Journal of Computer Science

Original Research Paper

Measuring Customers' Satisfaction Using Sentiment Analysis: Model and Tool

Ahmed Alqurafi and Tawfeeq Alsanoosy

Department of Computer Science, College of Computer Science and Engineering, Taibah University, Saudi Arabia

Article history
Received: 02-09-2023
Revised: 31-10-2023
Accepted: 30-11-2023

Corresponding Author:
Tawfeeq Alsanoosy
Department of Computer Science, College of Computer Science and Engineering, Taibah University, Saudi Arabia
Email: tsanoosy@taibahu.edu.sa

Abstract: Customer reviews are a valuable data resource for business owners and companies. Customers frequently write reviews on many platforms, such as eBay or Amazon. These reviews show, for instance, to what extent customers are satisfied with a service or a product. Therefore, these reviews can be used by companies and business owners to improve their services or products. However, numerous reviews make it difficult and time-consuming for companies to manually read, analyze, and classify every review. To tackle this issue, we proposed a sentiment analysis model that automatically analyses and classifies customer reviews. To build the model, six popular Machine Learning (ML) classifiers and a Deep Learning (DL) classifier were chosen. The six applied ML classifiers were implemented using three feature extraction techniques: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-grams. The aim was to determine the most efficient classifiers and feature sets for analyzing customer reviews. To train the model, we used a large, public, and real-world dataset that consisted of 4 million customer reviews. The results of this study confirmed some of the published results and showed some considerable improvements compared to some of the existing sentiment analysis models. Moreover, the findings indicated that applying N-grams revealed better accuracy of almost all ML classifiers. Among the selected ML classifiers, the higher accuracy was achieved at 91.3% when using the Support Vector Machines (SVM) with TFIDF and a combination of Unigram, Bigram, and Trigram. The worst accuracy was 77.3% when applying the Decision Tree (DT). However, Long Short-Term Memory (LSTM) showed the highest accuracy at 93.3%. We also utilized a web-based tool to deploy the sentiment analysis model so it would be freely accessible. Our tool will help companies and business owners analyze their customer reviews automatically and display a set of statistics effortlessly at a low cost, thereby measuring customer satisfaction.

Keywords: Natural Language Processing, Sentiment Analysis, Learning, Customer Reviews, Customer Satisfaction

Introduction

We live in an age of data deluge, where the amount of data is imposed to build novel tools to effectively extract, analyze, and understand the massive amounts of data. One of the most important types of data is customer reviews. Customer reviews are an essential and valuable resource for displaying customer satisfaction with a business or a product. Business owners and companies can use these reviews to further grow, enhance their services, and identify areas where they need to improve. For example, if a business receives many negative reviews about its customer service, it can use this feedback to train its staff and improve its customer service processes. Moreover, customers are more likely to trust reviews from other customers than they are traditional advertising. This is because the reviews are seen as more authentic and unbiased. However, the problem with this kind of data is that it is poorly structured and could reach thousands of reviews. Therefore, it is difficult for business owners to manually analyze, understand, or classify all customer reviews. As a result, there is a need to build a tool that analyses and organizes customer reviews automatically.

One of the ways to analyze and classify data is to use sentiment analysis. Sentiment analysis is a field of

natural language processing that analyses text to determine the author's emotional sentiment. Sentiment analysis is a rapidly expanding field of research with applications in several domains. For example, it can be used by businesses to understand customer feedback. This can be employed for a wide range of ends, such as understanding customer feedback, tracking social media sentiment, or identifying fake feedback. However, understanding the sentiment of a text requires more than just counting positive and negative words in a text; It is also essential to consider the context of the words and the overall tone of the review (Boukes *et al*., 2020; Nandwani and Verma, 2021).

Sentiment analysis is typically performed using Machine Learning (ML) and Deep Learning (DL) classifiers (Zhou *et al*., 2020; Hu *et al*., 2013; Medhat *et al*., 2014). Many ML and DL classifiers can be used for sentiment analysis. Consider of the most common ones include Support Vector Machines (SVM), Logistic Regression (LR), Decision Tree (DT), Multinomial N¨ıve Bayes (MNB), Bernoulli N¨ıve Bayes (BNB), Random Fores (RD), Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) (Ligthart *et al*., 2021; Dey *et al*., 2020; Pang *et al*., 2002; Rahman *et al*., 2020; Onan, 2020; Yuan *et al*., 2020). Thus, in this study, we will use ML and DL classifiers to build a sentiment analysis model that automatically analyses and classifies customer reviews.

Wankhade *et al*. (2022) conducted a detailed survey study on different sentiment analysis applications, methods, tools, and challenges. Four levels of sentiment analysis were discussed: Document level, sentence level, phrase level, and aspect level. The survey emphasizes several classification methods while discussing some of the necessary procedures in sentiment analysis. The authors found that NB and SVM classifiers are frequently employed as baseline performance measures. The authors also discussed and listed a few significant challenges faced such as methodological challenges. Also, Zhang *et al*. (2018) provided a detailed survey on how deep learning models are used in sentiment analysis. In a similar way, Birjali *et al*. (2021) conducted a survey study on sentiment analysis approaches, issues, and current directions. The authors recognized three major approaches to sentiment analysis: Lexicon-based, machine learning-based, and hybrid approaches. The authors then identified several challenges, such as dealing with sarcasm and irony, and mentioned several trends, such as using the increase of the DL and multimodal in sentiment analysis.

Contribution: In this study, we built a sentiment analysis model that automatically analyses and classifies customer reviews as positive and negative to help business owners measure customer satisfaction automatically. The aim of this study is to evaluate and select the most efficient classifiers and feature sets for analyzing customer reviews. To train the model, we used a large and real-world dataset that consisted of 4 million genuine customer reviews. We built the model by applying six ML classifiers (LR, SVM, DT, RF, MNB, and BNB) and one DL classifier (LSTM). The applied ML classifiers were experimented with using three feature extraction techniques: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-grams. The higher accuracy reached 91.3% when using the SVM classifier with TF-IDF and a combination of Unigram, Bigram, and Trigram. However, the LSTM classifier showed the highest accuracy at 93.3%. The results of this study confirmed some of the published results and showed some considerable improvements compared to some of the existing sentiment analysis models. We also employed a web-based tool to deploy the sentiment analysis model so it would be accessible to companies or business owners to analyze their customer reviews automatically and display statistics to measure customer satisfaction. The results of our work might also assist companies and business owners in improving their services, personalizing the customer experience, and providing real-time insights into customer satisfaction.

