Contextual Sentiment Analysis of Airbnb Customer Reviews in Albania: An Investigation of BERT-based Classification

Blerina Zanaj
Agriculture University of Tirana
Rr. "Pajsi Vodica"
Koder-Kamez, Tirane, 1001

Serxhio Kurti
Agriculture University of Tirana
Rr. "Pajsi Vodica"
Koder-Kamez, Tirane, 1001

Elma Zanaj
Polytechnic University of Tirana
Bul. "Deshmoret e Kombit"
Sheshi. "Nene Tereza", Tirane

ABSTRACT
This study evaluates sentiment analysis techniques for assessing tourist experiences with Airbnb accommodations in Albania by comparing textual reviews with conventional numerical ratings. Since star ratings may not fully capture the contextual and subjective aspects of customer experiences, Natural Language Processing (NLP) techniques are applied to extract sentiment from textual feedback. The study compares a BERT-based Transformer model with traditional machine-learning methods, including Support Vector Machine (SVM) and Naive Bayes. The experimental results indicate that BERT provides higher sentiment-classification performance, demonstrating its ability to capture contextual and semantic information in tourist reviews. The findings highlight the potential of BERT-based sentiment analysis to complement numerical ratings and provide more detailed insights into tourist satisfaction and accommodation service quality.

Keywords
Sentiment analysis, BERT, NLP, transformer, data preprocessing, removing noise, truecasing, stopwords, lemmatization, text normalization.

1. INTRODUCTION
The rapid expansion of digital accommodation platforms such as Airbnb has transformed the tourism and hospitality sector by generating large volumes of textual customer feedback. In Albania, the increasing use of Airbnb provides a valuable source of information regarding tourists' experiences, perceptions, and satisfaction. Unlike numerical ratings, textual reviews can capture detailed and contextual aspects of the customer experience, including opinions about cleanliness, host communication, location, service quality, and overall satisfaction. However, the manual analysis of large collections of reviews is time-consuming and subjective, creating a need for automated approaches capable of extracting meaningful information from unstructured text.
Natural Language Processing (NLP), as a major field of Artificial Intelligence (AI), provides computational methods for processing and interpreting human language. Among its applications, sentiment analysis enables the automatic identification of opinions and emotional polarity expressed in textual data. This approach can transform unstructured customer feedback into actionable knowledge, allowing accommodation providers and tourism stakeholders to identify strengths, service deficiencies, and factors influencing tourist satisfaction.
Recent advances in NLP have introduced Transformer-based architectures that significantly improve contextual language understanding. In particular, Bidirectional Encoder Representations from Transformers (BERT) has demonstrated strong performance in text classification because it considers the contextual relationships between words in both directions. This capability makes BERT particularly suitable for sentiment analysis, where the meaning and polarity of an expression may depend strongly on its linguistic context.
This study investigates the application of BERT for sentiment analysis of a dataset derived from a survey of tourists' reviews and experiences with Airbnb accommodations in Albania. The research considers three sentiment categories—positive, neutral, and negative—and evaluates the ability of the model to automatically classify customer feedback. To improve the quality of the textual data, preprocessing techniques including truecasing, stopword removal, and lemmatization are applied. The main objective of this research is to assess the potential of AI and Machine Learning techniques, particularly BERT, to convert unstructured Airbnb reviews into meaningful information that can support service-quality improvement in the Albanian hospitality sector. The findings aim to demonstrate how automated sentiment analysis can complement traditional numerical ratings by providing a more detailed representation of tourists' perceptions and experiences. Furthermore, the study highlights the potential of NLP-based analytical methods to bridge the gap between large-scale textual feedback and practical decision-making in tourism and accommodation management.

2. DATA PREPROCESSING FOR SENTIMENT ANALYSIS
Data analysis is a key component when it is applied by using AI techniques especially involving NLP in sentiment classification from respondents' reviews. In fact, this field includes multidimensional complex data that might not be structured as they should be. Language includes some aspects that bring into light the importance of analysis and preprocessing. Like this, ambiguity can bring to multiple interpretations that depend on the reader comprehension and point of view. Moreover, depending on the tourism type of hosting, from the cultural and social background of the respondents, it can be extraordinarily variable and difficult to interpret. In tourism context in Albania this aspect becomes difficult from the cultural bias where the estimation with the rating stars can be influenced from the traditional Albanian hostility, but it can mask the real critics that are found in random text. The lack of rigorous structure in human language as it is in programing languages might complicate the task execution that includes Natural Language Processing. In this research work the data analysis is performed on a dataset created from the survey done by using Google Forms. The form was filled by the tourists that have stayed in Airbnb in Albania. This dataset contains English text data that were first classified

