VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master Thesis

Predicting Customer Satisfaction in Wix Product Reviews Using Machine Learning and NLP methods

„Wix“ naudotojų pasitenkinimo prognozavimas remiantis produktų atsiliepimais, naudojant mašininio mokymosi ir natūralios kalbos apdorojimo metodus

Agnė Griniūtė

Supervisor : Doc., Dr. Dmitrij Celov

Reviewer : Dr. Gražina Korvel

Vilnius
2025

Summary

In today's digital world, understanding online customer reviews is essential for evaluating user satisfaction and improving business products and services. This thesis explores the potential of machine learning (ML) and natural language processing (NLP) methods to analyse and predict customer satisfaction based on user reviews. Using data from Wix, a leading website-building platform, the research aims to identify key topics in customer feedback, predict review star ratings, and assess the effectiveness of sentiment analysis as an alternative measure of satisfaction.

The thesis begins by reviewing existing literature on customer satisfaction analysis, ML and NLP methods. The methodology combines sentiment analysis, topic modelling, and supervised ML models. Reviews are preprocessed using advanced NLP techniques, including tokenisation, lemmatisation, and vectorisation, followed by training models such as Logistic Regression, Support Vector Machine, and XGBoost to predict star ratings.

The results show that advanced machine learning models can effectively predict star ratings using only the text from reviews. Sentiment analysis also proved useful in measuring customer satisfaction, aligning closely with traditional star ratings. Topic modelling uncovered common themes in customer feedback, such as usability problems, feature requests, and customer support experiences, offering valuable insights for improving Wix's products and services.

This study contributes to the field of automated review analysis by demonstrating how traditional and modern NLP techniques can work together effectively. The results offer practical value for businesses that seek to use customer feedback to improve user experience and satisfaction.

Keywords: online reviews, customer satisfaction, machine learning, natural language processing, logistic regression, support vector machine, xgboost.

Santrauka

Šiandienos skaitmeniniame pasaulyje yra būtina suprasti internetinius klientų atsiliepimus, jei norime įvertinti vartotojų pasitenkinimą ir tobulinti verslo teikiamus produktus bei paslaugas. Šiame magistro darbe gilinamasi į mašininio mokymosi (MM) ir natūralios kalbos apdorojimo (NKA) metodų taikymą analizuojant ir prognozuojant klientų pasitenkinimą remiantis vartotojų paliktais atsiliepimais. Analizei naudotas duomenų rinkinys iš Wix – pirmaujančios svetainių kūrimo platformos. Tyrimo tikslas yra identifikuoti pagrindines temas klientų atsiliepimuose, prognozuoti atsiliepimų žvaigždučių įvertinimus ir įvertinti sentimentų analizės efektyvumą kaip alternatyvą klientų pasitenkinimo vertinimui.

Tyrimas pradedamas nuo literatūros apžvalgos apie klientų pasitenkinimo analizę, MM ir NKA metodus. Tuomet taikoma sentimentų analizę, atliekamas temų modeliavimas ir apmokomi MM modeliai. Teksto apdorojimui pasitelktos pažangios NKA technikos, tokios kaip: tokenizavimas, lematizavimas ir vektorizavimas, po to apmokomi mašininio mokymosi modeliai: logistinė regresija, atraminių vektorių klasifikatorius (SVM) ir XGBoost, siekiant prognozuoti žvaigždučių įvertinimus.

Rezultatai rodo, kad pažangūs mašininio mokymosi modeliai, naudodami tik tekstą iš atsiliepimų gali sėkmingai prognozuoti žvaigždučių įvertinimus. Sentimentų analizė taip pat pasirodė esanti naudinga priemonė vertinant klientų pasitenkinimą. Temų modeliavimas atskleidė dažnai pasikartojančias temas klientų atsiliepimuose, tokias kaip problemos susijusios su naudojimu, naujų funkcijų užklausos ar klientų aptarnavimo patirtis, taip suteikiant vertingų įžvalgų, padedančių tobulinti Wix produktus.

Šis tyrimas prisideda prie automatizuotos atsiliepimų analizės srities, parodydamas, kad tradicinės ir modernios NKA technikos gali duoti vaisingų rezultatų. Rezultatai turi praktinės vertės verslams, kurie siekia pasinaudoti klientų atsiliepimais ir taip pagerinti vartotojo patirtį.

Raktiniai žodžiai: internetiniai atsiliepimai, klientų pasitenkinimas, mašininis mokymasis, natūralios kalbos apdorojimas, logistinė regresija, atraminių vektorių klasifikatorius, xgboost.

List of Figures

List of Tables

Contents

List of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
BOW	Bag of Words
EDA	Exploratory Data Analysis
KNN	K-Nearest Neighbors
FN	False Negatives
FP	False Positives
ML	Machine Learning
NLP	Natural Language Processing
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
TN	True Negatives
TP	True Positives
Wix	Wix.com
XGBoost	Extreme Gradient Boosting

Introduction

Understanding customer satisfaction and needs is essential for business success and user loyalty in today's digital world. Nowadays, most of us are active social media users, for whom buying from online shops is a natural habit and expressing opinions online is essential. With so many users leaving feedback online, reviews have become a rich source of information that can help companies improve their products and services.

This thesis analyses the data provided by Wix – a popular website-building platform with millions of users worldwide. Every day, Wix receives countless reviews from its users, providing valuable insights into what is working and what needs to be improved about their products. However, manually analysing all these reviews would be an overwhelming task for a human. This is where technology – machine learning (ML) and natural language processing (NLP) techniques, could come into help. NLP and ML enable advanced text analysis that could detect the tone of customer reviews and point out specific details, like common complaints or new feature requests. This helps companies better understand how their customers are feeling and what they want and, in this way, make requested and needed improvements to their products and services. But despite the advances in ML and NLP, there is still a gap in utilising these technologies effectively for actionable customer satisfaction prediction, particularly in the context of platforms like Wix with large and diverse user bases.

This thesis addresses this gap by analysing and predicting customer satisfaction with Wix products using a combination of ML and NLP techniques. Specifically, it focuses on identifying common topics in user reviews and predicting star ratings based on textual content. By doing so, the study contributes to the broader understanding of how user feedback can be effectively leveraged to improve products and services.

The following research questions guide this study:

1. How accurately can machine learning models predict the star ratings of product reviews based only on textual content?

2. Can sentiment analysis of reviews match the effectiveness of traditional star ratings for determining customer satisfaction?

3. What are the most common themes or topics discussed in reviews of specific products?

4. How do different NLP techniques impact the performance of machine learning models in review analysis?

To answer these questions, the thesis employs a combination of sentiment analysis, topic modelling, and supervised machine learning. Customer reviews are preprocessed by using NLP techniques, after sentiment analysis and topic modelling are implemented. Finally, Logistic Regression, Support Vectors Machine, and XGBoost models are trained and evaluated to predict star ratings.

To sum up, by combining ML and NLP methods, this research aims to demonstrate the potential of automated review analysis to understand customer satisfaction and drive business improvements.

The thesis is structured as follows: Chapter 1 reviews the existing literature on customer satisfaction analysis, related work done in this field, NLP methods, machine learning models and their evaluation. Chapter 2 details the methodology, including data collection, the dataset, preprocessing, and model selection. Finally, Results and Conclusions summarise the findings, highlight limitations, and suggest ideas for future research.

1 Literature Review

1.1 Importance of customer satisfaction in business

Customer satisfaction plays an important role in the long-term success of any business. Studies have shown that satisfied customers are more likely to repurchase, recommend products to others, and contribute positively to a company's reputation. In contrast, dissatisfied customers can spread negative feedback to other potential users and damage brand image in this way [3, 26].

A strong relationship exists between customer satisfaction and business performance. Research by Fornell et al. [15] demonstrates that higher customer satisfaction leads to increased customer loyalty, reduced price sensitivity, and ultimately higher profitability. Customer satisfaction is especially crucial for companies that depend on recurring revenue, like subscription-based businesses. When customers are happy, they're more likely to keep renewing their subscriptions, which directly supports steady income for the company. Retaining satisfied customers is also much more cost-effective than constantly trying to attract new ones. Empirical studies show that it's five to seven times less expensive to keep an existing customer than to find a new one [20]. Therefore, high satisfaction not only strengthens customer loyalty but also boosts the company's financial stability by reducing churn and lowering customer acquisition costs. For digital platforms like Wix, where users can choose from a wide range of competitors, understanding and improving customer satisfaction is essential for maintaining Wix's market share.

In recent years, customer satisfaction has been used as a predictive metric for future user behaviour. A common principle in marketing is that past customer behaviour often predicts future behaviour [34, 35]. Studies by Anderson and Mittal [2] emphasise that high levels of satisfaction predict repeated purchases and customer loyalty, while low satisfaction is a strong indicator of churn. This predictive power makes customer satisfaction a key target for improvement efforts. Companies that can proactively predict satisfaction levels based on product reviews and other feedback are in a better position to enhance their offerings and strengthen customer relationships.

For businesses, especially subscription-based platforms like Wix, the ability to automatically predict customer satisfaction from review data could provide a huge advantage. By identifying satisfaction trends early, the company can respond to customer needs more quickly and prioritise areas that require immediate attention, leading to improved product offerings and a better customer experience overall.

1.2 Role of Online Reviews

Online reviews have become one of the most influential sources of information in shaping consumer decisions. With the growing usage of e-commerce and digital platforms, customers increasingly rely on reviews to evaluate the quality, reliability, and value of products and services before making a purchase. Research shows that more than 90% of consumers read online reviews before making a purchase decision, highlighting their importance in the buying journey [28]. Reviews not only provide potential buyers with helpful insights from real customers but also give valuable in-

formation to understand how satisfied other customers are and what they could expect from that company or brand.

For companies, online reviews are more than just ratings. They are detailed, real-time feedback that goes beyond the results of customer satisfaction surveys. In reviews, customers freely share what they loved or did not like about a product or service, which helps businesses see what's working and what might need improvement. Chevalier and Mayzlin [11] found that online reviews can significantly impact sales, as positive reviews tend to increase consumer trust and drive purchases, while negative reviews can have an even greater impact, but in the opposite direction. This means that reviews can have a direct effect on a company's reputation and even its sales and revenue.

Online reviews are also critical for customer retention. By noticing and analysing common themes in reviews, businesses can identify possible issues that may be driving customer dissatisfaction. Studies suggest that customers who feel heard and see improvements based on their feedback are more likely to remain loyal to the brand [3]. Additionally, effective analysing and responding to customer reviews helps to build stronger user-business relationships, improves loyalty, and reduces the likelihood of churn. In this way, companies can engage directly with customers, showing responsiveness and commitment to improvements, which can further strengthen customer loyalty and trust.

In summary, online reviews are really important because they give valuable feedback from customers and can strongly affect how a brand is seen and how people make buying decisions. For businesses, using this feedback well is key to keeping a good reputation and improving their products and services to meet customers' needs. In today's competitive online market, companies that can use insights from reviews effectively are more likely to boost customer satisfaction, build loyalty, and grow their business.

1.3 Related Work

In recent years, review rating prediction become a popular problem in machine learning. Most of the recent work related to review rating prediction relies on sentiment analysis to extract features from the review text. Qu et al. [23] in their research introduce a novel feature extraction method called bag-of-opinions, to predict numerical ratings from product reviews. This approach represents opinions as combinations of root words, modifiers, and negations. By applying the proposed methodology, the study demonstrates bag-of-opinions effectiveness in providing more accurate numerical ratings.

Nabiha Asghar [4] explores various methods of feature extraction and multi-class classification to predict users' star ratings based on their textual reviews. The study contains sixteen predictive models by combining four feature extraction techniques (Unigrams, Bigrams, Trigrams, Latent Semantic Indexing (LSI)) with four supervised learning algorithms (Logistic Regression, Naïve Bayes Classification, Perceptrons, Linear Support Vector Classification (SVC)). These models were trained and tested using the Yelp dataset of online reviews to determine the most effective combination for accurate rating prediction. Using the cross-validation scores, the best performing model was the Lo-

gistic Regression algorithm in combination with the top 10,000 uni-grams and bi-grams as features with 64% accuracy.

