ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 7, pp. 101-116
DOI:10.33168/JLISS.2024.0706

Utilization of Text Mining for the Classification of Complaint Tickets Using Naïve Bayes in the Banking Industry

Tuga Mauritsius, Wisra Hendri, Hanson Geraldi Pardede, Faris Febrianto

Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jakarta, Indonesia

tmauritsus@binus.edu, wisra.hendri@binus.ac.id, hanson.pardede@binus.ac.id, faris.febrianto@binus.ac.id

Abstract. This study applies text mining techniques, specifically the Naïve Bayes algorithm, to classify complaint tickets at Bank XYZ. The aim is to demonstrate how leveraging business intelligence principles can enhance the bank's operational efficiency and service quality. The CRISP-DM framework is followed, and the dataset comprises 11,371 complaint tickets. The results show that the Naïve Bayes algorithm achieves an accuracy of 100%, a Kappa score of 1.000, and a classification error of 0.00%. The findings highlight the potential of text mining in addressing challenges in complaint report handling and improving operational workflows in the banking industry. However, the study has limitations, and future research should explore the scalability and adaptability of text mining solutions in dynamic business environments.

Keywords: Text Mining, Customer Support System, Customer Complaint, Naïve Bayes, Rapidminer

1. Introduction

As the landscape of business competition intensifies, the significance of expedited and precise operational services provided by information systems amplifies. The integration of information systems within the banking sector is crucial to bolstering bank operations. Present-day information systems within banking have transitioned beyond a mere auxiliary function to become primary catalysts for operational functionalities. The absence of optimal support from information systems would impede the effective and efficient functioning of bank operations.

Bank XYZ, a traditional bank operating in Indonesia, currently relies on Core Banking System and Non-Core Banking System applications to facilitate customer operations. To ensure continual improvement in the quality and availability of these information system applications, Bank XYZ prioritizes maintenance efforts. Notably, the Technology and Digitalization Division Support Section has been established as a dedicated unit to oversee information system operations. Through the Customer Support System (CSS) application, this division receives complaints regarding information system issues from various branches. Operational offices utilize the CSS application to report encountered problems, providing details such as complaint type, time, originating office, problem description, and unique complaint ticket number allocation for each submission.

Presently, all complaints related to the CSS application are solely archived within the database, lacking any comprehensive analysis of recurring issues, human errors, or patterns of disruptions across branch offices. The Support Section within Bank XYZ's Technology and Digitalization Division has yet to conduct a thorough examination of this complaint data, thereby impeding the generation of recommendations and suggestions for enhancing the quality of information system application services. To address this gap, we propose the application of business intelligence principles to analyze complaint data within the CSS application. This initiative aims to enhance the performance of the bank's information system application and mitigate occurrences of human errors, ultimately optimizing the delivery of banking services.

The problems identified by the authors based on the background of the issue are outlined as follows:

a. There is evidence to suggest that while data on complaints regarding operational issues at the bank is being collected in a database, systematic data analysis to enhance the services of information system applications has not been undertaken.

b. There is a proposal to apply the Business Intelligence concept to guide decision-making processes aimed at optimizing the operational services of the bank's information systems.

Based on these, the authors formulate the problems as follows:

1. To what extent can text mining techniques applied to complaint tickets enable the classification of information system service disruptions at Bank XYZ?

2. What recommendations can be proposed for enhancing the application utilized by Bank XYZ, drawing upon the categorized types of disruptions identified?

3. How might the enhancement of Bank XYZ's application, informed by the classified types of disturbances, contribute to the improvement of services provided to its customers?

To provide a clear understanding of the issue, the author confines the investigation to the analysis of customer complaint data specifically obtained through the CSS application. This research utilizes the Naive Bayes algorithm to categorize report data from the Customer Support System (CSS) at Bank XYZ. Integral to this classification process are key tools such as Microsoft Power BI, Microsoft Excel, and RapidMiner, enabling the analysis of a dataset containing 11,371 data samples.

The ultimate goal is to reduce disruptions that occur, thereby enabling the provision of services quickly and accurately. The study significantly contributes to the classification of customer complaint tickets by providing essential contextualization of the business environment. It underscores the heightened competition in various industries, particularly within banking, where expedited and precise

operational services are paramount. By tracing the evolution of information systems in banking, from auxiliary to primary operational drivers, the paper illuminates the specific challenges faced by Bank XYZ. This contextualization not only frames the urgency of the problem but also positions the proposed solution within the broader technological landscape of the industry.

Moreover, the study identifies a critical operational gap at Bank XYZ: the absence of systematic analysis of customer complaint data related to information system issues. This practical problem serves as a focal point for proposing an innovative solution - the application of business intelligence principles to analyze complaint data within the CSS application. By emphasizing the practical implications of the research and confining the investigation to real-world data obtained through the CSS application, the study ensures relevance and applicability to operational challenges faced not only by Bank XYZ but potentially by other organizations in similar contexts. In offering both theoretical insights into business intelligence principles and practical strategies for enhancing operational efficiency, the study enriches both theoretical knowledge in information systems and offers actionable insights for improving operational performance in the banking sector.

The subsequent section will present a review of relevant literature and related works. Section 3 will outline the methodology of the study, adhering to the CRISP-DM (Cross-Industry Standard Process Model for Data Mining) (Schröer et al., 2021) process. Section 4 will present the results and ensuing discussion, while Section 5 will offer concluding remarks.

2. Literature Review

2.1. Business Model and Process Modeling

In this research, our aim is to systematically categorize each user-reported issue within the reporting application, aiming to unveil the challenges inherent in its utilization at Bank XYZ. Given the relatively limited integration of business intelligence practices at Bank XYZ, the thorough examination of user-reported issues becomes imperative. It is anticipated that the implementation of business intelligence in this context will offer valuable insights into potential shortcomings of the application, whether manifested through user interface complexities hindering customer adoption or internal operational inefficiencies leading to frequent human errors. Through an in-depth analysis of historical user-reported issues, pertinent stakeholders and application developers can discern areas warranting targeted improvements within the application framework utilized at Bank XYZ. Consequently, such enhancements hold the promise of augmenting customer satisfaction, enhancing internal operational efficiency, and potentially yielding favorable impacts on the bank's profitability by streamlining service delivery, thereby minimizing both time and costs expended.

2.2. CRISP-DM Framework and Algorithms

The CRISP-DM (Cross-Industry Standard Process Model for Data Mining) represents an industry-agnostic framework for data mining processes. Comprising six iterative phases spanning from comprehending business requirements to implementation, its steps are outlined as follows:

Phase	Short description
Business Understanding	In this initial phase, it is imperative to evaluate the existing business environment to ascertain both the available resources and those required. A crucial component of this stage involves establishing clear objectives for data mining endeavors.
Data understanding	During this phase, data is gathered from diverse sources and subjected to scrutiny to ensure its quality and reliability through comprehensive exploration and examination.
Data preparation	The choice of data is contingent upon the specific model being employed. Data selection necessitates the establishment of inclusion and exclusion criteria, followed by the implementation of data cleaning procedures.

