A Smart Model to predict the problems of telecommunication customers

Samar Mahmoud Ibrahim[1] Fahad Kamal Alsheref[2]
Riham Haggag[3]

Abstract

The proliferation of data on the internet has been greatly accelerated by the emergence of social media platforms over the past twenty years. These platforms serve as valuable sources of user-generated information, with Twitter particularly standing out as a popular microblogging platform that provides concise insights. However, analyzing informal expressions from such platforms presents significant challenges, especially in understanding and analyzing customer concerns within the telecom sector. Our research focuses on preprocessing natural language sentences to aid comprehension and analysis. We explored two distinct approaches: using the Universal Sentence Encoder and preprocessing models to prepare tweets for analysis. Additionally, we utilized algorithms such as BERT and regression after preprocessing.

This approach allowed us to test four distinct modules: preprocessing with BERT, preprocessing with regression, Universal Sentence Encoder with BERT, and Universal Sentence Encoder with regression. The categorized data is then leveraged to develop predictive models through machine learning techniques aimed at assessing public sentiment and anticipating customer issues within the telecommunications sector. This study seeks to advance preprocessing methodologies to improve decision-making processes and enhance customer satisfaction within the telecommunication industry segment.

Keywords: Social media analysis, telecom sector, customer problems, classification, machine learning, Customer churn prediction.

[1] Master's Researcher in Business Information Systems, Helwan University.
[2] Associate Professor at the Faculty of Computers and Artificial Intelligence, Beni-Suef University
[3] Lecturer in Business Information Systems Department, Faculty of Commerce and Business Administration, Helwan University.

نموذج ذكي مقترح للتنبؤ بمشاكل عملاء الاتصالات

الملخص

لقد تسارع انتشار البيانات على الإنترنت بشكل كبير بسبب ظهور منصات وسائل التواصل الاجتماعي على مدار العشرين عامًا الماضية. تعمل هذه المنصات كمصدر قيم للمعلومات التي يولدها المستخدمون، مع تميز تويتر بشكل خاص كمنصة تدوين مصغر شائعة توفر رؤى موجزة. ومع ذلك، فإن تحليل التعبيرات غير الرسمية من مثل هذه المنصات يمثل تحديات كبيرة، وخاصة في فهم وتحليل مخاوف العملاء داخل قطاع الاتصالات. يركز بحثنا على معالجة الجمل باللغة الطبيعية مسبقًا للمساعدة في الفهم والتحليل. لقد استكشفنا نهجين متميزين: استخدام Universal Sentence Encoderونماذج المعالجة المسبقة لإعداد التغريدات للتحليل. بالإضافة إلى ذلك، استخدمنا خوارزميات مثل BERT والانحدار بعد المعالجة المسبقة. سمح لنا هذا النهج باختبار أربع وحدات مميزة: المعالجة المسبقة باستخدامBERT ، والمعالجة المسبقة بالانحدار، و Universal Sentence EncoderمعBERT ، و Universal Sentence Encoderمع الانحدار. يتم بعد ذلك الاستفادة من البيانات المصنفة لتطوير نماذج تنبؤية من خلال تقنيات التعلم الآلي التي تهدف إلى تقييم مشاعر الجمهور وتوقع مشكلات العملاء داخل قطاع الاتصالات. تسعى هذه الدراسة إلى تطوير منهجيات المعالجة المسبقة لتحسين عمليات صنع القرار وتعزيز رضا العملاء داخل قطاع صناعة الاتصالات.

الكلمات المفتاحية: تحليل وسائل التواصل الاجتماعي، قطاع الاتصالات، مشاكل العملاء، التصنيف، التعلم الآلي، التنبؤ بانخفاض عدد العملاء.

1) Introduction

Telecommunications has become a key industry in developed countries, with heightened competition driven by technological advancements and an increase in operators. Businesses are employing various strategies to survive and thrive in this highly competitive market. Three primary techniques have been proposed to boost revenue: acquiring new clients, upselling to existing clients, and prolonging client retention [1]. The latter method, which emphasizes the lower cost of retaining an existing client compared to acquiring a new one, proves to be the most profitable strategy when considering return on investment value. Implementing this technique requires businesses to enhance customer loyalty by addressing their needs proactively before they even voice them. Understanding that consumers play a crucial role in sustaining the company's operations is essential for companies' long-term success. By being aware of customer concerns and taking steps to address them preemptively, telecom companies can better prevent customers from switching to competitors and develop effective retention strategies[2].

In the telecommunications sector, focusing on understanding customer needs through analytics and tailoring services accordingly is crucial for establishing a customer-centric business approach globally [3]. Offering exceptional service quality, flexibility in addressing consumer requirements, convenience-oriented services as well as pricing models tailored towards customer satisfaction are vital aspects for achieving higher profitability with current customers amidst intense market competition. Customer retention, loyalty, and satisfaction serve as significant intermediate goals for telecommunication network operators seeking superior economic success[4] [5].

The surge in the number of customers using the communication sector and companies has led to a heightened level of competition [6]. Understanding customer needs is crucial as they seek the latest trends at competitive prices, prompting businesses to vie for their loyalty [7]. Companies are focused on retaining existing customers due to their value and cost-effectiveness compared to acquiring new ones [8]. Customer retention strategies include fostering loyalty among customers who can also serve as brand ambassadors through word-of-mouth marketing [9]. It is imperative for companies to predict customer churn early on to mitigate revenue loss, with research highlighting machine learning's efficacy in this area [10][11].

Social media platforms like Twitter provide valuable data that offers insights into human psychology and consumer behavior[12], particularly important for telecommunications firms seeking a better understanding of customer issues[13]. Analyzing tweets using machine learning and natural language processing enables a deeper comprehension of individual or group sentiments, although it poses various challenges such as handling large-scale data collection efficiently[14][15].

Given social media's growing influence as an avenue for public expression and idea exchange, real-time sentiment analysis presents numerous opportunities[16] with its low-cost yet effective monitoring capabilities across different platforms like Facebook[17], Twitter, Reddit, Instagram, news forums etc[18]., making it relevant across diverse sectors globally [19][20].

Pre-training language models have demonstrated substantial success in learning universal language representations, significantly advancing the field of natural language processing (NLP). Among these models, BERT (Bidirectional Encoder Representations from Transformers) stands out as a pioneering approach, achieving state-of-the-art results in a wide array of

language understanding tasks. BERT's ability to capture bidirectional context in text has enabled it to outperform previous models, establishing new benchmarks in tasks such as question answering, sentiment analysis, and language inference. This success underscores the importance of pre-training in developing versatile and powerful language models [30].

2) literature review

This section provides an overview of research on predicting and resolving customer issues using machine learning and deep learning. The aim is to give a comprehensive overview of existing developments in this field. As businesses strive to enhance customer satisfaction and loyalty, efficiently understanding and addressing customer issues has become paramount. This review will explore various methodologies and models previously proposed for predicting and resolving customer issues, highlighting their strengths and limitations.

2.1 Customer churn in telecommunication

Essam Abou el Kassem, Shereen Ali Hussein, Alaa Mostafa Abdelrahman, and Fahad Kamal Alsheref conducted a study on "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content." The study explored two methods for predicting customer churn. The first method involved building a dataset from practical questionnaires and analyzing it using machine learning algorithms such as Deep Learning, Logistic Regression, and Naïve Bayes. The second method analyzed user-generated content (UGC), including comments, posts, and reviews, using sentiment analysis to assess text polarity (positive or negative). The findings showed that while the algorithms had similar accuracy, they varied in how they weighted attributes in the decision-making process. However, a limitation of the study is that it did not consider additional

attributes that could potentially improve the models' performance. [24]

Maryani and Riana conducted a study on customer clustering and profiling using the RFM method for generating recommendations in customer relationship management,2017. This study used transaction data from sales, consisting of 326 records, as the dataset. The research followed several stages: data collection, pre-processing, clustering using the k-means method, cluster validity testing, classification, and customer profiling. Recommendations for Customer Relationship Management (CRM) were then developed based on the results of clustering and classification. The study proposed a customer mapping application to assist companies in decision-making. Future research is encouraged to incorporate additional attributes beyond RFM, explore alternative algorithms, and integrate economic theories for further improvements. [2]

S. K. Alifah and N. A. Windasari, "Unlocking Loyalty beyond Connectivity: A Customer-Centric Approach through B2B Customer Experience Management in the Digital Telco Company," This study investigated the needs and challenges of SMEs in Indonesia by conducting in-depth interviews with nine respondents from a range of industries and professional roles. The research aimed to enhance the Customer Value Proposition by analyzing two user personas along the Customer Journey. Qualitative data was collected through 30 to 60-minute interviews with SME customers who had encountered service failures or requested support. Participants were selected through criterion and convenience sampling, and the data was analyzed using Nvivo software. The findings revealed that larger SMEs with higher-level customer entities tend to rely on relational partnerships and require tailored products and services.[3]