The paper [36] applies Naive Bayes and Support Vector Machines algorithms along with different feature selection methods (TF-IDF) and higher-order n-grams) in order to perform sentiment analysis on movie reviews. The study's results indicate that the Linear SVM classifier achieves greater accuracy than the Naive Bayes classifier. Also, the experiments show that the Term frequency-inverse document frequency (TF-IDF) scheme gives maximum accuracy for linear SVM. The paper highlights that hybrid techniques, coupled with effective corpus usage and feature selection, can significantly enhance sentiment analysis outcomes.

Another study [1] by Alzami et al., proves high results of SVM along with TF-IDF achieved 87.3% accuracy in classifying customer review polarity in sentiment analysis. The authors used the Amazon food review dataset with ratings from 1 to 5, but the scores 4 and 5 were combined together to be positive as well as 1 and 2 to be negative, and 3 left to be neutral. The research highlights preprocessing steps such as removing punctuation and special characters and applying stemming for feature standardisation. Parameters optimisation was done by brute-force search, which is finding the best combination of parameters. Other used methods of ML include Random Forest, KNN and Naive Bayes. For feature extraction BOW and Word2Vec also was used.

Bampounis et al. [6] explored predicting product review ratings using machine learning models on Amazon reviews dataset. They applied preprocessing techniques such as tokenization and TF-IDF, together with word embeddings like Word2Vec and GloVe. A comparison of different machine learning algorithms for classification showed that the best-performing method in all metrics was Logistic Regression. The authors also investigated how class imbalance affects the model by using the oversampling approach. As a result, the weighted F1-score increased in the Naive Bayes classifier from 0.41 to 0.5 but did not have a significant impact on other classifiers. Moreover, the study compared the effect of using different embeddings as features. Among GloVe, Word2Vec, Doc2Vec and Tf-Idf weighted word counts. Finally, Doc2Vec embeddings perform better than both "GloVe" and Word2Vec embeddings, but still worse than the Tf-Idf weighted word counts. This study highlights the importance of preprocessing and feature engineering for rating prediction but relies primarily on classical machine learning models.

A different approach is used in [19]. As it is 1-5 stars Amazon films reviews dataset, 4-star reviews were excluded from the dataset and two classes were formulated - low (1, 2, or 3 stars) and high (5-stars) reviews. For feature selection, the authors experimented with various strategies, including the top-500 and top-900 words ranked by TF-IDF, as well as the top-200, top-600, top-900, and top-1000 words identified through information gain. Additionally, they evaluated the effectiveness of sentiment words with a frequency greater than five as features. These features were tested in various combinations using Support Vector Machine (SVM) and Naive Bayes classifiers. The best results showed the SVM classifier using the top-600 information gain words as features with an accuracy of 78%.

The study presented in this paper builds upon previous research by incorporating advanced techniques in natural language processing to predict review ratings from textual data. It extends

earlier work by focusing on more refined feature selection methods and testing a variety of machine learning models.

1.4 Natural Language Processing in Text Analysis for Customer Insights

Natural Language Processing (NLP) has become an essential tool for understanding customer feedback shared in an unstructured form of data such as online reviews, social media posts or survey responses. While traditional feedback analysis focused on structured data, like ratings or simple yes/no satisfaction responses, unstructured data could provide much deeper insights. Customers often express their experiences, preferences, and needs in more detailed and revealing ways than numbers alone. By using NLP techniques, companies can extract valuable information from this type of data, which could help them understand customer feelings and identify specific areas for improvement. This allows businesses to make well-informed decisions to improve customer experience and satisfaction.

1.4.1 Sentiment Analysis

One of the fundamental applications of NLP in customer feedback analysis is sentiment analysis, which is used to identify if the emotional meaning of the message is positive, negative or neutral. It is also known as "opinion mining". Sentiment analysis explains how a person feels about a particular topic [13]. As noted by Liu [21], sentiment analysis is valuable for businesses because it helps them quickly spot weak areas needing improvement and proactively respond to negative feedback. In this context, sentiment analysis aims to extract useful insights from vast amounts of unstructured text data and transform raw feedback into actionable business decisions.

In this study, two sentiment analysis methods will be used: TextBlob and RoBERTa. TextBlob is a straightforward, rule-based tool that assigns a sentiment polarity score to text, categorising it as positive, negative, or neutral. It is easy to use, computationally efficient and works well for basic sentiment analysis. However, due to its simplicity, it may not always capture more complex or subtle emotions in text. Despite these limitations, TextBlob is useful as a baseline method for sentiment classification in this study.

In contrast, RoBERTa is a more advanced machine learning model built on the transformer architecture, which is known for its ability to understand language context more effectively. It has been trained on large, diverse datasets, allowing it to recognise more nuanced and complex expressions of sentiment. RoBERTa is particularly well-suited for situations where understanding the context and subtle variations in sentiment is important. By using RoBERTa, this study aims to achieve more accurate sentiment classification, especially when dealing with complex or ambiguous expressions of sentiment.

Another NLP method, Aspect-based sentiment analysis (ABSA), takes sentiment analysis a step further by linking customer sentiments to specific parts of a product or service, like quality, value, or customer support. For example, a review may convey opposing sentiments (e.g., "Its performance is ideal, I wish I could say the same about the price") or objective information (e.g., "This one still

has the CD slot") for different aspects of an entity [31]. This approach gives companies a clear view of where they are performing well and where they could improve. By breaking down feedback into these detailed aspects, ABSA helps product and customer service teams understand how customers feel and what exactly is causing those feelings. While ABSA is not the focus of this study, it offers valuable potential for more detailed insights and could be explored in future research.

In this thesis, the main focus will be on general sentiment analysis using TextBlob and RoBERTa. Customer reviews will be classified as positive, negative or neutral.

1.4.2 Topic Modeling

In addition to analysing sentiment, NLP can extract key topics and themes from large datasets of textual data, such as customer reviews. Topic models learn topics (sets of words) automatically from unlabeled documents in an unsupervised way. This is an attractive method to bring structure to otherwise unstructured text data, but Topics are not guaranteed to be well interpretable. Therefore, coherence measures have been proposed to distinguish between good and bad topics [24].

Using techniques like Latent Dirichlet Allocation (LDA), companies can identify common themes in customer feedback without needing to define categories in advance [7]. This approach lets businesses uncover hidden patterns and determine which parts of their products or services customers care about the most. For example, it might reveal recurring complaints about usability or highlight features that people really love. Insights, in this case, play a crucial role when talking about companies' business strategies and goals.

Machine learning adds even more power to text analysis when combined with NLP. For instance, supervised learning models can categorise reviews by customer intent, satisfaction level, or even the likelihood of customer churn [17]. With enough historical feedback data, these models can even predict satisfaction trends, allowing companies to stay a step ahead and manage customer experience proactively. This predictive insight is especially valuable in competitive fields, where keeping customers and meeting their expectations is essential for lasting success.

Overall, NLP offers a wide range of methods that could turn unstructured text data such as social media posts, online reviews or tweets into actionable insights and improve businesses. Using sentiment analysis, topic modelling, ABSA, and machine learning, companies can better understand customers' needs, improve their products or services, and build stronger customer relationships.

1.5 Machine Learning Models for Predicting Customer Satisfaction

In recent years, machine learning (ML) usage for various tasks, such as customer behaviour or satisfaction prediction, has quickly grown and will potentially take a bigger and bigger place in the future by helping companies better understand their customers. By analysing past customer feedback and behaviour data, ML models allow businesses to gain insights into how the users feel and identify areas for improvement.

1.5.1 Classification vs. Regression in Predicting Customer Satisfaction

Machine learning models are broadly categorised into supervised and unsupervised learning paradigms, with supervised learning being the most relevant for predicting customer satisfaction. Within supervised learning, tasks are generally divided into classification and regression, both of which have distinct characteristics, methodologies, and applications. This section provides a comparative analysis of classification and regression in the context of predicting customer satisfaction from Wix product reviews.

Classification is a supervised learning task that predicts discrete labels or categories based on input features. For customer satisfaction analysis, classification models are often employed to assign reviews to predefined classes such as “positive”, “neutral”, or “negative”sentiments. These labels can also extend to specific satisfaction levels, such as a Likert scale from 1 to 5.

Some of the key characteristics of classification models include:

- **Discrete Outputs**: Predictions belong to one of the predefined categories.
- **Evaluation Metrics**: Common metrics, described in subsection 1.5.5, include accuracy, precision, recall and F1-score.
- **Algorithm Suitability**: Popular algorithms include Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and advanced deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).
- **Application in Sentiment Analysis**: Classification models are well-suited for text-based tasks where the primary goal is to identify the sentiment or satisfaction polarity of a review. For instance, a classification model might predict whether a review conveys satisfaction or dissatisfaction.

In contrast, regression is a supervised learning task designed to predict continuous numerical values. When applied to customer satisfaction, regression models might predict a continuous satisfaction score, such as a rating on a scale from 0 to 10.

The distinguishing features of regression models include:

- **Continuous Outputs**: Outputs are real-valued predictions, making them suitable for tasks requiring fine-grained estimation.
- **Evaluation Metrics**: Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared are used to measure model performance.
- **Algorithm Suitability**: Algorithms commonly used for regression include Linear Regression, Ridge and Lasso Regression, Support Vector Regression (SVR), and advanced neural network architectures.
- **Application in Customer Satisfaction**: Regression models are particularly useful when the goal is to predict exact satisfaction scores, which can provide more nuanced insights than categorical labels.

When deciding between classification and regression for predicting customer satisfaction, the choice depends on the problem formulation and the nature of the output variable. Classification is more appropriate if the goal is categorising reviews into distinct satisfaction levels. However, regression offers a more suitable framework if the objective is to predict a precise satisfaction score.

A study discussed in an article by Towards Data Science [30] compared classification and regression for predicting 1-to-5 star ratings. The study found that classification achieved better results than regression on all three datasets (see Table 1). The difference seems to be quite significant, reaching almost 6% on the Amazon Musical Instruments Reviews dataset.

Dataset	Classification	Regression
Coursera's Course Reviews Dataset	78.73%	75.84%
Amazon Musical Instruments Reviews	67.92%	62.09%
Trip Advisor Hotel Reviews	62.72%	59.27%

Table 1 *Performance comparison of classification and regression approaches on different datasets.*

In the context of Wix product reviews, classification models may be advantageous for tasks such as sentiment analysis and satisfaction categorisation, while regression models can be applied to predict average customer satisfaction scores based on textual review data. Both approaches can be enhanced using Natural Language Processing (NLP) techniques to extract meaningful features from textual data, including sentiment scores, keyword frequencies, and topic modelling outputs.

1.5.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) are a set of supervised learning algorithms commonly used for classification tasks. It is highly preferred by many as it produces significant accuracy with less computation power. The objective of the support vector machine algorithm is to find a hyperplane in N-dimensional space (N – the number of features) that distinctly classifies the data points (Figure 1). To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The main goal is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximising the margin distance provides some reinforcement so that future data points can be classified with more confidence [29].

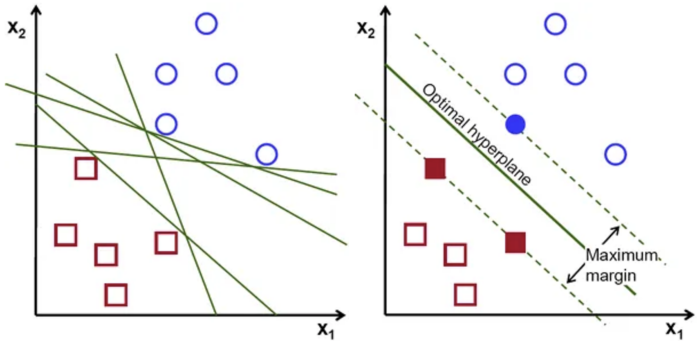


Figure 1 *Possible hyperplanes [29]*

When it is a binary classification problem, SVM transforms the input features into a higher-dimensional space, where a linear hyperplane can effectively separate the two classes. This is done using the kernel trick, which applies a kernel function like the Radial Basis Function (RBF) to handle non-linear boundaries effectively.

When dealing with multi-class classification, as in this research, two primary strategies are commonly employed: One-vs-One (OvO) and One-vs-All (OvA), which is also known as One-vs-Rest. Each approach has its own methodology and application scenarios, making them suitable for different types of classification problems.

The OvA approach involves training a single binary classifier for each class. In this method, each class is treated as the positive class, while all other classes are grouped together as the negative class, while OvO creates binary classifiers for each possible pair of classes [16]. These strategies allow SVM to classify more than two classes, making it suitable for tasks such as review classification (e.g., predicting positive, negative, and neutral reviews).