Modeling	During the data modeling phase, various modeling techniques are considered and implemented, encompassing the selection of appropriate methodologies, formulation of test scenarios, and development of models. The range of data mining techniques available may be employed, with selection typically guided by the specific business challenge at hand and the nature of the data under examination.
Evaluation	During the evaluation phase, the outcomes are assessed in comparison to predefined business objectives.
Deployment	This could manifest as either a conclusive report or as a software component. The deployment phase entails comprehensive planning, monitoring, and maintenance, as elucidated in the user guide.

In the realm of business analytics, diverse methodologies are employed, among which descriptive analytics holds significance. Descriptive analytics involves the characterization of historical data. Within descriptive analytics, two primary techniques are utilized: data visualization and data analysis. Data visualization entails the representation of data through graphical forms, while data analysis encompasses a range of statistical methodologies, including measures such as mean, median, standard deviation, and range, among others (Liu et al., 2023) .

Naive Bayes serves as an algorithm integral to both data and text mining endeavors. This algorithm operates on the assumption of attribute independence within classes, simplifying the computation required for probabilistic inference. As a result, Naive Bayes presents a relatively straightforward approach to data analysis (Foo et al., 2022).

2.3. Related Works

A study by (Rahman, 2023) indicates that the implementation of business intelligence within the banking sector in Bangladesh has significantly enhanced operational efficiency, thereby contributing to increased profitability. Through the utilization of business intelligence, it becomes feasible to analyze and identify areas within work units or branches where improvements in efficiency can be made. Similarly, another study (ARNET ZITHA, 2023) underscores the potential of business applications in the banking industry to streamline operational processes, thereby fostering sustainable cost reduction, and augmenting the capabilities and expertise of personnel. In line with these findings, a study by (Al Aqasrawi & Alafi, 2022) advocates for the adoption of business intelligence within contemporary organizations, given the rapid evolution of digital technology. The ability of organizations to swiftly adapt to market dynamics is deemed crucial.

Furthermore, (Al-Okaily et al., 2023) emphasizes the correlation between user satisfaction levels and the utilization of business intelligence systems, underscoring its implications for organizational profitability. It is emphasized that banks should assess system efficacy to discern strengths and weaknesses, thereby optimizing investments and ensuring tangible returns. Additionally, (Candra & Nainggolan, 2022) posits that business intelligence and analytical systems are instrumental in facilitating informed decision-making processes based on accurate and high-quality data. The capacity of organizations to furnish pertinent information enables decision-makers to make real-time decisions.

Despite these insights, there remains a dearth of research on the impact of business intelligence on bank performance, as noted in research by (Nithya & Kiruthika, 2021), making this area ripe for investigation. Findings from this research underscore the influence of business intelligence analytics on various facets of bank performance, including growth factors, internal processes, customer relations, and profitability. Moreover, research of (Bany Mohammad et al., 2022) underscores the importance of leveraging business intelligence to integrate customer information across banks, thereby enhancing efficiency and service quality, while also capitalizing on diverse data sources and applications to bolster business value, productivity, and competitiveness.

In the realm of business intelligence application, several tools facilitate data analysis and visualization to derive analytical insights tailored to an organization's strategic objectives. A study of (Ozdemir et al., n.d.) identifies various business intelligence tools favored by companies in Turkey, with Power BI emerging as the most preferred option. Each tool possesses distinct advantages and limitations, tailored to the specific requirements of the employing organization. For instance, Power BI offers an array of visualization tools and supports multiple data sources, including PostgreSQL, Microsoft SQL, Oracle SQL, and CSV files, enhancing its appeal for diverse analytical needs.

F BOZYİĞİT et.al. (BOZYİĞİT et al., 2022) develops accurate categorization of customer complaints about packaged food products in Turkish using various machine learning algorithms. Experimental results reveal XGBoost with TF-IDF achieves an 86% F-measure score, indicating its superior performance compared to word2vec-based classifiers, and feature selection through Chi Square further enhances prediction accuracy to 88%.

Serhat Peker (Peker, 2022) explores predicting firms' performances in online customer complaint management, crucial for fostering long-lasting customer relationships amid globalized competition, utilizing machine learning algorithms. Analysis of data from Turkey's prominent online complaint platform reveals random forests outperform other classifiers in performance prediction.

Muammar Nasser Saleh (Mohammed, 2020) Mohammed study how to optimize complaint management processes, particularly for entities like RTA, through the integration of artificial intelligence and machine learning algorithms. By leveraging AI-powered text classification, the aim is to automate complaint categorization and assignment, reducing the need for manual intervention and enhancing overall efficiency in managing customer complaints. Ultimately, the goal is to improve customer satisfaction and loyalty by implementing a responsive and intelligent complaints-handling system.

Anagun (Anagun et al., 2022) develop an intelligent customer complaint management system (CCMS) tailored for financial services organizations, aimed at accelerating complaint handling through automated categorization and routing. By implementing a pre-processing technique specifically designed for Turkish agglutinative language using deep learning algorithms, the study seeks to improve the performance of text classification, ultimately achieving a 96% accuracy score. The proposed method not only enhances text classification utility for a wider range of complaints but also demonstrates superiority over existing state-of-the-art strategies.

Ying Yang (Yang et al., 2018) address the classification imprecision inherent in handling customer complaints by developing a decision support system (DSS) equipped with an evidential reasoning (ER) rule-based classifier. This DSS aims to automate complaint handling processes by effectively classifying customer complaints, particularly when dealing with uncertain information in narratives. Through combining textual and numeric features to generate evidence and applying ER rules to classify complaints into categories with probabilities, the research provides telecommunication companies with a robust and data-driven approach to systematically and automatically manage customer complaints.

The common theme among these studies is the application of advanced technologies, particularly business intelligence, machine learning, and artificial intelligence, to optimize various aspects of organizational operations. Specifically, they focus on improving customer complaint management processes within different sectors such as banking, food packaging, and telecommunications. These studies highlight the importance of leveraging data analytics, automation, and decision support systems to enhance efficiency, customer satisfaction, and overall organizational performance.

In the context of customer complaint management, there is a gap in the research regarding the application of advanced technologies such as machine learning and artificial intelligence to automate complaint categorization, assignment, and handling processes, especially in Indonesian Company context.

3. Methodology

3.1. Data Understanding

In this study, data sourcing revolves around the Bank XYZ Customer Support System (CSS) application database. Acquisition of requisite data is facilitated through SQL querying of the CSS database, yielding an output in Excel format. This dataset encompasses complaint ticket records from users of information system applications across Bank XYZ's branch offices, spanning the past year. Specifically, the dataset for the year 2023, comprising user complaint data over the preceding year, serves as the training data for machine learning algorithms. The training dataset utilized for 2023 comprises 10,596 rows of complaint tickets, which have undergone prior categorization. This dataset is instrumental for both training and subsequent categorization tasks involving January 2024 and future data. The retrieval of data from the CSS database is conducted through the utilization of the SQL Yog database management system tool. Key data fields obtained include idlog, tgl_surat, ktr_asal, perihal, and bahasan, as presented in Table 1. In conducting the tests, the fields idlog and bahasan are primarily utilized. The description of the fields is given in Table 2.

