through programming. A limited number of works have been investigated on the advancement of e-commerce using data mining analysing tool TF-IDF and Word2Vec. It can be observed from the literature review although very little work has been conducted on taking various analysis models of data mining, no analysis and segmentation work have been highlighted.

3. METHODOLOGY

In this chapter, the work on e-commerce advancement by using data mining is further extended, including full descriptions of test methodologies. Following this, several data mining procedures are briefly discussed to enhance online technology. This is done with the aim of showing data analysis and categorisation can be precisely done by two robust models of data mining. Apart from that, a brief discussion of the specifications is also provided.

In this thesis, after a comprehensive literature review of recent advancements in data mining and analytics have helped propel the growth of the electronic commerce sector. Our goal was to look into the classification of e-commerce products after data mining into different categories. In this study, we have used open-source data indexed in OpenAIR, published on July 31st, 2019[21]. As mentioned in the dataset description, four classes of "Electronics," "Household," "Books," and "Clothing & Accessories" are in this classification-based E-commerce text database account which includes about eighty percent of all online shops. The ".csv" file contains two columns: the first identifies the class, and the second consists of the corresponding data point. The product and description from the e-commerce website serve as the data point. Datasets are collections of information that have been prearranged and labeled for analysis. Elements are the data points that make up a dataset. The goals of a dataset and the issues it is meant to determine its size and composition. Table 1 lists the characteristics of the data set.

Table 3.1: Elements of The Dataset That Were Used for This Analysis.

Aspects of Data Sets	Variable
Number of Occurrences	50425
Number of Courses	4
Specialization	C.S
Characteristics of Attribute	Genuine
Attributes Number	1
Tasks Associated	Categorizing
Characteristics of Data Set	Multivariate

E-commerce, or electronic commerce, is the use of the internet to buy, sell, and trade products and services (or electronic commerce). Although it may also make use of other technologies, such as email, for, at minimum, some of the transaction life cycles, it typically uses the web. Purchasing physical or digital commodities (like novels from Amazon) or services is the most common form of online commerce (like digital distribution via the iTunes Store, where music can be downloaded digitally).

Online shopping, digital marketplaces, and auction sites are the three pillars that support the e-commerce structure. E-commerce can't exist without the electronic business. Electronic commerce (E-commerce) aims to facilitate the buying and selling of goods and services through the internet, which has the dual benefits of reducing transaction costs for buyers and sellers and promoting greater mobility in the marketplace.

A sort of economic taxonomy known as "product categorization" or "product classification" refers to a set of categories that a group of items would fall under. Two distinct tasks are involved in product categorization:

a. Create, maintain, and expand the catalogue structure for the products being sold.

b. Tagging goods with the appropriate features and categories.

While the first activity does not have much of a place for machine learning, it is conceivable to automate the second process, which is relatively time-consuming and labor-intensive. The categorization of products supplied on e-commerce platforms based on the descriptions of the products listed therein is the issue under consideration in this notebook. Such categorization aims to improve user experience and produce better outcomes from external search engines. By using the website's search engine or browsing the catalogue, visitors can find the things they need easily.

3.1 TEXT CLASSIFICATION

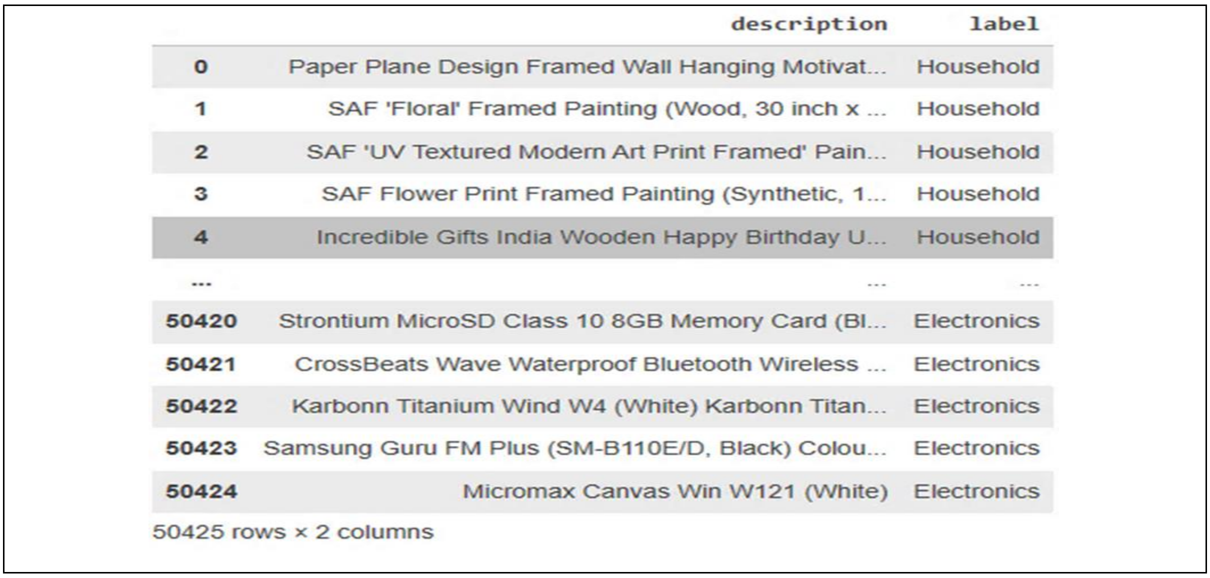
Natural language processing (NLP) tasks for text classification are frequently used in many commercial issues. The task entails selecting an appropriate category from a list of pre-defined categories to apply to a statement or document. The categories are chosen based on the dataset of choice. Applications for text classification include the automatic tagging of customer inquiries, news classification, the classification of emotions, and spam email detection.

In the current issue, the categories are Electronics, Household, Books, and Clothing & Accessories, and the statements are the product descriptions.

Python was used in this project via Google colab notebook. A coding environment similar to a notebook online called Google Colab is ideal for data analysis and machine learning. It allows GPU utilization and has a variety of machine-learning libraries. Data scientists and ML engineers are the key users. The CSV file, including all datasets, was uploaded to google colab, and then the codes were written there to get the results. The coding file has been provided as supplementary data.

For different categories of data, the labels were encoded with the numbers of 0 for Electronics, 1 for Households, 2 for Books, and 3 for Clothing & Accessories. The memory usage was calculated as 0.7695 MB, and the dataset was shaped in a (50425, 2) matrix. Table 2. demonstrates the loaded and subsequently customized data after calling it from google colab. Training models for machine learning is an example of a computationally intensive task that required access to Colab's GPUs and TPUs to complete. Individuals who don't have recourse to a robust local computer or who have jobs that would take too long to do on a local system will find this feature especially helpful.

Table 3.2: Dataset with Description and Labels.



	description	label
0	Paper Plane Design Framed Wall Hanging Motivat...	Household
1	SAF 'Floral' Framed Painting (Wood, 30 inch x ...	Household
2	SAF 'UV Textured Modern Art Print Framed' Pain...	Household
3	SAF Flower Print Framed Painting (Synthetic, 1...	Household
4	Incredible Gifts India Wooden Happy Birthday U...	Household
...	...	...
50420	Strontium MicroSD Class 10 8GB Memory Card (Bl...	Electronics
50421	CrossBeats Wave Waterproof Bluetooth Wireless ...	Electronics
50422	Karbonn Titanium Wind W4 (White) Karbonn Titan...	Electronics
50423	Samsung Guru FM Plus (SM-B110E/D, Black) Colou...	Electronics
50424	Micromax Canvas Win W121 (White)	Electronics