The decision on which strategy to use depends on several factors, such as the number of classes, the size of the dataset, etc. OvO approach is better when the number of classes is smaller, as it can provide higher accuracy through focused pairwise comparisons. OvA is more suitable for scenarios with a large number of classes.

While SVM is a really powerful and simple method to use, there can be some challenges when adopting it. Firstly, it may be scalability, as the number of classes increases, the complexity of training and prediction can grow significantly, particularly with the OvO approach. Also, the OvA approach can be sensitive to class imbalance, where the positive class is way smaller than the negative class.

1.5.3 Logistic Regression

Logistic regression is another widely used model in the prediction of user satisfaction, particularly to understand how different factors can affect it. For example, Liu et al. (2018) [22] used logistic regression to analyse how various aspects of service quality can affect customer satisfaction. Their analysis allowed them to break down and evaluate the influence of specific service qualities on the likelihood of customer satisfaction.

Logistic regression can be used for both binary and multi-class classification tasks. The model estimates the probability of a binary outcome based on input features, using a logistic function (also known as the sigmoid function). The logistic function transforms a linear combination of the input features into a probability value between 0 and 1. The binary classification equation is expressed as follows:

$$\mathbb{P}(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}},$$

where:

- $P(y = 1|X)$ is the probability that the instance belongs to class 1.
- β_0 is the intercept (bias term).

- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the predictor variables $X_1, X_2, \ldots, X_n$.
- e is Euler's number, the base of the natural logarithm.

If the classification problem needs to distinguish more than two classes, logistic regression needs to be extended to handle multiple categories. In multi-class classification, the goal is to predict the probability of an instance belonging to one of several classes (rather than just two). There are two primary approaches to extending logistic regression for multi-class problems: split the multi-class classification dataset into multiple binary classification datasets and fit a binary classification model on each. Two different examples of this approach are the One-vs-Rest and One-vs-One strategies [9]. Another way is to use multinomial logistic regression.

OvO and OvA approaches work in the same principles as it was explained before in the SVM subsection.

For multinomial logistic regression, the sigmoid function is replaced with the softmax function. In this method, a single model is trained to simultaneously predict probabilities for all classes.

$$\phi(y^i) = \frac{e^{y^i}}{\sum_{j=1}^{k} e^{y_k^i}},$$

where:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{i=0}^{n} w_i x_i = \mathbf{w}^\top \mathbf{x}.$$

This softmax function computes the probability of the feature $x(i)$ belonging to class j. Given the weight and net input $y(i)$. So, the probability ϕ for each class label in $j = 1, \ldots, k$ is computed [33].

Logistic regression can also be adapted to handle imbalanced data problems by using the `class_weight='balanced'` parameter. This setting adjusts the weight of each class based on its frequency, giving more importance to underrepresented classes. This is particularly useful when dealing with classes that are not equally represented, such as when there are two times more 1-star ratings than 5-star ratings.

Overall, choosing between One-vs-One (OvO), One-vs-All (OvA), and multinomial logistic regression, which uses the softmax function, depends on the specifics of the classification problem. OvO is ideal when dealing with fewer classes, but it can be slow if there are too many classes. OvA is more efficient with larger numbers of classes and is simpler to implement, making it a good choice when class imbalances exist or the classes are not closely related. On the other hand, multinomial logistic regression is the best option when class relationships are significant, and there is a need for a single model that predicts probabilities across all classes, though it can struggle with imbalanced datasets.

1.5.4 XGBoost (Extreme Gradient Boosting)

XGBoost, which stands for Extreme Gradient Boosting, is a powerful machine learning algorithm based on gradient boosting. It utilises decision trees as base learners and employs regularisation

techniques to enhance model generalisation [37]. XGBoost is known for its high performance and scalability, in particular, has been effective in satisfaction prediction tasks because of its speed and ability to handle large datasets [10].

The main idea of gradient boosting is to iteratively train decision trees in such a way that each subsequent tree corrects the errors made by the previous one. In each iteration, the model minimises the residual error from the previous iteration by fitting a new decision tree to the negative gradient of the loss function (Figure 2). This makes XGBoost highly effective in capturing complex patterns in data [27].

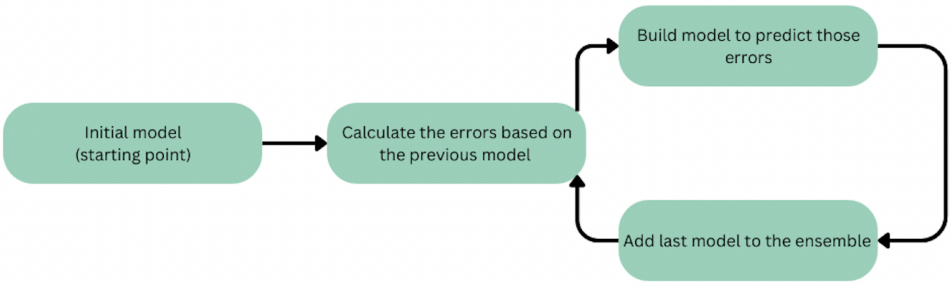


Figure 2 *Workflow of an iterative boosting algorithm*

When having a multi-class classification problem, XGBoost provides two objective functions: `multi:softmax` and `multi:softprob`. When using `multi:softmax`, the algorithm assumes that there are more than two classes to sort the data into. This objective function only outputs the class with the highest probability, instead of the probability of each class. When using `multi:softprob` objective function, it gives not the final predicted class as output, but the probability of the data belonging to each class. This is useful when you want to know not just the final decision, but also how confident the algorithm is in its decision [14].

XGBoost also provides several advantages, such as regularisation (L1 and L2), which helps prevent overfitting and handles missing data gracefully. It is often considered one of the best algorithms for tabular data and has been successful in numerous machine-learning competitions. On the other hand, XGBoost requires careful tuning of hyperparameters, also the model can be computationally expensive for very large datasets. Another disadvantage of XGBoost usage is that it is less interpretable than other, simpler models like logistic regression or SVM.

In summary, the discussed methods appear well-suited for a review ratings prediction model. By presenting diverse use cases along with their respective advantages and disadvantages, these methods will be thoroughly compared in the Methodology section of this research.

1.5.5 Classification Model Evaluation

Evaluating the performance of classification models is essential to ensure their reliability and effectiveness. Several metrics are commonly used to assess the quality of a classifier, including accuracy, precision, recall, and F1-score.

A confusion matrix (Table 2) is a table that summarises the predictions of a classification model. It compares the predicted labels with the true labels and includes four key values:

Actual / Predicted	Positive	Negative	Total
Positive	TP	FN	Actual Positives
Negative	FP	TN	Actual Negatives
Total	Predicted Positives	Predicted Negatives	Total Samples

Table 2 *Confusion Matrix for Classification Model*

- True Positives (TP): Correctly predicted positive samples.
- True Negatives (TN): Correctly predicted negative samples.
- False Positives (FP): Incorrectly predicted positive samples (Type I error).
- False Negatives (FN): Incorrectly predicted negative samples (Type II error).
- **Accuracy**: Measures the proportion of correctly classified samples. It is a valuable metric when data is well-balanced:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$
- **Precision**: Measures the proportion of true positive predictions among all positive predictions. For example, how many of the predicted patients to have a cancer, really had a cancer:
$$\text{Precision} = \frac{TP}{TP + FP}.$$
- **Recall (Sensitivity)**: Measures the proportion of true positives identified out of all actual positives. For example, how many of those patients who had a cancer were predicted to have it?
$$\text{Recall} = \frac{TP}{TP + FN}.$$
- **F1-Score**: The harmonic mean of precision and recall, balancing both metrics. Its values are between 0 and 1:
$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

In cases where class distributions are imbalanced, accuracy may not be an appropriate metric, as it could be biased toward the majority class. Metrics like precision, recall and F1-score provide better insight into model performance for minority classes [18].

2 Methodology Part

This section describes in depth the methodology of the proposed work. This section is further divided into various subsections.

2.1 Tools used

This thesis's practical part was mainly implemented using Python programming language. All calculations were performed in the Google Colab Notebooks environment, which is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and CPUs. Visualisations were made using Tableau Desktop, a business intelligence and analytics software.

2.2 Data Scraping

For data collection, a website scraping technique was selected. Python programming language offers several frameworks to ease this process. Specifically for this task, user reviews were collected from different Wix products, and the Python framework Selenium was used. Unlike traditional scraping libraries like BeautifulSoup, Selenium allows dynamic interaction with websites by handling JavaScript-rendered content. Using Selenium allows for more precise control over page elements, which is essential when handling complex websites with dynamic content, such as user-generated reviews [25]. By simulating user actions like scrolling and clicking, Selenium can effectively capture customer feedback from multiple pages and create a well-prepared dataset for further analysis.

Data was scraped from five different Wix products – Wix Stores, Wix Events and Tickets, Wix Bookings, Wix Online Programs and Wix FAQ. Products were selected by popularity, the number of reviews and the average ratings. It is important to note that data scraping adhered to Wix's terms of use. Only publicly available user reviews were collected, ensuring ethical and legal data extraction.

2.3 The Dataset

After scraping, 9,007 reviews were collected from five different Wix products. The reviews were written in several languages, such as English, Spanish, French, German, and Russian. Since the analysis uses NLP tools that work best with English, reviews in other languages were excluded.

To clean the data, irrelevant entries like spam or meaningless text were removed. Additionally, about 2,000 reviews had no descriptions and were also excluded. This left a final dataset of 6,242 reviews. The dataset contains reviews from March 2015 to September 2024. In Table 3 initial fields, their types and explanations are presented.

Field	Type	Explanation
review_id	int	Unique ID assigned for each review
user_name	string	Name of the comment author
review_date	date	Date when the review was posted
title	string	Title of the review
description	text	Main text content of the review
rating	int	Star rating of the review

Table 3 *Initial dataset description*

2.4 Data Preprocessing

Data preprocessing is an important step in the data modelling and sentiment analysis process, significantly impacting the effectiveness and accuracy of NLP models. This section describes the data preprocessing methodology applied to review the dataset before starting any modelling.

2.4.1 Text Cleaning

Text cleaning involves removing irrelevant content and noise from the data. Common steps include:

1. **Removing Punctuation, Special Characters, Numbers and Hyperlinks**:

   While punctuation and special characters are used to clarify the meaning of the text, punctuation often does not contribute to semantic meaning and should usually be eliminated to simplify the dataset.

   In any dataset, hyperlinks lose their significance and are only functionally useful. In this review dataset, the main focus is on sentiment. Therefore, it is essential to remove any hyperlinks from the dataset. Additionally, it is important to consider how numbers are treated – they may bring important information to the text, but they can also skew the sentiment analysis if not handled properly. As a result, cleaning the dataset involves eliminating both hyperlinks and unnecessary numerical data to ensure accurate sentiment representation.

2. **Lowercasing**:

   In customer reviews, people often write without following standard grammar rules, using a mix of upper and lower case letters. This can be problematic because many methods for sentiment analysis are case-sensitive. To avoid possible issues, the entire dataset text is converted into lowercase.

3. **Replace Contractions**:

   Replacing contractions is an important preprocessing step in NLP, which involves converting contracted forms of words into their expanded versions. Contractions are commonly used in everyday language (e.g., “don’t” becomes “do not”, “it’s” becomes “it is”), and they can introduce ambiguity in text analysis if not properly handled. Expanding contractions can improve

tokenisation by reducing the number of unique tokens that models must recognise. This simplification can lead to better performance and accuracy in NLP tasks.

2.4.2 Tokenization

1. **Word Tokenization**:

   Because of the unstructured data which we have in this review dataset, it would be hard to use this dataset for ML and NLP tasks. So, one of the primary reasons for tokenisation is to convert textual data into a numerical representation that can be processed by machine learning algorithms. With this numeric representation, we can train the model to perform various tasks, such as classification, sentiment analysis, or language generation. [5]

2.4.3 Normalization

1. **Stemming**:

   The goal of stemming is to simplify and standardise words, which helps improve the performance of information retrieval, text classification, and other NLP tasks. By transforming words to their stems, NLP models can treat different forms of the same word as a single entity, reducing the complexity of the text data. Some studies show that the stemming technique performed best in terms of computational speed when compared to other preprocessing techniques. [32]

   Stemming reduces a word to its root form by cutting off prefixes or suffixes. For example, stemming would reduce the words “running”, “runner”, and “runs” to their stem “run”. This allows the NLP model to recognise that these words share a common concept or meaning, even though they have different forms. [12]

2. **Lemmatization**:

   Lemmatisation is quite similar to stemming. Both are NLP techniques for reducing words to their root or base forms, but they do so in different ways and with distinct goals. Lemmatisation reduces a word to its “lemma,” which is its meaningful base form found in the dictionary. This method is more accurate but slower compared to stemming.

