THE IMPORTANCE OF MANAGERIAL RESPONSES - PREDICTING THE RATINGS OF CUSTOMER REVIEWS

IVAYLO PAPAZOV

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2051866

COMMITTEE

dr. Javad Pourmostafa Roshan Sharami
prof. dr. SanchezMelchor

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

7927

ACKNOWLEDGMENTS

THE IMPORTANCE OF MANAGERIAL RESPONSES - PREDICTING THE RATINGS OF CUSTOMER REVIEWS

IVAYLO PAPAZOV

Abstract

The following thesis paper provides a model comparison of three different models - Random Forest as a baseline, as well as two transformer models - BERT and roBERTa, on a multi-label classification task. Related to the importance of consumer reviews, the problem statement that this paper addressed will be related to whether or not the mentioned models can predict the review ratings (ranging from 1 to 5) of subsequent consumer reviews, given the textual information of the customer review itself, as well as the managerial review (MR) related to it. Furthermore, multiple sub-research questions are formed to inspect various techniques, such as data balancing in the form of weight balancing and random oversampling, and the addition of extra features related to MR, in order to increase the effectiveness of the models further. After extensive pre-processing and hyperparameter tuning, the models were compared firstly on the evaluation matrix, and secondly on error graphs, presented in the form of confusion matrices. After the described analysis, it was concluded that the BERT model, along with the random oversampling and the inclusion of additional features, yielded the best results in terms of accuracy and F_1 score. This study suggests that further research should be done on the topic, with a bigger range of hyperparameter values to be tested, as well as a bigger dataset in order to fine-tune the discussed models better.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The dataset used in this thesis is provided by an Yan et al. (2023) and Li et al. (2022). The author of this thesis complies with the terms and conditions set by the owners of the data, which retain full ownership for the whole duration and after the completion of this paper. The author of

this thesis acknowledges that they do not have any legal claim on the used data.

The code used in this thesis is publically available at the following link: Predicting the Rating of Customer Reviews - Thesis Paper . All figures and images were created by the author. Additional softwares for paraphrasing and spell-checking were used (Thesaurus, Grammarly). Parts of a publicly available code for the creation of transformers have been used by this paper, available on the Hugging Face site. This project does not involve the collection of data from human or animal participants.

2 INTRODUCTION AND PROBLEM STATEMENT

2.1 *Introduction*

Nowadays, everyone uses the Internet to make decisions. In the past, people relied mostly on word of mouth (WOM) from friends and family in order to acquire information and take action. Nowadays, however, since it is so easy to access anything online, electronic word of mouth (eWOM) has become the way of communicating what is good and what is not (Nieto et al., 2014). As mentioned by Melián-González et al. (2013), the key difference between WOM and eWOM is the huge difference in volume - as well as its publicity, making it available to everyone. And since one would want to see the experience of others before checking something for themselves, reviews are the first means of eWOM that the average person will turn to when assessing something new. What is more, this holds especially true when it comes to tourism, as studies have shown that both the valence (meaning whether or not the ratings and reviews were positive or negative) and the volume of the reviews can pose a significant impact on the mentioned business (Chen et al., 2019; Melián-González et al., 2013). Furthermore, reviews have shown to be of great importance when it comes to conversion rate, as a study by Askalidis and Malthouse (2016) discovered a 260% increase in conversion rate following the introduction of (positive) reviews.

Nonetheless, one thing that remains to be explored further, is what this thesis will refer to as managerial responses (MR). Ba and Pavlou (2002) carried out a study, showing that social media and web evaluation, such as reviews and ratings, bear significant importance to consequent consumer decisions, and is something that businesses and managers should definitely consider looking into. Thus, there are two popular actions that businesses can take. One of them, mentioned in the study by Chen et al. (2019), is censoring - deliberately hiding posted negative reviews. However, the effectiveness of this method is questionable and it can be applied only to

reviews hosted by the retailer, so this thesis will stray away from it. The other one, respectively, will be the MR that this paper will focus on. In the context of this specific research, we refer to the MR as the text carried out by the corresponding business, as a response to a customer/consumer review. Having mentioned all of that, this thesis will assess whether or not MR influence subsequent customer reviews and ratings.

When it comes to the scientific relevance that this paper will provide, the main thing that it will be concerned with is the machine learning approaches to answering this question. There have been many articles that explore sentiment analysis and ratings prediction, both using traditional machine learning approaches (Noori, 2021; Jain et al., 2021; Tripathy et al., 2016), as well as pre-trained Natural Language Processing (NLP) models (Boluki et al., 2023; Joshy and Sundar, 2022; Lehečka et al., 2020). However, little research has been conducted on also taking managerial responses into account when predicting the rating. Using pre-trained NLP models such as roBERTa and BERT, the following thesis paper will aim to address this gap in the literature. Furthermore, there have been studies that adopted the use of pre-trained NLP models, however, specifically in the area of sentiment analysis and rating prediction, there has not been a paper that has used traditional machine learning algorithms as the baseline for such a comparison.

Furthermore, the societal relevance stemming from this research will be related to the answer to the main research question. The study will aim to find out whether or not managerial responses have an impact on the rating of subsequent customer reviews, which can be directly related to the expansion of the respective business, therefore contributing to the economic growth of the country it is situated at (Elwalda & Lu, 2016). Moreover, a study carried out by (Sparks et al., 2016), examining various managerial responses to negative customer reviews, has shown that responses increase trustworthiness and decrease concern, which can once again be related to the success of a business.

2.2 *Problem Statement*

To predict the rating of the subsequent customer reviews after a managerial response, the following three machine learning algorithms will be used - a Random Forest classifier, BERT, and roBERTa transformers. The dataset that will be used to conduct this challenge is provided by Yan et al. (2023) and Li et al. (2022) and contains information from Google Local Data (2021), providing the reviews of customers and managerial responses to the respective businesses.

1.Main RQ - To what extent can using BERT, Random Forest, and roBERTa models (and their evaluation) predict the rating of subsequent customer reviews based on managerial responses?

To answer this question, some exploratory data analysis needs to be done on the data. To find if the MR are a strong predictor for subsequent customer rating, a new feature needs to be created, namely `'month_after'`, reflecting whether or not the review happened a month after the response. This, combined with the timeline provided of all the reviews and responses, will serve as the basis for the answer to this main research question.

2.Sub-RQ 1: Which data balancing method yields the best results for the transformers?

The first sub-research question this thesis will tackle will be related to the previously mentioned problem of imbalanced data. To do that, 2 different methods will be used - class weighting and random oversampling. The study by Madabushi et al. (2020) will serve as a basis for the oversampling method, while the class weighting alternative, using the 'imbalanced-learn' package (Lemaître et al., 2017) is inspired by Younes and Mathiak (2022).