The authors recommend that future research should expand the sample size to better represent various company sizes and customer roles, including micro and large businesses, to gain deeper insights into B2B customer experience management (CXM).[3]

A. Amin, B. Shah, A. M. Khattak, F. J. Lopes Moreira, G. Ali, A. Rocha, and S. Anwar, "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods," International Journal of Information Management, 2019.The study utilized two datasets: Dataset-1 (target company) and Dataset-2 (source company), with Dataset-1 having fewer samples and attributes. Dataset-2 included 15,760 non-churn and 2,240 churn customers, while Dataset-1 had 2,850 churn and 483 non-churn customers. The research applied four data transformation methods—log, rank, box-cox, and Z-score—to develop a CCCP predictive model based on multiple classifiers, including Naive Bayes, KNN, Gradient Boosting Trees, and others from various machine learning families. The study aimed to address challenges in the telecommunications sector, especially for companies lacking historical data or facing data loss. The approach could also be adapted for other domains in the future.[33]

2.2 Universal Sentence Encoder in Prediction:

S. B. Majumder and D. Das, "Detecting Fake News Spreaders on Twitter Using Universal Sentence Encoder," Notebook for PAN at CLEF 2020. In this paper the study employed Google’s pre-trained sentence embeddings and an LSTM-based deep learning framework to identify fake news spreaders. The model, evaluated in both English and Spanish, achieved 64% accuracy for English and 80% for Spanish. After initial preprocessing, tweets were embedded and processed through the LSTM. While

the English model exhibited overfitting issues, indicating a need for further refinement, the Spanish model performed more effectively. Future research may explore alternative embeddings such as BERT, despite its limitation of handling only 512 words, with plans to address these constraints. The dataset comprised 100 tweets from each of 300 authors for both languages.[9]

Asgari-Chenaghlu, M., Nikzad-Khasmakhi, N., & Minaee, S.. "Covid-Transformer: Detection of Popular Subjects on Twitter Utilizing Universal Sentence Encoder." The study proposed a model using the Universal Sentence Encoder to detect key topics in tweets. Data was first obtained from Twitter using the Twitter API, followed by a data cleaning process. By deriving semantic representations of tweets and clustering them with k-means, the model groups similar tweets based on sentence-level similarity. A deep learning-based text summarization technique is then applied to each cluster to identify the underlying topics. The model was evaluated on a dataset of tweets and demonstrated its ability to effectively uncover informative topics. Additionally, the analysis aimed to detect trending topics and major concerns of Twitter users, enabling a better understanding of the situation and facilitating improved planning. The framework is also adaptable to other social media platforms and contexts beyond COVID-19, with no restrictions on data distribution. [28]

A. M. Qamar, S. A. Alsuhibany, and S. S. Ahmed, "Sentiment classification of Twitter data belonging to Saudi Arabian telecommunication companies," *Int. J. Adv. Comput. Sci. Appl.*, 2017. The paper examines English-language tweets from three telecommunications companies in Saudi Arabia—STC, Mobily, and Zain—aggregating a total of 1,331 tweets. The analysis utilizes the k-Nearest Neighbor (kNN) algorithm

for machine learning and data collection was performed using the InfoGainAttributeEval tool in WEKA. Future research could expand the dataset by including sentiment analysis of tweets written in Arabic, thereby increasing the dataset's size.[37]

Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in the big data platform. J Big Data 6, 28 (2019). This research aimed to develop a customer churn prediction system for SyriaTel, utilizing various types of telecom-related datasets including customer services and contract data, tower and complaint databases, network logs, call detail records (CDRs), and mobile IMEI information. The model employed machine learning techniques on a big data platform, emphasizing innovative feature engineering and selection. The model's performance was evaluated using the Area Under the Curve (AUC) metric, achieving a high AUC score of 93.3%, demonstrating its effectiveness in predicting customer churn [1].

2.3 Preprocessing in text classification and prediction:

M. Mali and M. Atique, "The Relevance of Preprocessing in Text Classification," in Proceedings of Integrated Intelligence Enable Networks and Computing, K. K. Singh Mer, V. B. Semwal, V. Bijalwan, and R. G. Crespo, Eds. Algorithms for Intelligent Systems. Springer, Singapore, 2021. This study explored the impact of preprocessing techniques such as stemming and lemmatization on the classification accuracy of unstructured data into predefined categories. The experiment was conducted using the 20-newsgroups dataset, which consists of approximately 20,000 documents classified into 20 different newsgroups. Python was used for the implementation, with the scikit-learn API employed to retrieve the data. A key challenge

in achieving high classification accuracy post-preprocessing is the high dimensionality of the feature space. This issue can be mitigated by applying dimensionality reduction techniques, such as genetic algorithms, to further enhance performance [16].

Muhittin and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 28, no. 3, article 15, 2020. This study investigated the impact of various text preprocessing techniques on classifier performance in predicting fine-grained review rating stars. The experiments focused on a five-class review rating system, evaluating how different preprocessing methods influenced classification accuracy. Techniques such as simple stopword elimination, lowercasing, and the removal of common words, combined with 1-to-3 n-grams, were found to enhance classification performance more effectively than other methods. These preprocessing steps significantly improved accuracy when predicting review ratings. The datasets used in this study were sourced from the Yelp Dataset Challenge, available at Yelp Dataset. The first dataset consists of full review text data, including the user ID of the reviewer and the business ID being reviewed, while the second dataset contains business-related data, including location, attributes, and categories [8].

D. Ramachandran and R. Parvathi, "Analysis of Twitter Specific Preprocessing Technique for Tweets," Procedia Computer Science, vol. 165, pp. 245-251, 2019. The objective of the experiment is to compare and assess the impact of two different preprocessing techniques. First, tweets posted during disasters are extracted from Twitter. These tweets are then processed using both techniques, which include steps such as tokenization, Part-Of-Speech tagging, and stopword removal.

The preprocessed tweets are classified into Disaster and Non-Disaster categories using machine learning algorithms. The performance of each technique is evaluated based on classification accuracy, and the results are analyzed to determine the effectiveness of the preprocessing methods. [23]

2.4 Using Machine Learning to keep user satisfaction in the telecom sector:

Alshamari MA. Evaluating User Satisfaction Using Deep-Learning-Based Sentiment Analysis for Social Media Data in Saudi Arabia's Telecommunication Sector. The research method involved preprocessing the data to remove irrelevant information and then applying several deep learning models—CNN, LSTM, BiLSTM, GRU, and CNN-LSTM—to compare classification outcomes. While the methodology can be adapted to various languages, the preprocessing steps differ by language. Using a dataset from Twitter, the study demonstrated the effectiveness of these models in predicting customer satisfaction from Arabic tweets. It also highlighted how social media platforms allow customers to share feedback, aiding businesses in enhancing service quality and fostering customer loyalty.[35]

S. Aftan and H. Shah, "Using the AraBERT model for customer satisfaction classification of telecom sectors in Saudi Arabia," *Brain Sciences*, vol. 13, no. 147, 2023. The study utilized a raw dataset of customer reviews from Saudi Arabian telecommunications companies, specifically AraCust, which includes STC, Zain, and Mobily. Arabic sentiment analysis was employed to assess customer satisfaction with these companies based on tweets. The research evaluated customer satisfaction using three methods: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and AraBERT. Future

work aims to extend the current research model to various applications, including NLP analysis, time series prediction, and classification of diverse datasets, leveraging AraBERT and other deep learning techniques [36].

A. M. Qamar, S. A. Alsuhibany, and S. S. Ahmed, "Sentiment classification of Twitter data belonging to Saudi Arabian telecommunication companies," *Int. J. Adv. Comput. Sci. Appl.*, 2017. The paper examines English-language tweets from three telecommunications companies in Saudi Arabia—STC, Mobily, and Zain—aggregating a total of 1,331 tweets. The analysis utilizes the k-Nearest Neighbor (kNN) algorithm for machine learning and data collection was performed using the InfoGainAttributeEval tool in WEKA. Future research could expand the dataset by including sentiment analysis of tweets written in Arabic, thereby increasing the dataset's size [37].

Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in the big data platform. J Big Data 6, 28 (2019). This research aimed to develop a customer churn prediction system for SyriaTel, utilizing various types of telecom-related datasets including customer services and contract data, tower and complaint databases, network logs, call detail records (CDRs), and mobile IMEI information. The model employed machine learning techniques on a big data platform, emphasizing innovative feature engineering and selection. The model's performance was evaluated using the Area Under the Curve (AUC) metric, achieving a high AUC score of 93.3%, demonstrating its effectiveness in predicting customer churn [1].

2.5 Research Gap:

This study aims to assist telecommunication companies in retaining customer loyalty by predicting potential issues at an early stage, preventing customers from switching to competitors. It also focuses on advancing preprocessing methodologies to enhance decision-making processes and improve customer satisfaction within the telecom sector. By refining these techniques, the study seeks to provide a solid framework for analyzing customer feedback, enabling companies to respond promptly and effectively to customer needs.

3) BACKGROUND OF TECHNIQUE

3.1 Introduction

Machine learning and deep learning, once long-standing concepts, have recently undergone a dramatic transformation due to the surge in big data and advancements in computing power. Their growing popularity across sectors such as healthcare, finance, and retail stems from their capability to analyze extensive datasets, deliver precise predictions, and reveal insights that were previously beyond reach. [32-33]

In recent years, the field of artificial intelligence has rapidly advanced, primarily driven by the development of machine learning and deep learning. The exponential increase in data and improvements in computational power have highlighted the potential of these technologies to revolutionize various industries and reshape our world. As data continues to expand, the capacity of machine learning and deep learning to foster innovation and provide unparalleled insights becomes increasingly apparent [31].

3.2 Machine Learning Algorithm

Machine learning enables computer systems to carry out tasks by identifying patterns and making inferences, without relying on explicit programming. The goal is to allow computers to learn from data and generate predictions or decisions. This approach involves creating models that can generalize from the training data to new, unseen data, allowing systems to improve their task performance over time.[39]

Machine learning involves several distinct approaches: [40]

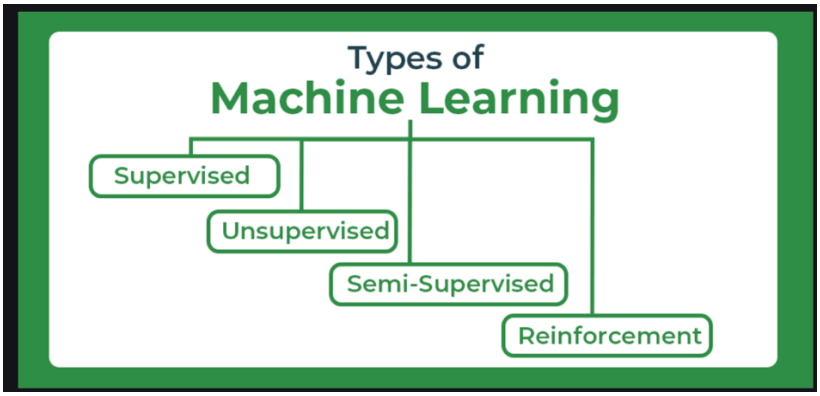


Figure 1 Types of ML [41]

 Supervised Learning: This approach uses labeled data to train models for making predictions. It includes: [40]

- **Classification**: Determines categorical outcomes (e.g., filtering spam from non-spam emails).
- **Regression**: Predicts continuous outcomes (e.g., estimating property values based on various attributes).

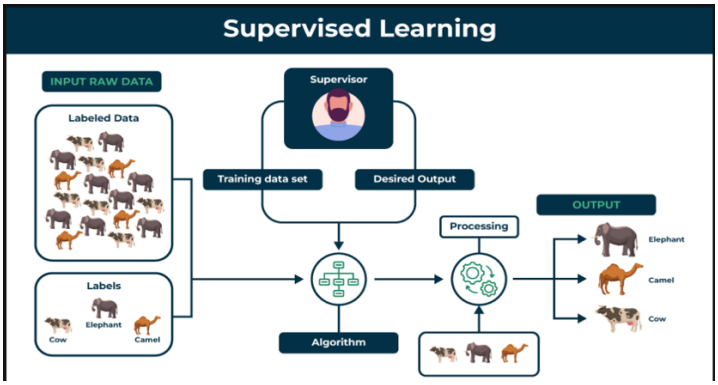


Figure 2 Supervised learning (ML types) [41]

 Unsupervised Learning: This method uncovers patterns in data without predefined labels. It includes: [40]

Clustering: organizes data into groups based on similarity (e.g., categorizing customers)

Dimensionality reduction: reduces the number of features in the data while preserving its essential structure (e.g. through principal component analysis).

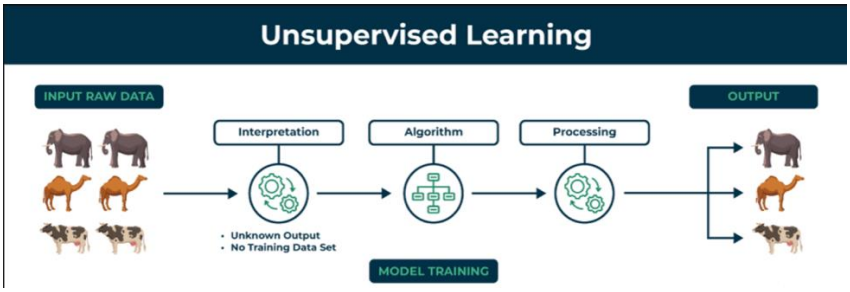


Figure 3 Unsupervised learning (ML types) [41]

Reinforcement Learning: This approach learns by interacting with an environment and receiving feedback, aiming to maximize cumulative rewards based on actions taken. [40]

3.3 Deep Learning Algorithm

DL, a subset of ML, is primarily focused on the development of artificial neural networks that can process vast amounts of data to make predictions or informed decisions. These networks are designed to mimic the human brain's processing capabilities, enabling the analysis of complex patterns and relationships within data [36]. On the other hand, NLP, which has seen growing interest in recent years, centers on the interaction between computational systems and human language. It aims to enable computers to understand, interpret, and generate human language in a way that is both meaningful and useful, making it a crucial area for advancements in AI and human-computer interaction.

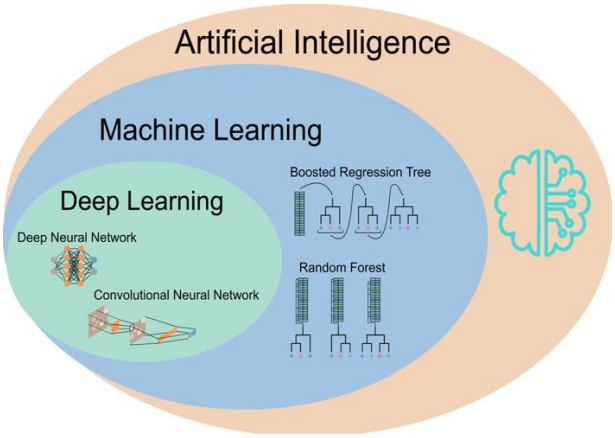


Figure 4 machine learning and deep learning [38]

Deep learning involves the utilization of networks to teach machines how to carry out tasks independently. [25] The "Sklearn model selection" comprises essential tools and functions for enhancing and assessing machine learning models [26]. A fundamental principle in machine learning is dividing the dataset into various segments for training and testing. Scikit-learn offers a convenient method, train_test_split(X, y, test size=0.25), designed to aid in this process. Here, X represents the design matrix or predictor dataset while y denotes the target variable [27].

Machine learning and deep learning are two separate domains within artificial intelligence. Machine learning handles tasks like image recognition, natural language processing, and predictive modeling. In contrast, deep learning is applied to more intricate tasks, such as object recognition, speech recognition, and self-driving cars. Both machine learning and deep learning have the potential to transform multiple industries, including healthcare, finance, and retail [31].

3.4 Universal Sentence Encoder

The Universal Sentence Encoder model transforms text input into high-dimensional vectors. These encoder models are designed to capture the meaning of sequences of words rather than just individual sentences. They have been trained and refined for processing content that extends beyond single words, including sentences, phrases, or paragraphs as illustrated in Figure 5 [17] [29].