Related Works

Singh *et al*. (2022) analyzed Amazon reviews using DL with different word embedding approaches, such as BERT, Glove, Elmo, and Fast Text. The model is based on a Recurrent Neural Network (RNN) trained on a customer reviews dataset. It also evaluates and investigates critical grammatical sections. The authors find that the Multichannel CNN model with a fast text classifier offers the highest accuracy at 79.83%. In our work, the proposed approach achieved a high classification accuracy of 91.3% when applying SVM and 93.3% when applying LSTM.

Rahman *et al*. (2020) collected customer reviews from several Android apps to identify users' opinions on these Android apps. The authors applied five ML classifiers (k-nearest neighbors, RF, SVM, DT, and NB) and adopted three teachings: BoW, N-grams, and TF-IDF. The highest accuracy reached 88.9% using SVM. In our work, six ML classifiers and a DL classifier were adopted and the highest accuracy reached 93.3%. Also, we developed a cloud-based tool to allow companies and business owners to analyze reviews easily.

Laksono *et al*. (2019) applied TextBlob, a lexicon-based sentiment analyzer, and NB classifier to analyze customer reviews from the TripAdvisor dataset. The results showed 72.06% accuracy using NB and 69.12% using TextBlob. Also, Hasan *et al*. (2020) conducted a sentiment analysis on restaurant reviews in Dhaka city. The dataset included 338 restaurants with a total of 7280 reviews, 4079 reviews were labeled as positive and 3201 reviews were labeled as negative. The authors only

applied one classifier (SVM) and used BoW and TF-IDF as feature analysis along with N-gram. The highest accuracy reached 91.53%. In this study, we adopted two feature extraction techniques BoW and TF-IDF along with N-grams, and applied six ML classifiers and one DL classifier. Our model achieved a high accuracy reaching 93.3%.

Taecharungroj and Mathayomchan (2019) used TripAdvisor reviews to develop a sentiment model. 65,079 reviews of tourist attractions in Phuket, Thailand, were analyzed. They applied latent Dirichlet allocation to extract dimensions of attractions and the NB algorithm to analyze the occurrence of each term in positive and negative appraisals. The accuracy of the model was achieved at about 70%. In our work, the proposed approach achieved a high classification accuracy of 91.3% when applying SVM and 93.3% when applying LSTM.

Pang *et al*. (2002) used movie reviews as a dataset from Internet Movie Database (IMDb) to classify documents by sentiments. They extracted 700 positive and 700 negative reviews randomly. They used the BOF framework with eight features most importantly unigram, bigram, and unigram features. Their text processing only included negation handling with the unigram feature, in which they appended NOT as a tag between the negation character and the closest punctuation mark to each word, although they admitted that this technique had a negligible effect on performance. The authors did not use stop word removal or stemming in the preprocessing stage. After that, the authors trained their model using NB, Maximum Entropy, and SVM. The accuracy achieved 78.7% with unigram using NB, 77% with bigram, and 82.7% with unigram and bigram using SVM. Our model achieved higher accuracy reaching 91.3% when applying SVM.

Materials and Methods

The proposed methodology to build the sentiment analysis model was divided into five steps:

Step 1: Dataset selection
Step 2: Dataset preprocessing
Step 3: Feature extraction
Step 4: Model training
Step 5: Model evaluation

Each step is briefly explained in the subsections.

Step 1: Dataset Selection

In this research, we used a large, public, and real-world dataset that consisted of 4 million customer reviews. The dataset is available online at Kaggle[1]. The dataset is titled 4 million Amazon reviews. It is divided into two files: Training and testing. The training file contains 3.6 M reviews and the testing file contains 400 K reviews. Each file consists of two columns: One is for the reviews and the other is for the sentiment labeling (2 for positive and 1 for negative). To train the model, a sample of 250 K reviews was randomly selected from each category. We used the random state parameter with an arbitrary seed to generate the same sample each time we compiled the code to duplicate the results.

Step 2: Dataset Preprocessing

Dataset pre-processing is a crucial step for reducing errors, making better predictions, and improving accuracy. To ensure the text is ready for further analysis, we used the following text preprocessing techniques:

1. Convert reviews to lowercase: Lowercase was applied to all reviews to maintain consistency among all tokens and reduce the complexity as case sensitivity increases model complexity (Uma *et al*., 2022; Alsanoosy and Alqarni, 2023). This can cause problems for machine learning models, which often rely on matching words and phrases to each other
2. Expand abbreviations: Each abbreviation was expanded to the corresponding meaning. Abbreviations and slang words were expanded using a dictionary that contains 246 abbreviations and slang words[2]. For example, the abbreviation “lol” was expanded into laughing out loud, where the space is replaced with an underscore, so the abbreviation is considered as one token
3. Converts emojis and emoticons: Emojis were converted to text using a Python library called emoji[3]. Also, emoticons such as :), :(or ;) are converted to text using a library called emot[4]
4. Stop words and single/double-letter words are removed: We removed stop words which are words that are meaningless such as a, an, is, are, and then, from the reviews except for negation words
5. Remove special characters: Emails, links, character repetition, HTML tags, numbers, and punctuation were removed to reduce the feature space by using a Python library called re

[1]https://www.kaggle.com/datasets/nabamitachakraborty/amazonreviews
[2]https://www.kaggle.com/code/nmaguette/up-to-date-list-of-slangsfor-text-preprocessing
[3]https://pypi.org/project/emoji/
[4]https://github.com/NeelShah18/emot