Table 1. Example Dataset Illustration

idlog	tgl_surat	ktr_asal	perihal	bahasan
20230800001	8/1/2023 7:46	1900	PENDAFTARAN OL	Kepada Yth.Pemimpin
20230800019	8/2/2023 11:57	1017	PENCETAKAN HEA	Dengan hormat,Sehub
20230800020	8/1/2023 10:22	600	PERMINTAAN UPI	MOHON BANTUAN BP
20230800021	8/1/2023 10:16	7202	PERUBAHAN KOLE	Kepada Yth, Bapak Per
20230800022	8/1/2023 10:21	600	PERMINTAAN UPI	MOHON BANTUAN BP
20230800025	8/1/2023 11:28	600	PERMINTAAN UPI	MOHON BANTUAN BP

This modeling endeavor aims to forecast the ticket category according to the description provided by the user. We employ the Rapidminer application tool to conduct text mining classification on the "bahasan" field using the Naive Bayes algorithm. Excel will serve as the platform for preprocessing our data to align it more effectively with our requirements. Broadly speaking, CSS data encompasses two distinct categories, outlined as follows:

1. Data pertaining to requests for rectifying operational issues within the Bank XYZ information system application:

a. Issues related to the Core Banking System Application, covering functionalities such as Savings, Current Accounts, Deposits, and Credit management applications.

b. Interference encountered with Non-Core Banking System Applications, including Online Payment System (OPS) Applications for Regional Taxes, PDAM (Water Utility Company), PLN (State Electricity Company), Credit, MPN (National Health Insurance), Tuition Fees, and School Fees.

2. Data concerning requests for the activation of application access by users.

Table 2. Fields Description

No	Field	Description
1	Idlog	Id log complaint ticket
2	tgl_surat	Date Entry ticket
3	ktr_asal	Office of origin of complaint
4	perihal	disturbance
5	bahasan	description of the problems encountered

3.2. Data Preprocessing

During the data preprocessing phase utilizing the Rapidminer application tool, several procedural steps are executed, includes: tokenization, case transformation, and stopword removal. The rationale

behind these steps is to prepare the text data for analysis by reducing noise and irrelevant information while preserving essential content. Tokenization breaks down text into smaller units (tokens), allowing for granular analysis and identification of patterns. Case transformation standardizes text to a consistent case, facilitating comparison and eliminating redundancy. Stopword removal removes common words that carry little meaning, reducing data dimensionality and focusing analysis on significant terms. These steps improve the accuracy, efficiency, and relevance of text mining results.

1. Initially, the obtained dataset undergoes cleansing, involving the removal of redundant fields while retaining only those deemed essential. Specifically, the requisite fields comprise Idlog, category, and discussion, which will serve as the training data for the forthcoming model development endeavor. The categorization of tickets is conducted manually utilizing CSS data for the year 2023, facilitated by the Excel application. Table 3 presents samples of dataset after Step 1 process.

Table 3. Fields obtained after Step 1 process

idlog	Kategori	bahasan
20230300595	AKSES WEB MENDAGRI	APLIKASI WEB PORTAL KEMENDAGRI PADA KAN
20230601071	AKSES WEB MENDAGRI	APLIKASI KEMENDAGRI DI CAPEM UNAND PADA
20230601146	AKSES WEB MENDAGRI	MOHON BANTUAN BAPAK/ IBU UNTUK PENGA
20230601227	AKSES WEB MENDAGRI	
20230601278	AKSES WEB MENDAGRI	untuk akses webportal dengan data id address

2. Text Clasification process
   During this stage, various operator functions are employed, including the filter example, set role, and nominal to text operators, aimed at facilitating the classification of textual attributes (Figure 1).

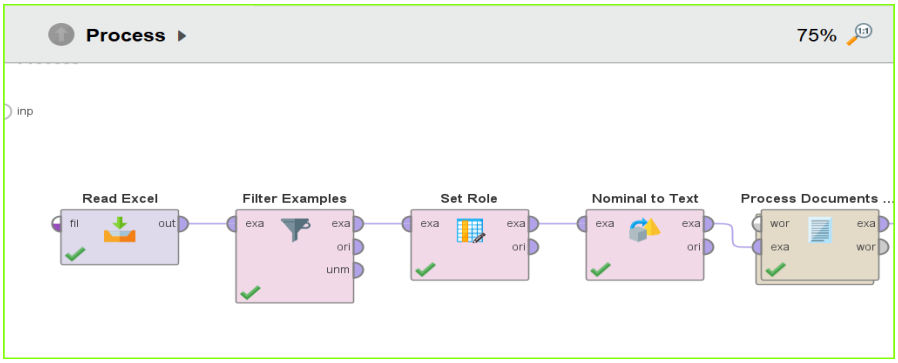

Fig.1: Text Preprocessing Workflow

3. Tokenization
   The tokenization stage serves to segment the text into individual words. This process entails utilizing spaces as delimiters to separate each word within the text. As each text typically comprises multiple interconnected words delineated by spaces, it becomes necessary to segment them to facilitate further text processing.
4. Transform Case
   The transformation of case feature facilitates the automatic conversion of all text characters to either lowercase or uppercase. In this investigation, all characters were converted to lowercase, given that the predominant text format comprises lowercase characters.
5. Stopword Filtering
   During the stopwords removal phase, the text corpus that has undergone tokenization proceeds to eliminate stopwords. Each word within the text undergoes scrutiny, with conjunctions, prepositions, pronouns, or irrelevant terms for sentiment analysis being identified and subsequently removed.
6. Token Filtering

Token filtering entails the extraction of significant words from the generated tokens based on specified character thresholds. In this process, words are filtered based on parameters such as minimum character length (min chars = 4) and maximum character length (max chars = 25).

Fig.2: Document Processing Workflow

3.3. Modelling

The process of data modeling utilizing Rapidminer incorporates the utilization of the Naïve Bayes text classification algorithm. In the study by (Ting et al., 2011), Naïve Bayes is employed as a document classifier owing to its simplicity and efficacy in document and text classification tasks. Through supervised training, Naïve Bayes adeptly discerns patterns among words and categories, enabling the classification of new words or documents into appropriate categories based on the highest posterior probability. In this instance, Naïve Bayes is applied to categorize complaint tickets from information system application users at Bank XYZ branch offices, aligning with previous research indicating its superiority as a document classifier. Nonetheless, to enhance precision, researchers advocate for a comprehensive preprocessing stage, which may prolong the duration of training and model execution. In a separate study (Ning et al., 2019), the Naïve Bayes algorithm is utilized for spam message classification, with the Bayesian learning approach proving effective. Although its performance can be compared with decision trees and neural network algorithms, Naïve Bayes offers a lower computational complexity.

The hyperparameters employed during the modeling phase adhere to the default settings provided by Rapidminer 10. These parameters encompass several aspects:

-Laplace Smoothing: By default, this feature is enabled with a value of 1, thereby implementing Laplace smoothing to mitigate zero probabilities for unseen features.

- Use Attribute Frequencies: Also enabled by default, this setting utilizes raw frequencies for computing likelihoods.