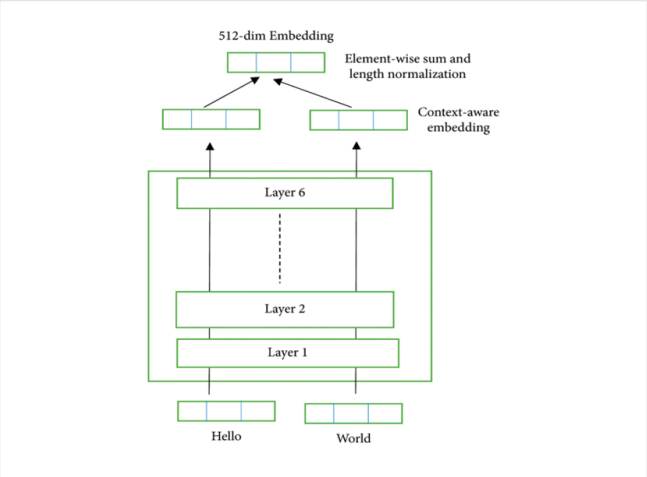


Figure 5 Transformer encoder model [17]

3.5 BERT (Bidirectional Encoder Representations from Transformer)

The BERT model offers several notable benefits. Firstly, it has demonstrated superior performance compared to traditional NLP approaches across various scenarios, highlighting its effectiveness in tackling average NLP problems more efficiently than classical methods. Empirical evidence supports BERT's advantages, reinforcing its status as a robust choice for NLP tasks. Additionally, BERT's effective use of transfer learning, enabled by its pre-training, proves particularly beneficial for smaller datasets, as it leverages previously acquired knowledge to enhance performance. Despite its strengths, there is potential for further improvement in BERT's performance through hyperparameter tuning, specifically using Bayesian optimization, to refine its application for new NLP tasks [34].

3.6 Evaluation metrics

We used some metrics to evaluate our models as Accuracy, Recall, and Precision:

3.6.1 Accuracy:

a key metric in model evaluation reflects the model's ability to correctly classify instances. It quantifies the ratio of accurately predicted instances to the total predictions made.

3.6.2 Precision:

another crucial metric, measures the ratio of correctly predicted positive observations to the total predicted positives. It is particularly useful in scenarios where the cost of false positives is high.

3.6.3 Recall:

another vital metric, measures the ratio of correctly predicted positive observations to all actual positives, indicating the model's ability to capture all relevant instances.

4) MATERIAL AND METHOD

This section outlines the materials and methodologies utilized in our study, encompassing the dataset and procedures of the envisioned model.

- **Dataset**

A total of 1,000 tweets were collected over six months using Twitter Archiver, from October 9th, 2021, to March 6th, 2022. Tweets were filtered by hashtags like "Vodafone," "telecom," and "Etisalat," and stored in CSV format, consisting of 21 attributes (see Figure 6).

Date	Screen Name	Full Name	Tweet_Text	Tweet ID	Link(s)	Media	Location	Retweets	Favorites	App	Followers	Follows	Listed	Verfied	User Since	Location	Bio	Webs	Timezone	Profile Image
17/02/2	@IRT_BCom	b-com / soo	soon at #Mobile	1.49E+18	https://	pic.twitter.com/nwK\		0	0	Twitt	4343	533	602		5/5/12	France	French li	https://b-com.com		View
17/02/2	@TeletekSi	Teletek Struc	The #women of	1.49E+18	https:// https://	https://pbs.twimg.co		0	1	Twitt	229	340	7		4/26/15	Woolwich,	Teletek S	http://www.teletek		View
17/02/2	@HtBTweets	Hans ten Be	[#5GRealtimeEr	1.49E+18	https://	https://pbs.twimg.co		0	0	Pow	578	1222	48		3/1/09	Netherland	Hans ten Berge working at			View
17/02/2	@Business_Pro	Christian Lar	[#5GRealtimeEr	1.49E+18	https://	https://pbs.twimg.co		0	0	Pow	773	135	1004		6/10/16	Ile-de-Fran	Project Management, Busir			View
17/02/2	@Business_Pro	Christian Lar	[#5GRealtimeEr	1.49E+18	https://	https://pbs.twimg.co		0	0	Pow	773	135	1004		6/10/16	Ile-de-Fran	Project Management, Busir			View
17/02/2	@depeekii	Darryl E. Pee	Edge cloud play	1.49E+18	https://google.smh.re/0bp_			1	0	Sma	1978	4847	10		5/23/13	Washingtor	Husband	https://cloud.goo		View
17/02/2	@Phedro_L	Pedro L.	[#5GRealtimeEr	1.49E+18	https://	https://pbs.twimg.co		0	1	Pow	188	1348	0		9/25/13	México				View
17/02/2	@IndianTech_1	Piyush Bhas	OnePlus TV Y1s	1.49E+18	https:// https://	https://pbs.twimg.co https://pbs.twimg.co		0	1	Twitt	631	98	5		4/5/20	Maharashtr	18y/o | L	http://linktr.ee/Indi		View
17/02/2	@MSafe369	MoneySafe3	#Telecom live e	1.49E+18		https://pbs.twimg.co		0	0	Twitt	7309	7714	5		4/19/20	Italy	Not havi	https://finviz.com/		View
17/02/2	@capacitymedi	Capacity Me	#CapacityNews	1.49E+18	https://	https://pbs.twimg.co		0	1	Sprii	10265	4646	308		2/21/11	London	Your ess	http://www.capac		View
17/02/2	@AugeMelanie	Mélanie Aug	[#5GRealtimeEr	1.49E+18	https://	https://pbs.twimg.co		0	0	Pow	129	163	85		4/20/15					View
17/02/2	@FSRComsMe	FSR ComsM	#ThrowbackTh	1.49E+18	https://tinyurl.com/4h3ky6vb https://tinyurl.com/yu8nt4ch			0	0	Twitt	959	783	30		6/8/12		Florence	http://fsr.eui.eu/cc		View

Figure 6 Data Set Example

- **Data preprocessing**

We utilized Scikit-learn, a widely used Python package offering a diverse set of utilities for applying different machine-learning methods and strategies. Before using Scikit-learn, we conducted preprocessing on the tweets including the elimination of noise by removing stop words, URLs, mentions, HTML tags, special characters, and punctuation. Additionally, we converted the text to lowercase and applied stemming. Following this pre-processing step, sentiment analysis was performed on the tweets.

Table 1Summary of the data preprocessing process

operation	original	cleaned
Removing numbers, ID and mobile numbers	Vodafone Idea: The firm's board has approved a proposal to raise funds upto ,Çπ14500 crores through various means. #Vodafone	Vodafone idea firm board approved proposal raise funds upto crores various means.
Removing symbols, hashtag and dollar sign	#BREAKING #Vodafone mobile network does not work in #Berdyansk, said the head of the #Zaporizhzhia region	Mobile network not work said head region.
Remove extra white spaces, punctuation and apply lower casing	RIP Vodafone network, #Vodafone #vodafoneidea	Rip vodafone network.

- ***Data Labeling***

Data labeling, also known as data annotation, is the process of assigning meaningful tags or labels to raw data. This labeling is typically done to create labeled datasets, which are used to train machine learning models. In data labeling, we assign labels to data points based on predefined criteria or guidelines. These labels can represent various attributes or characteristics of the data, such as "services, network, internet, or no issues". Data labeling is a crucial step in supervised learning, as it provides

the ground truth labels needed for training and evaluating machine learning algorithms.

1) Network → This tweet indicates a problem related to network connectivity.
2) Service → This tweet suggests there may be a problem with the services provided by a telecommunication company.
3) Internet → This tweet implies there might be an issue with the internet connection.
4) No → This tweet indicates that everything seems to be working fine, without any reported issues.

5) Proposed models

In this section, we present four proposed models for predicting the various issues faced by customers. These predictive models aim to assist the company in understanding and addressing customer problems more efficiently. The models are based on different combinations of pre-processing techniques and machine learning algorithms, including BERT and regression methods. Specifically, the models are:

Model 1: Preprocessing and Regression
Model 2: Preprocessing and BERT
Model 3: Universal Sentence Encoder (USE) and Regression Algorithm
Model 4: Universal Sentence Encoder (USE) and BERT
Detailed Steps for Model 1 and Model 2:
Data Collection from Twitter:

Twitter serves as a platform where users regularly share their opinions on a wide range of topics, from personal experiences to reactions to current events. This real-time and open exchange of viewpoints makes Twitter a valuable resource for collecting public opinion data. The dataset was sourced from Twitter and

includes tweets from countries across Europe, Africa, Asia, America, and the Middle East. The initial classification organizes these tweets by sentiment, categorizing them as either positive or negative based on regional perspectives.