6. Remove contractions: Words like didn't, couldn't, or wouldn't are converted to did not, could not, and would not using the contractions library
7. Stemming: Reviews were reduced words to their root forms by removing suffixes and prefixes. We used snowball stemmer over Porter and Lancaster stemmers because snowball stemmer strikes a balance between Porter and lacaster stemmers, it is not as aggressive as Lancaster and it is faster than Porter stemmer (Singh and Gupta, 2016; Gupta and Arora, 2022)

To increase transparency and remove any biases, we provide a sample of the reviews in Figs. 1-2. Figure 1 shows a sample of the 500 K Amazon reviews before the preprocessing steps and Fig. 2 shows a sample of the result after the preprocessing steps. Moreover, a sample of the 20 most frequent positive and negative reviews is provided to understand a review's overall sentiment and to detect any bias (Figs. 3-4). The time taken to preprocess the 500K sample reviews was approximately 511 sec.

	Review	Label
0	Excellent: When the subject of conversation is a Bob Greene book you hear many cliche's. Finger on the pulse of America aside, he's just a damn fi...	2
1	It rocks!: It has awesome songs that sound just like the entrances of the real wrestlers. I like The Rock's music the best. It is real cool.\n	2
2	Sounds great. Much better than my PC mic: I resisted USB Mic's for a LONG time but finally tried one. Sound quality is much better (sounds better ...	2
3	Academic's Delight: After studying Urban Planning and then switching to Literature, I find myself continually concerned with the ways our spatial ...	2
4	Terminator vs furminator: I have 2 cats, one long-haired and one short-haired. My long-haired one used to leave huge chunks of hair laying around ...	2
...	...	...
499995	Good for light jobs only.: I have owned this thing for 3 years, and use it to cut the weeds that grow seasonally on the hill behind my house. I go...	1
499996	Proves one thing.....: .....that Scott Stapp wasn't the only thing that stunk about Creed!!! That guitar sound? Horrid. Those arrangements.....	1
499997	SP-D1700 no better than SP-D1600: So. I get this new fresh from the factory SP-D1700 and it works fine, with very light use, for about 7 or 8 mont...	1
499998	Protection Software Include?!: After purchasing this game Brand New, after it was completely installed, an additional program popped up asking me ...	1
499999	This is just not the magazine for me: I enjoy an occasional Oprah show. I am impressed with her generosity as a human being. However, this is not ...	1

500000 rows × 2 columns

Fig. 1: A sample of reviews before the preprocessing steps

Label	Tokens
2	[excel, subject, convers, bob, green, book, hear, clich, finger, puls, america, asid, damn, fine, writer, bob, green, simpli, man, trust, right, s...
2	[rock, awesom, song, sound, like, entranc, real, wrestler, like, rock, music, best, real, cool]
2	[sound, great, better, mic, resist, usb, mic, long, time, final, tri, sound, qualiti, better, sound, better, mxl, dtk, mic, thermal, background, n...
2	[academ, delight, studi, urban, plan, switch, literatur, find, continu, concern, way, spatial, environ, affect, interest, mckittrick, book, natur,...
2	[termin, furmin, cat, long, hair, short, hair, long, hair, leav, huge, chunk, hair, lay, produc, good, size, hairbal, furmin, time, liter, shrank,...
...	...
1	[good, light, job, own, thing, year, use, cut, weed, grow, season, hill, hous, got, distanc, far, electr, outlet, not, want, hassl, gas, driven, u...
1	[prove, thing, scott, stapp, not, thing, stunk, creed, guitar, sound, horrid, arrangementsar, guy, grow, guitar, soloshorrid, tast, not, ounc, sub...
1	[better, new, fresh, factori, work, fine, light, use, month, understand, light, use, mean, mayb, dvd, month, like, day, stop, work, continu, load,...
1	[protect, softwar, includ, purchas, game, brand, new, complet, instal, addit, program, pop, ask, instal, softwar, protect, system, hesit, instal, ...
1	[not, magazin, enjoy, occasion, oprah, impress, generos, human, not, magazin, contain, advertis, rip, card, subject, matter, not, interest, order,...

Fig. 2: A sample of reviews after the preprocessing steps

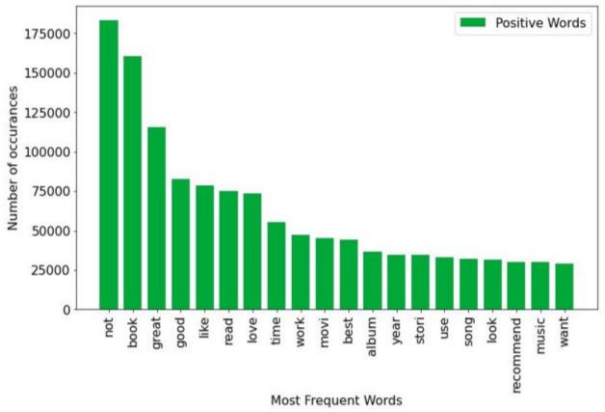

Fig. 3: Bar chart for 20 most frequent positive words

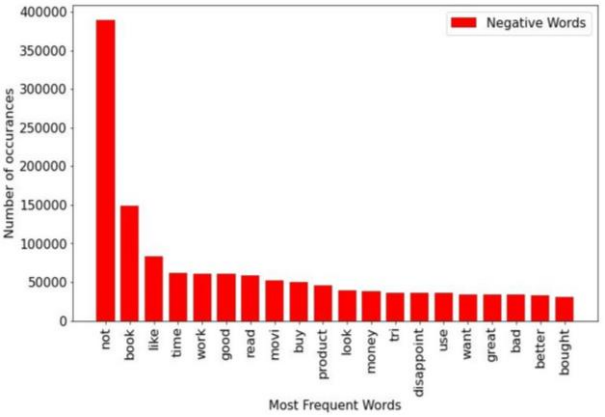

Fig. 4: Bar chart for 20 most frequent negative words

Step 3: Feature Extraction

After the preprocessing steps, we applied three feature extraction techniques. Feature extraction is the process of converting texts into vectors. We used the following three techniques:

1. Bag of Words: BoW model represents the frequency of a word occurrence in the text as a feature (Qader *et al*., 2019). It is used to extract features from text. BoW ignores word order, its semantic meaning, and its grammatical structure. Instead of counting the frequency of the individual words, BoW can apply a slightly more sophisticated approach with N-grams, adding a meaningful meaning to a feature
2. Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF is one of the most commonly used techniques for feature selection in text information processing. It is composed of two aspects: Term Frequency (TF) which refers to the frequency of occurrence of a feature term in the text set and Inverse Document Frequency (IDF) which measures a term's importance within a document (Ramos, 2003; Christian *et al*., 2016). One of the problems of BoW is that it may give a high score to a word that appeared frequently in a document, while it may not contain relevant information to the document. *TF-IDF* solves this problem by giving weight to the words

The *TF-IDF* mathematical equation can be represented as shown in (1):

$$TF - IDF = W(w,C) = TF(w)C \; x \log_{CF}^{N}(T) \quad (1)$$

where:

$TF(w)C$ = Number of a word (w) in reviews (C)
N = The total number of reviews in the dataset
$CF(t)$ = Number of documents containing the term (w)

3. N-grams: N-grams are sequential sequences of words in a document (Huston *et al*., 2011). N-grams are able to capture information about the structure of words and phrases in a text dataset, which has been shown to be effective. It can be applied to both BoW and *TF-IDF*. The *N* stands for how many words are in the sequence. Unlike the BoW, the N-grams take the order into consideration. For example, it would be better to use bigram on "Abu Dhabi" instead of separating them into single words because they carry a meaning together. If they were separated, the meaning would change. We extracted unigram, bigram, and trigram to examine their effectiveness in sentiment classification. Dave *et al.* (2003); Pang *et al.* (2002) found that using N-grams as features would enhance the performance of some models in classifying reviews into positive and negative classes

Step 4: Model Training

In this project, we selected the most popular ML classifiers (Damopoulos *et al*., 2012; Talukdar *et al*., 2020). We applied six ML classifiers: Support vector machines, logistic regression, decision tree, multinomial Na¨ıve Bayes, Bernoulli Na¨ıve Bayes and random forest, and one DL classifier, which is long short-term memory. The dataset has been split into 80% for training and the rest for testing. To shed light on the findings, the next section provides a presentation and analysis of the results. In this subsection, a brief description of each used classifier is given.

Support Vector Machines (SVM): SVM is a supervised ML classification algorithm that draws a decision boundary line called a hyperplane to separate and classify data (Noble, 2006). In sentiment analysis, SVM is used to classify text, for example, into positive, negative, or neutral sentiment. It excels at both predicting continuous values (regression) and distinguishing between discrete categories (classification). The hyperplane is chosen to maximize the margin between the

two classes, the distance between the hyperplane, and the closest data points from each class. SVM is a popular choice for sentiment analysis because they are relatively simple to implement and can achieve high accuracy on a variety of datasets (Wankhade *et al*., 2022).

N¨ıve Bayes (NB): NB is a probabilistic classifier that uses the Bayes theorem to calculate the probability of a text belonging to a particular sentiment class (Rish, 2001). NB is a simple and efficient algorithm. It ranks data on the basis of probabilities. In text classification, two common variants of NB are used: Bernoulli N¨ıve Bayes (BNB), which is used with features that have a binary outcome (e.g., positive and negative) and Multinomial N¨ıve Bayes (MNB), which is used with discrete features such as word frequency (Sammut and Webb, 2011; Rennie *et al*., 2003; Dey *et al*., 2016). MNB assumes that the words in a text document are independent of each other, given the class of the document. BNB and MNB are simple and efficient algorithms often used for text classification tasks (Abbas *et al*., 2019; Rabbimov and Kobilov, 2020). In this study, both BNB and MNB will be used.

Logistic Regression (LR): LR is a popular and essential ML classifier that relies on probabilities. It is used in binary response variables. In sentiment analysis, LR can be used to classify text as positive or negative sentiment. LR is a simple but effective algorithm for sentiment analysis. It is easy to implement and interpret and can be used to achieve better results in developing a sentiment analysis model (Wankhade *et al*., 2022). LR tries to find the relationship between the independent variables and the dependent variable. Once the model is trained, it can estimate the chances of new data points having one of two possible outcomes.

Decision Tree (DT): DT works well for both classification and regression. It analyses data in a tree-like fashion. DT algorithm first extracts features from the data then the algorithm builds a decision tree by repeatedly splitting the data into smaller subsets based on the values of the features. The algorithm splits data based on sentiment and chooses the feature by analyzing specific details. In sentiment analysis, DT performs classification by splitting the tree into branches where each branch represents a subset of the dataset with a standard feature. The splitting continues until it reaches the last level, where all the data is from one class (Priyanka and Kumar, 2020).

Random Forest (RF): RF is a practical algorithm for classification and regression. It is an ensemble method that combines multiple DTs, where each tree is constructed based on a random subset of features and training samples (Kirasich *et al*., 2018). The predictions from each tree are then combined to provide the outcomes. RF is often used in image classification, fraud detection, and sentiment analysis tasks (Patel *et al*., 2023; Adugna *et al*., 2022). It works by constructing a large number of DTs, each trained on a different random subset of the training data. The algorithm also introduces randomness into the tree construction process by randomly selecting a subset of features at each split. This helps to reduce overfitting and improve the generalization performance of the model.

Long Short-Term Memory (LSTM): LSTM is a type of Recurrent Neural Network that is specifically designed to handle sequential data, such as time series, speech, and text. Neural Network is a simulation of the human brain. Its architecture comprises three layers: An input layer, followed by hidden layers, and an output layer. Each layer is made up of nodes or artificial neurons. These nodes connect and have different weights or thresholds. The most standard type of neural network is the feed-forward neural network, wherein the connection between the neurons does not form a cycle (Staudemeyer and Morris, 2019). Recurrent Neural Networks are networks with cyclic connections or recurrent connections, meaning the output of a layer can be fed back as an input. However, the fed-back signal of the Recurrent Neural Networks is limited, which gives rise to a problem called the long-term dependencies. This problem happens when the gap between the relevant input data is large, LSTM solves this problem (Yu *et al*., 2019). Thus, LSTM will be applied in this research.