- Use Equal Probabilities: This setting remains disabled by default, whereby probabilities for unseen features are determined based on their frequency in the training data.

- Use Absolute Discounting: Not utilized in the default configuration, this hyperparameter remains disabled by default.

- Weight Threshold: By default, no explicit threshold is set; thus, all attributes are considered during classification without applying any threshold to attribute weights..

These default hyperparameters are chosen to provide a balanced approach to classification while avoiding overfitting and ensuring reasonable performance across different types of datasets.

While Naive Bayes is a popular and effective method for text mining, it has several limitations:

1. Assumption of Independence: Naive Bayes assumes that all features (words) are independent of each other given the class label. This is often not true in text data where words may be correlated or

have complex relationships. However, despite this simplifying assumption, Naive Bayes can still perform well in practice.

2. Inability to Capture Contextual Information: Naive Bayes treats each word as an independent feature, ignoring the context in which words appear. This means it may struggle to capture subtle nuances in language or understand the meaning of phrases or sentences.

3. Sensitive to Irrelevant Features: Naive Bayes can be sensitive to irrelevant features (words) in the dataset. Stopword removal and feature selection techniques can help mitigate this issue, but it still requires careful preprocessing and feature engineering.

4. Imbalanced Class Distributions: Naive Bayes may not perform well on datasets with imbalanced class distributions, where one class is much more prevalent than the others. It tends to favor the majority class and may struggle to accurately classify instances from minority classes.

5. Limited Model Complexity: Naive Bayes is a relatively simple classifier that cannot capture complex relationships between features. While this simplicity can be advantageous in terms of computational efficiency, it may also limit the model's ability to accurately represent the underlying data distribution, especially in more complex text mining tasks.

6. Zero Frequency Problem: If a word appears in the test data but not in the training data, Naive Bayes assigns it zero probability, which can lead to inaccurate predictions. While Laplace smoothing is often applied by default to mitigate this issue, it may not completely eliminate it and could still affect the classifier's performance.

Despite these limitations, Naive Bayes is often used successfully in text mining tasks, especially when the dataset is large, the classes are well-separated, and the features are relatively independent.

These are the procedural steps conducted within Rapidminer. Initially, in the phase of text classification utilizing the Naïve Bayes algorithm on the training dataset depicted in Figure 3, Excel files comprising 10,596 report data entries for 2023 and 775 for January 2024 were imported using the Excel read operator. Categorization was performed for the 2023 reports, while the January 2024 reports lacked categories. Subsequently, the examples filter operator was employed to extract data with assigned categories by filtering the category column with the condition "is not missing". Following this, the set role operator was applied to designate the attribute "category" as the target role "label", and "idlog" as the target role "id". The nominal to text operator was then used to convert all nominal attributes into text format. Subsequent to this, the process documents operator (Figure 5) was employed, incorporating tokenization, case transformation, and stopwords removal using an Excel file containing terms to be excluded (e.g., "YTH", "Kepada", "Assalamualaikum wr wb"), followed by filtering tokens based on a minimum of 4 characters and a maximum of 25 characters. The Naïve Bayes operator was subsequently utilized, wherein the model results were directed to the model store operator, and the example results were stored in the training data store operator.

Subsequently, the model application process ensues, involving the prediction of categories for the January data, as illustrated in Figure 3. Excel files containing 10,596 report entries for 2023 and 775 entries for January 2024 were initially imported utilizing the Excel read operator, with the 2023 reports categorized while those for January 2024 lacked categories. Following this, the examples filter operator was utilized to extract data lacking categories, enabling prediction using the category column filter and the "is missing" command. Subsequent operators included set role, nominal to text, and process documents (Figure 4), mirroring the preceding process. Moreover, the retrieve store operator for the training data was merged with the union operator to identify intersections and merge word columns. Further, the examples filter operator was employed once more to retrieve uncategorized data, facilitating prediction using the category column filter and the "is missing" command. The replace missing values operator was subsequently applied to alter the default to 0, resolving instances where word columns did not intersect, thereby assigning a value of 0. The retrieve store model operator was then utilized to access results from the training model. Subsequently, within the apply model operator, missing values

were incorporated, and the retrieved store model was integrated into the model. Finally, to visualize prediction outcomes in Excel format, the write excel operator was employed.

The quantity of categories can additionally be represented through a descriptive visualization to ascertain the most frequently occurring categories (refer to Figure 6). Moreover, the categorized groupings' outcomes can serve as input for the enhancement and advancement of information systems within Bank XYZ.