3.Sub-RQ 2: To what extent do additional managerial characteristics enhance the predictive capabilities of the transformer models ?

The second and final research question of this paper will take a deeper look at the meaning behind the MR. Using SpaCy (Honnibal et al., 2020)- a widely popular NLP library, the managerial responses will be classified as either personal or general. SpaCy is a perfect fit for this research as it works exceptionally well with pre-trained NLP models such as BERT and roBERTa. Moreover, the mentioned NLP library excels at named entity recognition location and extraction from unstructured text, such as pronouns and personal names, which is precisely what will classify a response as either personal or general in this paper (TYAGI et al., 2023). Additionally, two other variables related to the properties of the MR will also be taken into consideration to see whether or not the transformer models will yield a better result. Finally, one additional variable, specifically related to the price range of the business, will be implemented, inspired by Lei and Law (2015).

After conducting the above-mentioned analyses, it was concluded that the BERT transformer model, along with the random oversampling balancing technique and the addition of extra variables, yielded the most

promising results, with the roBERTa model sharing the same characteristic being closely second.

3 LITERATURE REVIEW

3.1 *Introduction*

The following literature review will first address the question relevant to this paper, followed by providing the three different machine learning and NLP methods that will be performed to get to the prediction. After providing their strengths and weaknesses, supported by previous literature, the gaps that this thesis will fill will and its connection to the research questions will also be addressed. Furthermore, any methodology that is different from the papers provided will also be addressed and further explained. Additionally, since no literature explores MR and their effect on subsequent customer ratings and reviews through machine-learning methods, the following literature section will instead reference papers that explore prediction models based solely on the response of the customer. The fact that the target variables are both textual data allows for this comparison.

3.2 *Random Forest*

To begin, the first machine-learning approach this thesis will implement is the Random Forest (RF) (Breiman, 2001) classification method. As explained by Hossain et al. (2021), this algorithm is based on supervised learning and can be used for both classification and regression problems, making it quite popular and flexible. Furthermore, in their study, when comparing RF along with two other traditional machine learning approaches (namely XG Boost and Logistic regression), the RF algorithm yielded the highest accuracy, as well as precision and recall. Furthermore, this method is also great for addressing class imbalances in the dataset, as well as data with high dimensionality, as shown by previous studies (Hossain et al., 2021; Rodriguez-Galiano et al., 2015).

Furthermore, as this thesis will be working with textual data due to the reviews and MR, the respective vectorizer options should also be addressed. Although a number of papers (Suryaningrum, 2023; Parida et al., 2021) have shown that, when comparing the use of CountVectorizer against the TF-IDF Vectorizer, the latter indicated better results, the following thesis paper will instead use CountVectorizer, as both were tested and it performed slightly better.

Finally, the knowledge gap that this thesis will cover with the use of Random Forest as a model will be explained. Firstly, when it comes to literature concerning sentiment analysis with the use of transformers, no studies have compared the more advanced NLP models to traditional ensemble methods. That is why, for this study, the Random Forest classifier will be used as the benchmark model. Furthermore, most studies related to sentiment analysis that enable the use of this ensemble model do not provide information about the hyperparameterization tuning that they employ. That is why this paper will also address those particular details in the methodology section.

3.3 *BERT Transformer*

Furthermore, this paper will also use two pre-trained NLP models to perform the sentiment analysis needed to get the predicted sentiment. The first algorithm to be addressed is BERT(Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2018). Developed from Google Research, this model is pre-trained on a huge dataset and is able to be fine-tuned specifically for the selected task at hand - in this situation the sentiment analysis, making it far better at predicting sentiment than standard machine-learning models (Joshy & Sundar, 2022).

To begin with, the advantages of the model will first be established. BERT is able to understand language so efficiently due to the model being trained on two specific mechanisms - Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Acheampong et al., 2021). As explained by the paper, MLM refers to taking a random sentence and masking some of the words, after which the model tries to construct the masked words from the content around them, while its ability to input two sentences at once and successfully determining their order is what NSP is referring to. These two mechanisms in combination makes BERT excellent at keeping information and word relationships through long distances text-wise (Acheampong et al., 2021). Additionally, a separate research by Lehečka et al. (2020), which implements the BERT pre-trained model for sentiment analysis, also explained another large advantage of the model, namely the explainability of its prediction, since it can track down with ease its sentiment back to the tokens it extracted from the reviews. Finally, as mentioned previously, the ability of the model to be fine-tuned down to the smallest detail makes it an exceptional method for more complicated tasks such as sentiment analysis, which is what this thesis paper will tackle.

However, there are also a number of disadvantages that should be mentioned before proceeding to the next section. Because of the fact that the model is trained on such vast amounts of data, the use of such transformer

models is usually much more computationally expensive compared to their traditional ML counterparts. Furthermore, as mentioned in the paper by Acheampong et al. (2021), BERT is also limited as it is limited to monolingual classification, preventing it from being used on datasets with multiple languages. To conclude, the same article also highlights the length of the input sentence and the model drawing pragmatic inferences as further limitations that should be taken into account.

A study by Alaparthi and Mishra (2020), which conducted a sentiment analysis by employing both traditional ML approaches such as Logistic Regression, advanced deep learning models such as LSTM, and compared them to a BERT-transformer. After the classification results, the transformer yielded the highest result for all metrics used on the study. Furthermore, in another paper by Joshy and Sundar (2022), after applying 3 different pre-trained models, all based on the BERT model, on a sentiment analysis task, the standard BERT task yielded the highest accuracy on all three sets - train, test, and validation respectively. What is more, this study also focuses on reviews, showing that the model is suitable for this thesis. Because of this, specifically, the BERT base model will be used as one of the complex NLP models for this study. Finally, employing this model will also help with the establishing of the scientific relevance, as there is limited research on the comparison of transformer models with the use of a traditional ML as a benchmark, specifically in the context of sentiment analysis on customer reviews.

3.4 *roBERTa Transformer*

The third and final model to be addressed will be the roBERTa (A Robustly Optimized BERT Approach) , initially proposed by Y. Liu et al. (2019) transformer model - a newer, optimized version of the BERT. Although both models have relatively similar architectures, there are a number of differences that need to be addressed in this section before continuing further.