Step 5: Model Evaluation

We evaluated the performance of the model by calculating the accuracy in terms of Accuracy (2) Precision (3) Recall (4) and F1-score (5): The following formulas were used to determine the performance of each classification model:

$$Accuarcy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5}$$

where, *TP* is True Positive, *TN* is True Negative, *FP* is False Positive and *FN* is False Negative.

Results and Discussion

This section presents the results of our experiments. The goal of this project was to evaluate the performance of different ML and DL classifiers to build a sentiment

analysis model. We used a variety of ML classifiers and feature extraction techniques to nominate the best classifiers and feature sets for analyzing customer reviews. In our experiments, we used six ML classifiers, one DL classifier, and three feature extraction techniques. The selected six classifiers were: LR, SVM, DT, RF, MNB, and BNB. The DL classifier was LSTM. The feature extraction techniques were BoW and TF-IDF, N-grams. We applied different combinations of N-grams such as unigram, unigram + bigram, and unigram + bigram + trigram. We used the random state parameter with an arbitrary seed to generate the same sample each time we compiled the code to duplicate the results.

In the first set of experiments, we applied BoW with different combinations of N-grams. In the second set of experiments, we applied TF-IDF with different N-gram combinations. Lastly, we applied the LSTM with no feature extraction technique. First, the results of the set of BoW experiments with different sets of N-grams will be presented (Table 1). Then, the results of TF-IDF experiments with different N-gram combinations will be presented (Table 2). Table 3 shows the model performance results of the highest achieved classifiers in terms of accuracy, precision, recall, and F1-score.

In the first set of BoW experiments, we applied BoW with Unigram. The results indicate that the LR achieved the highest accuracy of 88.1% and RF accuracy achieved slightly less than the LR with a score of 87%. However, the DT scored the least accuracy at 78.5%. Table 1 shows the results of using BoW with Unigram.

We then conducted an experiment using BoW with Unigram and Bigram. Results showed an improvement in the accuracy of all models, with LR leading the scores with an accuracy of 90.9% better than BoW using Unigram with an improvement of 2.8%. The SVM was the second best, achieving an accuracy of 89.7%. However, DT still performed the worst accuracy at 80.2% (Table 1).

In the last experiment with BoW, we applied unigram, bigram, and trigram. The results showed an improvement in the accuracy of all the models (Table 1). LR again outperformed all other models with an accuracy of 91.2%, followed by SVM achieving an accuracy of 90.4%. Figure 5 shows a bar chart that compares the models' performance using BoW feature extraction with different N-grams.

Table 1: Accuracy results of LR, SVM, DT, RF, MNB, and BNB classifiers using Bag of Words with different combinations of N-grams

Models	Unigram %	Unigram+ Bigram %	Unigram+ Bigram+ Trigram %
LR	88.1	90.9	91.2
SVM	86.7	89.7	90.4
DT	78.5	80.2	80.4
RF	87.0	88.1	88.2
MNB	83.9	86.9	87.5
BNB	84.4	87.2	87.5

Table 2: Accuracy results of LR, SVM, DT, RF, MNB, and BNB classifiers using TF-IDF with different combinations of N-grams

Models	Unigram %	Unigram+ Bigram %	Unigram+ Bigram+ Trigram %
LR	88.6	91.2	91.3
SVM	88.2	91.1	91.3
DT	77.3	79.0	78.9
RF	87.0	88.1	88.1
MNB	83.0	87.4	80.0
BNB	84.4	87.2	87.5

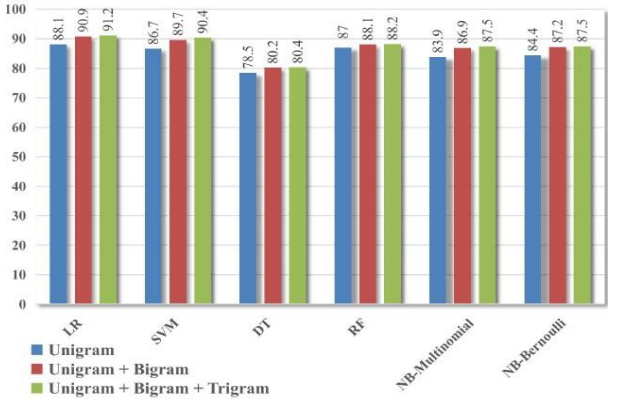

Fig. 5: Models' accuracy using BoW feature extraction with different N-grams

In the second set of experiments, we applied TF-IDF with different combinations of N-grams. We first applied TF-IDF with unigram. Results indicate an improvement in LR and SVM over BoW with Unigram from 88.1 and 86.7% to 88.6 and 88.2% and a decline in the accuracy of DT and MNB from 78.5 and 83.9% to 77.3 and 83%. RF and BNB achieved the same accuracy of 87 and 84.4% when using BoW with Unigram. Table 2 shows the results of using TF-IDF with unigram.

Then, we used TF-IDF with Unigram and Bigram. Again, LR achieved the best accuracy of 91.2% which was achieved using BoW with Unigram, Bigram, and Trigram. SVM followed this with an accuracy of 91.1%, almost the same as LR. As a result, it showed better improvement with 2.6% for LR and 2.9% for SVM, compared to TF-IDF with Unigram. Other models also had an improvement that ranged from 1.1-3.4% over TF-IDF with Unigram (Table 2).