50425 rows × 2 columns

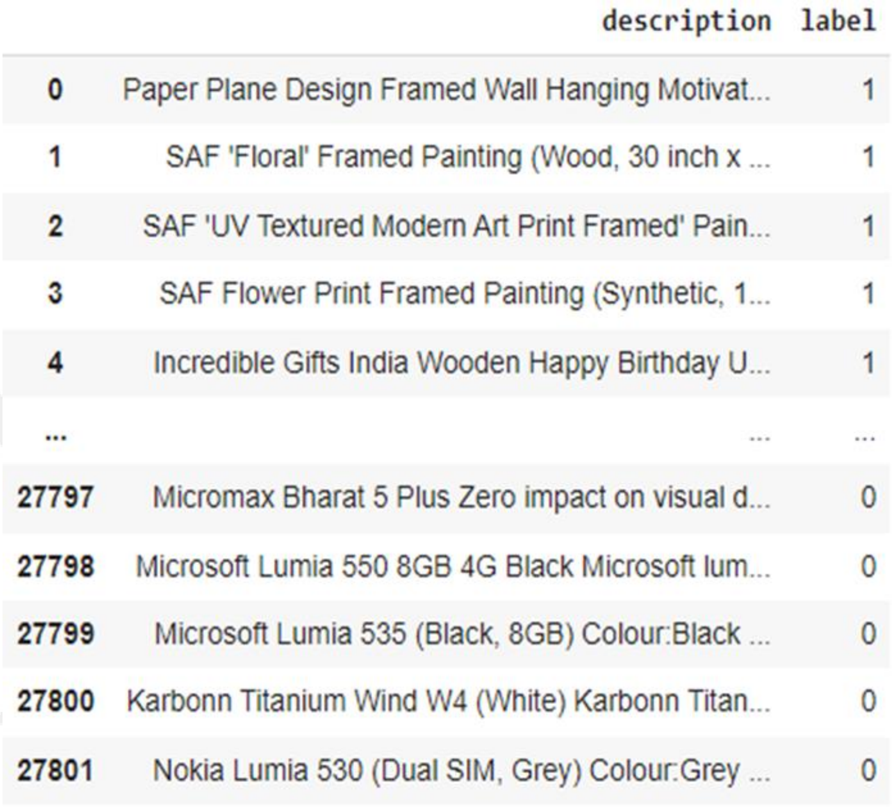
In the field of data mining, a database mainly composed, and labeling is a collection of information that contains both descriptive data and labels that specify the categories or classes that each piece of information fits into. Data mining routinely makes use of these kinds of collections to teach machine learning models how to categorize and predict outcomes by analyzing the data's attributes and patterns. Here data set were customized after calling it from google colab. Shared access to Google Colab facilitates collaborative development by allowing numerous developers to access and modify the same repository at once. This can be helpful while working in groups or when asking people for input or assistance.

Because of its compatibility with Google Drive, Colab allows several users to simultaneously make changes to the same notebook. This is helpful for group projects since it permits people to work on separate but related aspects of the project at the same time. Sharing a link to or publishing the notebook directly to the web allow others to access the information contained within a Colab notebook hosted on Google Drive. Sharing the work with more people or asking for input or assistance from others around you can both benefit from this. We started by transferring the dataset to our Google Drive in preparation for using it with Google Colab. The dataset(s) were uploaded to Google Drive by dragging and dropping the files or by clicking the "Upload" button in the Google Drive interface. After dropping the dataset into Cloud Storage, the following method was used to retrieve it in a Colab notes:

```
from google.colab import drive

drive.mount('/content/gdrive')

# Replace '/path/to/dataset' with the path to the dataset in Google Drive

dataset_path = '/path/to/dataset'
```

This code will mount the Cloud Storage to the Colab context, allowing the user to view the information from within the notebook. The dataset file(s) can then be read and manipulated using a standard Programming language. Based on the product's description that is accessible on the e-commerce platform, The focus of this research is to categorize a product into the four categories of Electronics, Household, Books, and Clothing & Accessories and check the accuracy of this classification. Also, the encoded access the latest data set version in Table 3.

Table 3.3: Encoded and Sorted Dataset.



	description	label
0	Paper Plane Design Framed Wall Hanging Motivat...	1
1	SAF 'Floral' Framed Painting (Wood, 30 inch x ...	1
2	SAF 'UV Textured Modern Art Print Framed' Pain...	1
3	SAF Flower Print Framed Painting (Synthetic, 1...	1
4	Incredible Gifts India Wooden Happy Birthday U...	1
...	...	...
27797	Micromax Bharat 5 Plus Zero impact on visual d...	0
27798	Microsoft Lumia 550 8GB 4G Black Microsoft lum...	0
27799	Microsoft Lumia 535 (Black, 8GB) Colour:Black ...	0
27800	Karbonn Titanium Wind W4 (White) Karbonn Titan...	0
27801	Nokia Lumia 530 (Dual SIM, Grey) Colour:Grey ...	0

Below is the histogram showing the frequency of electronics, households, books and clothing & accessories. The aim of the study is to categorize these four things and analyze them by discussing models. That's why the following categorization is important to analyze the precise data.

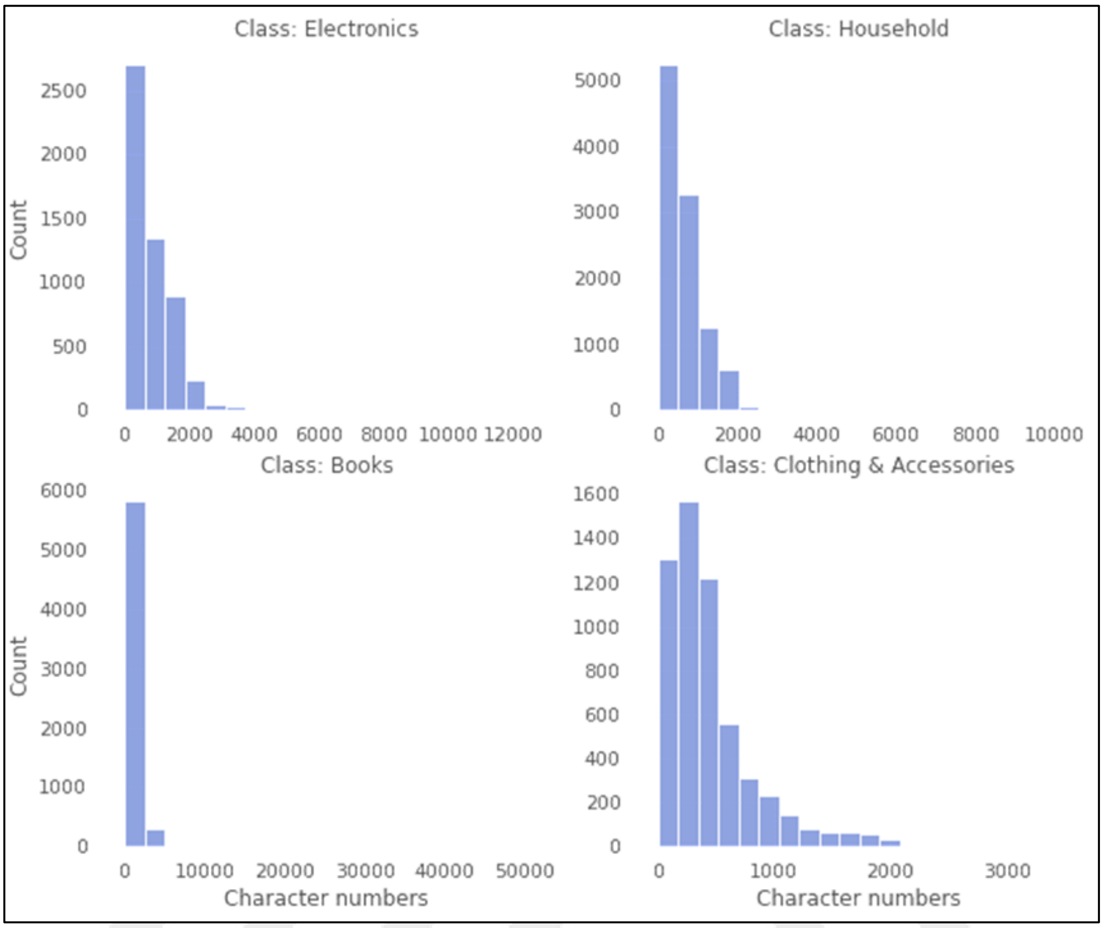


Figure 3.1: The Histogram Charts Show the Character Number in The Description.

Histograms could have been used to represent data in a variety of ways. Still, above, the method is used to display the frequency or count of four categorizations Electronics, Household, Books, and Clothing & Accessories with a specific number of characters. The histogram's x-axis shows the character count, while the y-axis shows how many descriptions have that length. Bars reflecting the frequency or count of descriptions with that character count are presented in the histogram, with the height of each bar denoting that value.