2.5 Exploratory Data Analysis

For a better data understanding, before starting the modelling part, it is important to make an exploratory data analysis and have a wider look into the dataset.

After data cleaning, the final dataset consists of a total of 6,242 reviews. Most of these reviews are for Wix Stores, which has 2,692 reviews, while Wix Online Programs has the fewest, with only 463 reviews. The number of reviews for each app is visualised in Figure 3.

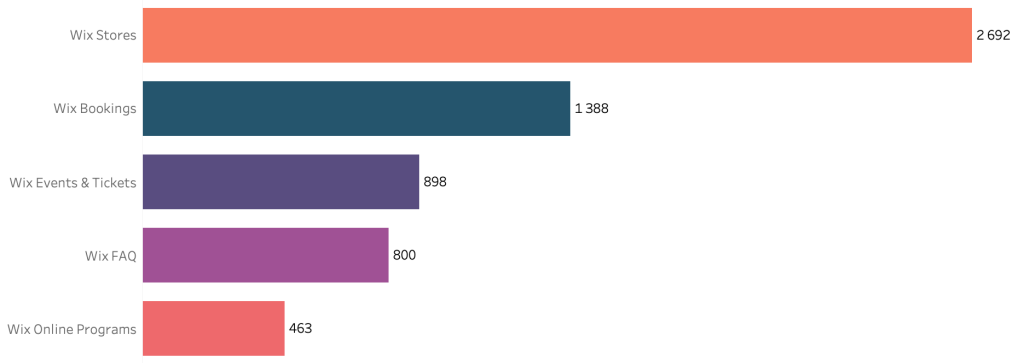


Figure 3 *Count of reviews per product*

When investigating the reviews for each product (Figure 4), it is clear that Wix Stores leads with 50% of reviews being 5 stars, followed by Wix Online Programs with 32% of 5-star reviews. In contrast, Wix FAQ has the highest proportion of 1-star reviews at 32%. Wix Bookings and Wix Tickets & Events show relatively good results, with 35% and 38% of reviews being 4 stars, respectively.

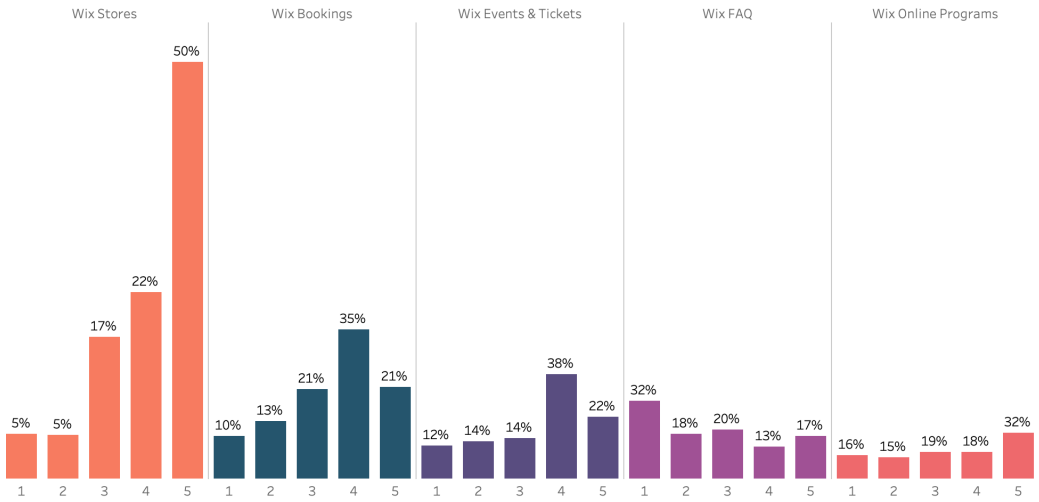


Figure 4 *Rating Distribution by Product*

When looking into the average rating distribution by-product (Figure 5), it shows some of the same patterns as the previous one (Figure 4) – Wix Stores has the biggest rating (4.07), and Wix FAQ has the lowest (2.66). The ratings for the other three products are quite similar, with Wix Online Programs having an even lower rating than both Wix Events & Tickets and Wix Bookings.

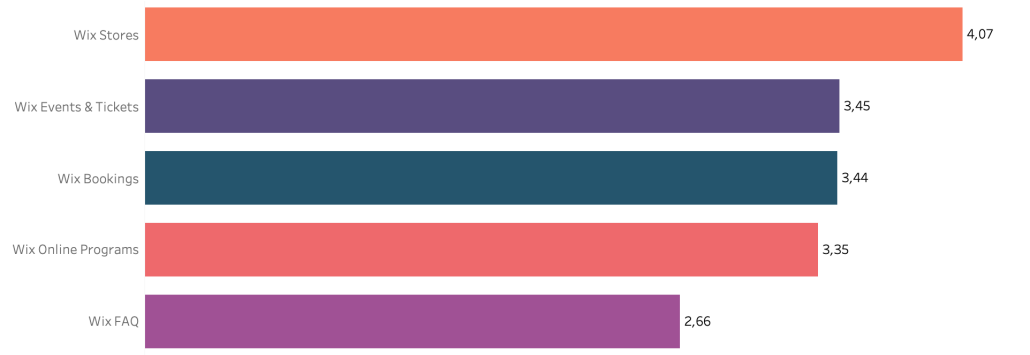
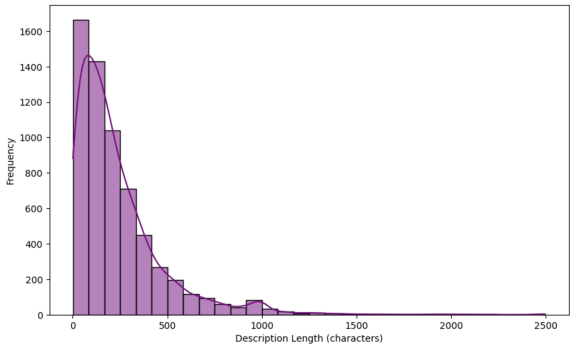
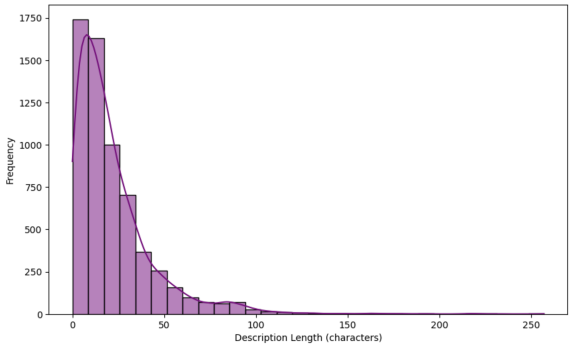


Figure 5 *Average Rating Distribution by Product*

In Figure 6, the graphs present the review description length (by characters) before data cleaning and after. Before, the mean description length was 236 characters per review, and after, it was only 27 characters. Data preprocessing drastically reduced the number of characters, which should improve the performance of the ML and NLP models significantly.



(a) *Description Length Distribution before Data Preprocessing*



(b) *Description Length Distribution after Data Preprocessing*

Figure 6 *Description Length Distribution*

When looking at the Table 4 below, we can see the tokens most often used by ratings. This table does not give much insight because it has a lot of recurring words related to the product (”wix”, “app”, “page”, etc.). To have a deeper analysis, a topic modeling analysis will be done further.

1-Star Ratings	2-Star Ratings	3-Star Ratings	4-Star Ratings	5-Star Ratings
wix	book	would	would	wix
app	app	book	app	app
page	wix	wix	book	easy
work	event	add	event	use
use	page	app	option	would
book	option	option	like	great
add	one	need	add	love
time	add	like	wix	add
one	would	one	one	store
event	time	customer	use	like

Table 4 *Frequently appeared tokens by star ratings*

Overall, the exploratory data analysis provided a deeper understanding of the dataset and the distribution of reviews, laying a solid foundation for feature engineering and subsequent model development.

2.6 Topic Modeling

In order to have a briefer look into the topics dominating the reviews and to discover hidden patterns, topic modelling was applied to the dataset. Using the Latent Dirichlet Allocation (LDA)

method, the most popular probabilistic method for identifying topics, different perspectives to explore data were exploited.

Firstly, to decide the number of topics, the coherence score was calculated (Figure 7) for each number of topics (from 2 to 15 topics). For the coherence score calculation, the C_v measure was selected. The C_v metric combines statistical measures with semantic similarities using a sliding window approach and cosine similarity. This makes it more robust than purely statistical measures like perplexity. However, the model showed the best results with only three topics.

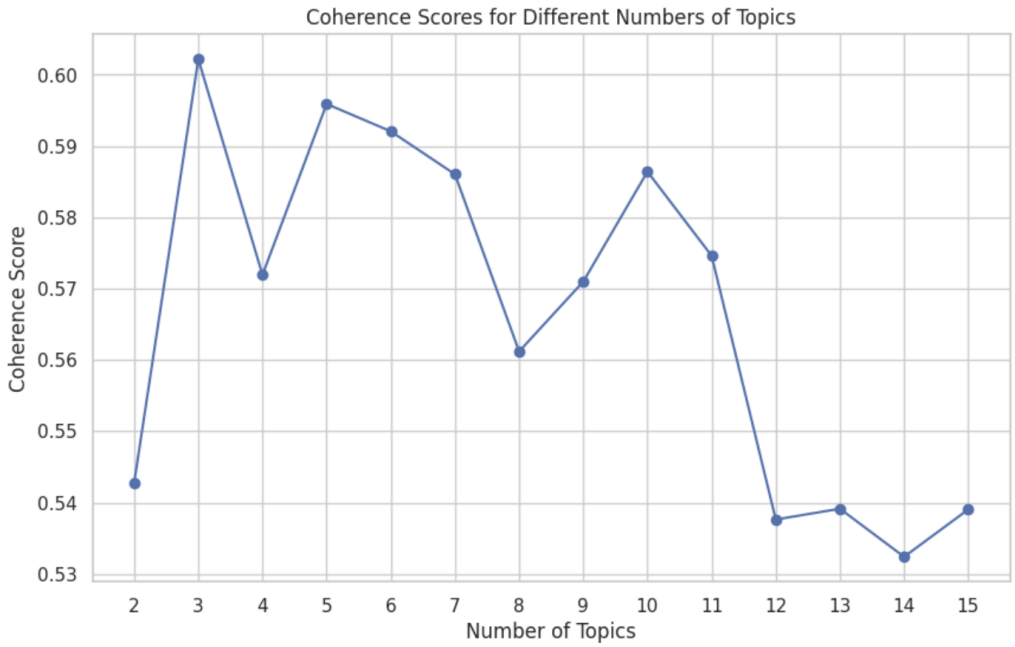


Figure 7 *Coherence Scores Across Different Numbers of Topics*

2.6.1 Topic Modeling by Product

For each product, 3 topics were generated (Table 5). Each topic is a collection of words that co-occur frequently in user's feedback or data related to the specific product.