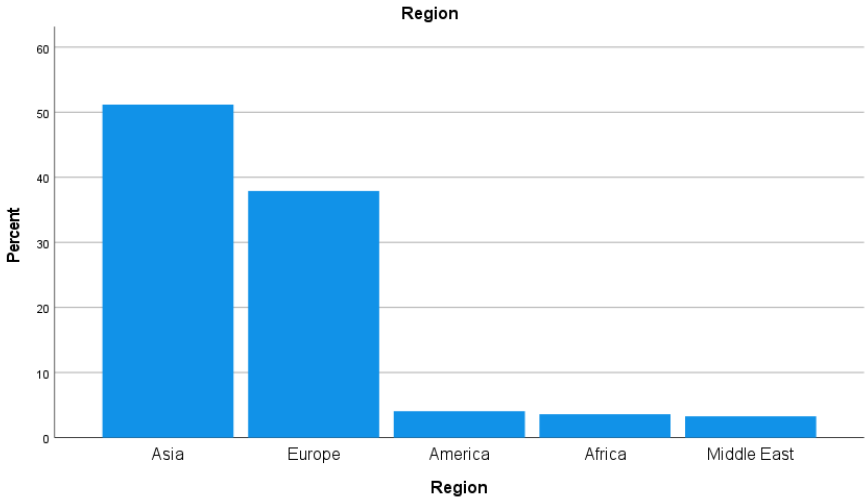


Figure 7 regions classification

Table 2 region classification percentage

			Opinion		
			Negative	Positive	Total
Region	Asia	Count	122	206	328
		% within Region	37.2%	62.8%	100.0%
	Europe	Count	100	143	243
		% within Region	41.2%	58.8%	100.0%
	America	Count	11	15	26
		% within Region	42.3%	57.7%	100.0%
	Africa	Count	13	10	23
		% within Region	56.5%	43.5%	100.0%
	Middle East	Count	9	12	21
		% within Region	42.9%	57.1%	100.0%

Total	Count	255	386	641
	% within Region	39.8%	60.2%	100.0%

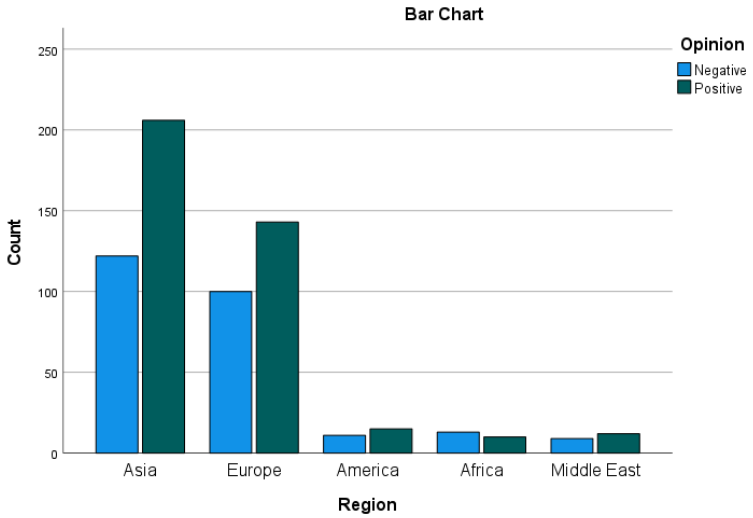


Figure 8 regions classification (positive and negative opinions

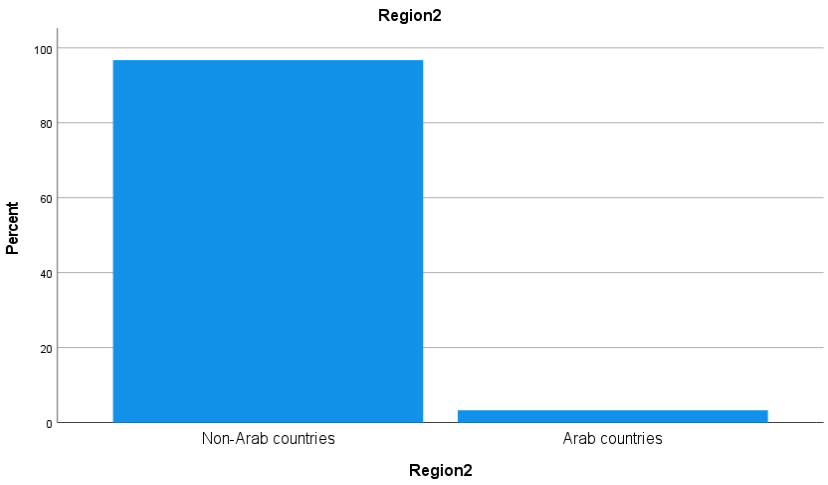


Figure 9 Arab and non-Arab classification

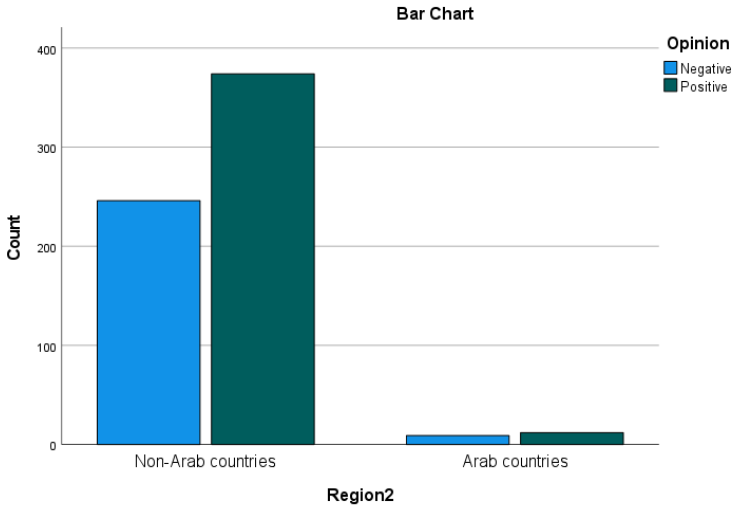


Figure 10 Arab and non-Arab classification (positive and negative opinions)

Table 3 percentage of Arabic and non-Arabic

			Opinion		
			Negative	Positive	Total
Region2	Non-Arab countries	Count	246	374	620
		% within Region2	39.7%	60.3%	100.0%
	Arab countries	Count	9	12	21
		% within Region2	42.9%	57.1%	100.0%
	Total	Count	255	386	641
		% within Region2	39.8%	60.2%	100.0%

Data Compilation and Storage:

After retrieving user data and opinions from Twitter, we meticulously compiled the information into a structured format. Subsequently, we stored the organized data in CSV format, as shown in Figure 6, for efficient management and analysis. This approach facilitated easy access and streamlined processing of the vast dataset collected from the platform.

Data Pre-processing:

Following the data collection phase, the next step involves pre-processing the acquired data per the methods outlined in the

data pre-processing section (4.b). This process entails various steps such as cleaning, tokenization, and normalization to ensure the data is in a standardized and analyzable format. By adhering to these pre-processing procedures, we aim to enhance the quality and reliability of the dataset, enabling more accurate analysis and interpretation of the collected information.

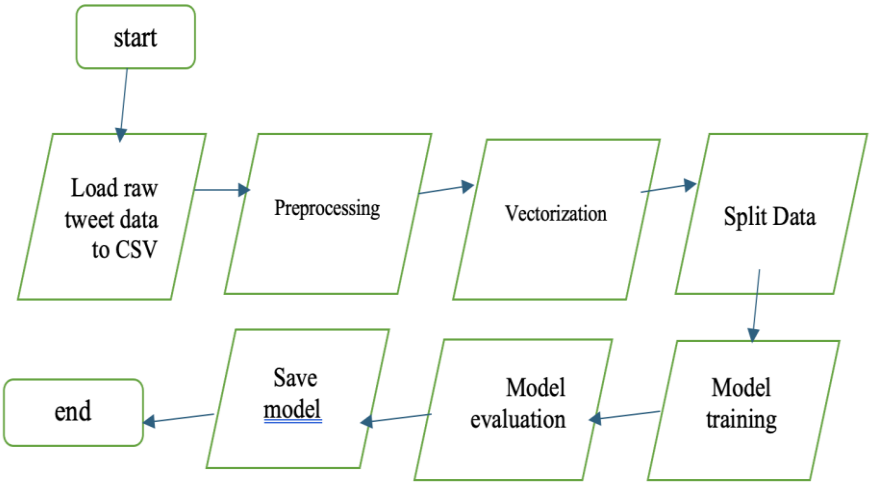
Model 1: Pre-processing and Regression:

In this model, after preprocessing the collected data, then followed by applying regression algorithms from Sklearn to predict customer issues.

Model 2: Pre-processing and BERT:

In this model, the data is pre-processed as described above. Following this, and then BERT (Bidirectional Encoder Representations from Transformers) is used for further encoding. The encoded data is then fed into Sklearn's model selection pipeline for prediction.