Firstly, the advantages of roBERTa, compared to BERT, will be explained and underlined. As illustrated by Y. Liu et al. (2019), the team behind the creation of roBERTa, the newer model is trained longer and with bigger batches, and on a new dataset, compared to its previous counterpart. Another significant difference between the transformers is the type of masking they implement. Z. Liu et al. (2021) and Y. Liu et al. (2019) both explain in their respective papers the difference in the mask implementation - while BERT employs static masking, uniformly masking all sensitive tokens everywhere, roBERTa instead uses a dynamic approach - it first establishes the sensitivity of the tokens, after which it tries to retain as much

of its information as possible. This can be of great use when analyzing reviews, as this thesis does, as it can capture important information from single words, rather than making everything uniform. Another study that implements roBERTa for sentiment analysis, specifically an aspect-based one, also showcased the power of the model, as it yielded the best performance matrix compared to other similar transformer models (Liao et al., 2021).

In the previously mentioned replication studies, when comparing BERT and roBERTa, both Z. Liu et al. (2021) and Y. Liu et al. (2019) concluded that the latter indeed yielded better results on most tasks. However, it should still be underlined that there can be many factors for this finding, which is why the study at hand will carry out a comparison of both BERT-based models. To name a few, roBERTa completely removes the next-sentence prediction objective (Y. Liu et al., 2019). What is more, as further mentioned by Acheampong et al. (2021), due to its resource-intensive nature, the computational intensity is much higher compared to its older counterpart, taking significantly more time to execute. Finally, using this transformer model also further contributes to the novelty of this thesis, as there is few research on the use of roBERTa for the sentiment classification of MR to reviews of customers.

3.5 *Conclusion*

To conclude, three models will be used for the implementation of this sentiment analysis - a Random Forest classifier, which will most likely serve as the benchmark model, as well as two NLP pre-trained models - BERT and roBERTa, respectively. The goal of these algorithms is to classify the customer reviews by predicting their rating, using information related to the managerial reviews, such as the length and timing of the response so that the importance of the MR can be highlighted. Furthermore, this study will later also explore whether or not the type of MR (personal vs general), as well as the price range of the business itself, additionally affects the predicted rating.

4 METHODOLOGY AND EXPERIMENTAL SETUP

4.1 *Introduction*

The following section will introduce the following topics in order - to begin, an explanation of the dataset will be provided, along with visualizations and explanations of the features relevant for the subsequent analyses. Finally, the necessary pre-processing and splitting will be explained.

4.2 *Dataset description*

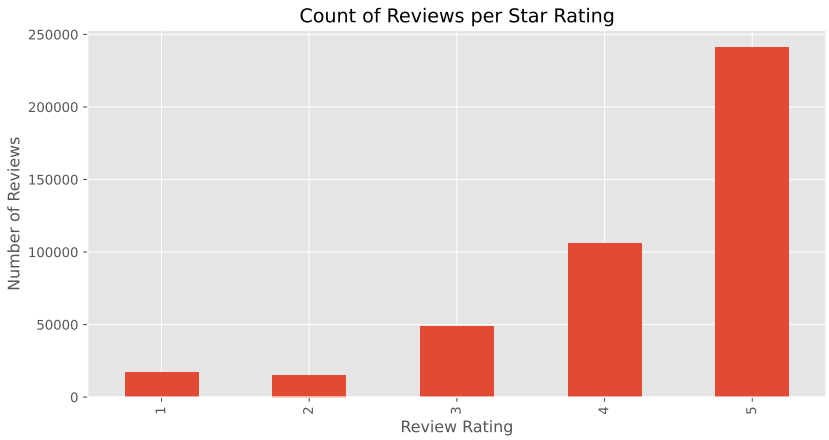

Figure 1: Counts of Reviews per Star Rating

In this thesis, two datasets, both provided by Yan et al. (2023) and Li et al. (2022) containing data from Google Local Data (2021) reviews and managerial responses will be used. The first dataset contains 3080115 entries of 7 features related to the review of the customer, such as the text of the review, the presence of pictures, the time and the rating of the review, as well as whether or not it received a managerial response. The second dataset contains 46286 entries of 14 features, mostly related to the business receiving the review and rating, such as opening hours of the business, its average rating, and number of reviews, as well as other related information. It is important to note that the variable `"resp"` is a dictionary containing the time and the content of the MR - which are of great importance of this study, so they need to be further unpacked to conduct the analysis. Furthermore, another feature, namely `"price"`, specifically important for the first sub-research question of this thesis, needs explanation. The variable divides business on four different price categories, signified by `"$"` signs. The categories range from inexpensive (under 10 dollars), moderately expensive (10 to 25 dollars), expensive (25 to 45 dollars), and very expensive (50 dollars and up), increasing along with the number of dollars signs. Finally, the two datasets are merged on the `"gmap_id"` variable which reflects the location of the business. Furthermore, it should also be noted that the dataset is imbalanced, therefore the appropriate evaluation metrics, which are the f1 score, precision and recall in this particular situation, should be taken into consideration.

Figure 2: Number of Reviews That Got a Managerial Response

4.3 *Data Pre-processing*

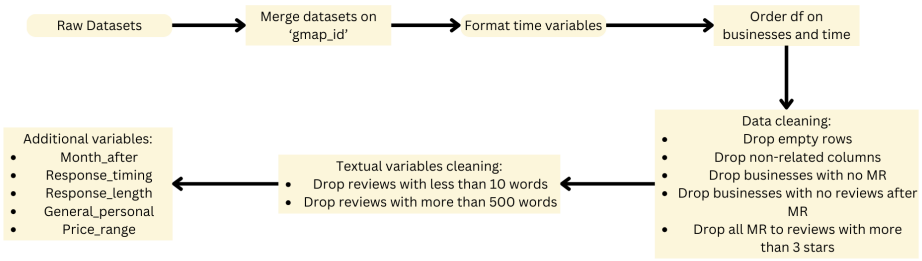

Figure 3: Data Pre-Processing

The following section will describe all the steps taken to pre-process the data in the desired way for the models to predict the rating based on the textual data from both the reviews of the customers and the MR, as well as the other non-textual variables employed in the models.

As mentioned previously, there are two datasets used for the performance of this machine learning experiment. Therefore, the first step of the data pre-processing was to merge the datasets on the variable `"gmap_id"`.

This make it possible to work with all of the variables simultaneously, as multiple pre-processing steps need to be made to both of the datasets.

Secondly, the time variables of the new dataset were formatted using the pandas package, in order to make further pre-processing easier.

Thirdly, as this thesis will be exploring how MR affect subsequent customer reviews for each individual business, the data was ordered firstly on the variable `"name_business"`, and on the variable "time" afterward.

Furthermore, a number of data cleaning steps were performed to rid the dataset of unnecessary information. To begin, all rows containing missing information related to the key variables in this experiment have been dropped, as they will be of no use later. Moreover, the same stands for columns unrelated to this thesis paper.