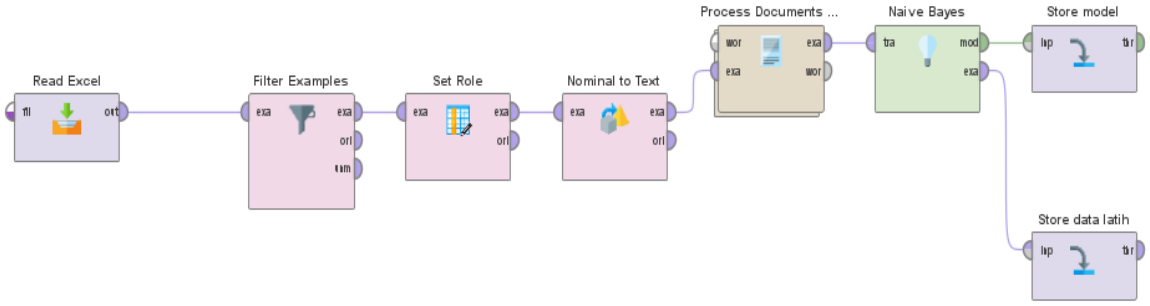

Fig.3: Procedure of Text Classification Utilizing the Naïve Bayes Algorithm

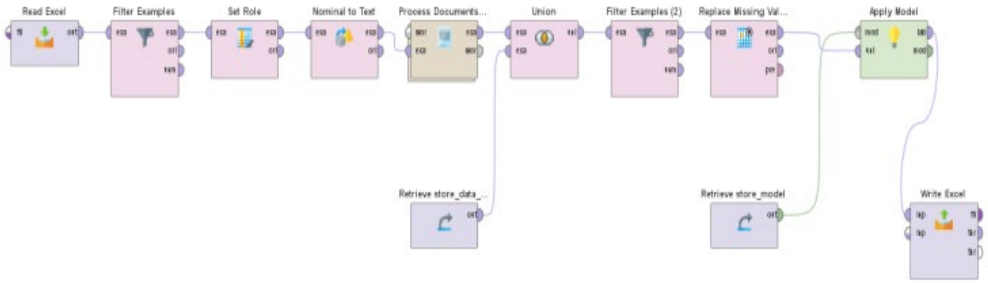

Fig.4: Workflow for Applying the Model

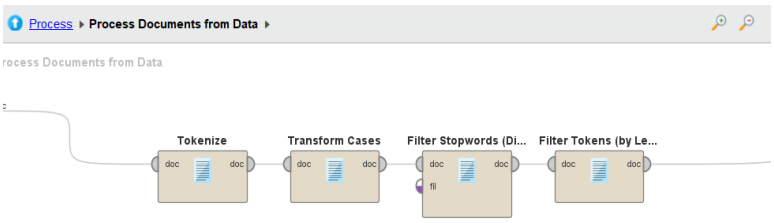

Fig.5: Internal Processes within the Document Processing

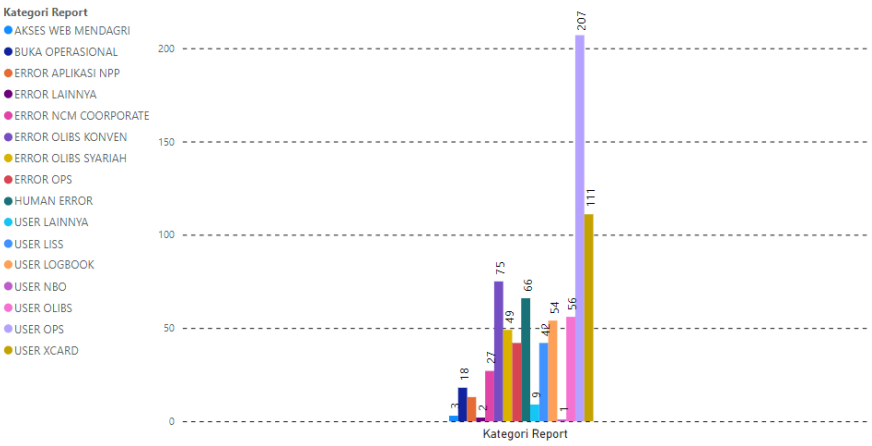

Fig.6: Utilization of Descriptive Graphic Representation with Power BI

4. Results and Discussion

4.1.Data Exploration

	idlog	text	Label
2	20240100003	divisi teknologi informasi mohon dibantu memindahkan user raihan firdaus cabang pariaman cabang bukittinggi bantuan kerjasama divisi teknologi informasi ucapkan	USER LOGBOOK
3	20240100005	pemimpin divisi teknologi informasi bank nagari kelancaran operasional cabang utama mohon mengaktifkan logbook card nagari back office data nama nino yulova kewenangan petugas administrasi capem indarung bantuan ucapkan terima ka	USER LISS
4	20240100006	pemimpin divisi teknologi digitalisasi bank nagari padang sehubungan cotnpapemimpin operasional bank nagari cabang syariah batusangkar deviasi pemimpin tanggal januari januari mohon bantuan mengaktifkan menu aplikasi core banking	USER OLIBS
5	20240100001	sehubungan nelly vindawati pensiun dana payakumbuh masuk kantor mohon bantuannya mengaktifkan menu aplikasi olibs liss nelly vindawati pensiun dana payakumbuh bantuan kerjasamanya ucapkan terima kasih	USER LISS
6	20240100007	pemimpin divisi mohon aktivasi user vanesha putri seldi nominatif adress bantuannya arapkan	USER OPS
7	20240100008	pemimpin divisi mohon pendaftaran user olibs user vanesha putri seldi teller cabang pariaman bantuannya ucapkan	USER OPS
8	20240100010	pemimpin divisi teknologi informasi bank nagari kelancaran operasional cabang utama mohon mengaktifkan logbook card nagari back office data nama vera agnella kewenangan pemimpin capem klas kemenag bantuan ucapkan terima kasih p	USER XCARD
9	20240100012	pemimpin divisi teknologi informasi bank nagari kelancaran operasional cabang utama mohon mengaktifkan logbook card nagari back office data nama yanti medona kewenangan pemimpin capem klas alai bantuan ucapkan terima kasih pemi	USER XCARD
10	20240100011	mohon bantuan pemindahan user teller muhammad reza pahlevi nomor nominatif teller mobil unit	USER OPS
11	20240100013	jakarta januari pemimpin divisi teknologi digitalisasi bank nagari padang mohon bantuannya mendaftarkan mengaktifkan user aplikasi user management kantor cabang jakarta shanty dewi fauzy pimpinan operasional bantuan terima kasih ban	USER LISS
12	20240100014	mohon bantuan pendaftaran user data nama raihan firdaus bantuan kerjasamanya terima kasih	USER LOGBOOK
13	20240100016	pemimpin divisi teknologi digitalisasi bank nagari padang sehubungan errita user mohon mengaktifkan meneset user adrian faisal pimkas linusina bukittinggi aplikasi xcard logbook dana address bantuannya ucapkan terima kasih rocladi delfi	USER XCARD
14	20240100017	pairun januari pimpinan divisi padang sehubungan penggantian teller payment point samsat tanggal desember sabtu mohon bantuan mengembalikan user nama nama ramadhan fauzi nominatif nama anggia dewi pratama nominatif nama nad	USER OPS
15	20240100015	sehubungan masuknya teller kantor capem klas nsad mohon bantuan pemindahan user teller alfa fatha rani nominatif	USER LOGBOOK
16	20240100019	mohon bantuan divisi menghapus data fase padang sicova	ERROR APLIKASI NPP
17	20240100022	pemimpin divisi teknologi digitalisasi bank nagari padang sehubungan gagal transaksi transfer virtual account aplikasi olib rekening terdebet nomor tujuan nominal tanggal nomor rekening nomor rekening virtual account tujuan mohon penyeli	ERROR NCM COORPORATE
18	20240100024	nomor lubuk gadang januari perihal memindahkan user logbook pemimpin divisi teknologi digitalisasi bank nagari surat mohon bantuan divisi teknologi informasi memindahkan user logbook data nama romi putra hasan user jabatan customer	USER OPS
19	20240100025	memindahkajati surati bupati kabupaten dharmasraya nomor diskusyendag tanggal desember perihal undangan pameran kebudayaan expo unbin bazar rakyat samsat kabupaten dharmasraya bank nagari menyediakan mobil samsat telling aca	USER OPS
20	20240100027	pemimpin divisi mohon aktivasi user vanesha putri seldi nominatif adress bantuannya ucapkan	USER OPS
21	20240100018	pemimpin divisi teknologi digitalisasi bank nagari padang sehubungan rotasi customer service lingkungan kantor bank nagari cabang mentawai mohon bantuan pemindahan user data nama yelsa perian manyah jabatan customer service capem	USER OLIBS
22	20240100028	mohon bantuan reset password xcard nama zulelfi jabatan wakil pemimpin	USER XCARD
23	20240100029	mohon perubahan status bayar perlist dimana kesalahan input sekolah dimana nama sicova nama dimi markus fahrizal sicova fahrizal mohon perubahan status bayar sicova keterangan sicova nama sicova dimi markus terinput periode desember	ERROR OPS
24	20240100030	pemimpin divisi teknologi digitalisasi padang assalamualaikum kelancaran operasional kantor cabang syariah solok mohon bantuan divisi teknologi digitalisasi reset password olibs user riyanto pemimpin seksi dana apapat bantuan perhatian	USER LISS
25	20240100031	pemimpin divisi mohon penambahan pendaftaran terminal komputer vanesha putri seldi adress bantuannya ucapkan	USER OPS
26	20240100034	pairun januari pimpinan divisi padang sehubungan dibayarkan transaksi pembayaran billing retribusi pengujian kendaraan bermotor tanggal januari dinas perhubungan meminta membatalkan transaksi lampiran data nama ramadhan fauzi no	ERROR OPS
27	20240100009	assalamualaikum kelancaran operasional kantor cabang syariah padang mohon bantuan mendaftarkan olibs management user data nama yeni oktavia user jabatan pemimpin seksi dana bantuan ucapkan wassalamualaikum	USER LISS
28	20240100035	mohon bantuan memproses cofiaga terkait bisanya export excel cetak rekening koran menu perorangan data rekening nama rekening universitas adzkia tanggal transaksi dicetak desember bantuannya ucapkan	ERROR OLIBS SYARIAH
29	20240100037	assalamualaikum sehubungan lupa password liss mohon bantuan divisi reset password data dibawah user nama donny empriedna jabatan pemimpin seksi kredit bantuan kerjasama divisi ucapkan wassalamualaikum	USER LISS
30	20240100038	pemimpin divisi teknologi digitalisasi bank nagari padang assalamualaikum sehubungan kesalahan intup rate hasil jadwal angsur olibs syariah pembiayaan nomor rekening nama nafur rahmi kode produk back back kons sliding tanggal desem	ERROR OLIBS SYARIAH
31	20240100039	pemimpin divisi teknologi digitalisasi bank nagari padang sehubungan pensiun dana ulul aeni menjalani cuti tahtunag tanggal januari digantikan angga desita pisa pemimpin seksi kredit cabang mentawai kelancaran operasional mohon bantua	USER LISS
32	20240100040	sehubungan fitria buswir wakil pemimpin payakumbuh menjalani cuti tahunan tanggal januari mohon bantuannya memindahkan menu aplikasi pimpinan operasional fitri taheni pemimpin payakumbuh bantuan kerjasamanya ucapkan terima k	USER LOGBOOK
33	20240100042	pemimpin divisi teknologi digitalisasi mohon bantuannya memindahkan user rizki oktaviana capem aasar menu teller bantuannya ucapkan terima kasih cabang utama padang	USER OPS