The details of our first two proposal are presented in Figure 12.



Process Flow for Model 1: Preprocessing and Regression

المصدر : الباحثة

Data Pre-processing Model Steps:

a) Load tweets data

The dataset was collected entirely from Twitter, with 96% of the samples coming from non-Arabic countries. A large portion of these samples originates from Asia, particularly India, followed by several European countries. Most reported issues are service-related, specifically concerning the services offered by telecommunications companies. These are followed by problems related to internet connectivity and, lastly, network issues.

b) Preprocessing as shown in figure 11

i. Remove NaN values and duplicates.
ii. Clean tweet text (remove URLs, mentions, hashtags).
iii. Remove stop words.
iv. Remove extra white spaces and punctuation.
v. Convert text to lowercase.

An example for a tweet that makes some preprocessing steps:

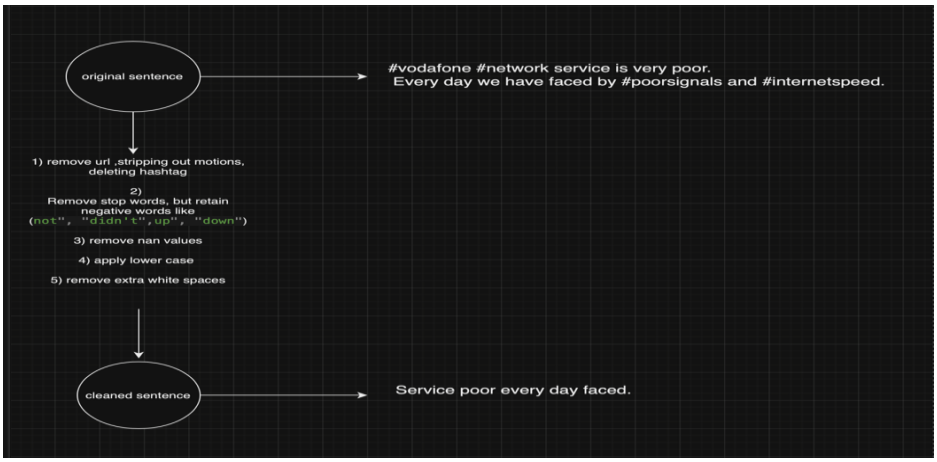


Figure 11 pre-processing steps.

c) Vectorization

i. Initialize CountVectorizer

ii. Fit and transform text data to create Bag-of-Words representation
iii. Save CountVectorizer for future use.

d) **Split data** (*Split dataset into training and testing sets.)*

To partition the dataset into training and testing sets, the train_test_split function from the scikit-learn library was utilized. This method splits the data into two subsets: 80% of the data is allocated to training, while the remaining 20% is used for testing. The test_size=0.2 parameter defines this split ratio. The training set is used to fit the machine learning model, allowing the model to learn patterns in the data. The testing set is kept separate and serves as an independent evaluation set, ensuring that the model is tested on unseen data to assess its generalization performance.

The parameter random_state=42 was specified to make the split deterministic, ensuring that the results are reproducible. This means that every time the code is run, the data will be split in the same way, enabling consistent comparisons across experiments. By using this function, both the input features (x) and the target labels (y) are divided into x_train, x_test, y_train, and y_test, ensuring that the model can be trained on one set and evaluated on another, which is crucial for preventing overfitting and ensuring robust model performance.

e) Model training

i. Initialize Multinomial Naïve Bayes classificatier.
ii. Train the classifier on the training data.

f) Model Evaluation:

i. Predict labels for test data.

ii. Calculate accuracy, precision, and recall scores.
iii. Print evaluation metrics.

g) Save Model:

i. Save the trained model for future use.

In this step, we train a Logistic Regression model from the scikit-learn library using the pre-processed data obtained from the previous step. The data has been manually labeled according to sentiment. By utilizing this labeled data in the training process, we aim to optimize the performance of the Logistic Regression model. This training phase allows the model to learn from the labeled examples, enhancing its ability to accurately classify sentiments in new, unseen data.

Finally, after completing the training process with the scikit-learn's model selection framework, we obtain a trained algorithm ready for use in prediction tasks. This trained algorithm has learned from the labeled data provided during the training phase and is equipped to make predictions on new, unseen inputs. With this trained algorithm, we can now deploy it in real-world applications to analyze text data, classify information, or make predictions based on the learned patterns, thereby facilitating various decision-making processes and generating valuable insights.

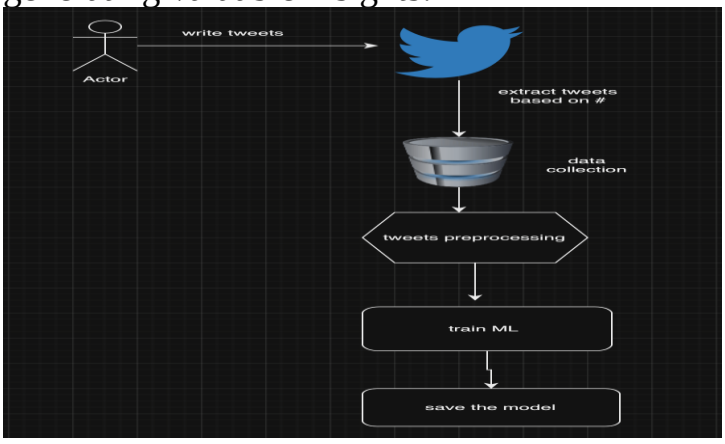


Figure 12 Proposed Prediction Model

Detailed Steps for Model 1 and Model 2:

On Twitter, people talk about everything under the sun. From sharing personal stories to discussing the latest news, it's like a giant chat room where everyone's opinions are on full display. This makes Twitter a goldmine for understanding what people think about various topics.

After retrieving user data and opinions from Twitter, we carefully gathered all the details and neatly organized them into a structured layout. Afterwards, we saved this organized data in CSV format, just like how it's illustrated in Figure 6.

We carefully went through each tweet, reading and understanding its message. This took a lot of time and effort, but it helped us categorize them accurately into labels such as 'service,' 'network,' 'internet,' and 'no issue.'

Then, we used a tool called the Universal Sentence Encoder (USE) from Google. It helps us turn tweets into easy-to-understand codes without needing to do a lot of extra work upfront. These codes retain all the important meanings from the tweets and are versatile, making our process faster and simpler.

For the third model, we combined the Universal Sentence Encoder with a Logistic Regression algorithm. This approach allows for a straightforward yet effective application of machine learning techniques on the encoded tweet representations, leading to efficient sentiment classification.

For the fourth model, we utilized these encoded tweets in combination with BERT (Bidirectional Encoder Representations from Transformers). This model leverages the comprehensive contextual information captured by BERT to further enhance the classification performance.

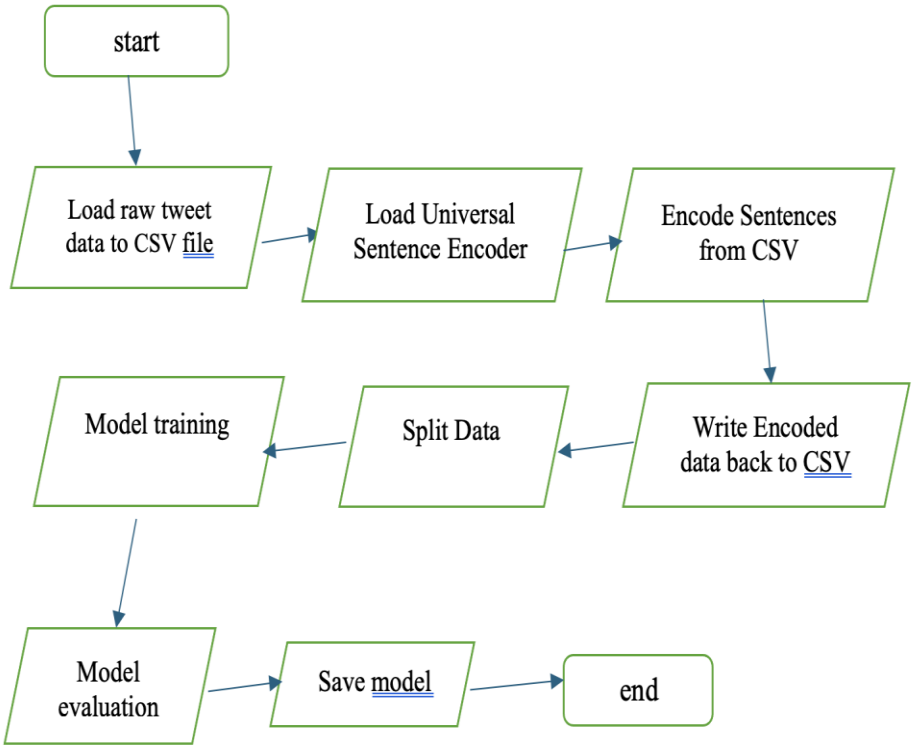
Model 3: Universal Sentence Encoder (USE) and Regression:

In this model, the USE is used to encode the textual data, and regression algorithms from Sklearn are applied to predict customer issues.