Finally, we applied TF-IDF with Unigram, Bigram, and Trigram. Based on the results, LR and SVM achieved the best results of all the previous experiments, with an accuracy of 91.3% for both models (Table 2). DT had its worst result with an accuracy of 78.9%. Figure 6 shows a bar chart that compares model performances using TF-IDF feature extraction with different features to provide more insight, we compare the BoW and TF-IDF of the applied ML using the

N-grams combination (Figs. 7-9). Overall, the results indicated that using LR or SVM achieved the best accuracy among all models. The best accuracy level was recorded by applying TF-IDF with Unigram, Bigram, and Trigram, which achieved a rate of 91.3% in both LR and SVM models. Also, the difference between the results of BoW and TF-IDF with the same feature classes was minutes and sometimes negligible. One reason for the success of LR and SVM for sentiment analysis is that can learn complex relationships between features and labels (Jadav and Vaghela, 2016; Ramasamy *et al*., 2021; Tyagi and Sharma, 2018). LR uses a probabilistic model to predict the probability that a review is positive or negative, given the features of the review. SVM uses a hyperplane to separate the positive and negative reviews in the feature space (Jadav and Vaghela, 2016). Both classifiers extract meaningful text features, such as word presence, the sentiment of individual words, and the overall structure of the review. LR and SVM are also well-suited for sentiment analysis because they are robust to noise in the data (Bertsimas and King, 2017; Agarwal and Mitra, 2014). Amazon reviews often contain misspellings, grammatical errors, and other forms of noise. LR and SVM are able to handle this noise effectively and still produce accurate predictions.

We lastly applied the LSTM with no feature extraction technique. To train the LSTM, we used the same dataset used with the ML. We split it into 80% training data and 20% testing data. The result we achieved with the LSTM model outperforms all the results of ML models with an accuracy of 93.3%, better accuracy than the SVM and LR using TF-IDF with unigram, bigram, and trigram by 2% accuracy (Table 3). Thus, the highest result we achieved was 93.3% by employing LSTM.

We then tested our model with unseen data to avoid biased results using two different data sets. We used the IMDb movie reviews dataset from Stanford. The dataset is available online[5]. The dataset consists of 50K movie reviews categorized as positive and negative and our model achieved an accuracy of 88%. We used the IMDb movie reviews dataset used by Pang *et al*. (2002). The dataset is available online[6]. Compared to Pang's work, our model achieves a higher accuracy of 84.2% slightly better than Pang's result, which is 82.7%. When we tested the LSTM model with unbiased data, the LSTM model achieved an accuracy of 86% with the IMDb dataset and 78% with Pang's dataset. Also, our results outperformed the model proposed by Singh *et al*. (2022). Based on our work, we could say that LSTM might achieve better results than using Multi-channel CNN (fast text). Thus, our model shows better preference. We also employed a web-based tool to host the model using Next JS and Flask, a Python web framework. Figures 10-11 show snapshots of the developed web-based tool. The snapshots show the results of the analysis of the IMDb movie reviews dataset used by Pang *et al*. (2002).

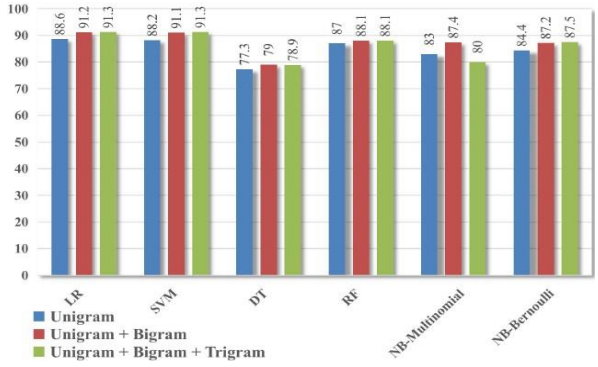

Fig. 6: Models' accuracy using TF-IDF feature extraction with different N-grams

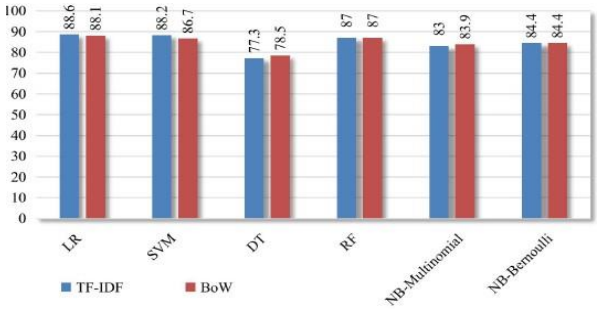

Fig. 7: Comparison between BoW and TF-IDF using Unigram

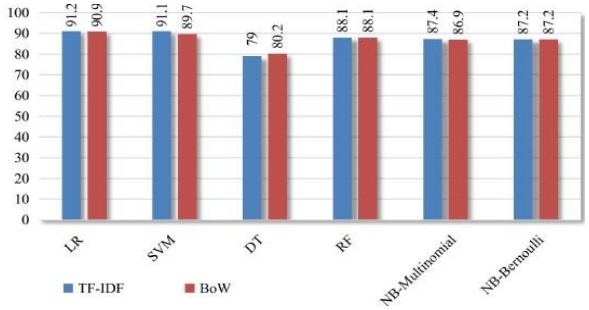

Fig. 8: Comparison between BoW and TF-IDF using Unigram and Bigram

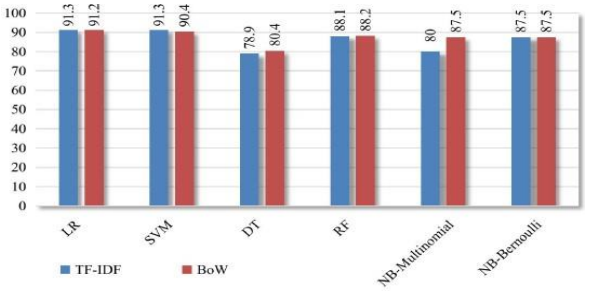

Fig. 9: Comparison between BoW and TF-IDF using Unigram, Bigram, and Trigram Table 3. Models performance of the highest achieved classifiers in terms of accuracy, precision, recall, and F1-score