Fig.7: Outcome Generated by the Naive Bayes Algorithm

From the outcomes of the data modeling process, the classification of Bank XYZ report tickets was derived (refer to figure 7). All records for January 2024 have been categorized based on the issues encountered in each complaint ticket, employing the Naive Bayes algorithm. This algorithm predicts the categories of complaint tickets for label attributes that were previously unassigned by Bank XYZ. Each IDlog corresponds to a ticket number along with the year of its creation. The text attribute encapsulates the grievances expressed by customers and Bank XYZ branch offices. It was previously observed that the 2023 dataset encompasses 19 categories, each associated with various issues. The results of the implemented model indicate the association of each idlog and text entry with its corresponding category, as determined by the Naive Bayes analysis of the 2023 dataset. This demonstrates the algorithm's capability to link attributes with the designated target variable, specifically the problem category in each ticket.

4.2.Evaluation Results

Following the conducted assessment, the Naïve Bayes algorithm has demonstrated notable accuracy. Throughout the evaluation phase, we employed various metrics including Accuracy, Precision, Recall, F1-score, Kappa, and classification error to assess the predictive performance of Naïve Bayes concerning the categorization of application user complaint tickets at Bank XYZ. Refer to Figure 8 for the evaluation process flow utilized in this study. To ensure attribute congruence with their designated roles, we incorporated the set role operator in RapidMiner. Employing a training data ratio of 90% and testing data ratio of 10%, the obtained results revealed a remarkable accuracy rate of 100% for the algorithm. Additionally, the Kappa Score reached the maximum value of 1.000, with a classification error rate of 0.00%.

The comprehensive delineation of classification outcomes is observable through the confusion matrices presented in Figures 9, 10, and 11. The collective findings reveal that the recall and precision metrics across all categories attain a 100% rate, denoting a flawless performance in prediction. Each affirmative classification accurately corresponds to its true positive counterpart, without any occurrence of false positives or false negatives. Such outcomes render the model highly dependable and credible in its predictive capabilities. This underscores its proficiency in correctly recognizing all pertinent instances, as indicated by recall, while simultaneously ensuring that all identified positives are genuinely relevant, as reflected in precision. Given that the F1 score amalgamates precision and recall into a singular metric, the model achieves a perfect F1 score of one, signifying an optimal equilibrium between precision and recall, devoid of any misclassification errors.

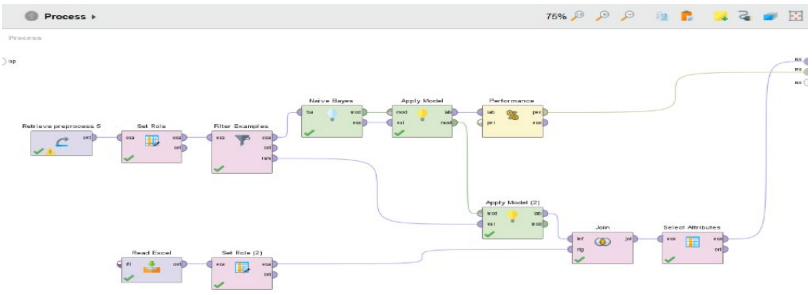

Fig.8: Evaluation workflow in RapidMiner

Table View Plot View

accuracy: 100.00%

	true ERR...	true USE...	true USE...	true USE...	true USE...	true BUK...	true USE...	true ERR...	true ERR...	true ERR...	true USE...	true
pred. ER...	3	0	0	0	0	0	0	0	0	0	0	0
pred. US...	0	20	0	0	0	0	0	0	0	0	0	0
pred. US...	0	0	4	0	0	0	0	0	0	0	0	0
pred. US...	0	0	0	8	0	0	0	0	0	0	0	0
pred. US...	0	0	0	0	5	0	0	0	0	0	0	0
pred. BU...	0	0	0	0	0	3	0	0	0	0	0	0
pred. US...	0	0	0	0	0	0	2	0	0	0	0	0
pred. ER...	0	0	0	0	0	0	0	12	0	0	0	0
pred. ER...	0	0	0	0	0	0	0	0	3	0	0	0
pred. ER...	0	0	0	0	0	0	0	0	0	3	0	0
pred. US...	0	0	0	0	0	0	0	0	0	0	11	0
pred. HU...	0	0	0	0	0	0	0	0	0	0	0	2
pred. US...	0	0	0	0	0	0	0	0	0	0	0	0
pred. ?	0	0	0	0	0	0	0	0	0	0	0	0
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100