Here at the start of the character number, the number of descriptions is the highest in the first three categories. The highest descriptions are visible for the books, and the amount is almost 6000. Then it decreases a little bit for the number of households. Though the amount is nearly the same, books number of books outweighs the household number. The descriptions number for the household is more than 5000 but is slightly less than the books. Now, if we can elucidate the electronics, it is visible that the description number is almost half of that household. It is the third most in that bar diagram.

On the contrary, Clothing & Accessories hold a minor position in this category. The description number of this segment is too low. The amount is almost one-fourth of the number of books. The description number of Clothing & Accessories is only more than 1500.

As the character's count increases, the description number of all categories decreases. It is following a downward trend. When the count number increases to 500, both the categories' electronics and household reduce to almost half. While it is nearly one-third to ten thousand for households and it is half of the amount for electronic things. The exception occurs at Clothing & Accessories, which increases when the count increase and the count number is 1600.It is necessary to count the number of characters in each description before creating a histogram of the character count in the set of descriptions. The descriptions are then sorted into groups based on the number of words used to describe them and the overall number of words used to describe each set.

In figure 3.2 the histogram represents the word number in the description. Word number is essential for word vectorization, which is conducted through word embedding to capture and analyze the data category. In natural language processing, word number is the main factor in getting a good result.

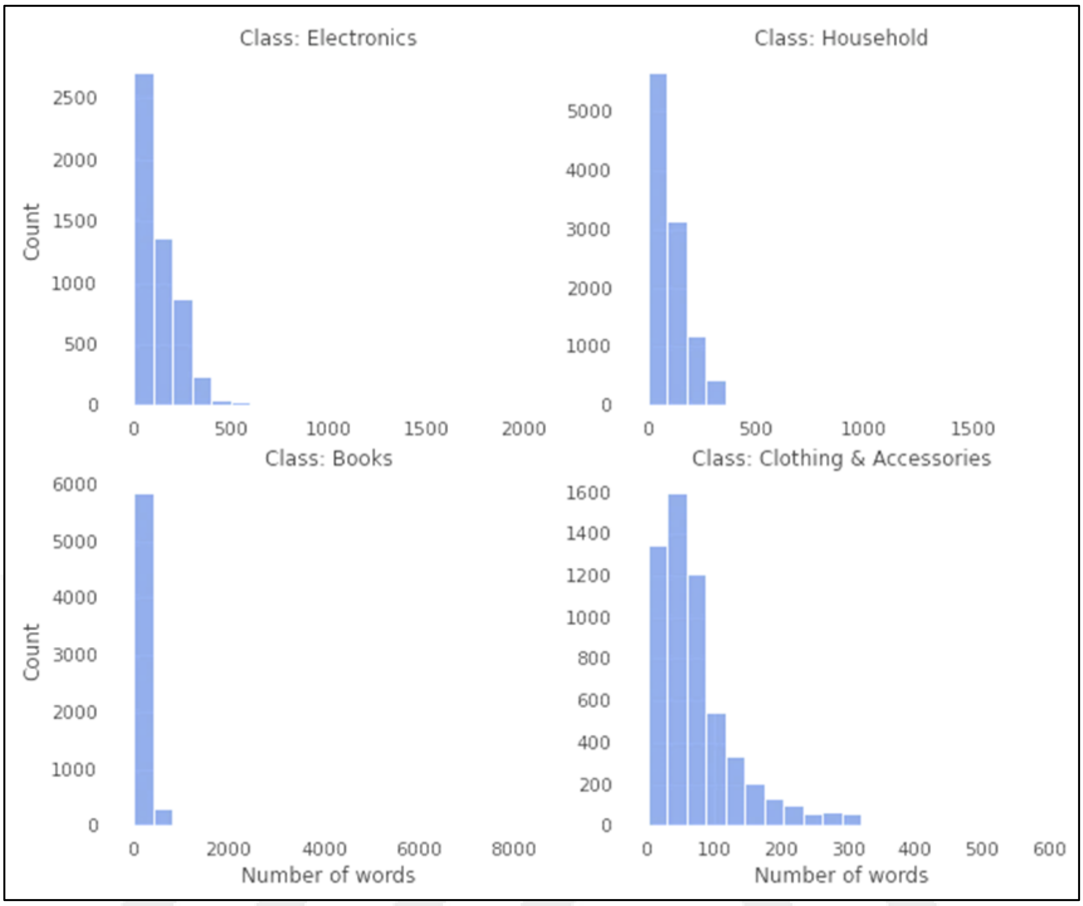


Figure 3.2: The Histogram Charts Show the Word Number in The Description.

In order to show how often each of the four groups—Electronics, Home, Books, and Clothes and Accessories—occur, a method is utilized to format the data. On a histogram, the x-axis represents the total number, while the y-axis represents the relative frequency. The histogram plots a series of bars, where the elevation of each bar represents the total number of occurrences of a certain descriptor in the data set.

The first three groups all have more descriptions than this point in the word number. Nearly 7,000 books have their full descriptions shown. After then, there is a significant drop in the number of homes. After the huge decrease in books, there are far more households than people have to read. The total number of home descriptions is around 3,000, which is double the number of electronics while the word number increases. Amazingly, Clothes and Accessories outreach the word number of electronics. It's the second highest on that particular bar chart.

Contrarily, books are a somewhat early finished sub-category here. This section's characterization number is inadequate. The sum is around half of the total households. Fewer than 6500 unique word numbers are used for books.

The total number of descriptions for each group drops as the word count rises. Currently, it's trending down. At 500, the percentages of both electronics and domestic goods drop below 50%. Clothing & Accessories is the only category where an increase happens when the count is more than 1600.

Prior to creating a histogram of all the characters in the collection of descriptions, it is important to count the number of words in each description. Next, we group the descriptions by how many characters they have in common and count how many words fall into each category.

The number of symbols in a word is called its length, whereas the kind of data being represented or stored is called its character length. When discussing data mining, the term "data character" is used to describe the specifics of the information being studied. Average word length was determined by dividing the total number of characters by the total number of words in the text. Through this method, we were able to determine the typical length of a word. In order to find trends and patterns in the data, data mining techniques were utilized to analyze these characters.