Additionally, some cleaning was also done related to the problem at hand. Foremost, businesses that do not carry out MR and businesses with no subsequent reviews after MR have been removed from the dataset, as they would bring no value for the answer of the research question. Additionally, MR to reviews with more than 3 stars were also dropped, as this study will not count them as responses to a negative review.

To proceed, some filtering related to the textual variables 'text' also needs to be done. Specifically, reviews with less than 10 words and more than 500 words have been dropped out for this thesis, as they have been deemed unnecessary for the analysis.

Lastly, to conclude the pre-processing part of this thesis paper, the additional features needed for the answer of the research questions need to be addressed. Firstly, the variable `'month_after'` has been created, signifying whether a review has been made a month after a MR. This is done to see whether the presence of one affects subsequent ratings. Additionally, as described previously in the problem statement section of this study (2.2), a variable `'personal_general'` has been created to also reflect the nature of the MR itself, using the 'Spacy' (Honnibal et al., 2020) library to detect named entities, such as pronouns, names of people and names of organizations. In addition to that, two other variables related to the MR themselves were addressed, namely the timing of the MR (`'response_timing'`) and the length of the response (`'response_length'`). Finally, one more variable, specifically related to the business itself rather than the MR has been added, addressing the price range of the business. The addition of this variable was inspired by the study conducted by Lei and Law (2015), who discovered that expensive establishments did not yield high customer satisfaction, compared to their cheaper counterparts. The variable is divided into four categories, previously described in section 4.2.

4.4 *Data split*

The final version of the pre-processed and cleaned dataset contains 14075 instances. In order to proceed with the machine learning models, the previously mentioned data was split accordingly - 60% for the training set, 20% for the validation and 20% for the train set. The following split was inspired from previous research that is closely related to the literature gap that this thesis aims to fill. The resulting splits consisted of 8445, 2815 and 2815 samples, respectively. Both of the split used a stratify method in order to represent the original intact data accordingly. The training set was used to teach the model, while the validation and test sets were used in the implementation and evaluation of the three models, respectively.

5 EXPERIMENTAL PROCEDURE

The following section will explain the experimental procedure that this study has implemented to reach the results it has. Firstly, the specific hyperparameter tuning for each model, including the baseline, will be explained and presented. Furthermore, the oversampling and addition of MR-related features will also be addressed, as well as the pre-processing related to it. This section will conclude by presenting all the software used, along with their version, with the goal of transparency and replicability, as well as the evaluation metrics used to assess the effectiveness of each model.

5.1 *Hyperparameter tuning*

5.1.1 *Random Forest*

Firstly, the baseline of this study and its corresponding parameters will be explained. Regarding the vectorizer used to pre-process the textual data from the variables related to the reviews and MR, as mentioned previously in the literature review section of this thesis, the CountVectorizer method from sklearn will be implemented, using 1000 as the parameter of the max features. Furthermore, in order to deal with unnecessary information, stopwords were also removed during the implementation. During the first and initial run, as previous literature suggested, a random forest classifier with 100 as the number of estimators was used in order to get the initial results.

Furthermore, random search was implemented as the hyperparameter tuning of this baseline. The selected hyperparameters to be tested, along with their corresponding ranges of values, were inspired from the

comprehensive study by Probst et al. (2019), where they tested various hyperparameters on the most traditional ML methods, on more than 20 datasets. Those ranges, along with the best value for this specific study are presented in table 2 in Appendix A (page 31) at the end of this paper. After running the required code, the best hyperparameters were defined as the following:

1. Number of trees - 1000
2. Bootstrapping - True
3. Max samples - 0.5
4. Max features - sqrt
5. Using out-of-bag samples - True
6. Minimum samples to be at a leaf node - 1

The mentioned hyperparameters provided a validation F_1 score of 0.5278, and were later used to run the model on the test set in order to assess the effectiveness and generalizability of the model on unseen data.

5.1.2 *BERT*

Foremost, the first instance of the BERT transformer model was used to run the simpler model, inspected in the main research question. Regarding the pre-processing of the data, firstly the dependent variable of this study, namely the rating of the subsequent reviews, needed to be presented in an ordinal fashion to the model. This was done by mapping the ratings as integers from 0 to 4. Furthermore, both textual variables needed to be turned to strings, concatenated, and encoded using the BERT tokenizer. Finally, data loaders for the train and validation sets were created for the training of the transformer. Both the initiation of the model and the tokenizer were from the checkpoint for BERT, provided by the checkpoint table (table 5 at the end of Appendix A (page 31).

1. Learning rate - 3e-05
2. Batch size - 32
3. Epochs - 4
4. Dropout rate - 0.0

5. Accumulation steps - 4

The best trial ended by providing a train loss of 0.9588, validation loss of 0.8385, a validation F_1 score of 0.6397 and an accuracy score of 0.6352. The above-mentioned hyperparameters were utilized for the predictions made by the model later.

5.1.3 *roBERTa*

Likewise, the roBERTa model used in this study will follow similar steps to the other transformer. Firstly, the pre-trained model called from the checkpoint will be trained once again on just the initial dataset consisting only of the rating, reviews and MR (once again, see Appendix A(page 31) for the checkpoint (table 5). Like its BERT counterpart, this transformer model will also be initialized with the train and validation dataloaders and an Adam optimizer.

To conclude this section, the hyperparameters need to be addressed once again. The table 4 can be referred to for the ranges of the hyperparameters used for the search space. Just like BERT, this model will also use random search with 10 trials to get the best hyperparameters. After the results, the best ones were the following:

1. Learning rate - 3e-05
2. Batch size - 32
3. Epochs - 3
4. Dropout rate - 0.1
5. Accumulation steps - 4

The hyperparameters yielded a training and validation losses of 0.8003 and 0.8018, respectively, a validation F_1 score of 0.6431, and an accuracy score of 0.6345.

5.2 *Data Balancing Techniques*

As mentioned in the first sub-research question from the problem statement, this study aims to explore two different data balancing techniques in order to make up for the imbalance of the previously shown rating distribution.

5.2.1 *Class weighting*

The first method used in this thesis is class weighting. The following technique, inspired by Younes and Mathiak (2022), is used to balance the training data used to train all models, specifically by assigning weights to each class (in this case rating) individually, according to the relative frequencies in the set. The performed adjustments are used mainly to put more emphasis on the minority classes, which in this specific case are the lower ratings, therefore decreasing the effect of class imbalance on the model training.

Specifically, this thesis employs the `'compute_class_weight'` function available in the scikit-learn library (Pedregosa et al., 2011), which calculates the inverse of the frequency of each rating along with the 'balanced' mode. The finalized classes are later incorporated into the loss function of both transformer models, altering the calculation the model performs.