Fig.9: Accuracy of the model

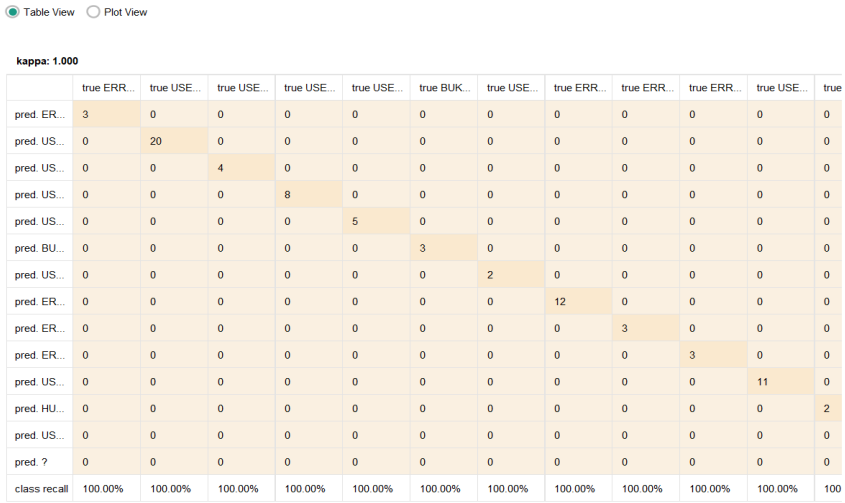

Table View Plot View

kappa: 1.000

	true ERR...	true USE...	true USE...	true USE...	true USE...	true BUK...	true USE...	true ERR...	true ERR...	true ERR...	true USE...	true
pred. ER...	3	0	0	0	0	0	0	0	0	0	0	0
pred. US...	0	20	0	0	0	0	0	0	0	0	0	0
pred. US...	0	0	4	0	0	0	0	0	0	0	0	0
pred. US...	0	0	0	8	0	0	0	0	0	0	0	0
pred. US...	0	0	0	0	5	0	0	0	0	0	0	0
pred. BU...	0	0	0	0	0	3	0	0	0	0	0	0
pred. US...	0	0	0	0	0	0	2	0	0	0	0	0
pred. ER...	0	0	0	0	0	0	0	12	0	0	0	0
pred. ER...	0	0	0	0	0	0	0	0	3	0	0	0
pred. ER...	0	0	0	0	0	0	0	0	0	3	0	0
pred. US...	0	0	0	0	0	0	0	0	0	0	11	0
pred. HU...	0	0	0	0	0	0	0	0	0	0	0	2
pred. US...	0	0	0	0	0	0	0	0	0	0	0	0
pred. ?	0	0	0	0	0	0	0	0	0	0	0	0
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100

Fig.10: Kappa Score

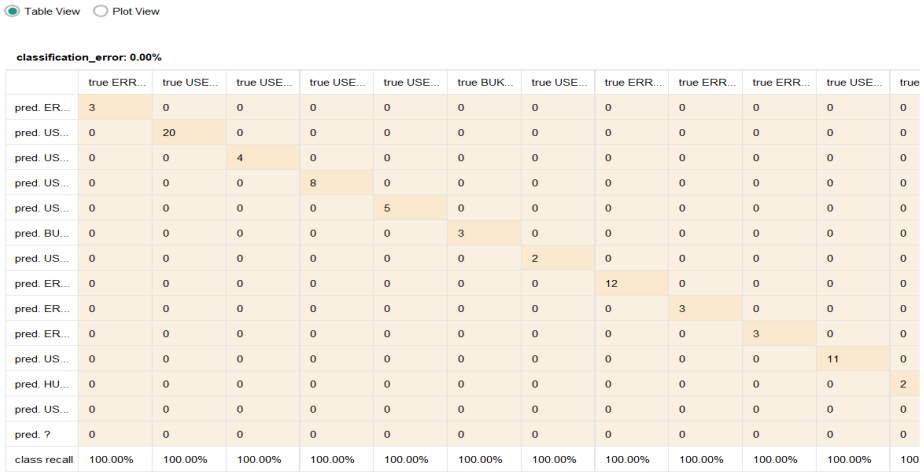

Table View Plot View

classification_error: 0.00%

	true ERR...	true USE...	true USE...	true USE...	true USE...	true BUK...	true USE...	true ERR...	true ERR...	true ERR...	true USE...	true
pred. ER...	3	0	0	0	0	0	0	0	0	0	0	0
pred. US...	0	20	0	0	0	0	0	0	0	0	0	0
pred. US...	0	0	4	0	0	0	0	0	0	0	0	0
pred. US...	0	0	0	8	0	0	0	0	0	0	0	0
pred. US...	0	0	0	0	5	0	0	0	0	0	0	0
pred. BU...	0	0	0	0	0	3	0	0	0	0	0	0
pred. US...	0	0	0	0	0	0	2	0	0	0	0	0
pred. ER...	0	0	0	0	0	0	0	12	0	0	0	0
pred. ER...	0	0	0	0	0	0	0	0	3	0	0	0
pred. ER...	0	0	0	0	0	0	0	0	0	3	0	0
pred. US...	0	0	0	0	0	0	0	0	0	0	11	0
pred. HU...	0	0	0	0	0	0	0	0	0	0	0	2
pred. US...	0	0	0	0	0	0	0	0	0	0	0	0
pred. ?	0	0	0	0	0	0	0	0	0	0	0	0
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100

Fig.11: Classification Error

The model can be utilized in various ways across the Bank XYZ. This includes:

1. Automated Routing: Upon receiving a complaint, the system categorizes it based on its content, keywords, or other features. It then automatically routes the complaint to the appropriate department or individual for resolution.

2. Prioritization: Complaint categorization allows the Bank to prioritize issues based on severity or impact. High-priority complaints, such as those related to safety or security concerns, can be flagged and addressed urgently, while lower-priority issues can be handled in due course.

Intelligent customer complaint categorization systems offer organizations a range of benefits, including improved efficiency, better customer service, enhanced compliance, and valuable insights for continuous improvement. By leveraging advanced technology and data analytics, organizations can optimize their complaint management processes and drive positive outcomes for both customers and the business.

The limitations of our study attribute to: attain a precision and recall rate of 100% implies robust performance; however, it is imperative to scrutinize this outcome within the contextual framework encompassing the task, dataset, and inherent limitations. We recognize the constraints and possible implications associated with the dataset's attributes. Its diminutive size and potential divergence from real-world distributions warrant careful consideration. The attainment of perfect precision and recall

may not carry the same significance under these circumstances, potentially suggesting an overreliance on dataset memorization rather than genuine predictive capability.

One promising area for investigation involves the application of more sophisticated machine learning algorithms beyond traditional approaches like Naïve Bayes, logistic regression or decision trees. Deep learning models offer potential for enhanced feature extraction and pattern recognition, which could lead to more accurate and nuanced complaint categorization. Moreover, ensemble methods like Random Forests or Gradient Boosting Machines could be investigated to further improve classification performance by leveraging the strengths of multiple models. Furthermore, the incorporation of additional data sources holds significant potential for enriching the complaint categorization process. By embracing these avenues of inquiry, researchers can strive towards more accurate, interpretable, and context-aware complaint categorization systems, ultimately enhancing customer satisfaction and organizational efficiency.

4.3.Analysis of Findings

Several studies share a similar focus with this article. One such study conducted by Goel, Gautam, & Kumar in 2016 involved sentiment analysis of films through Twitter. The authors utilized a dataset comprising 1.6 million tweets. Their findings revealed an accuracy rate of merely 58.40% when employing Naive Bayes (Goel et al., 2016).