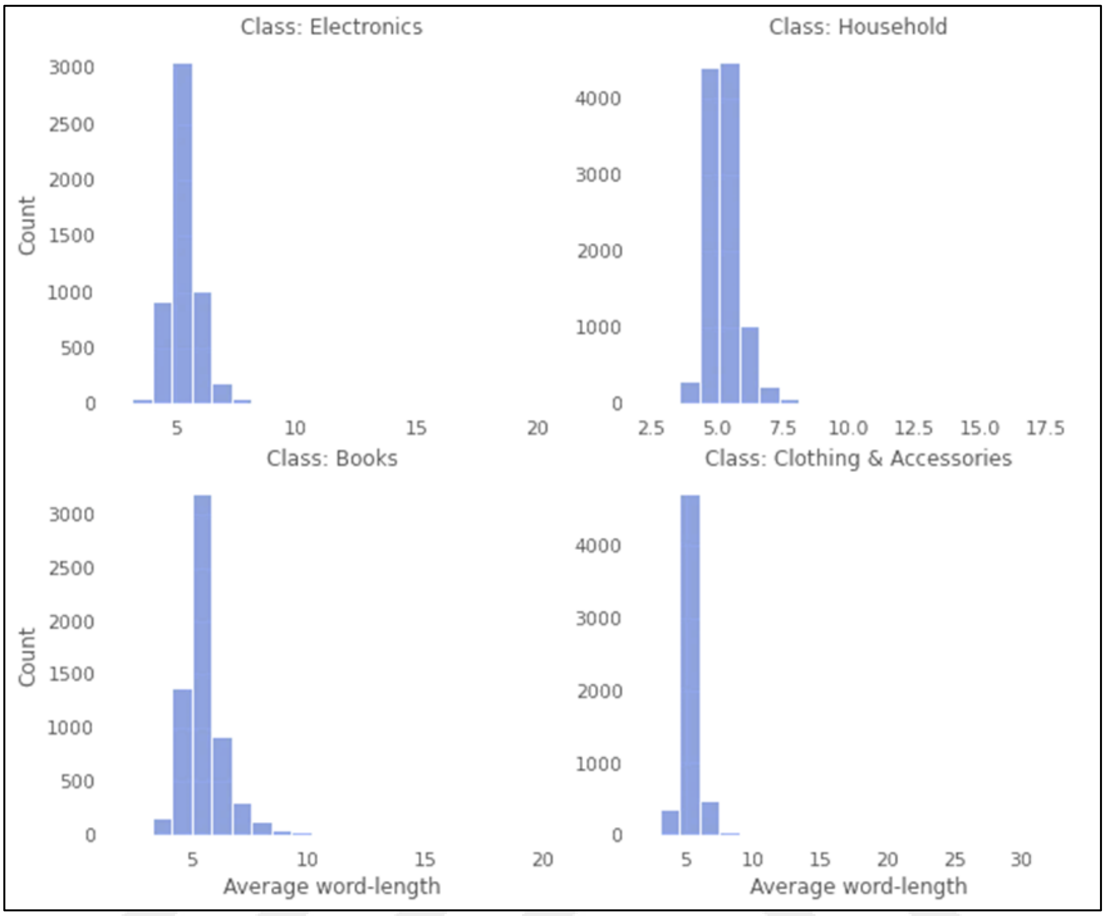


Figure 3.3: The Histogram Charts Show the Average Word Length in The Description.

When text is normalized, it is standardized into one authoritative version in natural language processing. We take into account several text-normalizing procedures. We consolidate a few procedures into a single function after the section and use it in the product descriptions.

Term frequency-inverse document frequency, or TF-IDF for the purposes of data mining, is a numerical statistic used to signify a word's relevance to a set of documents (a corpus), used as a ranking component in numerous types of searches, including text mining, data mining, and user modeling. The frequency with which a word appears in the text affects its TF-IDF value, which appears to be the case by the overall frequency with which that term is used. By how many papers in the corpus contain the word. The terms T.F. and IDF make up its two components.

The relative frequency of a phrase inside a particular text is known as term frequency (T.F.). It is calculated by dividing the overall word count in the text by counting how many times a term appears there. In mathematics, it is represented by

$$TF(t,d) := \frac{f_{t,d}}{\sum_{t\in d} f_{t,d}}$$

Where frequency is $f_{t,d}$ (number of times) that the word "t" appears in document "d". Remember that Word count (d) is the denominator, and each search term occurrence is added to that amount.

The relative frequency of a phrase is its appearance in a document divided by the total number of words. The frequency of a term was determined by dividing the number of occurrences of the phrase by the grand total of words in the text, and thereafter the result was multiplied by 100 to represent it as a percentage. In the fields of natural language analysis and text analysis, the relative frequency was utilized to determine which words and phrases were the most prevalent. It was also utilized to analyze the occurrences of various phrases both within and between publications.

A word's inverse document frequency (IDF) gauges how often or seldom it appears in each document. The inverse percentage of documents that contain the phrase can be calculated by taking the logarithm of the ratio of the total number of records to the total number of papers that have used the term. A word's frequency in a given corpus of documents can be quantified using the metric known as inverse document frequency (IDF). Word frequency was calculated using this method to rank how frequently a given word appeared in a set of papers. The term frequency-inverse quantity and variety (TF-IDF) measure is a statistical tool for determining how significant a given word is to a given document inside a given collection or corpus, and the document term frequency is commonly employed as a weighting function in the calculation of this measure. In mathematics, it is symbolized by

$$IDF(t,D) := log\frac{N}{|\{d \in D : t \in d\}|}$$

here $|S|$ is the significance of cardinality of the set S, the number of documents, denoted by "N," in total, i.e., N=|D|, and the denominator |{d∈D:t∈d}|is the total number of sources containing t as a keyword. If t is absent from the corpus, then IDF(t, D) will become indeterminate. Frequently, the denominator is changed to 1+|{d∈D:t∈d}to avoid this. TF-IDF is the product of the following terms T.F. and IDF, i.e.

$$TF - IDF(t,d,D) := TF(t,d) \times IDF(t,D)$$

It considers the fact that certain words are used more often than others when assessing a word's relevance to a text within a collection of texts. An element set's cardinality is equal to its size. An essential topic in computer science and mathematics, cardinality measures how large or complex a collection is. An individual's interpretation of a set's cardinality may change based on the specific application. The complexity and practicality of operations on a set are significantly impacted by its cardinality.

In this study, firstly, the texts were vectorized, and then two different models of TF-IDF and Word2Vec were utilized to analyze and classify the dataset.

4. RESULTS AND DISCUSSION

This chapter focuses on experimentally quantifying the accuracy of two models, which were utilized to analyse the database and categorise all the data sets. The first part of the chapter gives an overview of the properties of text vectorization, TF-IDF base modelling, TF-IDF Hyperparameter Tuning, word2vec model, and word2vec Hyperparameter Tuning. Importantly in this chapter, a large volume of test results is presented to derive the accuracy of the categorization and analysis of the data set using two models. Finally, test results are compared and discussed concerning the main variables.

4.1 TEXT VECTORIZATION

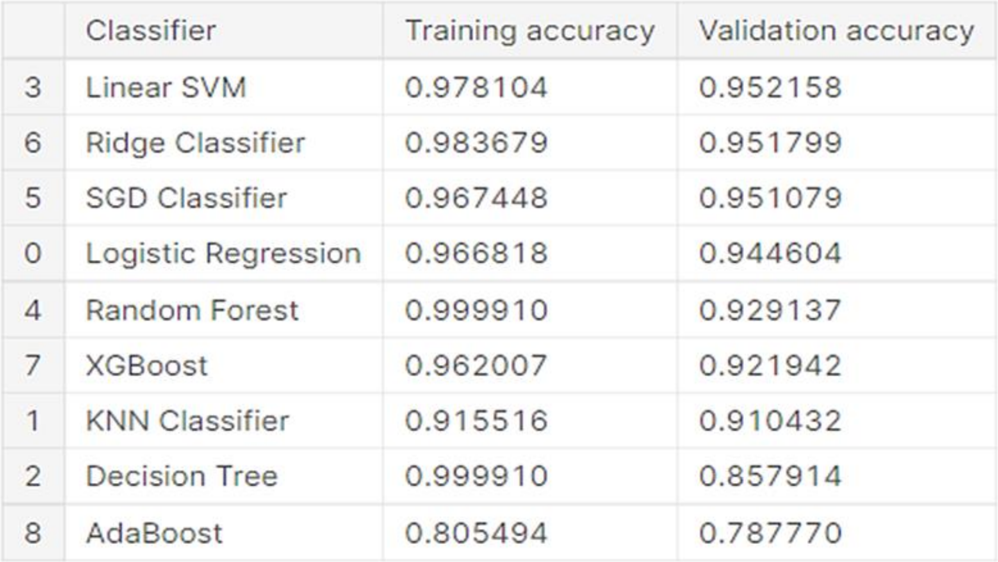
Getting text data into a numerical representation that machine learning algorithms can understand and work with is a crucial stage in the data mining process. When not converted to vectors, text input stays in its unstructured form, making it difficult for machine learning algorithms to evaluate and process. Bag-of-words, term frequency, one-hot encoding, inverse document frequency (TF-IDF), and word embeddings are only a few examples of the many text vectorization methods available. The type of text data and the data mining project's objectives will determine the technique(s) chosen, each of which has advantages and disadvantages. Textual data processing in applications for natural language processing must be vectorized. By transforming the textual contents into useful numerical representations, vectorization helps the machines to comprehend the textual contents. Rao et al.[9] also proved that text vectorization helps the machine to understand the textual content of the data. Another study of Wu et al.[7] also utilized text vectorization method to train the machine learning model for understanding the textual content. The suggested study aims to find news items that cannot be combined to conduct multi-document summarization. In this study word embedding was used training on a huge corpus of text to provide a continuous, low-dimensional representation of each word in the lexicon. This study need to convert the texts into vector representations so that machine learning may be applied to text data. Text vectorization is the process of turning words, phrases, or even larger text data units into numerical vectors in natural language processing.