Table 5 *Topic Distribution by App Name*

App Name	Topics
Wix Bookings	**Topic 0** – book, one, would, app, like, service, add, time, want, client
	Topic 1 – book, class, would, add, service, option, need, hour, wix, time
	Topic 2 – book, app, would, service, client, time, wix, use, option, appointment
Wix Events & Tickets	**Topic 0** – event, would, app, page, add, one, ticket, like, option, great
	Topic 1 – event, ticket, app, use, would, form, option, need, like, wix
	Topic 2 – event, ticket, app, wix, email, work, need, great, button, add
Wix FAQ	**Topic 0** – faq, app, add, page, would, site, wix, work, one, use
	Topic 1 – work, like, app, faq, color, category, option, even, one, use
	Topic 2 – question, faq, app, page, wix, answer, fix, add, please, work
Wix Online Programs	**Topic 0** – program, app, page, course, wix, would, online, need, feature, add
	Topic 1 – course, would, like, wix, online, one, use, add, option, app
	Topic 2 – step, program, add, course, able, like, quiz, challenge, member, need
Wix Stores	**Topic 0** – customer, order, product, need, add, ship, one, wix, possible, page
	Topic 1 – great, easy, use, good, product, website, wix, awesome, would, far
	Topic 2 – store, wix, product, would, app, add, option, customer, order, make

By examining these word clusters, recurring themes that reflect user concerns, needs, or desires were identified. The main principles used for insight generation were looking at the most common words, finding repeating topics, grouping similar terms to understand what users want, connecting those findings to the product features, and turning those patterns into useful suggestions

Wix Bookings:

- Topic 0 – the words “book”, “service”, “time”, and “client” suggest that users are focused on the functionality of booking services, managing schedules, and client interaction. Adding more flexibility to customise booking options could enhance user satisfaction.
- Topic 1 – emphasises scheduling and class-related needs (”class”, “hour”, “time”). This may indicate a focus on managing classes or scheduling-specific features.
- Topic 2 – highlights app functionality for bookings, with a focus on appointments and options (”appointment”, “use”, “option”). This suggests users might want enhanced appointment-handling features in the app.

Wix Events & Tickets:

- Topic 0 – addresses event creation and ticketing features (”event”, “ticket”, “page”, “option”). Users likely want flexibility in managing event pages and ticketing options.
- Topic 1 – discusses usage and functionality for events and tickets (”use”, “form”, “need”). This might reflect a desire for better ticketing workflows and customisation.

- Topic 2 – focuses on email and button-related features for events (”email”, “button”). Users might expect improved communication or CTA options for events.

Wix FAQ:

- Topic 0 – deals with the general functionality of FAQs on websites (”faq”, “add”, “site”). Users might be discussing the integration and usability of FAQ pages.
- Topic 1 – suggests interest in customisation and appearance of FAQ sections (”color”, “category”). Users may want more control over FAQ aesthetics.
- Topic 2 – centres on FAQ-specific content like questions and answers (”question”, “answer”, “fix”). Users might be highlighting technical issues or feature requests related to FAQ management.

Wix Online Programs:

- Topic 0 – discusses online programs, courses, and needed features (”program”, “course”, “online”). Users may want expanded functionality to offer and manage online programs.
- Topic 1 – focuses on course-related specifics and options (”course”, “option”). Indicates potential interest in course customisation and usability.
- Topic 2 – suggests advanced features like quizzes and challenges (”quiz”, “challenge”). Users might be asking for tools to enhance interactivity and member engagement.

Wix Stores:

- Topic 0 – highlights customer and order-related concerns (”customer”, “order”, “ship”). Users may want improvements in handling customer orders and shipping processes.
- Topic 1 – suggests positive feedback and general usability of Wix Stores (”great”, “easy”, “awesome”). Highlights user satisfaction and usability of the platform.
- Topic 2 – discusses store customisation and product options (”store”, “product”, “option”). Reflects user interest in making stores more customisable and functional.

Overall, when looking into the full picture, words like “app”, “add”, “option”, and “wix” often appear, indicating a common desire for more features, customisation, and smoother app integration. Also, it is not surprising that each app category’s topics align with specific user needs (e.g., booking services, managing events, customising FAQs, enhancing e-commerce).

2.6.2 Product Topic Modeling by Ratings

To evaluate the situation from another angle, a topic modelling was performed only for one product – Wix Tickets & Events was made. For each app rating, 3 topics were generated (Table 6).

Table 6 *Topics associated with ratings for Wix Events & Tickets*

Rating	Topics
1-Star	**Topic 0** – event, get, app, page, site, use, day, wix, work, money
	Topic 1 – event, wix, page, app, fee, use, one, editor, make, get
	Topic 2 – event, ticket, wix, add, need, date, ability, one, change, use
2-Star	**Topic 0** – event, want, use, option, change, time, show, button, one, wix
	Topic 1 – event, ticket, view, box, sell, mobile, page, app, two, look
	Topic 2 – event, page, would, option, button, like, wix, need, app, day
3-Star	**Topic 0** – event, page, add, app, change, would, please, ticket, payment, one
	Topic 1 – event, would, ticket, page, button, add, option, need, use, like
	Topic 2 – event, app, would, feature, like, time, change, ticket, show, use
4-Star	**Topic 0** – event, would, app, one, page, ticket, option, great, two, like
	Topic 1 – event, would, add, app, ticket, option, like, need, use, page
	Topic 2 – event, would, app, change, could, great, one, make, like, ticket
5-Star	**Topic 0** – event, would, app, show, add, ticket, option, great, one, need
	Topic 1 – event, app, add, would, great, one, guest, page, wix, feature
	Topic 2 – event, ticket, app, would, need, option, love, registration, email, feature

1-Star Ratings (Low Satisfaction)

- Topic 0 – users express dissatisfaction with basic functionality, referencing issues with events, pages, and the app itself. Words like “work” and “money” suggest frustrations with app reliability and perceived value.
- Topic 1 – complaints revolve around “fees”, “page”, and difficulties with the “editor”, indicating usability issues and hidden costs.
- Topic 2 – the focus is on the ticketing system, words “add”, “need”, “ability”, “change” and “ticket” express the need for more features and abilities for tickets.

2-Star Ratings (Low to Moderate Satisfaction)

- Topic 0 – users mention “option”, “change”, “time”, “button” which suggest the need of button related changes, which may save time for users.
- Topic 1 – mentions of “sell” and “mobile” suggest struggles with e-commerce or mobile accessibility.
- Topic 2 – repeated mentions of “option”, “need”, “button” and “like” may indicate interface issues or dissatisfaction with customization options.

3-Star Ratings (Moderate Satisfaction)

- Topic 0 – word “payment” suggests that users are asking for payments improvements. The mention of “change” and “please” indicates requests for modifications.

- Topic 1 – keywords reflect a mix of neutral sentiments, with an emphasis on “would” and “add”, suggesting a desire for additional functionality and improved usability.
- Topic 2 – words like “feature”, “time” and “would” show appreciation for existing functionality but also a desire for new features and more efficiency.

4-Star Ratings (High Satisfaction)

- Topic 0 – positive mentions such as “great” and “like” shows the user satisfaction, but minor room for improvement exists as words “option” and “would” mentioned.
- Topic 1 – mentions of “like”, “need”, and “add” indicate requests for further enhancements, especially around ticketing and customization.
- Topic 2 – positive terms like “great” appear alongside improvement requests (”could”, “make”).

5-Star Ratings (High Satisfaction)

- Topic 0 – the app is highly appreciated for its event management, with keywords like “show”, “option”, and “great” indicating satisfaction with existing capabilities.
- Topic 1 – users are satisfied with app features, especially guest management, because of keywords “guest”, “feature”, “love”.
- Topic 2 – mentions of “registration”, “email”, “feature” and “love” suggest that users are impressed with registration to events and email functionalities.

In conclusion, across all ratings, some words such as “event”, “ticket” and “app” repeat frequently but it is not surprising because of the topic. Summarising low ratings, users are asking for new features for improved usability, expressing the need for more customisation abilities and suggesting dissatisfaction with hidden fees. On the other hand, high ratings frequently emphasise terms like “great” and “love”, which indicates user satisfaction, but still, there is some space for improvement.

2.6.3 Topic Modelling by Rating

The third approach to look deeper into topics was by analysing only ratings. By having 3 topics for each 1-5 rating, it was possible to identify specific themes and concerns associated with different levels of user satisfaction (Table 7).

Table 7 *Topics associated with ratings*

Rating	Topics
1-Star	**Topic 0** – wix, app, work, page, book, need, one, customer, use, get
	Topic 1 – app, faq, wix, question, page, site, use, work, add, time
	Topic 2 – event, wix, app, page, use, add, site, time, book, one
2-Star	**Topic 0** – event, page, faq, app, add, one, change, work, wix, show
	Topic 1 – book, service, time, app, client, wix, option, one, add, feature
	Topic 2 – event, wix, page, app, option, would, need, make, use, like
3-Star	**Topic 0** – product, wix, store, customer, feature, app, need, add, option, payment
	Topic 1 – book, add, wix, page, app, one, option, would, use, need
	Topic 2 – would, event, book, app, one, like, add, service, option, customer
4-Star	**Topic 0** – would, app, like, add, product, wix, customer, option, page, store
	Topic 1 – book, would, like, option, time, customer, one, great, client, app
	Topic 2 – event, app, one, use, would, wix, option, add, great, ticket
5-Star	**Topic 0** – page, one, create, event, would, good, option, like, add, time
	Topic 1 – easy, wix, use, store, great, love, website, book, would, amaze
	Topic 2 – app, add, would, amaze, one, option, feature, great, need, customer

From the first look, looks like Topics matches star ratings pretty well. In low ratings (1-2 starts), we see potential issues with usability, as words like “need”, “add”, “work” and “use” appear frequently. Mid ratings (3 stars) indicate more balanced feedback, which is expected. We still see some “add”, and “need” words, but we can also find some positive mentions as “like”, and users discuss features and options, but there is still room for improvement in usability or added functionalities. High ratings (4-5 stars) focus on more positive feedback – ease of use, specific features that users enjoy, and the platform’s ability to fulfil customer expectations. Words like “great”, “love”, and “amaze” highlight satisfaction and appreciation.

1-Star Ratings:

- Topic 0 – dissatisfaction centres around basic functionality (”wix”, “work”, “need”), indicating either technical issues or poor usability. Terms like “customer” and “get” suggest customer service challenges.
- Topic 1 – this topic highlights frustrations with the FAQ app (”faq”, “question”) and general navigation or content issues (”page”, “site”).
- Topic 2 – feedback about events seems negative, with mentions of additions and usability (”add”, “use”, “time”).

2-Star Ratings:

- Topic 0 – dissatisfaction with event management is prevalent (”event”, “page”, “change”), potentially pointing to limited customisation or adaptability of event features.

- Topic 1 – the focus here shifts to service-related apps, like bookings. Issues with time management (”time”) and feature options (”option”, “add”) are noted.
- Topic 2 – users mention general improvements needed for the apps (”would”, “need”, “make”), pointing to usability or workflow inefficiencies.

3-Star Ratings:

- Topic 0 – discussions centre on Wix Stores (”product”, “store”, “customer”, “payment”) and highlight mixed reviews of store-related features.
- Topic 1 – users want improvements in bookings (”book”, “add”, “use”) and feature options.
- Topic 2 – events and services show potential but remain inconsistent. Users highlight a desire for more robust features (”add”, “like”) and customisation.

4-Star Ratings:

- Topic 0 – the focus is on productivity improvements (”app”, “add”, “like”) and store/customer management. Users are generally satisfied but may want additional customisation.
- Topic 1 – bookings and customer interactions are appreciated (”client”, “customer”), but mentions of time and options suggest opportunities for further enhancements.
- Topic 2 – event-related satisfaction rises here, with appreciation for ticketing and options (”great”, “use”), although there’s room for improvement.

5-Star Ratings:

- Topic 0 – users praise the ease of use and the ability to create effectively (”create”, “option”, “add”). The words “good” and “time” suggest users find the platform productive.
- Topic 1 – the store functionality and website tools are highlighted. Words like “love”, “amaze”, and “great” indicate high satisfaction and excitement about the features.
- Topic 2 – app-specific features and customisation capabilities drive delight. Keywords like “amaze”, “great”, and “feature” underline appreciation for the platform’s flexibility.

Summarising the results about topics by ratings, across 1-2 star ratings, users often mention technical difficulties or missing features, particularly in events, bookings, and FAQs. Complaints often arise because of inflexibility, navigation challenges, and support gaps. At 3 stars, users appreciate the available features but call for more options and improved workflows, particularly in-store and event management. 4–5-star ratings highlight ease of use, customisation capabilities, and specific success stories with bookings, stores, and events. Positive emotions are reflected in terms like “great”, “amaze”, and “love”.

2.7 Sentiment Analysis

Sentiment analysis was applied to extend the interpretation of customer reviews. By analysing the sentiment of review text together with star ratings, we can gain deeper insights into customer satisfaction. A review's sentiment might not always align with overall satisfaction. For example, customers might appreciate some aspects of a product (e.g., ease of use) while criticising others (e.g., customer support). In this section, the results of two sentiment analysis models will be compared: TextBlob and RoBERTa. Two methods were selected to compare the performance of a rule-based approach (TextBlob) and a more advanced transformer-based approach (RoBERTa).