Additionally, Wongkar and Angdresey (Wongkar & Angdresey, 2019) conducted a study in 2019 focusing on analyzing public sentiment towards presidential candidates during the 2019 election. The researchers utilized a dataset sourced from Twitter consisting of 443 tweets spanning from January to May 2019. Three algorithms, namely Naive Bayes, Support Vector Machine (SVM), and K-NN, were employed with the assistance of Python software. The findings of this investigation revealed that Naive Bayes achieved an accuracy of 75.58%, SVM attained 63.99% accuracy, and K-NN exhibited an accuracy of 73.34% (Wongkar & Angdresey, 2019).

These two research endeavors demonstrate advancements in the efficacy of the Naive Bayes algorithm concerning text classification. In our study, we employed various preprocessing filters, including tokenization, a tailored stopword filter (utilizing a dictionary), case transformation, and tokenization based on length. Through this meticulous process, we achieved notable outcomes with Naive Bayes, notably achieving a 100% accuracy rate. Furthermore, we incorporated several additional metrics to evaluate the performance of Naive Bayes, thereby ensuring the precision of the conducted predictions.

This study endeavors to implement text mining techniques to enhance the operational workflows conducted at Bank XYZ. Prior investigations have underscored the notable outcomes achieved through the utilization of the Naive Bayes algorithm. Our aim is to augment these outcomes by incorporating additional processes for tokenization, preprocessing, and evaluation. The dataset utilized in this study comprises information maintained by Bank XYZ spanning from the year 2023 to January 2024. The extensive nature of the acquired dataset affords us the opportunity to conduct in-depth exploration, thereby facilitating the attainment of meaningful outcomes.

The outcomes of the conducted data modeling affirm the capability of Naive Bayes to precisely categorize complaint ticket categories from XYZ Bank customers. Furthermore, these findings are reinforced by exceptionally high evaluation scores obtained from the Naive Bayes algorithm. In the forthcoming stages, there exists a strong likelihood for text mining to transition into the deployment phase, whether undertaken by other researchers or directly by Bank XYZ itself. Through the adoption of text mining, Bank XYZ stands to mitigate the challenges encountered during the receipt of complaint reports. However, it is imperative for Bank XYZ to ensure the availability of adequate infrastructure and resources to facilitate the optimal implementation of text mining initiatives.

5. Conclusion

This study demonstrates the effectiveness of text mining techniques, particularly the Naïve Bayes algorithm, in classifying complaint tickets at Bank XYZ. The findings highlight the potential of leveraging business intelligence principles to enhance operational efficiency and service quality in the banking industry. The study contributes to the existing literature by showcasing the application of text mining in a real-world setting and providing insights into the challenges and opportunities associated with its implementation. However, the study has limitations, and future research should explore the long-term impact and scalability of text mining solutions in dynamic business environments. Bank XYZ and other financial institutions should ensure sufficient infrastructure and resources for successful text mining implementation to reap the benefits of improved complaint handling and operational workflows.

Looking ahead, there are promising areas for further investigation, such as exploring more sophisticated machine learning algorithms beyond Naive Bayes. Deep learning models and ensemble methods hold potential for improving complaint categorization accuracy. Additionally, integrating additional data sources could enrich the categorization process, leading to more accurate and context-aware systems. Embracing these avenues of inquiry can ultimately enhance customer satisfaction and organizational efficiency in the banking sector.

References

Al Aqasrawi, I. S., & Alafi, K. K. (2022). Impact of Business Intelligence on Strategic Entrepreneurship: The Mediating Role of Organizational Agility. *International Review of Management and Marketing*, *12*(5), 12.

Al-Okaily, A., Teoh, A. P., Al-Okaily, M., Iranmanesh, M., & Al-Betar, M. A. (2023). The efficiency measurement of business intelligence systems in the big data-driven economy: a multidimensional model. *Information Discovery and Delivery*, *51*(4), 404–416.

Anagun, Y., Bolel, N. S., Isik, S., & Ozkan, S. E. (2022). DEEP LEARNING-BASED CUSTOMER COMPLAINT MANAGEMENT. *Journal of Organizational Computing and Electronic Commerce*, *32*(3–4), 217–231.

ARNET ZITHA, D. R. (2023). A MODEL FOR THE BUSINESS INTELLIGENCE SYSTEM ACCEPTANCE IN THE SOUTH AFRICAN BANKING SECTOR. *Journal of Theoretical and Applied Information Technology*, *101*(8).

Bany Mohammad, A., Al-Okaily, M., Al-Majali, M., & Masa'deh, R. (2022). Business intelligence and analytics (BIA) usage in the banking industry sector: an application of the TOE framework. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(4), 189.

BOZYİĞİT, F., Doğan, O., & KILINÇ, D. (2022). Categorization of customer complaints in food industry using machine learning approaches. *Journal of Intelligent Systems: Theory and Applications*, *5*(1), 85–91.

Candra, S., & Nainggolan, A. (2022). Understanding business intelligence and analytics system success from various business sectors in Indonesia. *CommIT (Communication and Information Technology) Journal*, *16*(1), 37–52.

Foo, L.-K., Chua, S.-L., & Ibrahim, N. (2022). Attribute Weighted Naïve Bayes Classifier. *Computers, Materials & Continua*, *71*(1).

Goel, A., Gautam, J., & Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. *2016 2nd International Conference on next Generation Computing Technologies (NGCT)*, 257–261.

Liu, S., Liu, O., & Chen, J. (2023). A review on business analytics: definitions, techniques, applications and challenges. *Mathematics*, *11*(4), 899.

Mohammed, M. N. S. (2020). *Customer Complaints Auto-assignment using Machine Learning Algorithms*.

Ning, B., Junwei, W., & Feng, H. (2019). Spam message classification based on the Naïve Bayes classification algorithm. *IAENG International Journal of Computer Science*, *46*(1), 46–53.

Nithya, N., & Kiruthika, R. (2021). Impact of Business Intelligence Adoption on performance of banks: a conceptual framework. *Journal of Ambient Intelligence and Humanized Computing*, *12*(2), 3139–3150.

Ozdemir, M., ÜLKÜ, E. E., & YILDIZ, K. (n.d.). Analysis and Comparison of Business Intelligence Tools Most Preferred by Companies in Turkey. *Balkan Journal of Electrical and Computer Engineering*, *11*(2), 144–155.

Peker, S. (2022). Predicting Firms' Performances in Customer Complaint Management Using Machine Learning Techniques. *International Conference on Intelligent and Fuzzy Systems*, 280–287.

Rahman, M. M. (2023). The effect of business intelligence on bank operational efficiency and perceptions of profitability. *FinTech*, *2*(1), 99–119.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534.

Ting, S. L., Ip, W. H., & Tsang, A. H. C. (2011). Is Naive Bayes a good classifier for document classification. *International Journal of Software Engineering and Its Applications*, *5*(3), 37–46.

Wongkar, M., & Angdresey, A. (2019). Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 1–5.

Yang, Y., Xu, D.-L., Yang, J.-B., & Chen, Y.-W. (2018). An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications. *Knowledge-Based Systems*, *162*, 202–210.