for the sentiment evaluation through a newly constructed variable the mean value of six quantitative evaluations (for example cleanliness, communication, check-in, correctness of photo published and the real accommodation conditions, location and the quantity of money paid). It is important to explore and to understand the data characteristics because every possible information might be useful for the task of sentiment analysis. It analyzed different preprocessing techniques, and it was checked on how they might effect in the results' metrics. Given the complexity of sentiment analysis, preprocessing techniques are essential for capturing relevant textual cues. Features such as capitalization, emojis, and stopwords may convey important information about the emotions and attitudes expressed in customers' Airbnb reviews and should therefore be carefully evaluated during preprocessing.

2.1 Categorized Dataset

The dataset was created from responses collected through a Google Forms survey designed to evaluate guests' experiences with Airbnb accommodations in Albania. The quantitative variables assess key aspects such as cleanliness, host communication, check-in procedure, photo accuracy, location, and value for money. The "Rate" variable represents the average score of six rating-based questions, providing an overall measure of guest satisfaction.

In addition to the numerical ratings, respondents were asked to briefly describe their Airbnb experience in Albania. These comments provide qualitative information that complements the quantitative ratings and may highlight aspects not directly covered by the structured questions. The textual feedback is classified into three sentiment categories positive, neutral, and negative to provide a more comprehensive understanding of guest experiences and to identify the main strengths and weaknesses of Airbnb accommodations in Albania.

2.2 Sentiment Field Elaboration: Rate Classification in Labels

To obtain an accurate classification of the sentiment based on Likert scale (1-5), the average of quantitative evaluation is mapped into three clear categories. This process of transformation of quantitative estimation into qualitative labels is achieved through as listed below:

- Negative Sentiment: it is attributed to reviews where the average of the estimation is in the interval between 1 and 2.9. This category is identified with the experience where the services offered has not fulfilled the visitors' expectations.
- Neutral Sentiment: it is attributed to the reviews where the average of the estimation is varied between 3 and 3.9. This category represents an average experience, where the visitors have expressed themselves to have had a moderate satisfaction level where there are obvious improvements that might be made.
- Positive Sentiment: it is attributed to reviews where the average of the estimation is the interval between 4 and 5. This category reflects a high level of satisfaction where the service has been more than what was expected from the visitor.

This clear standardization allows an objective division of the dataset with trustworthy labels that serves for structuring the survey done with Google Forms.

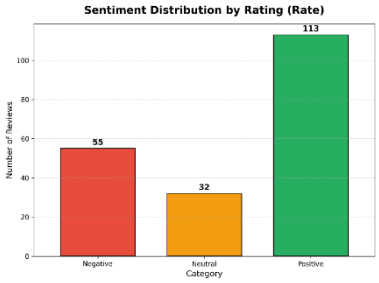

Fig1. Sentiment distribution by rate

Data shows a positive look but in presence of challenge reviews considerations:

- Positive Sentiment: With 113 reviews this group is dominating the sample where it shows that most of the visitors have had a satisfactory experience or that it has fulfilled their expectations.
- Negative Sentiment: with 55 reviews, this segment makes a considerable part of the samples that shows that independent of positive dominating part, there are still some lacks or dissatisfactions that are present and that requires the attention to amend them.
- Neutral Sentiment: with 32 reviews this group represents those visitors that have had an uncertain experience which means that it has not been completely disappointing and not entirely brilliant.

Positive sentiment was the most dominant category, indicating that Airbnb users in Albania generally reported a good level of satisfaction. However, positive reviews were approximately twice as frequent as negative reviews, suggesting that service quality was not uniform across all experiences. The presence of neutral and negative reviews indicates areas where service improvements may increase visitor satisfaction.

3. UNCATEGORIZED DATASET: RESOURCES OF OPEN-ENDED TEXT

The focus of this analysis stands on the open-ended responses available in the column "Briefly describe the experience of your stays in Airbnbs in Albania:" and on their corresponding sentiment labels that are set at the "Verified manually" column of the dataset. These labels were assigned manually to each interpretable review like: Positive, Neutral, or Negative and those are used as reference sentiment for the subsequent analysis.