4.2 TF-IDF BASELINE MODELING

The statistical measure TF-IDF (term frequency-inverse document frequency) can be used to determine how significant a term is to a document within a corpus. It is a common metric in text mining and information retrieval, where it helps determine which words are the most significant. To put it another way, If a term is pretty frequent in one document but not in the majority of other documents in the corpus, TF-IDF will attribute more importance to that document than it would to a word that often appears in textual records and databases. A set of documents was acquired for analysis in order to develop a TF-IDF baseline model. Next, for every word in the corpus, we determined its TF and IDF. Calculating a word's TF is as simple as counting how many times it appears in a manuscript and then dividing that number by the overall number of words contained in the document. Words' IDFs are determined by dividing the ratio of documents contained in the collection that include the word by the overall number of the document in the collection. Specifically, here's how it operates:

Table 4.1 demonstrates the training accuracy and validation accuracy comparison with different Classifier. Training accuracy ranges from 80.54% to 98.36%. On the other hand, validation accuracy ranges from 78.77% to 95.21%. Ridge classifier is the highest among all the training accuracy, whereas Adaboost stands the lowest. On the other side, Linear SVM's validation accuracy is top of the list while Adaboost is again at the bottom. Kamba and Hang also [26] showed that Linear SMV shows the highest accuracy among other classifier. Wang et al.[36] have showed different results regarding the accuracy level of classifier where they confirmed Ridge classifier showed the highest accuracy.

Table 4.1: Training Accuracy and Validation Accuracy of Different Classifiers Used In This Study.



	Classifier	Training accuracy	Validation accuracy
3	Linear SVM	0.978104	0.952158
6	Ridge Classifier	0.983679	0.951799
5	SGD Classifier	0.967448	0.951079
0	Logistic Regression	0.966818	0.944604
4	Random Forest	0.999910	0.929137
7	XGBoost	0.962007	0.921942
1	KNN Classifier	0.915516	0.910432
2	Decision Tree	0.999910	0.857914
8	AdaBoost	0.805494	0.787770

The occurrence count of a word in a given text is known as its term frequency (T.F.). If take the logarithm of the ratiob was taken between how many pages make up the corpus containing the word, its inverse document frequency (IDF) can be found. An individual word's TF-IDF document score is calculated by multiplying its T.F. and IDF values.

4.2.1 TF-IDF Hyperparameter Tuning

Modifying a model's hyperparameters for better results is known as "hyperparameter tuning." Several hyperparameters of a TF-IDF model can be tweaked to improve the model's accuracy. A description of the normalizing method employed while determining TF values. Several methods exist for adjusting the TF values to be more comparable, such as log standardization, dual normalization, and augmented normalization. How smoothing was used to the IDF numbers when calculating them. With smoothing, uncommon words won't be over-emphasized in the model. Smoothing methods range from linear interpolation to Jelinek-Mercer smoothing to Dirichlet smoothing. On the baseline model that performs the best, we adjust the hyperparameters. The results of hyperparameter tuning are as follows:

Gridpoint #1: {'C': 0.1, 'kernel': 'linear'}

```
Training accuracy: 0.9357043298412842, Validation accuracy: 0.9251798561151079, Runtime: 3m59s

 Gridpoint #2: {'C': 1, 'kernel': 'linear'}

Training accuracy: 0.9781035025403534, Validation accuracy: 0.952158273381295, Runtime: 2m30s

Gridpoint #3: {'C': 10, 'kernel': 'linear'}

Training accuracy: 0.9982015197158401, Validation accuracy: 0.9460431654676259, Runtime: 2m35s

Gridpoint #4: {'C': 100, 'kernel': 'linear'}

Training accuracy: 0.999370531900544, Validation accuracy: 0.939568345323741, Runtime: 2m29s

Best model: SVC(C=1, kernel='linear')

Best parameters: {'C': 1, 'kernel': 'linear'}

Best validation accuracy: 0.952158273381295
```

This study needed better or more optimization in the search method to get better results when testing each new machine. Even if the initial run on a brand-new computer achieves respectable performance, this optimization is still required. There are methods for hyperparameter optimization, and these include linear search and manual search. The best accuracy can be achieved with the use of this grid search by reviewing the parameters inside the servant list and comparing the results. However, not all of the findings in this implementation improve accuracy; some actually decrease accuracy. Grid search and manual search are the two most common methods used for hyperparameter tuning. In addition, grid search only supports a single hyperparameter, and the user must explicitly provide the names and values for each hyperparameter.

4.3 WORD2VEC MODEL

Word embeddings are used regarding NLP (natural language processing) to represent a word in terms of a vector of real numbers representing the semantics of a word, in the hope that close-by terms, whereas in the realm of vectors, may share semantic similarities. It can determine semantic and syntactic similarities, as well as other contextual relationships with other words in the document, and it can record the context of a word in a document. The word-embedding method known as Word2Vec makes use of a neural network model to learn word connections from a large text corpus. Following training, what the model can spot are related phrases suggesting additional words to complete a sentence. As its name indicates, word2vec maps each unique word to a vector. This assignment is made in a fashion that allows a simple mathematical operation on the vectors that the words are mapped to reflect word-to-word resemblance in terms of meaning (for instance, cosine similarity between the vectors).

4.3.1 Text Normalization in Part

Data mining text data requires a preprocessing procedure known as text normalization. Data mining can be more exact as a result and fruitful if the text data is reliable and consistent. Text normalization can be achieved by a number of different methods, such as case conversion, removing punctuation, removing stop words, stemming, and tokenization. These methods can be used to clean up the textual data by getting rid of redundant information and mistakes like mis-spelled words and inconsistent capitalization. When preparing text data for use as input in machine learning algorithms, normalization is often performed as a preparatory step before vectorization. Factoring in data dimensionality is reduced, and the quantity of unique tokens on the page is decreased when the text is normalized. When we have pre-trained embeddings, standard text normalization procedures like stemming, lemmatization, or the removal of stop words are not advised. This is because such preprocessing procedures result in the loss of important data that the neural network might utilize. Before supplying the tokenized words to the pre-trained model to get the embeddings, we will simply take a few specific text normalization techniques into consideration.

4.3.2 Word Embedding

Word embedding is a method for providing machine learning algorithms with a quantitative representation of textual data. A dense, continuous vector representation of each word in the text collection is learned. Word embeddings are a sort of vector representation that can be utilized in a wide variety of natural language processing (NLP) activities, including text classification, sentiment analysis, and machine translation, by capturing the semantic links between words.Word2vec, GloVe, and fastText are just a few examples of the many tools available for generating word embeddings. Most commonly, a machine learning model is used to learn word embeddings from a vast body of text data. The model's goal is to determine, from the surrounding words, the most likely setting in which a particular word will appear. During its training, the model discovers connections between words, which it then encodes as numerical vectors. By providing a concise and efficient representation of text data that can be utilized as a factor in the development of algorithms (machine learning), word embeddings are helpful in data mining. And unlike more standard methods like one-hot encoding or bag-of-words, they may capture intricate word connections. Word embeddings can be learned by these approaches from a huge dataset of words, such as a corpus of text or a set of user reviews, using a variety of different methodologies. Here word embedding is used by Word2vec to capture the contexts and semantics. For doing that, TF-IDF was used to get the result.