Model 4: Universal Sentence Encoder (USE) and BERT:

This model combines the Universal Sentence Encoder (USE) for initial text encoding, followed by further encoding using BERT. The final encoded data is then used for prediction using Sklearn's model selection.

By exploring these diverse approaches, we aim to determine the most effective method for predicting customer issues, thereby enhancing the company's ability to understand and address customer problems more efficiently.



Process Flow for Model 4: Universal Sentence Encoder and BERT

Data Encoder Model Steps:

a) Load tweets data

b) Load Universal Sentence Encoder: load the Universal Sentence Encoder (USA) module from TensorFlow Hub. This module is used to encode text data into fixed-dimensional vectors.

c) Encode Sentences from CSV:

- Read the CSV file containing the tweet data.
- Iterate over each row in the CSV file.
- Extract the text from the specified column ('Tweet_Text').
- Encode the text using the Universal Sentence Encoder.
- Append the encoded text to a list.

d) Write Encoded Data Back to CSV:

1) Update the header of the CSV file to include the new column ('encoded_text').
2) Write the updated rows (including encoded text) back to the same CSV file.

e) Update the CSV file with the encoded text by adding a new column ('encoded_text') containing the encoded vectors.

f) Write Encoded Data Back to CSV:

1) Update the header of the CSV file to include the new column ('encoded_text').
2) Write the updated rows (including encoded text) back to the same CSV

g) Split data

1) Split dataset into training and testing sets.

h) Model training

1) Initialize Multinomial Naïve Bayes classificatier.
2) Train the classifier on the training data.

i) Model Evaluation:
   1) Predict labels for test data.
   2) Calculate accuracy, precision, and recall scores.
   3) Print evaluation metrics.

j) Save model
   1) Save the trained model for future use.

After finishing the training with scikit-learn's model selection, we get a trained algorithm ready for predictions. This algorithm learned from labeled data during training and can now predict outcomes for new inputs. With this trained model, we can analyze text, classify information, and make predictions in real-world scenarios. One advantage of using the Universal Sentence Encoder model is its ability to capture the meaning of text accurately, making our predictions more reliable and insightful.

6) Evaluation Metrics

In this part, we employed several evaluation metrics to assess the performance of our proposed models. These metrics provide valuable insights into the model's effectiveness in making predictions and capturing the underlying patterns in the data.

a) Accuracy: In our assessment, we compared the accuracy of two primary models: the Encoder Model and the Pre-processing Model.

The table presents the accuracy of these models applied to a specific task, further categorized based on the use of two algorithms: Regression and BERT (Bidirectional Encoder Representations from Transformers).

- Encoder Model - Regression: This model, using a regression algorithm, achieves an accuracy of 85%.
- Encoder Model - BERT: With the BERT algorithm, this model attains an accuracy of 66%.

- Pre-processing Model - Regression: Applying the regression algorithm to pre-processed data, this model reaches an accuracy of 74%.
- Pre-processing Model - BERT: Using the BERT algorithm on pre-processed data, this model achieves an accuracy of 82%.

The data indicates that the Encoder Model with Regression has the highest accuracy (85%), while the Pre-processing Model with BERT also performs notably well with an accuracy of 82%. These results underscore the varying levels of effectiveness of the different model and algorithm combinations for the task at hand.

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \times 100$$

Where, TP = true positives (correctly predicted positives)

TN = true negative (correctly predicted negatives)

FP = False Positives (incorrectly predicted as positive)

FN = False Negatives (incorrectly predicted as negative) [32].

b) Recall:

Recall, another vital metric, measures the ratio of correctly predicted positive observations to all actual positives, indicating the model's ability to capture all relevant instances.

- Encoder Model - Regression: 67%
- Encoder Model - BERT: 25%
- Pre-processing Model - Regression: 65%
- Pre-processing Model - BERT: 56%

The Encoder Model with Regression has a recall of 67%, indicating a strong ability to identify positive instances, while

the Encoder Model with BERT has a lower recall of 25%. The Pre-processing Model with Regression has a recall of 65%, and the Pre-processing Model with BERT has a recall of 56%.

$$\text{Recall} = \frac{TP}{TP+FN},$$

Where, TP = true positives
FN = False Negatives [32].

c) Precision:

Precision, another crucial metric, measures the ratio of correctly predicted positive observations to the total predicted positives. It is particularly useful in scenarios where the cost of false positives is high. The precision values for the different models are as follows:

- Encoder Model - Regression: This model achieves a precision of 89%.
- Encoder Model - BERT: With the BERT algorithm, this model attains a precision of 16%.
- Pre-processing Model - Regression: Applying the regression algorithm to pre-processed data, this model reaches a precision of 60%.
- Pre-processing Model - BERT: Using the BERT algorithm on pre-processed data, this model achieves a precision of 62%.

$$\text{Precision} = \frac{TP}{TP+FP}.$$

Where, TP = true positives
FP = False Positives [32].

Table4 Evaluation Metrics results

Model	Accuracy	Precision	Recall
Encoder Model- Regression	85%	89%	67%
Encoder Model- BERT	66%	16%	25%
Pre-processing Model- Regression	74%	60%	65%
Pre-processing Model- BERT	82%	62%	56%

In this table:

- Model: Name or identifier of the model being evaluated.
- Accuracy: Percentage of correctly predicted instances out of the total instances in the dataset.
- Precision: proportion of true positive predictions out of all positive predictions made by the model.
- Recall: Proportion of true positive predictions out of all actual positive instances in the dataset.

Explanation:

a. **Encoder Model - Regression**:
   - **Accuracy**: At 85%, this model demonstrates strong overall performance in predicting correctly.
   - **Precision**: With an 89% precision score, it excels at correctly identifying true positives with minimal false positives.
   - **Recall**: The recall is moderate at 67%, indicating that while it performs well in many cases, it misses some true positives.

b. **Encoder Model - BERT**:
   - **Accuracy**: At 66%, this BERT-based encoder model has relatively low performance overall.
   - **Precision**: Precision is notably low at 16%, indicating that a large proportion of predicted positives are incorrect.

- **Recall**: With a 25% recall, this model also struggles to identify actual positives effectively, highlighting its limitations in both precision and recall.

c. **Pre-processing Model - Regression**:

- **Accuracy**: The model achieves 74% accuracy, reflecting reasonable performance.
- **Precision**: At 60%, its precision is decent, but it still allows for a higher number of false positives compared to the Encoder Regression Model.
- **Recall**: With a recall of 65%, it effectively captures the majority of true positives, making it relatively balanced across metrics.

d. **Pre-processing Model - BERT**:

- **Accuracy**: The BERT-based pre-processing model has a strong accuracy of 82%, indicating overall reliability in its predictions.
- **Precision**: At 62%, it shows better precision than the BERT encoder model, meaning fewer false positives.
- **Recall**: With a recall of 56%, it captures a decent proportion of true positives but misses a few, showing room for improvement.

Summary:

- The **Encoder Model - Regression** stands out with the highest precision and accuracy, making it the best at minimizing false positives and performing reliably overall.
- **BERT-based models** tend to underperform in precision and recall, especially in the encoder model, but the pre-processing BERT model achieves solid accuracy, indicating potential depending on the task.
- The **Pre-processing Model - Regression** balances accuracy, precision, and recall reasonably well, while the **Encoder Model - BERT** struggles with both precision and recall, showing significant limitations.

7) CONCLUSION

In our quest, a new model was proposed to be used in the classification of customers' problems. The most important source of income for communication companies is the customers. So, we build a model to help companies predict the problems of customers based on their comments.

So, we discovered the secrets of tweets, we enlisted the help of two tools: the Universal Sentence Encoder (USE) and a tweet-specific multi-preprocessing model.

So, what's the big thing about the Universal Sentence Encoder? It's like a Twitter superhero. Unlike normal helpers, USE understands tweets rather than simply tidying them up! It's like giving our model unique glasses for understanding the true meanings of words and sentences. This made our predictions highly accurate and enhanced our tweet understanding skills. On the other hand, our multi-preprocessing tool is good at dealing with hashtags, mentions, and messy text.