Using the previously described hyperparameters for both BERT and roBERTA, and utilizing the newly balanced data, the following hyperparameters for both models respectively yielded the best results:

1. Learning rate - 5e-05, 3e-05
2. Batch size - 32, 32
3. Epochs - 4, 5
4. Dropout rate - 0.5, 0.1
5. Accumulation steps - 3, 4

Using the balanced data, roBERTa performed significantly better on the validation set, with a validation F_1 score of 0.6334, compared to BERT - 0.6118. The same can also be said for the losses yielded, as BERT yielded a train and validation losses of 0.8096 and 0.8600, respectively, while its robust counterpart scored 0.7088 and 0.7759.

5.2.2 *Oversampling*

The second technique utilized for the balancing of the used data is the oversampling method. As explained in the study by Madabushi et al. (2020), oversampling is a commonly used data augmentation method that creates artificial minority class instances in order to match the size of the majority class. What is more, the same study implied random oversampling and yielded a 4% increase in the performance of BERT. Furthermore, in a study carried out by Younes and Mathiak (2022), oversampling also enhanced

both the BERT and roBERTa model effectiveness, with roBERTa outperforming BERT in all instances. A formula representing the mathematical equation for random oversampling, presented by Mahmoudi and Salem (2022) is illustrated in Appendix B (1).

To achieve the random oversampling, the 'RandomOverSampler' function from the 'imbalanced-learn' (Lemaître et al., 2017) library is used in order to generate additional instances of the minority classes in the used dataset, thereby enhancing the model's predictive capabilities on all classes.

Just like with the weight balancing technique, the same hyperparameters will be used for BERT and roBERTA. After evaluation, which was once again done using 10 trials of random searching using the 'Optuna' backaging, the following parameters, for BERT and roBERTa respectively, lead to the best-performing model with the least amount of overfitting:

1. Learning rate - 3e-05, 3e-05
2. Batch size - 32, 32
3. Epochs - 5, 5
4. Dropout rate - 0.0, 0.1
5. Accumulation steps - 4, 4

When comparing both pre-trained transformer models, BERT had considerably better evaluation metrics, with a validation accuracy and F_1 score of 0.6448 and 0.6545, respectively, compared to the 0.6333 and 0.6252 outputted by roBERTa. The train and validation losses for BERT also had better results compared to roBERTa, with values of 0.4203 and 0.4506, against 0.4670 and 0.480, for the train and validation losses, respectively.

5.3 *Additional Features*

The following subsection will address the additional variables related to the MR, mentioned in the second sub RQ of this paper. The four variables that will be used to create the more advanced model are previously described in section 4.3.

Firstly, since all additional features are numerical, it is important to underline the additional pre-processing that needs to be done for the model to be able to consider them during training and evaluation. Firstly, a custom sequence classificator needs to be built in order for both BERT and roBERTa. Furthermore, the embeddings of the model need to be concatenated with the additional data. Finally, after the custom classifier

has been built and initiated, along with other pre-processing on the initial data, including the more effective data balancing method from the ones described in subsection 5.2, the additional features need to be turned into tensors, and divided into batches.

Regarding the hyperparameter optimization of the model using the additional features, the ones yielding the highest F_1 were the following (for BERT and roBERTa, respectively):

1. Learning rate - 5e-05, 3e-05

2. Batch size - 16, 32

3.Epochs - 3, 5

4.Dropout rate - 0.1, 0.1

5.Accumulation steps - 3, 4

5. Weight decay - 0.08, 0.1

After running the random search for the final three models to be utilized in this model comparison study, BERT once again performed the best. Specifically, a validation accuracy of 0.6348 and a validation F_1 score of 0.6323 were the results of the above-mentioned hyperparameters for BERT, outperforming its robust counterpart's respective scores of 0.6303 and 0.6298. What is more, BERT also demonstrated better performance in both train and validation losses, outputting 0.4267 and 0.4602, respectively, in contrast to roBERTa's 0.4701 and 0.4805.

5.4 *Software*

For the creation of this thesis, Python version 3.11.9 was used. In order to gain additional processing power for the running of the transformer models, GPU4EDU, provided by Tilburg University, as well as Google Colab Pro, providing 15 GB of additional RAM, were both implemented. For the transformers used in this dataset, BERT (Devlin et al., 2018), using the pre-trained checkpoint 'bert-base-uncased', as well as roBERTa (Y. Liu et al., 2019), using the 'roberta-base', were both implemented. Furthermore, below are presented all libraries used in this study, along with their respective versions:

1. Scikit-learn for the baseline implementation, data splitting and model evaluation- version 3.11.9 (Pedregosa et al., 2011)

2. Transformers for integration of pre-trained transformer models - 4.39.3 (Vaswani et al., 2017)

3. Torch for tensor pre-processing and hyperarameter setup - 2.2.2+cpu (Paszke et al., 2017)

4. Pandas for loading and processing of data - 2.2.1 (pandas development team, 2020; McKinney, 2010)

5. Numpy for building graphs and data preparation - 1.26.4 (Harris et al., 2020)

6. Seaborn for visualization- 0.13.2 (Waskom, 2021)

7. Matplotlib for visualization- 3.8.4 (Hunter, 2007)

8 Imbalanced-learn for the balancing of the dataset - 0.12.2 (Lemaître et al., 2017)

9. Spacy for the creation of variables using named entity recognition - 3.7.4 (Honnibal et al., 2020)

9. Optuna for the creation of the search space needed for the BERT and roBERTa transformer models - 2.21.0 (Akiba et al., 2019)

5.5 *Evaluation metrics*

5.5.1 *Initial evaluation*

For the initial evaluation of all three models, the F_1 score and accuracy will be used. More specifically for the F_1 score, the 'average' parameter of the metric will be set to 'weighted' - this calculates the metric for each class individually, after which it averages them together. This method is particularly efficient in imbalanced datasets, as it helps combat the fact that one class has significantly more instances than the others. Furthermore, during the hyperparameter tuning of the models, the validation and training losses will also be presented, with the goal of assessing whether or not the model is overfitting or not.

5.5.2 *Final Evaluation*

After the previously described steps have been completed and the model has finished evaluation and hyperparameter tuning, the following paper

will provide two ways of evaluating the predictions. Firstly, the test accuracy and test F_1 score will be provided for each model, with the ones having the best performance in each group being bolded. Secondly, a confusion matrix will be provided for each of the models, with the goal of detecting which classes were predicted the best and which ones the model had considerable difficulty predicting correctly. Additionally, processing times for each procedure will be presented in order to provide ease of transparency and replicability for further studying.