[5]URL: https://ai.stanford.edu/ amaas/data/sentiment/

[6]URL: https://www.cs.cornell.edu/people/pabo/movie-review-data/

Table 3: Models performance of the highest achieved classifiers in terms of accuracy, precision, recall, and F1-score

Models	Feature	Accuracy %	Class	Precision	Recall	F1-Score
LR	BoW: Unigram + Bigram	90.9	Neg	0.92	0.90	0.91
			Pos	0.90	0.92	0.91
	BoW: Unigram + Bigram + Trigram	91.2	Neg	0.92	0.91	0.91
			Pos	0.91	0.92	0.91
	TF-IDF: Unigram + Bigram	91.2	Neg	0.92	0.91	0.91
			Pos	0.91	0.92	0.91
	TF-IDF: Unigram + Bigram + Trigram	91.3	Neg	0.92	0.91	0.91
			Pos	0.91	0.92	0.91
SVM	BoW: Unigram + Bigram + Trigram	90.4	Neg	0.91	0.90	0.90
			Pos	0.90	0.91	0.90
	TF-IDF: Unigram + Bigram	91.1	Neg	0.92	0.91	0.91
			Pos	0.91	0.92	0.91
	TF-IDF: Unigram + Bigram + Trigram	91.3	Neg	0.92	0.91	0.91
			Pos	0.91	0.92	0.91
LSTM	No feature extraction used	93.3	Neg	0.91	0.89	0.90
			Pos	0.89	0.92	0.90

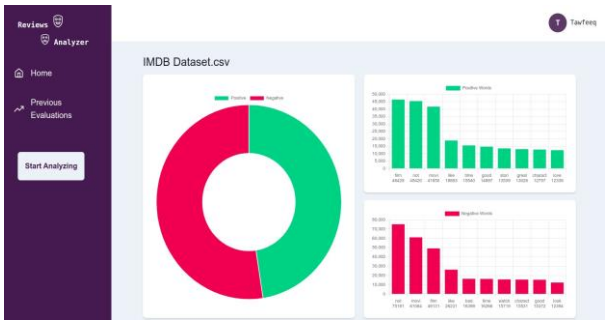

Fig. 10: Analysis of IMDb movie reviews dataset

Fig. 11: Analysis of IMDb movie reviews dataset

The potential benefits of our work include:

1. Analyses and classifies customer reviews automatically
2. Helps customer service to prioritize customer service tickets and to personalize the customer experience by recommending products that are likely to be of interest to the customer
3. Helps business owners to identify trends in customer satisfaction and develop targeted customer service strategies; and
4. Provides real-time insights into customer satisfaction
5. Helps product managers improve existing products/services

However, it is essential to be aware of the limitations of sentiment analysis models, such as poor understanding of the context for a given feedback, and therefore need to be used carefully.

Conclusion

Customer reviews are an essential and valuable resource for measuring customer satisfaction with products or services. Customers frequently write reviews, posting their experience with the provided services. Therefore, this data can be used by companies and business owners to improve their services or products. However, numerous reviews make it difficult and time-consuming for companies to manually read, analyze, and classify every review. The amount of data is imposed to build novel tools to effectively extract, analyze, and understand these massive amounts of data. As a result, it is beneficial to build a sentiment analysis model that analyses and classifies customer reviews automatically. We applied the popular six ML classifiers and one DL classifier to build the model. The applied ML classifiers were evaluated using three feature extraction techniques: BoW, TF-IDF, and N-grams. For the ML classifiers, the best accuracy achieved was 91.3% in both LR and SVM, when applying TF-IDF with Unigram, Bigram, and Trigram. The worst accuracy was achieved at 77.3% when applying the DT with TF-IDF and Unigram. However, using the LSTM outperformed all ML results with an accuracy of 93.3%.

We also developed a web-based tool to deploy the sentiment analysis model so it would be accessible by any companies or business owners to analyze their customer reviews automatically and display a set of

statistics to indicate customer satisfaction. The developed tool will assist companies and business owners to analyse customer reviews automatically and, therefore help them improve their services and achieve customer satisfaction. Moreover, the proposed model and tool can be used to identify the aspects of a product or service that customers are most or least satisfied with. This information can then be used to prioritize improvements and make changes that will likely have the most significant positive impact on customer satisfaction. Also, this information can be used to personalize the customer experience, such as by recommending products or services that are likely to interest the customer.

In future works, we aim to use an Arabic dataset in the future because it is more challenging. Some potential challenges and specific areas of improvement for building sentiment analysis models using Arabic datasets are sarcasm, morphological complexity, variety of dialects and slang, and the use of negation as it can be expressed in a variety of ways, both explicitly and implicitly. We also aim to improve the accuracy of our model by using different DL classifiers, such as Generative Adversarial Networks (GANs) and Recurrent Neural Networks (RNNs). We also would like to enhance the functionality of our tool by adding more analysis diagrams and features.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Ahmed Alqurafi: Proposed the idea, carried out some of the experiments, and wrote the manuscript.

Tawfeeq Alsanoosy: Carried out all the experiments and wrote the manuscript.

Ethics

The authors confirm that this article has not been published in any other journal. The corresponding author confirms that all the authors have read and approved the manuscript. Additionally, no ethical issues are involved in the manuscript or the dataset, and no conflicts of interest are involved.