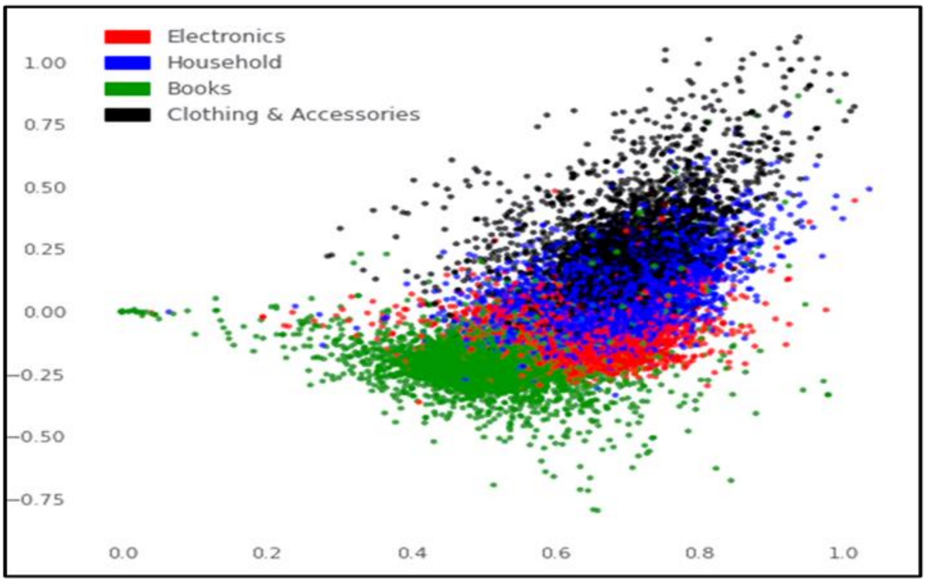
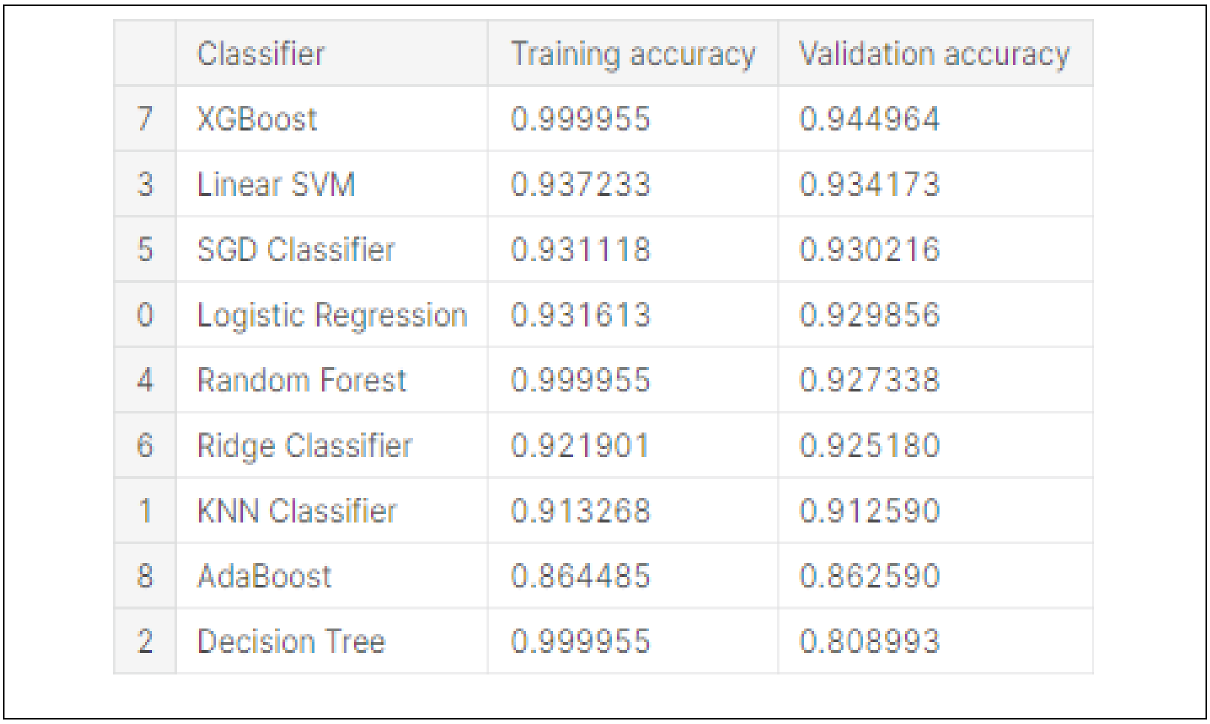


Figure 4.1: Word2Vec Baseline Modeling.

Word2vec relies on two primary learning algorithms—skip-gram and bag-of-words. Bag-of-words made its prediction based on the word's substance, and the historical sequence of the words had no bearing on the result. However, given the present word, skip-gram anticipated the surrounding words. Different bags of words with skip grams were employed to portray the context in a dispersed manner. It must be stressed that all word positions used the same weight matrix connecting the projection layer and the input.

Table 4.2 demonstrates the training accuracy and validation accuracy comparison with different Classifier. Training accuracy ranges from 86.45% to 99.99%. On the other hand, validation accuracy ranges from 80.89% to 94.49%. XGBoost classifier is the highest among all the training accuracy, whereas Adaboost stands the lowest. On the other side, Again, XGBoost validation accuracy is top of the list, while the Decision tree is at the bottom. Berry and linnoff [19] also showed thatXGBoost can perform more accurately than other classifier. They also showed that Adboost cannot outperform other classifier based on accuracy and data validation.

Table 4.2: Training Accuracy and Validation Accuracy of Different Classifiers for Word2Vec Model.



	Classifier	Training accuracy	Validation accuracy
7	XGBoost	0.999955	0.944964
3	Linear SVM	0.937233	0.934173
5	SGD Classifier	0.931118	0.930216
0	Logistic Regression	0.931613	0.929856
4	Random Forest	0.999955	0.927338
6	Ridge Classifier	0.921901	0.925180
1	KNN Classifier	0.913268	0.912590
8	AdaBoost	0.864485	0.862590
2	Decision Tree	0.999955	0.808993

4.3.3 Word2Vec Hyperparameter Tuning

The term "hyperparameter tuning" is used to describe the procedure of tweaking the values of hyperparameters in an algorithm for machine learning to boost its efficiency. When working with a Word2Vec model, there are a number of hyperparameters that may be tweaked to improve the model's accuracy. How many dimensions the vector space the words are embedded in is determined by the size of word embeddings. Even though a bigger size has the potential to capture more nuanced inter word interactions, it may also be more computationally intensive and prone to overfitting. Also, the hyperparameter tuning was conducted on the best-performing baseline model. The results are as follows:

Gridpoint #1: {'learning_rate': 0.03, 'min_child_weight': 0, 'n_estimators': 200, 'reg_lambda': 1, 'seed': 40}

Training accuracy: 0.9734724158086417, Validation accuracy: 0.9327338129496403, Runtime: 20m4s

Gridpoint #2: {'learning_rate': 0.03, 'min_child_weight': 0, 'n_estimators': 200, 'reg_lambda': 2, 'seed': 40}

Training accuracy: 0.9722584416168337, Validation accuracy: 0.9302158273381295, Runtime: 20m4s

Gridpoint #3: {'learning_rate': 0.03, 'min_child_weight': 10, 'n_estimators': 200, 'reg_lambda': 1, 'seed': 40}

Training accuracy: 0.96974056921901, Validation accuracy: 0.9316546762589928, Runtime: 19m33s

Gridpoint #4: {'learning_rate': 0.03, 'min_child_weight': 10, 'n_estimators': 200, 'reg_lambda': 2, 'seed': 40}

Training accuracy: 0.968886291084034, Validation accuracy: 0.9330935251798561, Runtime: 19m14s

Gridpoint #5: {'learning_rate': 0.3, 'min_child_weight': 0, 'n_estimators': 200, 'reg_lambda': 1, 'seed': 40}

Training accuracy: 0.999955037992896, Validation accuracy: 0.9485611510791367, Runtime: 20m39s

Gridpoint #6: {'learning_rate': 0.3, 'min_child_weight': 0, 'n_estimators': 200, 'reg_lambda': 2, 'seed': 40}

Training accuracy: 0.999955037992896, Validation accuracy: 0.947841726618705, Runtime: 20m37s

Gridpoint #7: {'learning_rate': 0.3, 'min_child_weight': 10, 'n_estimators': 200, 'reg_lambda': 1, 'seed': 40}

Training accuracy: 0.999955037992896, Validation accuracy: 0.9464028776978417, Runtime: 13m57s

Gridpoint #8: {'learning_rate': 0.3, 'min_child_weight': 10, 'n_estimators': 200, 'reg_lambda': 2, 'seed': 40}