5.6 *Experiment and Model Architecture*

The visualization presented below summarizes the steps that this thesis will go through in order to answer all research questions needed to fill the literature gap, as well as contribute to the societal and academic relevance in the respective fields.

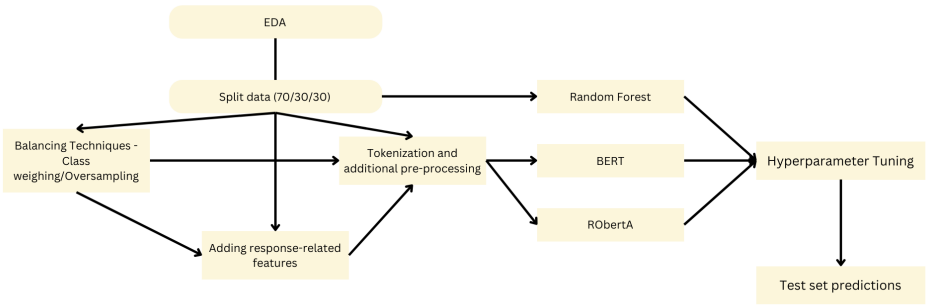

Figure 4: Experiment and Model Architecture

6 RESULTS

In the following section, the results of the previously described models will be presented. Firstly, the initial evaluation between the models, the accuracy and F_1 score specifically, will be looked at. Furthermore, the confusion matrices will be inspected in order to see any irregularities with the data. Finally, the processing time of all calculations so far will be presented, along with the most well-performing model chosen for this study.

6.1 *Model Results*

Below in table 1 are illustrated the previously mentioned evaluation metrics for all the models compared in this thesis study:

Table 1: All Model test accuracy and F_1 scores.

Models	**Accuracy**	F_1 **Score**
Random Forest - Simple	0.5517	0.5190
BERT - Simple	0.6156	0.6079
roBERTa - Simple	**0.6469**	**0.6345**
BERT - Weight Balanced	0.6224	0.6185
roBERTa - Weight Balanced	**0.6316**	**0.6212**
BERT - Oversampled	**0.6447**	**0.6458**
roBERTa - Oversampled	0.6395	0.6281
Random Forest - Additional	0.5620	0.5202
BERT - Additional	**0.6387**	**0.6424**
roBERTa - Additional	0.6352	0.6333

Firstly, it is important to interpret the results derived from the initial simple models containing only the textual information of the reviews of the customers and the MR. Unsurprisingly, the Random Forest classifier, being the baseline of this study, yielded the lowest evaluation metrics even after optimizing its hyperparameters and removing stopwords. Furthermore, as backed up by previous literature (Z. Liu et al., 2021; Y. Liu et al., 2019), roBERTA performed relatively better than its BERT counterpart,].

Secondly, the data balancing techniques need to be addressed. The first one, explained in subsection 5.2.1, utilizing the balancing of the weight of classes, had little impact on the effectiveness of the models. In fact, while it slightly increased the evaluation metrics of the BERT model, it lowered the effectiveness of the roBERTa model. Furthermore, the second technique used for this study in order to deal with the data imbalance, random oversampling, produced promising results. Although it did little to further increase the accuracy and F_1 score of the roBERTa model, it significantly improved the metrics for the BERT model, increasing it by 0.04 from the initial model offered in section 5.1.2.

To conclude the model results section, the final three models, employing the use of additional features to further boost the effectiveness of the models, will be analyzed. Firstly, regarding the baseline of this study - the Random Forest classifier, adding the numerical variables to the dataset did not yield significantly better results, although it did slightly improve the performance metric. However, regarding the BERT transformer model,

the introduction of additional features in fact lowered both performance metrics, although only slightly. Finally, regarding the roBERTa transformer, the extra data positively affected only the F_1 score of the model, while slightly lowering the accuracy metric.

6.2 *Confusion Matrices*

After the presentation of the evaluation metrics, the following subsection will delve further into the model comparison by focusing on the confusion matrices, introduced in subsection 5.5.2, with the goal of establishing whether or not the models did a sufficient job with predicting the labels (ratings) of the customer reviews. The error analysis conducted can be seen below, in figure 5, and presents the predicted labels on the X-axis, compared to the actual labels on the Y-axis.

Figure 5: Confusion Matrices of All Models.

To begin, the error analysis related to the initially offered simple models needs to be evaluated. Initially, it can clearly be seen that the baseline - Random Forest, managed to make the most correct predictions regarding the 5-star reviews - the dominant class of this study, compared to both transformer models. However, upon further inspection, it can also be seen that the model also made the most errors when classifying 4-star reviews, predicting almost two-thirds of them as 5-star instead. The reason for that is because, while effective for many tasks, including multi-class classification, Random Forest lacks an attention mechanism compared to the other two models implemented in this study, making the model significantly weaker in terms of complexity, and therefore serving as a baseline of this study. When it comes to roBERTa and BERT, the former manages to predict 3-star reviews significantly better than the latter, therefore achieving better performance metrics (see table 1).

Continuing further, the confusion matrices utilizing the two data balancing techniques need to be addressed. Firstly, it is clear both from the figures presented above, as well as from the results table mentioned in section 6.1 that the random oversampling technique, albeit marginally, demonstrates superior performance compared to the balancing of weight classes.

Thirdly, the models taking additional features into account will quickly be addressed. When it comes to BERT, the addition of the numerical variables, along with the previously mentioned oversampling, yields the best possible results out of all the models offered previously. roBERTa, on the other hand, indicates better handling of all classes, compared to its previous model counterparts.

To conclude, the confusion matrices displayed above show that the best-classified rating was the 5-star one (the dominant class), while the one the models had the most trouble with was the 4-star ratings, possibly since the content of the reviews is relatively similar. With the addition of data balancing techniques, both transformers showed reduction in minority class errors, indicating the effectiveness of their implementation. Finally, with the addition of MR-related features, all models demonstrated a significantly improved performance, indicating that the extra variables indeed contribute positively.

6.3 *Processing Times*

In the following subsection, the processing times of all models, as well as the hyperparameter tuning procedures are illustrated (see table 6 in Appendix B). The following information is present for transparency and ease of replication for further research. Furthermore, from the following

table can also be deduced that the transformer models take considerably more time for both the hyperparameter searching, as well as the actual training and prediction of the models. Finally, it can be concluded that from BERT and roBERTa, the former model takes less time to be run, compared to its former counterpart.