Dataset characteristics are like follows

- Data sources: Dataset is constructed through the data obtained from a survey done with Google Forms by guaranteeing a straightforward and authentic gathering of consumer attitudes.
- Volume: Initial dataset has 200 incomes that represent a database that is enough to get the feelings of user experiences in Albanian market with Airbnb.

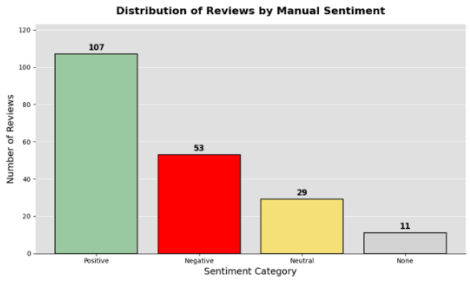

Fig2. Distribution of reviews by manual sentiment

The graphic shows the unbalanced spread of different sentiments obviously in negative and positive categories. The difference between the classes with bigger numbers and the lower one is approximately 54 answers as is shown in the above graphic, where it is possible to see exactly the number of sentences for each category (107 positives, 54 negatives and 29 neutrals)

Of the 200 collected responses, 189 contained interpretable text and were manually categorized as positive, neutral, or negative. The remaining 11 responses consisted mainly of unclear symbols, codes, numbers, or otherwise uninterpretable content (for example ***grgrgrgr, Really, ?????, Time 19:45, GRFERWXD*** etc.); therefore, they were excluded from the sentiment classification and retained an undefined (None) sentiment value. The BERT model classifies words that are not understood as neutral, and this could lead to an incorrect interpretation of the data by placing them in this category. Based on this classification error, the dataset was manually checked, and the words that had no meaning were assigned None values for the sentiment value.

This dataset is not just a list of comments, but it shows the reflection of real market perception. However, before these data are elaborated with the application of BERT model, those can be classified like "raw data". Due to the free nature of writing, those contain incorrect information, thermic noise and elements that ask for rigorous pre-processing (Text Preprocessing) to guarantee that the BERT model can analyze only the semantic content and not the spelling errors or some useless elements.

3.1 Text Preprocessing

Cleaning and text transforming from the raw data into a format that is more manageable is very important step in NLP pipeline. Application of data preprocessing like tokenization, removing the stopwords and noise reduction can increase significantly the analysis efficiency and particularly improve the metrics of chosen model performance [1].

3.1.1 Emoji Removal

In communication channels the usage of emojis is increasing, where it is believed that those might serve to add more emotion to the text. Their inclusion in such tasks as sentiment analysis might influence sentiment generalization of a sentence; so, considering them might bring a better accuracy to the results obtained [2].

To check if the emojis hold information that is crucial in the dataset in the initial phase of preprocessing, there is applied a conversion of them in text description (demojization). For doing it a specific Python library was used for emojis analyzing.

However, after the test it was observed that textual description of emojis only created "noise" so they were unnecessary for the BERT architecture. For instance, converting a single emoji in long textual sequence (like it can be seen in the demojization of "waste of $$$ 😡 😡 😤") it changes the semantic structure of the review, by making it more difficult for the model to isolate key words for describing the experience.

As BERT operates with higher optimality on the clean texts it was decided that on the last phase of preprocessing pipeline all the emojis will be removed completely from the dataset. This decision ensures that the model will be trained with a corpus as standardized as possible. By removing technical noise, it simplified the Tokenization process and increased the ability of the model to identify the main semantic specifications, by guaranteeing a more stable performance in sentiment predictions [2].

3.1.2 Text cleaning and Regular Expressions (Regex)

To transform the raw dataset into suitable format for BERT the cleaning process relies on the usage of regular expressions. This expression allows automatic manipulation of specific models of written texts from respondents.

Usage of Regex in the study includes:

· To preserve consistency, it turns all the letters into lowercase.

· Removes HTML tags by using BeautifulSoup to bring out only meaningful text.

· Eliminates the numbers and the quotation marks to reduce noise.

· It uses regex (\W+ and \s+) to remove the special characters and excess spaces.

The process of text cleaning has fundamentally transformed the raw comments of the respondents. The main changes include:

"Noise" reduction: Removing special words, emojis and URL links that do not add value to sentiment analysis.

Standardization: Converting the text into lowercase letters and by eliminating the quotation marks that were excessive to create an input more uniform for tokenization of the BERT model.

Simplification: Eliminating words or expressions that might cause confusion during the elaboration performed by the model, helping like this to conserve the essence of the message conveyed by the respondent.