2.7.1 TextBlob Sentiment Analysis

First, sentiment analysis was performed using TextBlob, a Python library for processing textual data. TextBlob is a simple and fast method that uses a built-in model to calculate two key aspects of sentiment: polarity and subjectivity. The polarity value ranges from -1 (negative) to 1 (positive), indicating the emotional tone of the text. The subjectivity value ranges from 0 to 1, where 0 is objective, and 1 is highly subjective. This helps to evaluate which reviews are opinion-based and which are fact-based. For each review, these values were calculated to determine the sentiment of the text.

To analyse sentiment trends across different applications and ratings, the dataset was grouped by app name and rating. For each group, the average polarity and subjectivity values were computed to provide insights into the overall sentiment of reviews per application and rating (Figure 8). This allowed to examine how the sentiment differed based on the rating given by users, as well as to explore sentiment patterns across various applications.

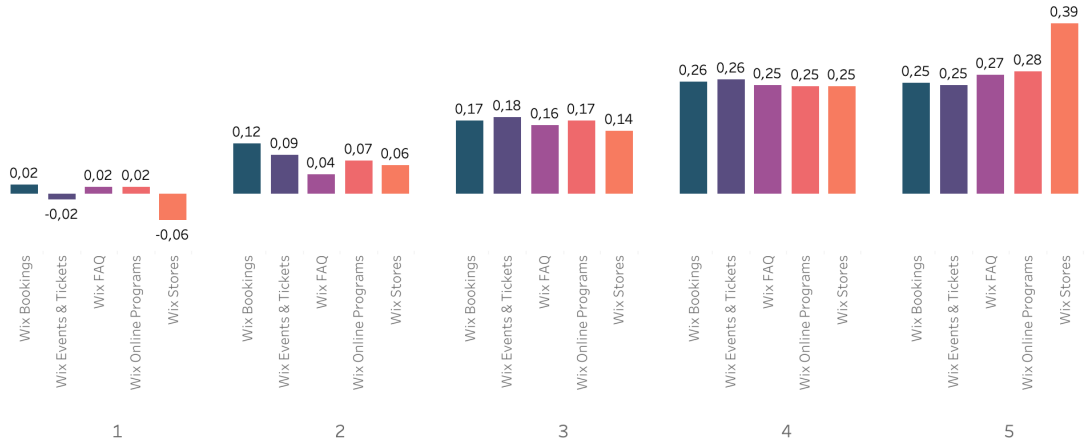


Figure 8 *Average polarity by ratings and product*

Figure 8 shows that, as expected, higher polarity values correlate with higher ratings, while lower polarity values are observed with lower ratings. Anyway, the reviews with even the weakest ratings have an average sentiment score of near 0, which indicates that it is neutral sentiment.

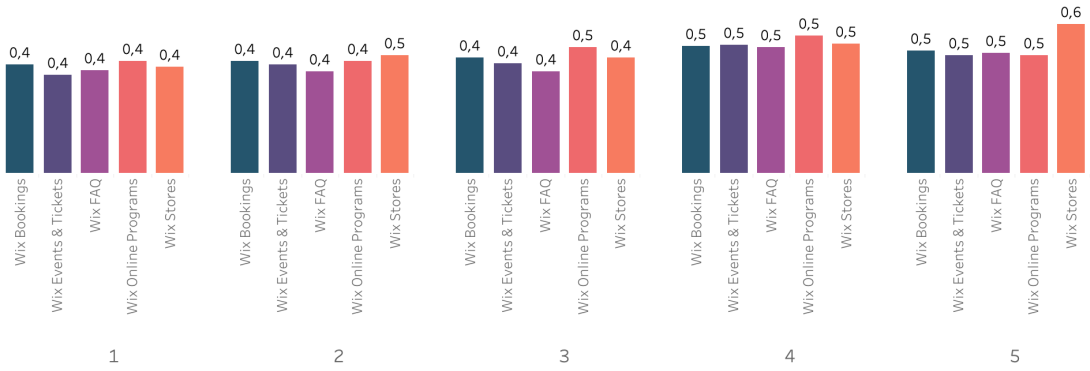


Figure 9 *Subjectivity by ratings and products*

Additionally, subjectivity analysis in Figure 9 revealed the degree to which users' opinions and feelings influenced the reviews. Subjectivity scores vary between 0.4 and 0.6, which indicates moderate subjectivity, meaning the reviews contain a mix of objective facts and subjective opinions.

2.7.2 RoBERTa Sentiment Analysis

In addition to TextBlob, the RoBERTa method was selected to analyse sentiment using a more advanced deep learning model. RoBERTa (Robustly optimised BERT approach) is a transformer-based model developed by Facebook AI, known for its high performance on NLP tasks, especially sentiment analysis.

Unlike TextBlob, which uses predefined rules, RoBERTa leverages deep learning to interpret text by understanding the context in which words are used. Transformer models like RoBERTa often perform better on raw data because they can learn directly from the natural structure of the text. For this reason, it was decided to apply this method to raw (i.e., unprocessed) data. The pre-trained RoBERTa model for sentiment analysis was applied to a labelled dataset of Wix product reviews to classify sentiment as positive, neutral, or negative.

Figure 10 clearly demonstrates a strong relationship between positive sentiment values and higher ratings. This suggests that as sentiment becomes more positive, there is a notable increase in associated ratings, highlighting the connection between expressed satisfaction in reviews and numerical rating scores.

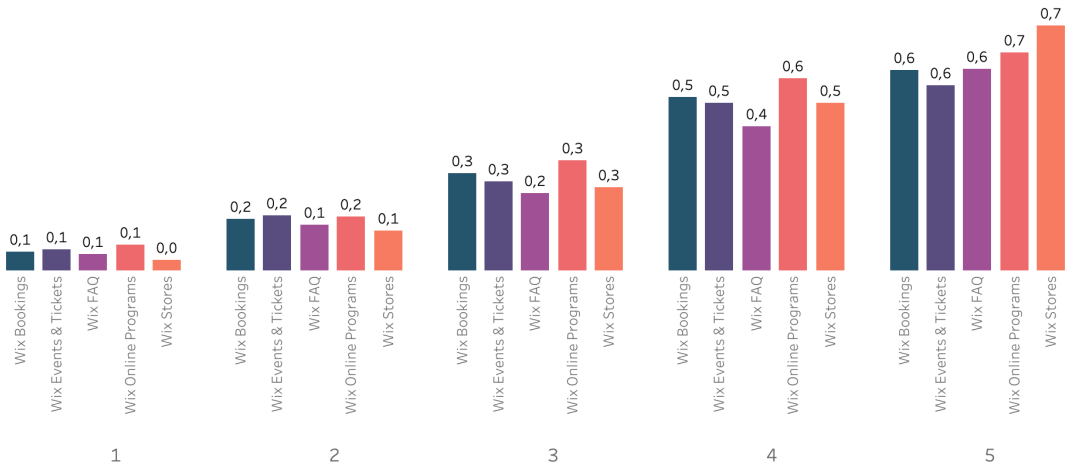


Figure 10 *Average positive sentiment values by ratings and products*

Figure 11 with average neutral sentiment scores shows a different trend. For 1 star reviews, the score is low (0.2), meaning that these reviews are mostly negative with little neutral content. In reviews of 2-4 stars, the score is higher (0.3–0.4), showing a mix of positive and negative feedback. For 5-star reviews, the score drops back to 0.2, as these reviews are highly positive with little neutral language. This suggests that neutrality is more common in mixed ratings, whereas extreme ratings are stronger in sentiment.

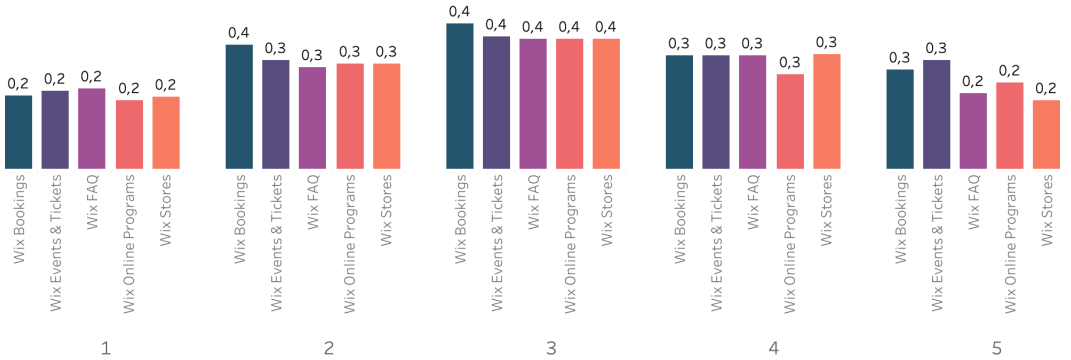


Figure 11 *Average neutral sentiment values by ratings and products*

The last, average negative sentiment values also correlate with ratings, but in the opposite direction, as ratings are growing, sentiment values are decreasing (Figure 12).

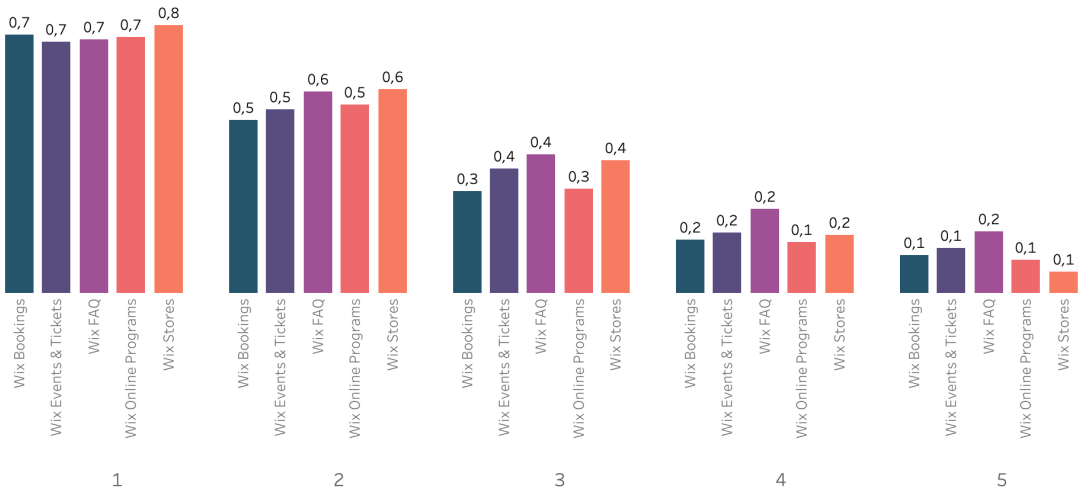


Figure 12 *Average negative sentiment values by ratings and products*

Both TextBlob and RoBERTa sentiment analysis methods produced similar and expected results, even though they operate differently. There were no significant outliers among the products, with all five showing consistent sentiment trends. This indicates that both models capture the sentiment similarly between products, highlighting the reliability of the analysis.

2.7.3 Sentiment Analysis Evaluation

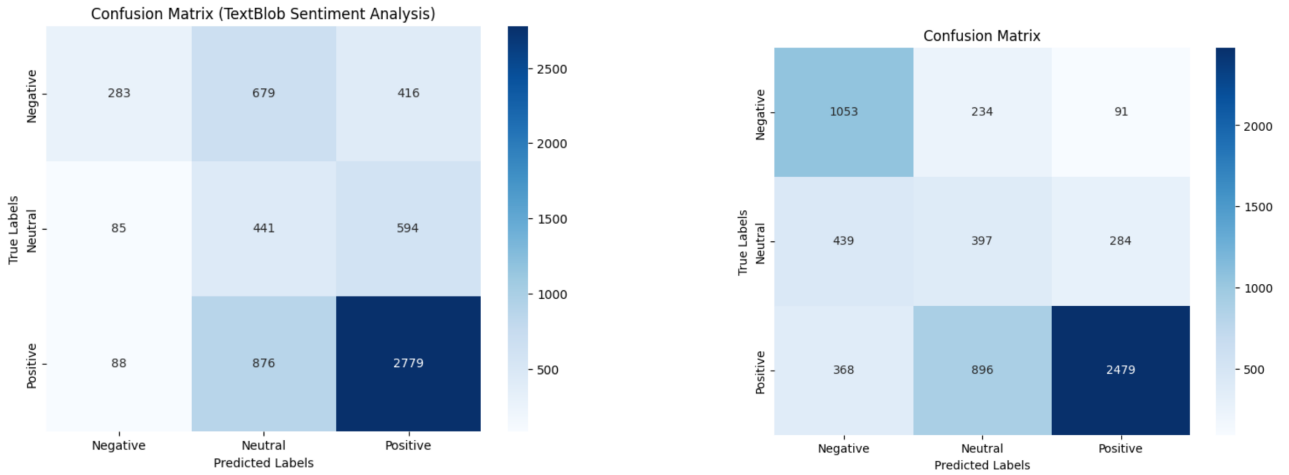
Additionally, the sentiment analysis models were compared using standard classification metrics, including precision, recall, F1-score, and accuracy, as shown in Table 8. In terms of accuracy, the RoBERTa model achieved the best performance with an accuracy of 63%. Both models performed well in classifying positive reviews (4-5 star ratings) but struggled with identifying neutral reviews (3-star ratings). The macro-averaged F1-score for RoBERTa was 0.57, reflecting overall improved balance across classes compared to TextBlob with a macro-averaged F1-score was 0.44.