6.4 *Final Selected Model*

To conclude the results section of this paper, the final model will be selected, according to both the initial and final evaluation presented above. As previously described in sections 6.1 and 6.2, the BERT model containing both an oversampled dataset, as well as the additional features, yielded the most promising results. The following conclusion was based on three criteria - balanced performance, accuracy and error reduction, provided by the confusion matrices. Particularly, the model showed significant improvement in the correct label prediction of 4-star customer reviews, being the most difficult class to predict, compared to its other counterparts. Therefore, it is recommended that for future research and usage, this model should be implemented and further studied.

7 CONCLUSION AND DISCUSSION

The following section will first summarise the main problem that this study will try to solve. Afterwards, the main findings related to each of the three questions presented in the problem statement will be explained, and a conclusion for each one will be drawn. A comparison to previous work will be provided in order to see similarities or differences related to the performance of the models. To conclude, the limitations of this study will be pinpointed, as well as further research to potentially enhance the effectiveness of the models.

7.1 *Goal of the Study*

To begin, it is important to once again shed light on the goal that this study is trying to accomplish, as well as remind the societal relevance behind the paper. Based on the paper by (Sparks et al., 2016), the following thesis focused on the content of managerial responses (MR), as well as the customer reviews the responses are made for, in order to predict the rating (sentiment) of subsequent reviews from the same business. There were three models used in this study - a Random Forest classifier in the role of baseline, as well as two transformer models - BERT and roBERTa.

For the sake of expanding the model comparison even further, two data balancing techniques, namely weight balancing and random oversampling, were implemented, as well as additional features, specifically related to the MR were considered.

The societal relevance stemming from this paper is closely related to the implementation of MR as an effective customer support strategy. Furthermore, enhanced customer support through the examined responses can potentially lead to both normal and electronic word-of-mouth (WoM), driving customer retention rates and therefore contributing to the expansion of the respective businesses. Finally, the success of individual businesses drives the economic growth of the country they are situated in, thus having an effect on an even larger scale (Elwalda & Lu, 2016).

7.2 *Findings*

The following subsection will provide a quick reminder of each research question addressed in this study, as well as the findings stemming from them.

7.2.1 *Main Research Question*

The main research question of this study was the following:

1.Main RQ - To what extent can using BERT, Random Forest, and roBERTa models (and their evaluation) predict the rating of subsequent customer reviews based on managerial responses?

As highlighted in the results section (6.1), with all three models together, the framework managed to predict 60% of the subsequent customer reviews on average, having the content of the customer reviews and MR in mind. Out of the initially offered 3 models at first, roBERTa performed the best in terms of the evaluation metrics.

7.2.2 *First Sub-Research Question*

Transitioning into the first sub-research question of this study:

2.Sub-RQ 1: Which data balancing method yields the best results for the transformers?

Two data balancing methods were tested in the following study. The first approach, inspired by Younes and Mathiak (2022), implemented class weighting using the 'imbalance-learn' package, with the goal of putting more emphasis on the minority classes. The second approach mentioned

by Madabushi et al. (2020), utilizes random oversampling intending to balance out the number of instances in the minority (low rating) and majority (high rating) classes by creating synthetic instances. Following the same conclusions that both previously mentioned studies got to, BERT indeed had a drastic increase in both accuracy and F_1 score when oversampling was conducted. Furthermore, although roBERTa had a reduction in misclassification with the introduction of both methods, it still could not surpass BERT.

One thing that is important to note about the weight balancing method is that the technique might require model-specific tuning, as it can be seen that its effectiveness is drastically different between both transformer models.

7.2.3 *Second Sub-Research Question*

The second and final sub-research question of this study explored the addition of MR-related variables:

3.Sub-RQ 2: To what extent do additional managerial characteristics enhance the predictive capabilities of the transformer models ?

Inspired by the study by Sparks et al. (2016), which also takes into account MR to negative customer reviews, 3 variables related to the length, timeliness, and personalization/generalizability of the response. Additionally, another variable related to the price range of the establishment itself was also added to the model, inspired by Lei and Law (2015). Unsurprisingly, and in alignment with the mentioned literature, the addition of the features yielded a positive effect for all three models, by both increasing the evaluation metrics, as well as balancing the performance in terms of label prediction (see again figure5).

The final conclusion, which diverges from the discoveries made by Z. Liu et al. (2021) and Y. Liu et al. (2019), was that the BERT model, complemented with the random oversampling and the previously mentioned additional features, performed the best and should be used by experts interested in this paper.

7.3 *Study Contribution*

This paper contributes to the respective literature in two significant ways. Firstly, it provides a predictive model that incorporates MR and their effect on subsequent customer reviews, taking also specificities of the responses into consideration in the process. Secondly, it provides a comprehensive model comparison between both a traditional ML approach, used as a

baseline, as well as two transformer models, which are highly effective in the context of textual data analysis.

7.4 *Limitations and Further Research*

Finally, as a conclusion to this thesis paper, the limitation of the study will be addressed, as well as some insights for future research in the field.

The following research has four big limitations that hinder the ability of the models to perform even better. The first one being the dataset constraint. The dataset from Yan et al. (2023) and Li et al. (2022), used for this study, as previously described, is heavily imbalanced, having many more rows without a MR, compared to the ones with. Therefore, although some balancing techniques have been implemented in this study, stemming from previous literature, it is important that other methods, such as the synonym replacement offered by Madabushi et al. (2020) and Wei and Zou (2019), are also tested on the dataset.

Further building up on the dataset constraints, it is important to note that the size of the data itself is also something that hinders the performance of the model. Therefore, it is recommended that future research implements other, larger datasets, with the ability to provide more instances on the train set, thus enhancing the predictive capabilities further.

Thirdly, the computational constraints need to be addressed. As this study works with models trained on large datasets, such as BERT and roBERTa, running and hyperparameter tuning takes a significant amount of resources and time. Consequently, three things are recommended for future research on this topic. Firstly, the hyperparameter search space should be changed from random search to something more consistent, such as a grid search. Secondly, once again due to resource constraints, only 10 trials per model were run in order to discover the optimal hyperparameters, which is a number that should further be increased. Thirdly, a wider range of hyperparameters needs to be taken into consideration, such as bigger batch sizes, as well as more epochs.

Finally, the last constraint that this paper will mention is closely related to the gap in literature this thesis aims to fill. As previously stated, few articles are made on the topic of MR effectiveness on subsequent customer reviews, therefore this research stems all its references and ideas from sentiment analyses and predictions purely on customer reviews. Therefore, it is important to take into account that this assumption might lead to some reduction in the model's effectiveness.

REFERENCES

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of bert-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

Alaparthi, S., & Mishra, M. (2020). Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*.

Askalidis, G., & Malthouse, E. C. (2016). The value of online customer reviews. *Proceedings of the 10th ACM Conference on Recommender Systems*, 155–158.