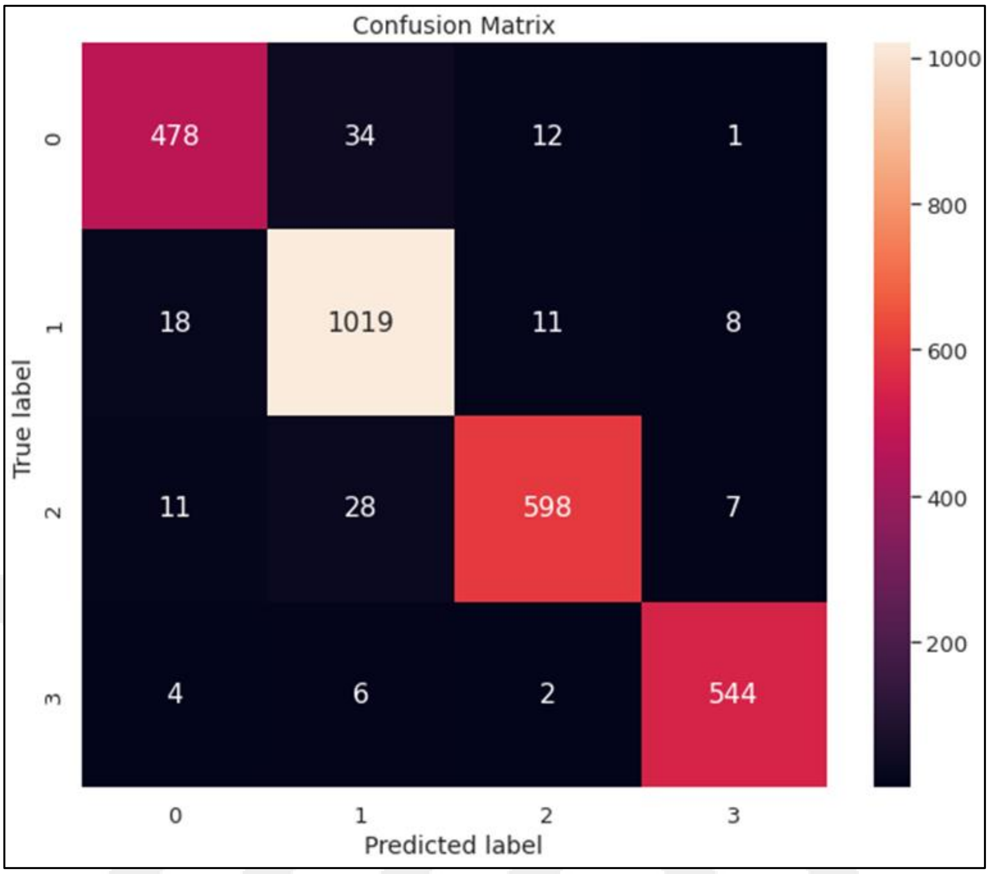
Training accuracy: 0.999955037992896, Validation accuracy: 0.9446043165467626, Runtime: 14m27s

Here in every gridpoint, there is a correlation between training accuracy and validation accuracy. We can get an idea of those accuracies from Tables 4.1 and 4.2. Their accuracy was shown using two distinct models.

The size of the window in which the context is displayed. The context window size controls how many words in either direction around a target word are used to make a prediction about that word. More information about the context around each word can be captured by using a bigger context window; however, this may increase the amount of data and processing power needed. The Word2Vec model's performance can be fine-tuned for a given application by modifying these hyperparameters. It is common practice to experiment with different hyperparameter values and then assess the model's performance with an appropriate evaluation metric (such accuracy or F1 score).

4.3.4 Performance Evaluation

An essential part of data mining, performance evaluation establishes how efficient and effective a data mining model or algorithm is in practice. The holdout method, cross-validation, bootstrapping, and numerous performance metrics, including accuracy, recall, F1 score, mean squared error (MSE), and root mean squared error, are just a few of the methods available for assessing a data mining model's efficacy (RMSE). In machine learning and data mining the effectiveness of a classifier can be measured by a table called a confusion matrix. It excels at binary classifier tasks, in which the classifier is asked to make a decision between two possible outcomes (e.g., positive or negative). To forecast the test observations' labels, we use the model with the greatest validation accuracy. The resultant test accuracy and confusion matrix are then reported. The number of correct predictions (TP), incorrect predictions (FP), incorrect predictions (FN), and correct predictions (TN) are all represented in the four cells of the confusion matrix. Data classes are shown in the rows and predicted classes are shown in the columns. To assess how well a classifier does in a binary categorization problem, one can look at the results of the task using a confusion matrix. Confusion matrices are useful for learning about the classifier's strengths and limitations and pinpointing places where it may be enhanced. Reliability, accuracy, memory, and F1 score are just few of the evaluation metrics that may be derived from these findings to provide a more in-depth picture of the classifier's efficacy. In the confusion matrix, we can observe the accuracy of that model. The "forecasted level" of a quantity is the value that has been estimated or predicted using a model. The "real level" of a quantity is its actual value. It is common practice to plot the expected level on the horizontal line of a confusion matrix, while the actual level is shown against it on the vertical.



	0	1	2	3
0	478	34	12	1
1	18	1019	11	8
2	11	28	598	7
3	4	6	2	544

Figure 4.3: Confusion Matrix Results.

As here, the upper row indicates the true positive, and it's almost 500; then this study can conclude that the classifier is successfully recognizing a high percentage of positive examples, and then the number of accurate positive predictions should be high. And the true negative{Abanda, 2016 #1} number is near 1000 so it can be said that the model successfully recognizes the inaccurate data.

The data mining project's goals and model selected will dictate the methodology and metrics used to gauge performance. Classification models can be assessed using precision, recall, and F1 score. So it can be summarized that the result of the confusion model is very much robust in strength according to its provided data.

5. CONCLUSION

E-commerce businesses that want to succeed in the modern marketplace cannot afford to ignore the importance of data mining. First, data mining helps e-commerce businesses in many ways by facilitating activities like product development, consumer behavior analysis, and sales forecasting, all of which put them ahead of the competition and increase income. Identifying spiders, Changing the shape of the data, and the capacity to scale data mining methods, clarity of data mining models for business users, support for gradually changing dimensions, and ease of data transformation are all topics that have been extensively discussed in the literature. Spider identification, data transformation, scalability of data mining algorithms, making data mining models comprehensible to business users, supporting slow-changing dimensions, and making data transformation and model building accessible to business users are all examples of the difficulties associated with using techniques from data mining to improve online business.

As an important asset, information on customers and their purchases must be handled strategically by e-commerce businesses. The provision of customer-focused services is crucial for such businesses, and data mining plays a significant part in this. In today's globally competitive market, it is essential for e-commerce businesses to employ data mining technologies.

This study also discusses cloud computing, a topic of considerable interest in the e-commerce field of data mining. Every day, thousands of businesses find themselves in need of data mining tools, but as their popularity increases, the challenges associated with integrating these tools into cloud computing get greater. Cloud computing's obvious benefits in e-commerce include enabling more efficient data mining and lowering associated costs.

REFERENCES

[1] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electronic Markets*, vol. 26, no. 2, pp. 173-194, 2016/05/01 2016, doi: 10.1007/s12525-016-0219-0.

[2] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge Discovery in Databases: An Overview," *AI Magazine*, vol. 13, no. 3, p. 57, 09/15 1992, doi: 10.1609/aimag.v13i3.1011.

[3] H. Yu, L. Cao, Y. Li, and Y. Yang, "Research of data mining in electronic commerce," in *2011 International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2011: IEEE, pp. 4323-4326.

[4] B. Ambulkar and V. Borkar, "Data mining in cloud computing," in *MPGI National Multi Conference*, 2012, vol. 2012.

[5] M. Ismail, M. Ibrahim, Z. Sanusi, and M. Cemal Nat, "Data Mining in Electronic Commerce: Benefits and Challenges," *International Journal of Communications, Network and System Sciences*, vol. 08, pp. 501-509, 01/01 2015, doi: 10.4236/ijcns.2015.812045.

[6] S. Iqbal, M. L. Mat Kiah, N. Anuar, B. Daghighi, A. Wahid, and S. Khan, "Service Delivery Models of Cloud Computing: Security Issues and Open Challenges," *Security and Communication Networks*, vol. 9, 08/01 2016, doi: 10.1002/sec.1585.