Table 8 *Results Comparison of TextBlob and RoBERTa Sentiment Analysis Models*

Model	Class	Precision	Recall	F1-Score
TextBlob	Negative	0.62	0.21	0.31
	Neutral	0.22	0.39	0.28
	Positive	0.73	0.74	0.74
Accuracy	56%			
RoBERTa	Negative	0.57	0.76	0.65
	Neutral	0.26	0.35	0.30
	Positive	0.87	0.66	0.75
Accuracy	63%			

The confusion matrix Figure 13 further shows that the TextBlob model heavily misclassified negative and neutral reviews, with a large number of neutral reviews predicted as positive (876 instances). This reflects a tendency for TextBlob to overestimate positivity in sentiment. The confusion

matrix of the RoBERTa model reveals that while the model correctly classified the majority of positive reviews, it also misclassified a significant portion of neutral reviews (439 instances as negative, 397 instances as positive). Similarly, a notable portion of negative reviews (234 instances) was misclassified as neutral but differently than TextBlobl, RoBERTa showed much higher correct classification of negative reviews (1053 instances vs. 283 instances).



***(a)** TextBlob Confusion Matrix* ***(b)** RoBERTa Confusion Matrix*

***Figure 13** Sentiment Analysis Confusion Matrix*

Overall, the RoBERTa model outperformed TextBlob across most metrics, particularly in accuracy and F1-score. While both models struggled with neutral sentiment classification, RoBERTa demonstrated stronger performance for negative and positive sentiment detection. The improvement can be attributed to the pre-trained nature of the RoBERTa model, which leverages deep learning and contextual understanding of language, as opposed to the lexicon-based approach used by TextBlob.

2.8 Feature Encoding for Numerical Representation of Textual Data

In order to effectively apply machine learning models to textual data, it is essential to convert raw text into a numerical format that algorithms can process. This process, known as feature encoding, involves transforming text data into a numerical representation.

One common approach for feature encoding used in this study is Term frequency-inverse document frequency (TF-IDF) which is a numerical score that can highlight the importance of a word in a document collection or corpus. The TfidfVectorizer is used to extract the top 5000 features based on their TF-IDF scores, as different feature numbers (1000 and 2500) were tested as well but model performed the best with 5000 features. This method ensures that words with high relevance are emphasized, while common or less informative words are down-weighted. The resulting feature matrix is sparse and can be directly used in models such as support vector machines or logistic regression.

Another selected method for encoding textual data is Word2Vec, which represents words as dense, continuous vectors in a high-dimensional space. This method captures semantic relationships between words by training on a corpus to produce word embeddings. The Word2Vec model is

trained on the tokenized sentences, with each word mapped to a 100-dimensional vector, different dimensions were tested, but 100-dimensional vector showed the best results. The sentence-level representation is then created by averaging the word embeddings of all words in a given sentence. This method is particularly useful for capturing contextual information in text, allowing the model to understand word similarities and relationships.

To improve the features, the code adds sentiment scores from a sentiment analysis model, RoBERTa, which includes negative, neutral, and positive values. These sentiment scores are combined with TF-IDF and Word2Vec features, providing a richer representation of each text sample. By merging these different feature types, the model gets a deeper understanding of the text, which helps improve its ability to make accurate predictions.

2.9 Model Selection and Training for Star Rating Prediction

To predict Wix products reviews ratings, classification methods were employed. Firstly, reviews were grouped into three categories – High ratings (4-5 stars), Medium ratings (3 stars) and Low ratings (1-2 stars). When three machine learning algorithms were selected: Support Vector Machines (SVM), Logistic Regression, and XGBoost. These models were chosen based on their effectiveness demonstrated in previous research.

After data preprocessing and feature encoding, the dataset was split into 80% training and 20% testing sets. Given the data imbalance—757 High ratings, 282 Medium ratings, and 210 Low ratings—the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the training data. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line [8].

For feature representation, two methods were evaluated: TF-IDF and Word2Vec embeddings. Sentiment scores, derived using the RoBERTa and TextBlob models, were combined with these features to enhance representation. The feature values were normalised using StandardScaler, which transforms the data to have a mean of 0 and a standard deviation of 1. Normalisation was necessary because both Support Vector Machines (SVM) and Logistic Regression are sensitive to the scale of the input features. Without normalisation, features with larger magnitudes could disproportionately influence the model's performance. To optimise model performance, Grid Search was employed to tune hyperparameters for each classifier, ensuring the best combination of settings for each model.

Table 9 summarises the performance of each model across precision, recall, and F1-scores for all three classes using both feature sets. XGBoost and SVM, by using TF-IDF features, achieved quite similar results and showed the highest performance, with an accuracy of 70%. For Word2Vec features, Logistic Regression showed comparable results, reaching an accuracy of 67%. Overall, models achieved better results when using TF-IDF than Word2Vec.

Table 9 *Classification Results*

Model	Class	Precision	Recall	F1-Score
TF-IDF				
SVM	High	0.80	0.84	0.82
	Medium	0.33	0.21	0.26
	Low	0.61	0.68	0.64
	Accuracy	0.70		
Logistic Regression	High	0.87	0.76	0.81
	Medium	0.33	0.38	0.35
	Low	0.61	0.73	0.67
	Accuracy	0.69		
XGBoost	High	0.81	0.84	0.82
	Medium	0.32	0.23	0.27
	Low	0.63	0.70	0.66
	Accuracy	0.70		
Word2Vec				
SVM	High	0.90	0.65	0.76
	Medium	0.28	0.51	0.36
	Low	0.64	0.73	0.68
	Accuracy	0.65		
Logistic Regression	High	0.87	0.71	0.79
	Medium	0.29	0.41	0.34
	Low	0.64	0.75	0.69
	Accuracy	0.67		
XGBoost	High	0.83	0.78	0.80
	Medium	0.29	0.32	0.30
	Low	0.59	0.63	0.61
	Accuracy	0.67		

The results indicate that TF-IDF features generally outperform Word2Vec embeddings across all models, particularly for High and Low ratings. However, the models struggle with Medium ratings, as shown by the lower precision and recall values in this category. This limitation highlights the challenge in distinguishing medium sentiment in review data.

An analysis of the confusion matrices shown in Figure 14 reveals that all models perform relatively well with the High class, although TF-IDF tends to yield better results compared to Word2Vec. The Medium rating is generally the most challenging to classify accurately, as it is often confused with both the High and Low classes. For the Low ratings, the performance is generally acceptable, with some slight improvements seen when using Word2Vec in certain Logistic Regression and SVM models. Overall, XGBoost demonstrates more consistent success than both Logistic Regression and SVM, highlighting its superior capacity for capturing complex relationships in the data.

Results and conclusions

Extensive literature research indicates that online customer reviews significantly impact business success and customer loyalty, as they influence purchasing decisions and brand perception. With the latest advancements in Machine Learning (ML) and Natural Language Processing (NLP), businesses have the opportunity to enhance their products and services by leveraging user feedback in a purposeful manner.

In this thesis, a dataset of customer reviews was scraped from Wix, a leading website-building platform. Reviews were selected from five diverse products to capture a variety of user experiences and better represent the overall popularity and usage trends of Wix's offerings. First, the dataset was preprocessed using techniques such as lemmatisation, tokenisation, and other standard text-cleaning methods to prepare the data for further analysis and predictions.

Subsequently, topic modelling was applied, which revealed specific areas for improvement, such as requests for additional customization features, better payment processing, and enhanced mobile accessibility. These insights could help Wix prioritize product development efforts to address key user concerns. Conversely, high ratings highlighted user satisfaction with the platform's ease of use, particularly praising the Wix-Stores product.

As the next NLP step, sentiment analysis was conducted on the customer reviews dataset to evaluate how well sentiment analysis aligns with actual star ratings. Two sentiment analysis methods were employed: the simpler and faster TextBlob method and the more advanced deep learning-based RoBERTa model. Both methods performed well in predicting positive reviews (4-5 star ratings); however, both struggled to correctly classify neutral reviews (3-star ratings), likely due to the ambiguous nature of the language used in these reviews, which often contains mixed sentiments. For negative reviews (1-2 star ratings), RoBERTa significantly outperformed TextBlob, achieving an F1-score of 0.65 compared to TextBlob's 0.31. Overall, the RoBERTa model outperformed TextBlob, achieving an accuracy of 63% in identifying the sentiment of customer reviews.

To train supervised machine learning models for predicting star ratings, textual data must be encoded. For this purpose, two feature encoding methods were selected: TF-IDF, which highlights the importance of words in the dataset, and Word2Vec, which represents words as numerical vectors. After encoding the features, three ML methods were employed for star rating prediction: Logistic Regression, Support Vector Machine (SVM), and XGBoost. For features, previously calculated sentiment scores were combined with TF-IDF and Word2Vec embeddings to capture both the semantic meaning of the reviews and their overall sentiment, improving the model's predictive power. Optimal hyperparameters for each model were determined using GridSearch.

Finally, the results showed that the best performance was achieved by XGBoost and SVM with TF-IDF features, both yielding an accuracy of 70%, which represents a significant improvement over baseline methods and aligns with the performance reported in similar studies. All models struggled to classify Medium ratings effectively, with F1-scores ranging between 0.26 and 0.36. This challenge likely comes from the ambiguous nature of neutral reviews, which may contain mixed sentiments and lack clear indicators of satisfaction or dissatisfaction. The results highlight the importance of

feature representation in text classification tasks. TF-IDF worked better than Word2Vec, especially for Medium and Low ratings, likely because it gives more importance to keywords like “great” or “disappointed” that show customer feelings. Word2Vec looks at word relationships, which can make it harder to focus on these important sentiment words.

The findings suggest that machine learning models combined with NLP techniques can provide valuable insights into customer satisfaction. By focusing on High and Low ratings, businesses like Wix can identify areas of strength (e.g., ease of use, Wix-Stores) and address weaknesses (e.g., customization features, mobile accessibility, payment processes).

Future Work

Future research should prioritize improving the classification of medium (3-star) ratings, particularly by utilizing advanced contextual embeddings, such as BERT or GPT-based models. In addition, incorporating additional features like review length and user demographics could further enhance the model’s performance. Exploring alternative models and experimenting with additional preprocessing techniques may also lead to improved outcomes. Furthermore, extending the dataset and ensuring a more balanced distribution of ratings could contribute to more robust and generalizable predictions.

Limitations

One of the downsides of the data was that the dataset was unbalanced – had much more positive reviews than negative ones. It could be solved by reducing the number of positive reviews, but then the dataset would be too small. Scraping more products could help, but it takes a lot of time. Another possible limitation is that sentiment models might struggle with understanding the context of customer feedback, such as sarcasm, irony, or mixed feelings. Also, some reviews might have ambiguous sentiments or conflicting opinions (e.g., praising a product’s features but criticizing its price). This can lead to challenges in classifying reviews accurately.

References and sources

[1] F. A. et al. “Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis.” In: *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control* 5.2 (2020), pages 235–242.

[2] E. W. Anderson, V. Mittal. “Strengthening the Satisfaction-Profit Chain.” In: *Journal of Service Research* 3.2 (2000), pages 107–120.

[3] E. W. Anderson, M. W. Sullivan. “The Antecedents and Consequences of Customer Satisfaction for Firms.” In: *Marketing Science* 12.2 (1993), pages 125–143.