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly*, 243–268.

Boluki, A., Pourmostafa Roshan Sharami, J., & Shterionov, D. (2023). Evaluating the effectiveness of pre-trained language models in predicting the helpfulness of online product reviews. *Intelligent Systems Conference*, 15–35.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chen, W., Gu, B., Ye, Q., & Zhu, K. X. (2019). Measuring and managing the externality of managerial responses to online customer reviews. *Information Systems Research*, *30*(1), 81–96. https://doi.org/10.1287/isre.2018.0781

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elwalda, A., & Lu, K. (2016). The impact of online customer reviews (ocrs) on customers' purchase decisions: An exploration of the main dimensions of ocrs. *Journal of customer Behaviour*, *15*(2), 123–152.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Hossain, M. I., Rahman, M., Ahmed, T., & Islam, A. T. (2021). Forecast the rating of online products from customer text review based on machine learning algorithms. *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 6–10.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer science review*, *41*, 100413.

Joshy, A., & Sundar, S. (2022). Analyzing the performance of sentiment analysis using bert, distilbert, and roberta. *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*, 1–6.

Lehečka, J., Švec, J., Ircing, P., & Šmídl, L. (2020). Bert-based sentiment analysis using distillation. *International Conference on Statistical Language and Speech Processing*, 58–70.

Lei, S., & Law, R. (2015). Content analysis of tripadvisor reviews on restaurants: A case study of macau. *Journal of tourism*, *16*(1).

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Li, J., Shang, J., & McAuley, J. (2022). Uctopic: Unsupervised contrastive learning for phrase representations and topic mining. *arXiv preprint arXiv:2202.13469*.

Liao, W., Zeng, B., Yin, X., & Wei, P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, *51*, 3522–3533.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Lin, W., Shi, Y., & Zhao, J. (2021). A robustly optimized bert pretraining approach with post-training. *China National Conference on Chinese Computational Linguistics*, 471–484.

Madabushi, H. T., Kochkina, E., & Castelle, M. (2020). Cost-sensitive bert for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*.

Mahmoudi, L., & Salem, M. (2022). Improving multi-class text classification using balancing techniques. *International Conference on Artificial Intelligence: Theories and Applications*, 264–275.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Melián-González, S., Bulchand-Gidumal, J., & González López-Valcárcel, B. (2013). Online customer reviews of hotels: As participation increases, better evaluation is obtained. *Cornell Hospitality Quarterly*, *54*(3), 274–283.

Nieto, J., Hernández-Maestro, R. M., & Muñoz-Gallego, P. A. (2014). Marketing decisions, customer reviews, and business performance: The use of the toprural website by spanish rural lodging establishments. *Tourism management*, *45*, 115–123.

Noori, B. (2021). Classification of customer reviews using machine learning algorithms. *Applied Artificial Intelligence*, *35*(8), 567–588.

pandas development team, T. (2020, February). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Parida, U., Nayak, M., & Nayak, A. K. (2021). News text categorization using random forest and naıve bayes. *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)*, 1–4.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. *NIPS-W*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818.

Sparks, B. A., So, K. K. F., & Bradley, G. L. (2016). Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management*, *53*, 74–85.

Suryaningrum, K. M. (2023). Comparison of the tf-idf method with the count vectorizer to classify hate speech. *Engineering, MAthematics and Computer Science Journal (EMACS)*, *5*(2), 79–83.

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117–126.

TYAGI, J., CHOUDHARY, V., & GOYAL, N. (2023). Enhancing personalized recommendations: A spacy-based hybrid approach for book recommendation systems.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Yan, A., He, Z., Li, J., Zhang, T., & McAuley, J. (2023). Personalized showcases: Generating multi-modal explanations for recommendations. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2251–2255.

Younes, Y., & Mathiak, B. (2022). Handling class imbalance when detecting dataset mentions with pre-trained language models. *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, 79–88.

APPENDIX A

Table 2: Random Forest Hyperparameters.

Hyperparameter	Explanation	Value
n_trees	Number of trees in the forest	200 to 2000 in steps of 10
Bootstrap	Bootstrap or not	True or False
max_samples	Max sample number to draw from train set	0.5, 0.6, 0.7, 0.8, 0.9, 1.0
max_features	Number of features to consider for best split	sqrt, auto, p, 0.257
oob_score	Using out-of-bag samples or not	True or False
min_node_size	Min samples to be at a leaf node	1, 5, 10, 15, 20

Table 3: BERT Hyperparameters.

Hyperparameter	Explanation	Value
Lr	The learning rate of the model	4e-5, 3e-5, 2e-5
Batch_size	Data loaders batch sizes	16, 32, 64
Epochs	Number of epochs for the model to run	3, 4, 5
Dropout	Dropout rate of the model	0.0 - 0.5
Accumulation steps	Gradient accumulation steps before parameter updating	1-4
Weight decay	Degree of L2 regularization	0.0 - 0.01
Early stopping patience	Early stopping to prevent overfitting	2

Table 4: roBERTa Hyperparameters.

Hyperparameter	Explanation	Value
Lr	The learning rate of the model	1e-5, 2e-5, 3e-5
Batch_size	Data loaders batch sizes	16,32
Epochs	Number of epochs for the model to run	3, 4, 5
Dropout	Dropout rate of the model	0.0-0.5
Accumulation steps	Gradient accumulation steps before parameter updating	1-4
Weight decay	Degree of L2 regularization	0.0 - 0.01
Early stopping patience	Early stopping to prevent overfitting	2

Table 5: BERT and roBERTa pre-trained checkpoints.

Transformer	Checkpoint
BERT	'bert-base-uncased'
roBERTa	'roberta-base'

APPENDIX B

$$R(X) = \frac{X_{\text{minority}}}{X_{\text{majority}}} \tag{1}$$

Table 6: All Models and Hyperparameter Tuning Process Times.

Models	**Time in Total**	**Time per Epoch**	**Hyperparameter Tuning Time**
Random Forest - Simple	~2 min	-	~17 min
BERT - Simple	~43 min	~10 min	~3 hrs
roBERTa - Simple	~51 min	~14 min	~4 hrs
BERT - Weight Balanced	~50 min	~12 min	~5 hrs
roBERTa - Weight Balanced	~1 hr	~15 min	~6 hrs
BERT - Oversampled	~1 hr	~12 min	~6 hrs
roBERTa - Oversampled	~1 hr	~15 min	~8 hrs
Random Forest - Additional	~2 min	-	~19 min
BERT - Additional	~1 hr	~14 min	~6 hrs
roBERTa - Additional	~1 hr	~16 min	~8 hrs