These changes might ensure that the BERT model will be concentrated exclusively on emotion content and text semantics, by increasing like this the accuracy of final classification of sentiment [3].

3.1.3 Text Normalization: Handling Redundancy and Slang

To improve the accuracy of BERT model the dataset underwent to a normalization process that addresses the informal language and slang used by the respondents. This process transforms the colloquial expressions into standard language, by reducing semantic "noise" that might prevent the model to identify the sentiment.

This process was done to guarantee an input with quality for BERT architecture:

Jargon translation: For the informal terms there was applied a

substitute dictionary. The abbreviations like "w" (with), "u" (you), "r" (are), and "pls" (please) were expanded. Also, the expression like "smh" (shake my head), "btw" (by the way), "tbh" (to be honest), "imo" (in my opinion), "fyi" (for your information), "lol" (laughing out loud) were standardized into their complete form.

Normalization eliminates language uncertainties, by offering to the BERT model a rich context and a stable one and it allows a more accurate classification of the sentiment based on whole context of the sentence [4].

3.1.4 Normalization through "Truecasing"

Transforming all letters into lower cases is an important role in sentiment analysis. In the context of social network, their usage might hold different meanings; for example, those might be preprocessed and it is fundamental for the research scenario of this paper. The approach that tries to avoid the equal treatment of each word is called Truecasing. However, before its usage the dataset was analyzed to understand the trends to use capital letters is also present within negative category. Through this process the original sentences are transformed into a professional format and by improving the readability like this. For Bert model this is fundamental: while the original text might create confusion because of the uppercase letters that are randomly distributed, the Truecased version assures that the model will be focused on the essence of the word. So Truecasing does not cause the lose information on the contrary it cleans it, so the input passed to BERT is accurate and more reliable for the sentiment classification [5].

3.2 Tokenization and data preparation

After the normalization and Truecasing the text has reached a grammatically structured form. However, to make it more understandable the BERT architecture, the text might undergo to a Tokenization process. This step is a technical conversion where the text is "translated" in a numerical language that allows the model to perceive and to divide the emotional payload (sentiment) hidden in the sentence. Through this process BERT does not only read the sentence but identifies the semantic models that distinguish a Positive experience (satisfaction), Neutral (objective), or a Negative (displeasure), by dividing those into units that the model might classify them accurately [6].

3.2.1 Dividing the dataset (Train/Test Split)

Before the tokenization the dataset is divided into two groups to ensure an objective evaluation of model ability to distinguish the sentiment:

Train Dataset: 151 inputs that allow the model to "learn" the difference between the sentiments through examples.

Test Dataset: 38 inputs that are used to prove if the model has really grasped the sentiments in new data.

After the data were divided those have undergone to tokenization through WordPiece. This model allows BERT to get the small morphological details that sometimes are the key to differentiating a negative comment from a positive one.

For applying this process in PyTorch framework it is used the class BertTokenizer with fixed length (padding) with 128 tokens for each sentence. The technical results of this process are listed as below:

Train Dimensions: torch.Size([151, 128])

Test Dimensions: torch.Size([38, 128])

This final process transforms the dataset from a simple list of comments into a complex numerical array, where each token holds necessary information for final sentiment classification, by allowing to BERT model to analyze with accuracy "sensation" or the client experience in Airbnb.

3.2.2 Removing stopwords

An important general step in a pre-processing pipeline is the removal of stopwords. In fact, these kinds of words might bring noise into a sentence, so their removal can help in grasping the specification of key words respectively and extract the main essence from a text. However, this approach is not always suggested, especially for some tasks where each word has a meaning importance, like the problems of machine translation, text summarization or question-answering systems. On the contrary, by removing the stopwords might increase the accuracy for certain classification tasks, like spam filtering, or the genres classification, by bringing into the light spot the main components of the sentences [6].

3.2.3 Lematization

Besides the removing of the stopwords an essential step in preprocessing of the data is Lemmatization. Differently from Stemming that cut the prepositions of the words, Lemmatization utilizes morphological structure to reduce each word in its root form or its "lemma". The purpose is the unification of the different versions of a word (i.e. its plural to singular form of it), by allowing the model to understand that those represent the same semantic concept and by reducing the unnecessary vocabulary sparsity.
In the dataset this process resulted in transformation of 287 words in their root format. This helps in consolidating the used terms from the clients, by avoiding the same concept handling as there were different items [6].