References

Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, *19*(3), 62. https://doi.org/10.13140/RG.2.2.30021.40169

Adugna, T., Xu, W., & Fan, J. (2022). Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images. *Remote Sensing*, *14*(3), 574. https://doi.org/10.3390/rs14030574

Agarwal, S., & Mitra, M. (2014). Lamb wave based automatic damage detection using matching pursuit and machine learning. *Smart Materials and Structures*, *23*(8), 085012. https://doi.org/10.1088/0964-1726/23/8/085012

Alsanoosy, T., & Alqarni, A. (2023). YouTube Sentiment Analysis: Performance Model Evaluation. In *Kids Cybersecurity Using Computational Intelligence Techniques* (pp. 269-282). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-21199-7_19

Bertsimas, D., & King, A. (2017). Logistic regression: From art to science. *Statistical Science*, 367-384. https://www.jstor.org/stable/26408297

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *14*(2), 83-104. https://doi.org/10.1080/19312458.2019.1671966

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285-294. https://doi.org/10.21512/comtech.v7i4.3746

Damopoulos, D., Menesidou, S. A., Kambourakis, G., Papadaki, M., Clarke, N., & Gritzalis, S. (2012). Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers. *Security and Communication Networks*, *5*(1), 3-14. https://doi.org/10.1002/sec.341

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 519-528). https://doi.org/10.1145/775152.775226

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. *arXiv preprint arXiv:1610.09982*. https://doi.org/10.48550/arXiv.1610.09982

Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020, February). A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)* (pp. 217-220). IEEE. 10.1109/IC3A48958.2020.233300

Gupta, S., & Arora, B. (2022). Stemming Techniques on English Language and Devanagari Script: A Review. *Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 1*, 541-550. https://doi.org/10.1007/978-981-16-8248-3_45

Hasan, T., Matin, A., & Joy, M. S. R. (2020, December). Machine learning based automatic classification of customer sentiment. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCIT51783.2020.9392652

Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 607-618). https://doi.org/10.1145/2488388.2488442

Huston, S., Moffat, A., & Croft, W. B. (2011, February). Efficient indexing of repeated n-grams. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 127-136). https://doi.org/10.1145/1935826.1935857

Jadav, B. M., & Vaghela, V. B. (2016). Sentiment analysis using support vector machine based on feature selection and semantic analysis. *International Journal of Computer Applications*, *146*(13). https://doi.org/10.5120/ijca2016910921

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9. https://scholar.smu.edu/datasciencereview/vol1/iss3/9/

Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019, July). Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes. In *2019 12th International Conference on Information and Communication Technology and System (ICTS)* (pp. 49-54). IEEE. https://doi.org/10.1109/ICTS.2019.8850982

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review*, 1-57. https://doi.org/10.1007/s10462-021-09973-3

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology*, *24*(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565

Onan, A. (2020). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, *28*(1), 117-138. https://doi.org/10.1002/cae.22179

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv Preprint cs/0205070*. https://doi.org/10.48550/arXiv.cs/0205070

Patel, A., Oza, P., & Agrawal, S. (2023). Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model. *Procedia Computer Science*, *218*, 2459-2467. https://doi.org/10.1016/j.procs.2023.01.221

Priyanka, & Kumar, D. (2020). Decision tree classifier: A detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246-269. https://doi.org/10.1504/IJIDS.2020.108141

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications and challenges. In *2019 International Engineering Conference (IEC)* (pp. 200-204). IEEE. https://doi.org/10.1109/IEC47844.2019.8950616

Rabbimov, I. M., & Kobilov, S. S. (2020, May). Multi-class text classification of uzbek news articles using machine learning. In *Journal of Physics: Conference Series* (Vol. *1546*, No. 1, p. 012097). IOP Publishing. https://doi.org/10.1088/1742-6596/1546/1/012097

Ramasamy, L. K., Kadry, S., Nam, Y., & Meqdad, M. N. (2021). Performance analysis of sentiments in Twitter dataset using SVM models. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(3), 2275-2284. https://doi.org/10.11591/ijece.v11i3.pp2275-2284

Rahman, M. M., Rahman, S. S. M. M., Allayear, S. M., Patwary, M. F. K., & Munna, M. T. A. (2020). A sentiment analysis based approach for understanding the user satisfaction on android application. In *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19* (pp. 397-407). Springer Singapore. https://doi.org/10.1007/978-981-15-1097-7_33

Ramos, J. (2003, December). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* (Vol. *242*, No. 1, pp. 29-48). file:///C:/Users/PC/Downloads/UsingTF-IDFtoDetermineWordRelevanceinDocumentQueries.pdf

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 616-623). https://cdn.aaai.org/ICML/2003/ICML03-081.pdf

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (Vol. *3*, No. 22, pp. 41-46). https://sites.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf

Sammut, C., & Webb, G. I. (Eds.). (2011). *Encyclopedia of Machine Learning*. Springer Science and Business Media. ISBN-10: 0387307680.

Singh, J., & Gupta, V. (2016). Text stemming: Approaches, applications and challenges. *ACM Computing Surveys (CSUR)*, *49*(3), 1-46. https://doi.org/10.1145/2975608

Singh, U., Saraswat, A., Azad, H. K., Abhishek, K., & Shitharth, S. (2022). Towards improving e-commerce customer review analysis for sentiment detection. *Scientific Reports*, *12*(1), 21983. https://doi.org/10.1038/s41598-022-26432-3

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv Preprint arXiv:1909.09586*. https://doi.org/10.48550/arXiv.1909.09586

Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, *75*, 550-568. https://doi.org/10.1016/j.tourman.2019.06.020

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations-A review. *Remote Sensing*, *12*(7), 1135. https://doi.org/10.3390/rs12071135

Tyagi, A., & Sharma, N. (2018). Sentiment analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology (UAE)*, *7*(2), 20-23.

Uma, R., Jawahar, P., & Rishitha, B. V. (2022, November). Support Vector Machine and Convolutional Neural Network Approach to Customer Review Sentiment Analysis. In *2022 1st International Conference on Computational Science and Technology (ICCST)* (pp. 239-243). IEEE. https://doi.org/10.1109/ICCST55948.2022.10040381

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, *31*(7), 1235-1270. https://doi.org/10.1162/neco_a_01199

Yuan, J., Wu, Y., Lu, X., Zhao, Y., Qin, B., & Liu, T. (2020). Recent advances in deep learning based sentiment analysis. *Science China Technological Sciences*, *63*(10), 1947-1970. https://doi.org/10.1007/s11431-020-1634-3

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253. https://doi.org/10.1002/widm.1253

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in neural NLP: Modeling, learning and reasoning. *Engineering*, *6*(3), 275-290. https://doi.org/10.1016/j.eng.2019.12.014