Evaluating models using many measures (accuracy, precision, and recall) provides a thorough picture of their performance. Here's an overview of our findings and conclusions:

- Encoder Model - Regression: This model exhibits the highest accuracy (85%) and precision (89%), along with a strong recall (67%). This indicates that it not only correctly classifies a high percentage of instances but also maintains a high ratio of true positive predictions. Therefore, this model is highly effective in scenarios where both accuracy and precision are critical.
- Encoder Model - BERT: With lower performance across all metrics (66% accuracy, 16% precision, and 25% recall), this model struggles with both correctly classifying instances and maintaining a high ratio of true positives. Its low precision suggests issues with false positives, and the low recall indicates it misses many relevant instances. This model may require further optimization or might not be suitable for applications where high precision and recall are essential.

- Pre-processing Model - Regression: This model shows balanced performance with 74% accuracy, 60% precision, and 65% recall. It provides a reliable option with moderate performance across all metrics, making it suitable for applications where a balance between these metrics is needed.
- Pre-processing Model - BERT: Demonstrating good performance with 82% accuracy, 62% precision, and 56% recall, this model is a strong contender. While its recall is slightly lower, the high accuracy and reasonable precision make it a viable choice for applications that prioritize accurate classification and positive prediction rates.

The Encoder Model with Regression outperforms the other models, notably in terms of accuracy and precision. However, the Pre-processing Model with BERT performs well and can be considered a feasible option, particularly when accurate and precise results are required. These findings underscore the need of employing a variety of criteria to select the best model for various tasks, ensuring that the chosen model meets the application's particular demands.

Funding statement no funding received.

Data availability statement the data supporting the present research can be obtained by reaching out to the corresponding author via email or through the contact information provided in the paper.

Declaration of competing interest the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8) References

1) Ahmad, A.K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6, 28.
2) Maryani, I., & Riana, D. (2017). Clustering and profiling of customers using RFM for customer relationship management recommendations. In 2017 5th International Conference on Cyber and IT Service Management (CITSM).
3) Alifah, S.K., & Windasari, N.A. (2024). Unlocking loyalty beyond connectivity: A customer-centric approach through B2B customer experience management in the digital telco company. American International Journal of Business Management, 7(7), 236-246.
4) Ciancarini, P., et al. (2013). Semantic annotation of scholarly documents and citations. In International Conference of the Italian Association for Artificial Intelligence.
5) Gerpott, T.J., Rams, W., & Schindler, A. (2001). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunications Policy, 25, 249-269.
6) DOAA, M.E., & MOHAMED, H. (2016). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 4.
7) Fares, N., Lebbar, M., & Sbihi, N. (2019). A customer profiling machine learning approach for in-store sales in fast fashion. In M. Ezziyyani (Ed.), Advanced Intelligent Systems for Sustainable Development (AI2SD'2018) (Vol. 915, pp. 602-611). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-11928-7_53.
8) Muhittin, & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. Turkish Journal of Electrical Engineering and Computer Sciences, 28(3), Article 15.
9) Majumder, S.B., & Das, D. (2020). Detecting fake news spreaders on Twitter using universal sentence encoder. Notebook for PAN at CLEF 2020.
10) Wei, C.P., & Chiu, I.T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications, 23(2), 103-112.

11) Amin, A., et al. (2019). Customer churn prediction in the telecommunication industry using data certainty. Journal of Business Research, 94, 290-301.
12) Qureshii, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., & Rehman, A. (2013). Telecommunication subscribers churn prediction model using machine learning. In Eighth International Conference on Digital Information Management (pp. 131-136).
13) Ullah, I., et al. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in the telecom sector. IEEE Access, 7, 60134-60149.
14) Umayaparvathi, V., & Iyakutti, K. (2016). A survey on customer churn prediction in the telecom industry: Datasets, methods, and metrics. International Research Journal of Engineering Technology, 3(4), 1065-1070.
15) Yu, W., Jutla, D.N., & Sivakumar, S.C. (2005). A churn-strategy alignment model for managers in mobile telecom. In Communication Networks and Services Research Conference (Vol. 3, pp. 48-53).
16) Mali, M., & Atique, M. (2021). The relevance of preprocessing in text classification. In K.K. Singh Mer, V.B. Semwal, V. Bijalwan, & R.G. Crespo (Eds.), Proceedings of Integrated Intelligence Enable Networks and Computing: Algorithms for Intelligent Systems (Springer, Singapore).
17) Eldesouky Fattoh, I., Alsheref, F.K., Ead, W.M., & Youssef, A.M. (2022). Semantic sentiment classification for COVID-19 tweets using universal sentence encoder. Computational Intelligence and Neuroscience, 2022, Article ID 6354543, 8 pages. https://doi.org/10.1155/2022/6354543.
18) Chaudhary, S., & Naaz, S. (2017). Use of big data in computational epidemiology for public health surveillance. In Proceedings of the 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN) (pp. 150–155). IEEE, Gurgaon, India.
19) Choi, S., Lee, J., Kang, M.G., Min, H., Chang, Y.S., & Yoon, S. (2017). Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. Methods, 129, 50–59.

20) Ali, K., Dong, H., Bouguettaya, A., Erradi, A., & Hadjidj, R. (2017). Sentiment analysis as a service: A social media-based sentiment analysis framework. In Proceedings of the 2017 IEEE International Conference on Web Services (ICWS) (pp. 660–667). IEEE, Honolulu, HI, USA.
21) Hirt, R., Kühl, N., & Satzger, G. (2019). Cognitive computing for customer profiling: Meta classification for gender prediction. Electronic Markets, 29, 93–106.
22) Zhang, T. (2018). Telecom customer segmentation and precise package design by using data mining (Master's thesis). ISCTE Business School.
23) Ramachandran, D., & Parvathi, R. (2019). Analysis of Twitter specific preprocessing technique for tweets. Procedia Computer Science, 165, 245-251.
24) Alharbi, M., Doncker, A.S., & Doncker, E. (2018). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research.
25) Abou el Kassem, E., Hussein, S.A., Abdelrahman, A.M., & Alsheref, F.K. (2020). Customer churn prediction model and identifying features to increase customer retention based on user-generated content.
26) Abdalla, R., & Esmail, M. (2018). WebGIS for disaster management and emergency response (p. 59). Springer.
27) Bisong, E. (2019). More supervised machine learning techniques with Scikit-learn. In Building Machine Learning and Deep Learning Models on Google Cloud Platform.
28) Asgari-Chenaghlu, M., Nikzad-Khasmakhi, N., & Minaee, S. (2020). Covid-Transformer: Detecting trending topics on Twitter using universal sentence encoder. arXiv.
29) Cer, D., Yang, Y., Kong, S.Y., et al. (2018). Universal sentence encoder.
30) Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings (Vol. 18, pp. 194-206). Springer International Publishing.

31) Sharifani, K., & Amini, M. (2023). Machine learning and deep learning: A review of methods and applications. World Information Technology and Engineering Journal, 10(07), 3897-3904.
32) Amini, M., & Rahmani, A. (2023). Machine learning process evaluating damage classification of composites. International Journal of Science and Advanced Technology, 9(12), 240-250.
33) Amin, A., Shah, B., Khattak, A.M., Lopes Moreira, F.J., Ali, G., Rocha, A., & Anwar, S. (2019). Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods. International Journal of Information Management, 46, 304-319.
34) Garrido-Merchan, E.C., Gozalo-Brizuela, R., & Gonzalez-Carvajal, S. (2023). Comparing BERT against traditional machine learning models in text classification. Journal of Computational and Cognitive Engineering, 2(4), 352-356.
35) Alshamari, M.A. (2023). Evaluating user satisfaction using deep-learning-based sentiment analysis for social media data in Saudi Arabia's telecommunication sector. Computers.
36) Aftan, S., & Shah, H. (2023). Using the AraBERT model for customer satisfaction classification of telecom sectors in Saudi Arabia. Brain Sciences, 13(147).
37) Qamar, A.M., Alsuhibany, S.A., & Ahmed, S.S. (2017). Sentiment classification of Twitter data belonging to Saudi Arabian telecommunication companies. International Journal of Advanced Computer Science and Applications.
38) Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution.
39) Murphy, K.P. (2012). Machine learning: A probabilistic perspective. Cambridge, MA: MIT Press.
40) Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer.
41) Types of machine learning. (2024). GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/types-of-machine-learning/. [Accessed: Sep. 1, 2024].