4. DEVELOPMENT AND VALIDATION OF THE MODEL

After extraction and data analysis a fundamental step in the machine learning pipeline is training and model evaluation. In AI there exist architecture and the algorithms that can be used for different purposes. Therefore, it is vital to understand which type of technology can most effectively influence the final assessment.

In this phase the model gets trained and evaluated for a labelled dataset, where the specific metrics (like Precision, Recall and F1-Score) are applied to determine which technique performs better.

In this section is given an overview of the tested approaches starting from the baseline models - TF-IDF/SVM to BERT model. BERT model will be elaborated upon a detailed manner as it represents the architecture that offers the best performance. It will be analyzed different architecture configurations to understand which of them offers the optimum performance in language context used in the reviews of Airbnb. As an overview, the development phase and the evaluation one includes a rigorous selection of the algorithms and metrics by defining like this the best possible configurations that guarantee an optimal result and stable one with new data.

4.1 Experiments with sentiment analysis models

In the first phase of model evaluation, the original dataset is used without any kind of prior modifications. The total number of elements in the dataset is 189. An alternative worth exploring concerns the very nature of the task. Generally, a standard

approach for classification problem is through the usage of Softmax layer to get the probability for each category. Another possible option is the handling of this problem, like a regression task. Categorial variables of sentiment can be translated into a numerical variable and to use the training of a regression model. In this way the output that the model gives is a real number and the value get classified through the rounding off of it to the nearest integer value.

There can be followed different ways for text representation. The approach of Bag-Of-Words that is based on TF-IDF representation is a methodology that is widely used. However, the sentiment analysis task and the order of the words and the compound words might be decisive to define the sentiment or emotion of the sentence. On the other side Word Embeddings (word vectorizing) allows to get the relationships between words and helps to understand better the semantic connection between words. It is possible to perform specific vectorizations or to use a pre-trained model to have them. For this evaluation phase the sentiment values that were converted by using the One-Hot Encoding that is offered from Sklearn library. This kind of encoding represents each sentiment with a vector, the length of it is equal to the number of possible sentiments. In every vector only one element has the value 1 and the others are 0. The position of the value 1 represents the respective category.

Table 1: One-Hot encoding for the sentiment categories

Sentiment	Vector (One-Hot)
Negative	[1, 0, 0]
Neutral	[0, 1, 0]
Positive	[0, 0, 1]

This class representation aims the usage of the Softmax function as a decisive activation layer in neural networks. In the scenarios of multi-class classification this function is fundamental as it converts the exits of the model with a normalized probability spreading. Through this mechanism the model does not generate only a binary decision but a probable evaluation that defines which sentiment category does the text analysed belong to. Model evaluation includes the dataset divided into specific data groups. To test these conditions, the dataset was divided into two distinguished subgroups: the training set and the test set. The division were performed in a specific way by allocating 80% of the data in the training group and 20% for the testing group. Furthermore, the data division were done in a stratified fashion for the purpose to guarantee that the testing group contains all the possible categories for the evaluation, by conserving the same spread of classes percentages like the original dataset.

Table 2. Number of elements for Training and Testing

Dataset	Number of elements in the training	Number of the elements in the testing	Total
Original dataset	151	38	189

In table 2 were shown the number of elements for the training set and those of the test set for each evaluated dataset.

4.2 Embeddings Evaluation

Evaluation process starts with the examination of different types of vectorization (embeddings), where can be included TF-IDF, Word Embeddings that were created through the usage of Keras library and the contextual vectorization with BERT. Performance metrics (Accuracy, Precision, Recall, F1-Score) were compared through these variations to identify the most efficient model for sentiment analysis with our data [7].

TF-IDF: Vectorizing methodology TF-IDF were applied by using the scikit-learn library. TfidfVectorizer object allows the transformation of cleaned texts (Truecased_Review) in a vectorial representation based on the frequency of the terms. Training procedure were exclusively applied with the data by creating a vectorial space where the most important discriminating words were saved into 5000(max_features). Vectorization obtained were used later on as input to the logistic regression model for sentiment analysis [6].

Word Embeddings with Keras: With the vectorization generated with Keras it is used a tokenizer to construct the vocabulary of the words from the column of Truecased_Review. This process mapped each word into an integer value, by allowing like this the transformation of the comments into numerical sequences with standard length through padding (100 tokens). Vectorization was generated through the Embeddings layer in neural networks which converted the input sequence in dense vectors with dimension equal to 16. This layer followed by GlobalAveragePooling1D allowed the dimension reduction and the extraction of the main features for the classification into three categories (Positive, Neutral and Negative) [7].