[7] W. Mingli, Z. Hui, and L. Yebai, "Data mining pattern valuation in apparel industry E-commerce cloud," in *2013 IEEE 4th International Conference on Software Engineering and Service Science*, 23-25 May 2013 2013, pp. 689-692, doi: 10.1109/ICSESS.2013.6615400.

[8] S. Madhumathi and R. Gomathi, "Data mining in Ecommerce platforms for product managers," *Research Journal of Engineering and Technology*, vol. 12, no. 1, pp. 1-7, 2021.

[9] T. K. R. K. Rao, S. A. Khan, Z. Begum, and C. Divakar, "Mining the E-commerce cloud: A survey on emerging relationship between web mining, E-commerce and cloud computing," in *2013 IEEE International Conference on Computational Intelligence and Computing Research*, 26-28 Dec. 2013 2013, pp. 1-4, doi: 10.1109/ICCIC.2013.6724234.

[10] M. L. Roberts, & Zahay, D. , *Internet Marketing: Integrating Online and Offline Strategies in a Digital Environment.* Cengage, , 2017, p. 512 p.

[11] V. Simakov, "History of formation of e-commerce enterprises as subjects of innovative entrepreneurship," *Three Seas Economic Journal,* vol. 1, no. 1, pp. 84-90, 2020.

[12] S. Malovichko, "Equifinal transformations in time and change of basic contours of electronic commerce of enterprises," *Economic space,* no. 98, pp. 25-34, 2015.

[13] O. Melnychuk, "Development of electronic commerce in the structure of the information economy of Ukraine," *Bulletin of the Taras Shevchenko National University of Kyiv. Economy,* no. 8, pp. 93-97, 2014.

[14] D. L. Banks and Y. H. Said, "Data mining in electronic commerce," *Statistical Science,* vol. 21, no. 2, pp. 234-246, 2006.

[15] O. I. Shaleva, *E-commerce / Textbook. way. Kyiv*. Center for Educational Literature, , 2011, p. 216 p.

[16] P. L. Carbone, "Expanding the meaning of and applications for data mining," in *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0,* 8-11 Oct. 2000 2000, vol. 3, pp. 1872-1873 vol.3, doi: 10.1109/ICSMC.2000.886383.

[17] P. Burrows, "The Era of Efficiency. ," *Business Week,* pp. 92-99, 2001.

[18] J. R. Reidenberg, "Yahoo and Democracy on the Internet," *Jurimetrics,* vol. 42, p. 261, 2001.

[19] M. J. Berry and G. S. Linoff, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2004.

[20] N. R. Srinivasa Raghavan, "Data mining in e-commerce: A survey," *Sadhana,* vol. 30, no. 2, pp. 275-289, 2005/04/01 2005, doi: 10.1007/BF02706248.

[21] Gautam., "E commerce text dataset (version - 2) [Data set].", 2019.

[22] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Foundations of Data Organization and Algorithms,* Berlin, Heidelberg, D. B. Lomet, Ed., 1993// 1993: Springer Berlin Heidelberg, pp. 69-84.

[23] L. Chung-Sheng, P. S. Yu, and V. Castelli, "HierarchyScan: a hierarchical similarity search algorithm for databases of long sequences," in *Proceedings of the Twelfth*

International Conference on Data Engineering, 26 Feb.-1 March 1996 1996, pp. 546-553, doi: 10.1109/ICDE.1996.492205.

[24] C. Rygielski, J.-C. Wang, and D. C. Yen, "Data mining techniques for customer relationship management," *Technology in Society,* vol. 24, no. 4, pp. 483-502, 2002/11/01/ 2002, doi: https://doi.org/10.1016/S0160-791X(02)00038-6.

[25] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining," *Data Mining and Knowledge Discovery,* vol. 2, no. 4, pp. 311-324, 1998/12/01 1998, doi: 10.1023/A:1009726428407.

[26] M. K. J. Han, J. Pei, *Data Mining: Concepts and Techniques*, 3rd edition ed. Morgan Kaufmann Publishers Inc., , 2011.

[27] S. Ann Keller, S. E. Koonin, and S. Shipp, "Big data and city living – what can it do for us?," *Significance,* https://doi.org/10.1111/j.1740-9713.2012.00583.x vol. 9, no. 4, pp. 4-7, 2012/08/01 2012, doi: https://doi.org/10.1111/j.1740-9713.2012.00583.x.

[28] D. Barton and D. Court, "Making advanced analytics work for you," (in eng), *Harv Bus Rev,* vol. 90, no. 10, pp. 78-83, 128, Oct 2012.

[29] L. M. Beskow, J. Y. Friedman, N. C. Hardy, L. Lin, and K. P. Weinfurt, "Developing a simplified consent form for biobanking," (in eng), *PLoS One,* vol. 5, no. 10, p. e13302, Oct 8 2010, doi: 10.1371/journal.pone.0013302.

[30] C. Beath, I. Becerra-Fernandez, J. Ross, and J. Short, "Finding value in the information explosion," *MIT Sloan Management Review,* vol. 53, no. 4, p. 18, 2012.

[31] T. H. Davenport, "Competing on analytics," *Harvard business review,* vol. 84, no. 1, p. 98, 2006.

[32] C. Boja, A. Pocovnicu, and L. Batagan, "Distributed parallel architecture for" big data"," *Informatica Economica,* vol. 16, no. 2, p. 116, 2012.

[33] I. Witten and E. Frank, "The Morgan Kaufmann Series on Data Mining Management Systems: Data Mining," ed: Publisher Morgan Kaufmann, San Francisco, 2014.

[34] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69-72.

[35] R. Kohavi and F. Provost, "Applications of Data Mining to Electronic Commerce," in *Applications of Data Mining to Electronic Commerce*, R. Kohavi and F. Provost Eds. Boston, MA: Springer US, 2001, pp. 5-10.

[36] B. Wang, S. Liu, K. Ding, Z. Liu, and J. Xu, "Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology," *Scientometrics,* vol. 101, no. 1, pp. 685-704, 2014/10/01 2014, doi: 10.1007/s11192-014-1342-3.

[37] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. 2015.

[38] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[39] D. Pyle, *Data preparation for data mining*. morgan kaufmann, 1999.

[40] R. Agrawal, K.-I. Lin, H. S. Sawhney, and K. Shim, "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," in *Very Large Data Bases Conference*, 1995.

[41] J. E. I. a. D. Pregibon, ". A statistical perspective on knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Know ledge Discovery and Data Mining," 1996.

[42] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," *Knowledge discovery in databases,* pp. 229-238, 1991.

[43] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," in *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1996, pp. 1-12.

[44] M. Lloyd-Williams, "Discovering the hidden secrets in your data-the data mining approach to information," *Information research,* vol. 3, no. 2, pp. 3-2, 1997.

[45] H. Liu and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*. Springer Science & Business Media, 1998.

[46] A. Birnik and C. Bowman, "Marketing mix standardization in multinational corporations: A review of the evidence," *International Journal of Management Reviews,* https://doi.org/10.1111/j.1468-2370.2007.00213.x vol. 9, no. 4, pp. 303-324, 2007/12/01 2007, doi: https://doi.org/10.1111/j.1468-2370.2007.00213.x.

[47] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 18-20 Dec. 2014 2014, pp. 1-6, doi: 10.1109/ICCIC.2014.7238499.

[48] H. D. William and R. M. Ephraim, "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update," *Journal of Management Information Systems,* vol. 19, no. 4, pp. 9-30, 2003/04/01 2003, doi: 10.1080/07421222.2003.11045748.

[49] S. Devaraj, M. Fan, and R. Kohli, "Antecedents of B2C Channel Satisfaction and Preference: Validating e-Commerce Metrics," *Information Systems Research,* vol. 13, no. 3, pp. 316-333, 2002. [Online]. Available: http://www.jstor.org/stable/23015740.

[50] E. Cherif and D. Grant, "Analysis of e-business models in real estate," *Electronic Commerce Research,* vol. 14, no. 1, pp. 25-50, 2014.

[51] R. Kimball and J. Caserta, *The data warehouse ETL toolkit*. John Wiley & Sons, 2004.