[4] N. Asghar. *Yelp Dataset Challenge: Review Rating Prediction*. https://arxiv.org/pdf/1605.05362. 2016.

[5] A. Ashraf. *Tokenization in NLP: All You Need to Know*. URL: https://medium.com/@abdallahashraf90x/tokenization-in-nlp-all-you-need-to-know-45c00cfa2df7 (viewed 2024-11-10).

[6] A. Bampakis, T. Spanoudis. *Products Review Rating Prediction from Users' Text Reviews*. https://github.com/stergiosbamp/nlp-review-rating-prediction/blob/main/paper/Products-review-rating-prediction-from-users-text-reviews.pdf. 2021.

[7] D. M. Blei, A. Y. Ng, M. I. Jordan. “Latent Dirichlet Allocation.” In: *Journal of Machine Learning Research* 3 (2003), pages 993–1022.

[8] J. Brownlee. *SMOTE Oversampling for Imbalanced Classification*. Accessed: 2024-12-20. 2017. URL: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[9] J. Brownlee. *One vs Rest and One vs One for Multi-Class Classification*. URL: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/ (viewed 2024-12-05).

[10] T. Chen, C. Guestrin. “XGBoost: A Scalable Tree Boosting System.” In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM, 2016, pages 785–794. https://doi.org/10.1145/2939672.2939785.

[11] J. A. Chevalier, D. Mayzlin. “The effect of word of mouth on sales: Online book reviews.” In: *Journal of Marketing Research* 43.3 (2006), pages 345–354.

[12] S. Cloud. *Stemming*. URL: https://saturncloud.io/glossary/stemming/ (viewed 2024-11-10).

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa. “Natural Language Processing (Almost) from Scratch.” In: *Journal of Machine Learning Research* 12 (2011), pages 2493–2537.

[14] M. Filho. *XGBoost Multiclass Classification in Python*. URL: https://forecastegy.com/posts/xgboost-multiclass-classification-python/ (viewed 2024-12-05).

[15] C. Fornell, M. D. Johnson, E. W. Anderson, J. Cha, B. E. Bryant. “The American Customer Satisfaction Index: Nature, Purpose, and Findings.” In: *Journal of Marketing* 60.4 (1996), pages 7–18.

[16] GeeksforGeeks. *Multi-Class Classification Using Support Vector Machines (SVM)*. URL: https://www.geeksforgeeks.org/multi-class-classification-using-support-vector-machines-svm/ (viewed 2024-12-05).

[17] M. Hu, B. Liu. “Mining and Summarizing Customer Reviews.” In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2004, pages 168–177.

[18] B. Juba, H. Le. “Precision-Recall versus Accuracy and the Role of Large Data Sets.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, HI, USA: AAAI Press, 2019, pages 4039–4048.

[19] M. Kavousi, S. Saadatmand. “Estimating the rating of the reviews based on the text.” In: *Data Analytics and Learning*. Springer, 2019, pages 257–267.

[20] P. Kotler, K. L. Keller. *Marketing Management*. 15th. Pearson Education, 2016.

[21] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[22] H. Liu, L. Lobschat, P. Verhoef. “Multichannel Retailing: A Review and Research Agenda.” In: *Foundations and Trends in Marketing* 12 (2018), pages 1–79.

[23] Q. Lizhen, I. Georgiana, W. Gerhard. “The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns.” In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, 2010, pages 913–921.

[24] S. Mifrah, E. H. Benlahmar. “Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID’19 Corpus.” In: *International Journal of Advanced Trends in Computer Science and Engineering* 2319 (4 2020), page 2020. https://doi.org/10.30534/ijatcse/2020/231942020.

[25] R. Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol, CA: O’Reilly Media, 2015.

[26] R. L. Oliver. *Satisfaction: A Behavioral Perspective on the Consumer*. New York: McGraw-Hill, 1999.

[27] OpenAI. *ChatGPT*. 2024. URL: https://www.openai.com/ (viewed 2024-12-06).

[28] S. Paget. *Local Consumer Review Survey 2022*. Accessed: 2024-11-02. 2022. URL: https://www.brightlocal.com/research/local-consumer-review-survey/.

[29] V. Parmar. *Support Vector Machine — Introduction to Machine Learning Algorithms*. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (viewed 2024-12-05).

[30] S. Poliak. *1 to 5 Star Ratings — Classification or Regression?* URL: https://towardsdatascience.com/1-to-5-star-ratings-classification-or-regression-b0462708a4df (viewed 2024-12-06).

[31] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar. "SemEval-2014 Task 4: Aspect Based Sentiment Analysis." In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Edited by P. Nakov, T. Zesch. Dublin, Ireland: Association for Computational Linguistics, 2014, pages 27–35. URL: https://aclanthology.org/S14-2004.

[32] S. Pradha, M. N. Halgamuge, Q. Vinh. "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data." In: *Proceedings of the 11th IEEE International Conference on Knowledge and Systems Engineering (KSE'19)*. Da Nang, Vietnam, 2019. https://doi.org/10.1109/KSE.2019.8919368.

[33] G. Satprayoon. *Multiclass Logistic Regression with Python*. URL: https://medium.com/@gilsatpray/multiclass-logistic-regression-with-python-2ee861d5772a (viewed 2024-12-05).

[34] P. Sheeran, S. Orbell, D. Trafimow. "Does the temporal stability of behavioral intentions moderate intention-behavior and past behavior-future behavior relations?" In: *Personality and Social Psychology Bulletin* 25.6 (1999), pages 724–734.

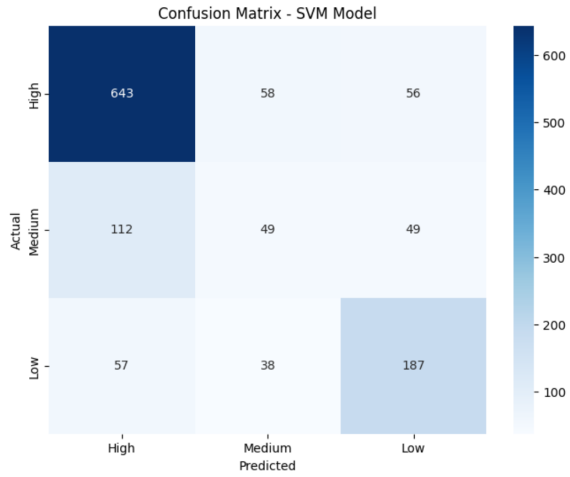
[35] M. Soderlund, M. Vilgon, J. Gunnarsson. "Predicting purchasing behavior on business-to-business markets." In: *European Journal of Marketing* 35.1/2 (2001), pages 168–181.

[36] G. Tripathi, N. S. "Feature Selection and Classification Approach for Sentiment Analysis." In: *Machine Learning and Applications: An International Journal (MLAIJ)* 2 (2015), pages 15–27.

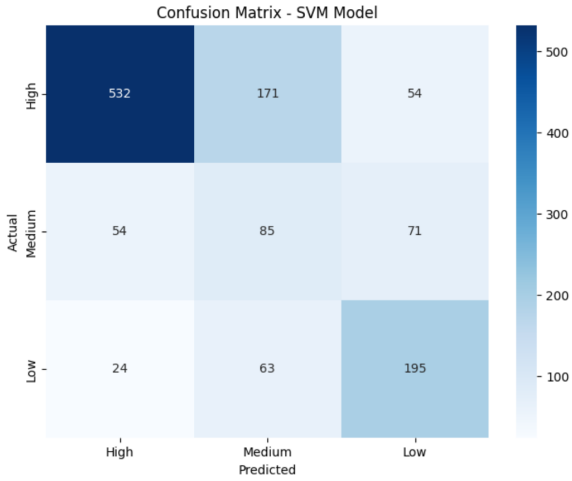
[37] A. Vidhya. *An End-to-End Guide to Understand the Math Behind XGBoost*. URL: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (viewed 2024-12-29).

Appendix 1. Confusion Matrices

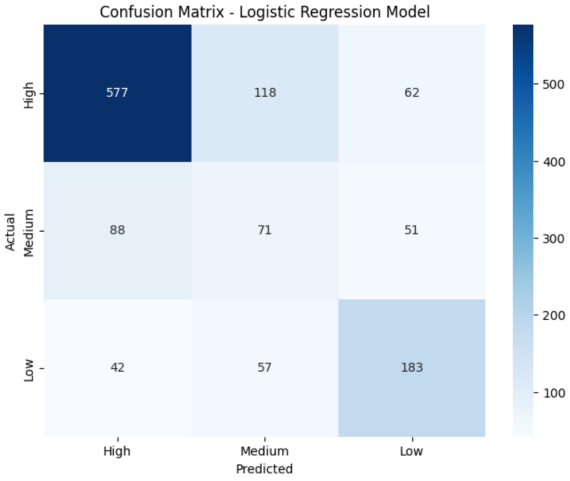
Confusion Matrices of different classification models.



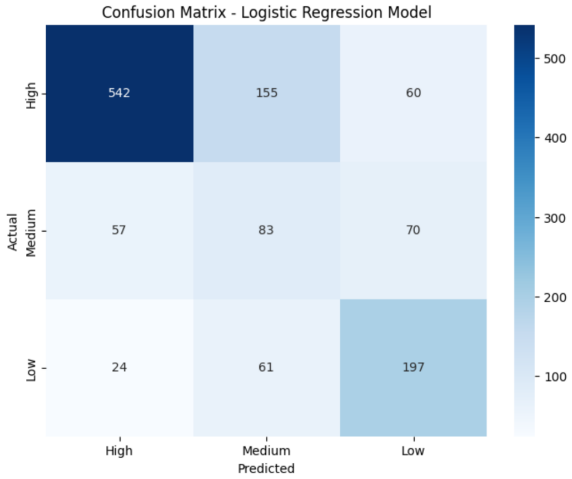
***(a)** SVM with TF-IDF*



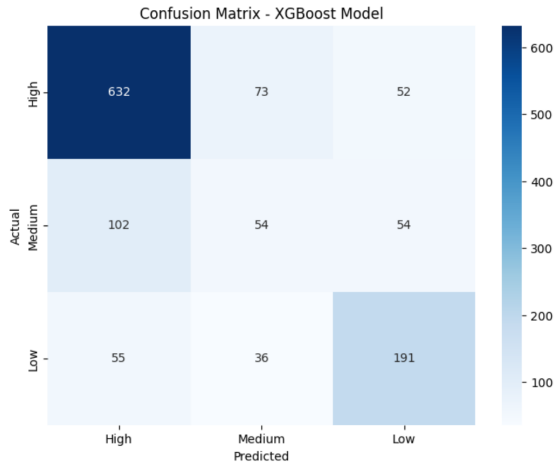
***(b)** SVM with Word2Vec*



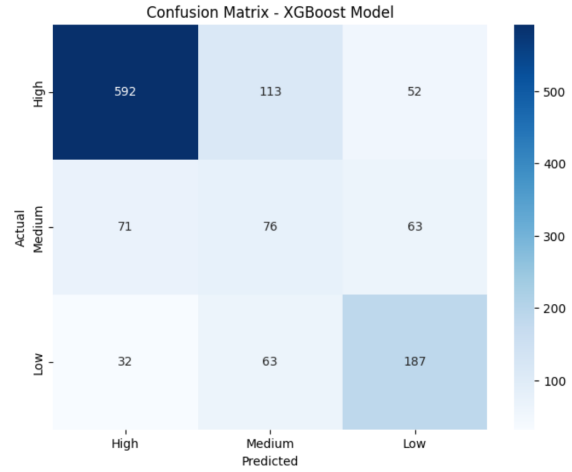
***(c)** Logistic Regression with TF-IDF*



***(d)** Logistic Regression with Word2Vec*



***(e)** XGBoost with TF-IDF*



***(f)** XGBoost with Word2Vec*

***Figure 14** Confusion Matrices*

Appendix 2. Declaration of Tool Usage

In the preparation of this thesis, the following tools were utilised to support the writing and editing process:

1. Grammarly – used for grammar, spelling, and style checks to enhance the clarity and readability of the text.
2. ChatGPT – used to improve language and check mistakes in text, brainstorm ideas and provide explanatory suggestions.

All content and conclusions presented in this thesis are the result of the author's original work, with the above tools serving as supplementary tools.

Appendix 3. Programming Code

The Python code developed for this thesis is available on GitHub and can be accessed via the following link: `https://github.com/AgneG25/MasterThesis`