BERT Embeddings: For contextual analysis it used the BERT architecture with transformers library. This model was chosen for its ability to grasp the structural and semantic nuances of the English text by overpassing the statistical methods. The process includes the usage of a ready pipeline for sentiment analysis that is trained over multilingual data, by allowing the comments classification of Airbnb based on their emotional evaluation. Model forecasting (initially expressed as evaluation 1-5 stars) were converted in target classes of the study (0, 1, 2) through a mapping function that were personalized, by assuring that BERT results to be comparable with the labels that were verified manually [8].

Table 3. Metrics value with different embeddings

Vectorization(Embeddings)	Accuracy	Precision	Recall	F1-Score
TF-IDF	0.642857	0.413265	0.642857	0.503106
Keras Embeddings	0.642857	0.413265	0.642857	0.503106
BERT	0.892857	0.935714	0.892857	0.903140

The results obtained are shown in table 3, the high values show the significant differences between the traditional and advanced models:

- TF-IDF and Keras Embeddings: Both these methods show an identical performance by achieving the Accuracy of 64.29% and F1-Score of 50.31%. This modest performance shows that the statistics models and those simple embedding can not elaborate fully and are incapable of bringing up all the language nuances and the emotional context of the users' comments.
- BERT Embeddings: BERT model overpassed

significantly the other techniques by achieving an Accuracy of 89.29% and F1-Score of 90.31%. This increase in accuracy with 25% compared the other models which confirms the contextual representation of BERT. The model could identify with high precision accuracy (Precision: 93.57%) the expressed sentiment in the analysed texts.

The empirical results confirm that BERT architecture is the optimal solution for research performed. The significant improvement in the metrics, especially the passing from accuracy value of 64.29% to 89.29% shows the ability of the model to get the natural language context which is decisive to offer reliable results in analysis or clients experiences in the accommodation platforms.

4.3 Regression Evaluation

Evaluation process included also the exploration of alternative approaches depending on the task topology. Through some specific operations it is possible that a classification problem can be treated as a regression problem.

To analyze it the categorical labels of Verified_manually were mapped to the integer values like this: 0 for negative class, 1 for the neutral one and 2 for the positive class. Each model used approximately the same number of trainable parameters. Moreover, the hyperparameters value like learning rate and epochs were kept consistent during all testing to assure comparability.

To obtain the results there were used standard loss functions and metrics that are applied in regression tasks, like the MSE (Mean Squared Error) function that is widely used in the scenarios where the distance measurement between the numerical forecasting and the real value is required.

At least the different vectorization representatives were tested (TF-IDF, Keras Embeddings and BERT) to determine their performance through this alternative approach, by confirming that BERT model remains the most stable solution even when it is treated like a regression problem [6].

Table 4. Metrics values for different vectorization techniques (Regression Approach)

Embeddings	Accuracy	Precision	Recall	F1-Score
TF-IDF	0.714286	0.802198	0.714286	0.645022
Keras	0.642857	0.413265	0.642857	0.503106
BERT	0.892857	0.935714	0.892857	0.903140

Table 4 presents the performance of the models under the regression-task approach. The results highlight clear differences in the predictive capabilities of the models:

- TF-IDF: This method showed a solid performance with Accuracy of 71.43% and F1-Score of 64.50%. Independent of the basic technique it showed stability by maintaining a high Precision value of 80.22%.
- Keras: this model showed the lowest performance in the testing with an Accuracy of 64.29% and F1-Score of 50.31%. The results suggest that the configuration used in this model asks for further optimizations to get better the variance of the data in the regression task.
- BERT: BERT model confirmed its lead in the approach by achieving the Accuracy of 89.29% and an impressive F1-Score of 90.31%. The high Precision of 93.57% confirms that BERT is the most reliable model to map the comments of the user in numerical values for the sentiment analysis by minimizing the error of classification.

This empirical comparison confirms that the regression approach when it is associated with contextual vectorization as in the BERT case assures stability and high accuracy, by transforming it in the optimal choice for the sentiment analysis for the dataset that is considered.

4.4 Binary Evaluation

An additional experiment that was done based on the possible number of classifications. In fact, in this specific case of the problem from a classification with multiple classes it turned to binary classification task. For this scenario the "Negative" categories were kept apart from the "Neutral" category and "Positive", where the last two were merged into a single category called "Positive" by bringing the dataset with only two categories. Each model used approximately the same number of trainable parameters. Moreover, the hyperparameters value like the learning rate and epochs were kept consistent with the previous experiments.

Also to obtain the results were used the different loss functions and metrics, specifically Binary Cross-Entropy and the activation function of Sigmoid, these are standard for binary problems. At last were tested different vectorization representatives (embeddings) to define their performance under the alternative approach as shown in the table 5.

Table 5. Binary Metrics for Binary Evaluation

Embeddings	Accuracy	Precision	Recall	F1-Score
TF-IDF	0.710526	0.710526	1.000000	0.830769
Keras Embeddings	0.710526	0.710526	1.000000	0.830769
BERT	0.921053	0.961538	0.925926	0.943396

From the data shown in table 5 it can be shown the results as listed below:

Performance of TF-IDF and Keras Embeddings: Both these two methods have shown identical results (Accuracy 71.05%, Recall 1.00). The maximal value of Recall (1.00) shows that the models have been able to identify all the positive cases (Neutral/Positive) but this comes with a high cost regarding Precision. This suggests that TF-IDF models and Keras in this case tend to "classify everything as Positive" not to lose any case but in this way it increases the number of "False Positives" (those are negative comments but by error get classified as positive).

BERT model superiority: BERT model shows an obvious superior performance with all the main metrics values as listed below:

- Accuracy (92.10%): BERT arranges the right classification over 92% of the comments, by overpassing like this the other models.
- Precision (96.15%): This result shows that when BERT forecasts a comment as "Positive" it has a higher degree of accuracy by reducing considerably the errors.
- F1-Score (94.34%): As F1-Score is the mean value of the harmonics between Precision and Recall, the high value of it confirms that BERT is the most balanced model and most trustworthy for this task of binary classification.

Binarization analysis shows that the Transformers, more specifically BERT model is more capable to understand the context and language nuance from the comments of Airbnb in Albanian language compared to the other statistical methods (TF-IDF) or the simple neural network (Keras Embeddings). This ability of BERT makes this model most suitable for practical applications in monitoring the client's sentiment.

5. Experiments Conclusions

As was observed in different scenarios that were tested BERT model always tries to give better results that is why it is essential to do further studies about it. The models developed in previous phases were relatively simple and they used different vectorization techniques (embeddings) and diverse approaches about the problem. BERT offers the fine-tuning possibility of a pre-trained model for a specific task, which makes assessing its full applicability in this field to be vital. The values obtained from the evaluation of dataset in the binarization experiment shows quite promising results by showing that the merging of the classes "Neutral" and "Positive" allows a much more accurate accuracy in identifying critical cases. However, the main project objective remains the accuracy of characterization and detailed one from the user's experience. By limiting the analysis in only two sentiments in the dataset might reduce the influence and the value of the final result, so it is necessary to keep the approach with multiple classes to get the nuances of clients' reactions in Albanian Airbnb.

On the other hand, the usage of simple statistical models like TF-IDF, even though they offer a base for comparison, did not offer excellent results as those achieved through the advanced approaches of BERT classification. Low results of Precision in traditional models compared to the high accuracy of BERT, confirmed that the architectures that are based on Transformer are essential in NLP and the research contest.

In machine learning pipelines baseline models are simple models that serve as a reference point for the performance evaluation of the more complex techniques but more powerful. The main idea is starting with a simple algorithm that is recognized and simpler to be applied with the purpose that it can be set to a minimal level of performance that can be achieved undoubtedly for a specific task. Basic models serve as benchmarks for the future more advanced models, as the evaluation completely depends on the specific use case that is analysed. Moreover, another key gain is that those allow us to better understand the data, by allowing the verification if this kind of data are suitable for the aimed task. In the section are discussed the basic models that were analysed for the task of sentiment classification. More specifically the attention will be concentrated on Naive Bayes, Support Vector Machine (SVM) and Decision Trees. These are the three algorithms that are tabulated as basic models that are fundamental for the sentiment classification of the text. For these basic models it is used the vectorization TF-IDF as input due to the simplicity and the performance compared with BERT, meanwhile that the vectorization generation with BERT appeared to be more challenging compared to the other methods. The models were trained for a classification task with multiple classes by imposing that in this phase it is not possible to perform binary evaluation or regression. The dataset used for the training of the basic models is the initial dataset. In table 6 is shown the sample spreading in this dataset.

Table 6. Initial Dataset

Sentiment category	Reviews Number
Negative	53
Neutral	29
Positive	107
Total	**189 (labeled)**

The dataset was divided into training set and test set where those makes 80% and 20% of the total dataset, for the basic model as well for the StarSpace approach.

Table 7. Metrics values for Naive Bayes

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.7250	0.7250	1.0000	0.8406

Table 8. Metrics performance of SVM model

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.8500	0.8286	1.0000	0.9062

5.1 Final Baseline Considerations

In Figure 3 is shown the comparison between the metrics of three models based that were analysed Naive Bayes, Support Vector Machine (SVM) and Decision Tree. About the computational efficiency, the Naive Bayes models and SVM have demonstrated a longer time for the construction of the hierarchical structure of the nodes.

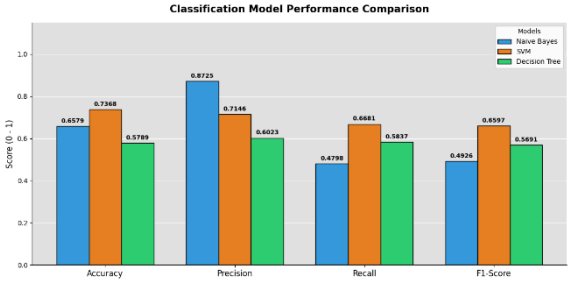

Fig.3 Comparison of the models performance metrics

As it can be observed the models perform differently as listed below:

- SVM remain as the best performer model with Accuracy value of 0.8500 and F1-Score value of 0.9062. This model could identify with an absolute accuracy (Recall = 1.0000) all the positive cases by

showing like this a high stability.

- Naive Bayes and Decision Tree have shown that the Accuracy were moderate with a value of 0.7250. Meanwhile Naive Bayes achieved the Recall whole value of (1.0000), its Precision is equal to its Accuracy, this thing demonstrated a trend of the model to classify a large number of cases as positives.
- Decision Tree offers a better balance between Precision (0.8000) and Recall (0.8276), by demonstrating that quite interpretative model independent of the low value of Accuracy.

In this specific case of usage, the purpose was to get a better identification of client's sentiment. High performance of SVM with F1-Score value greater than 0.90, so this model is more trustworthy to minimize the errors in classification.

By passing on the estimation it becomes clear that the basic models offer a minimal threshold for the performance. However, these models operate mainly over statistical bases of the word appearance. As consequence the Transformer integration like BERT will enable to grasp the contextual and semantic nuances of the users' comments, by overpassing the limitations of these classical approaches. In conclusion the basic models serve as a reference solid point by confirming even that the statistical approaches are effective the computation power of the advanced models is necessary for an intelligent analysis of the sentiment for the Albanian Airbnb.

6. REFERENCES

[1] Text Representation: A Simple Explanation Of Complex Techniques.[Online]. https://spotintelligence.com/2024/10/01/text-representation-a-simple-explanation-of-complex-techniques/

[2] Julia Rayz Byungkyu Yoo, Understanding Emojis for Sentiment Analysis. 2018. [Online]. https://journals.flvc.org/FLAIRS/article/view/128562

[3] Cuantum Technologies LLC, Natural Language Processing with Python. [Online]. https://www.perlego.com/book/5381587/natural-language-processing-with-python-master-text-processing-language-modeling-and-nlp-applications-with-pythons-powerful-tools-pdf

[4] From Slang to Standard: Text Normalization Techniques with Python. [Online]. https://medium.com/@codeasarjun/from-slang-to-standard-text-normalization-techniques-with-python-9a5ca834168b

[5] Abraham P Ittycheriah, Salim Roukos, Nanda Kambhatla Lucian Vlad Lita, tRuEcasIng., 2003. [Online]. https://www.researchgate.net/publication/2903066_tRuEcasIng

[6] Natural Language Processing with Python Updated Edition From Basics to Advanced Projects. [Online]. https://www.cuantum.tech/books/natural-language-processing-with-python-updated-edition?srsltid=AfmBOorvqQa_ghoiAuURT0mi4VO8oXNLVJccEo7P6PbnQyWcdro_OINo

[7] Felipe Almeida Geraldo Xex´eo, Word Embeddings: A Survey. [Online]. https://arxiv.org/pdf/1901.09069

[8] Inc, 2017 Ashish Vaswani et al. Attention is All you Need. Ed. by I. Guyon et al. Vol. 30. Curran Associates. [Online]. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf