Customer Sentiment Analysis in Hotel Reviews Through Natural Language Processing Techniques

Soumaya Ounacer[1], Driss Mhamdi[2], Soufiane Ardchir[3], Abderrahmane Daif [4], Mohamed Azzouazi[5]
Faculty of Sciences-Department of Information and Modelisation Technologies,
Hassan II University, Ben M'sik, Casablanca, Morocco[1, 2, 4, 5]
National School of Business and Management, Casablanca, Morocco[3]

***Abstract*—Customer reviews of products and services play a key role in the customers' decision to buy a product or use a service. Customers' preferences and choices are influenced by the opinions of others online; on blogs or social networks. New customers are faced with many views on the web, but they can't make the right decision. Hence, the need for sentiment analysis is to clarify whether opinions are positive, negative or neutral. This paper suggests using the Aspect-Based Sentiment Analysis approach on reviews extracted from tourism websites such as TripAdvisor and Booking. This approach is based on two main steps namely aspect extraction and sentiment classification related to each aspect. For aspect extraction, an approach based on topic modeling is proposed using the semi-supervised CorEx (Correlation Explanation) method for labeling word sequences into entities. As for sentiment classification, various supervised machine learning techniques are used to associate a sentiment (positive, negative or neutral) to a given aspect expression. Experiments on opinion corpora have shown very encouraging performances.**

***Keywords*—Topic modeling; aspect-based sentiments analysis; aspect extraction; sentiment classification; machine learning**

I. Introduction

In the last few years, the use of the internet and online interactions has grown tremendously. A significant quantity of data are generated daily via social media, forums, chats, and other sources that is primarily displayed as natural language text[1]. The way internet users behave online has also changed how the internet works. For instance, rather than being merely content consumers, internet users are becoming content creators[2]. One significant piece of information that is produced daily within the wide range of content produced by internet users is opinions[2].

Internet users have the ability to criticize or popularize a service or a product with a simple comment or review on the internet and in different fields[3]. Numerous enterprises and businesses have taken advantage of this pertinent data to offer the greatest services or goods for their clients. Among these areas, tourism which is a continuously developing industry and an important key industry for many regions and countries[4]. The opinions and reviews of tourists who visit touristic places every year are shared on various sites such as TripAdvisor, Booking and Yelp...etc[5][6]. Internet users do not have the ability to read, understand and summarize the large number of reviews available for a specific hotel. It is challenging for a simple user to make use of the information at hand to choose a comfortable hotel for his/her trip. The principle on which this work is based is to carry out an analysis of customers' opinions on hotels located in Marrakech in order to allow them to improve their services and focus more on the main obstacles that have an impact on the attractiveness of these hotels. In this article, a study and application of Aspect-Based Sentiment Analysis are carried out in the hotel and tourism industry. Specifically, opinions will be analyzed so as to determine the sentiment that is expressed towards certain characteristics of the hotel and the service delivered by its employees. The main goal is to produce results whose conclusions can provide directions that lead to improve the performance in sentiment analysis[7]. To achieve this goal, several objectives will be accomplished. The first objective is to use the various preprocessing steps available for text preprocessing. The second objective is using existing libraries like, TextBlob or Vader to label the dataset. The third aim is to use and compare multiple classification methods to classify the views so as to correct aspects and sentiments, i.e., classification of online comments into polarity (negative, positive, and neutral), and finally apply an Aspect-Based Sentiment Analysis on the product (hotel) features identified. In order to achieve these objectives through a clear and logical progression, this work will be presented according to the following structure. Section II will be devoted to the different related works linked to the Aspect-Based Sentiment Analysis. As for Section III, it will expose a background of Topic Modeling and Machine Learning model. The construction of the DataFrame, the methodology and the experimental results will be presented in Section IV. A summary of the experiment's results will be shown in Section V. Last but not least, Section VI will be devoted to the conclusion and future work.

II. Related Works

For aspect-based opinion classification, aspect extraction is a crucial task. The vast bulk of extractions techniques have recently been put forth for the tourism industry. These methods have employed a variety of mechanisms and techniques to extract crucial information from tourism reviews. These methods can be split into three primary groups: methods based on rules, seeds, and topic models. There are several works in the field of hotels and tourism that concern Aspect-Based Sentiment Analysis which will be described as the following:

″Pekar et al.″ [8] utilized TermExtractor to divide hotel reviews into terms. The terms were then trained in a lexicon. Finally, they manually extracted from the term lexicon the six most obvious characteristics (single nouns and multi-word

nouns). This proposed method is based on rules that allow to extract aspects from hotel reviews using aspect appearance on every review.

Similar preprocessing steps were used by ″Muangon et al. ″[9] and were supported by LexToPus. These steps are used to categorize all hotel reviews into features. These characteristics include polar words as well as aspects. They extracted all of the top-rated aspects using a prioritized method.

″Marrese taylor et al.″ [10] have suggested an algorithm with the goal of extracting aspects. Aspects from restaurant reviews can be extracted thanks to this algorithm. The authors converted the reviews into sentences and then used Part-Of-Speech tagger to extract nouns from the sentences.

A different method for aspect extraction was proposed by ″Hai et al.″ [11]. According to two criteria—domain specific and domain independent—the authors extracted aspects. They created a list of candidate aspects by first using syntactic dependency rules. Then, they determined the intrinsic domain relevance score (IDR) and extrinsic domain relevance score (EDR) for each specific domain and independent domain of each extracted candidate feature, respectively. And at the end, these candidate features are extracted from the list of candidates that have low IDR score and high EDR score.

An algorithm based on a bootstrapping approach, which has been proposed by ″Wang et al″[12], extracts the main aspects of the review. In this algorithm, each sentence was initially given an aspect based on the maximum of overlap between its words and the aspect. Then, to examine the relationship between the allocated aspect and the sentence words, they determine the basic dependencies between them. Finally, sentence words that have a strong relationship with the assigned aspect are added to the list of aspect keywords and are considered to be aspects.

BESAHOT, which is a system that has been presented by ″Walter Kasper et al.″ [13], performs analyzed comment processing for text segmentation, statistical polarity detection of text segments, and extraction of linguistic information from review topics and their aspects. It is a quality control support system for hoteliers that provide them with complete overviews and summaries of their hotel and how it is rated and commented by users on the web.

III. Aspect-based Sentiment Analysis

An essential task in the field of Sentiment Analysis is Aspect-Based Sentiment Analysis (ABSA) [14]. It involves assigning a polarity (positive, negative or neutral) to each aspect evoked in an opinion sentence. Aspect extraction and aspect-level sentiment analysis are often the two main tasks used to accomplish this.

Although traditional Sentiment Analysis is done using document and sentence level Sentiment Analysis techniques, the current trend is to move to a deeper level which is presented as ABSA (features). This latter [15] performs a deeper and a better analysis, as it directly examines the opinion itself. This domain is a deeper end in Natural Language Processing (NLP) [16] where it presents a richer problem for researchers.

One of the main features of NLP is Topic Modeling. Topic Modeling can be applied to any form of text: emails, tickets, feedbacks, etc. in order to have a global vision of customers' concerns.

A. Topic Modeling

Natural Language Processing (NLP) [17] includes Topic Modeling, which is used to train Machine Learning models. It entails identifying from a document or corpus of data the words of themes that are associated with a particular topic.

Topic Modeling is an unsupervised Machine Learning approach to discover topics in various text documents. It can find patterns of words and phrases and automatically cluster groups of words and associated phrases that best represent the whole [18]. It also provides a useful view of a large corpus in terms of the relationships between them and individual documents.

The figure above (Fig. 1) represents the ordinary workflow of the thematic modeling process. A set of text documents is introduced into the black box of the thematic modeling algorithm and the following results are obtained:

List of Topics: Topics are the key themes representing the entire collection of documents. Each topic consists of several words that occur at the same time. A word can belong to more than one topic because a word can have a different meaning in a different context.

Topic Definition: A topic is represented by the weighted frequency of words. Each topic can be interpreted as a theme.

Topic Distribution of Document: Each document is represented as a topic distribution where the weight of a topic defines the part of the document covered by that topic. In a way, it provides a "soft grouping" of the document.

Topic modeling is a method for selecting a set of topics from a group of documents that best summarizes the information in the group. To create topic models, numerous techniques are employed. One of the areas of interest is in: LDA, LSA, NMF, and Corex which will be discussed later on in this section.

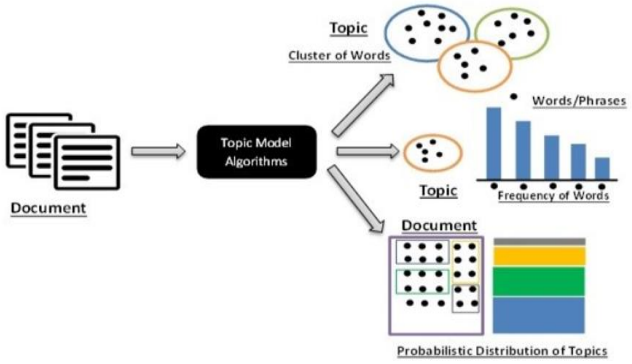

Fig. 1. Topic models process

1) LDA: Latent Dirichlet Allocation, [19] is a powerful learning algorithm for automatically and jointly classifying words in documents in mixtures of contexts. It has been successfully applied to model changes in scientific domains

over time. LDA is a probabilistic generative model. Based on the assumption that the order of documents in the collection and the order of words in a text are indifferent, LDA defines finite mixture models on sets of underlying topics to generate the collection. Each topic is being modeled as an infinite mixture on probabilities of the underlying topics. In an iterative procedure, these probabilities are computed several times, until the algorithm converges.

Advantages: Among the advantages of using the LDA method, the following are worth mentioning:

- LDA is easy to implement, understand and use.
- It maximizes inter-class scattering.
- It reduces intra-class scattering.

Disadvantages: Despite these advantages, a set of negative points still exist such as:

- LDA is costly in computation time.
- It is also costly in memory space.
- It renders poor results when the number of training images is large.
- It is hard to know when LDA is working. Metrics like perplexity are acceptable to check if learning is working, but there is a very poor indicators of overall model quality. For example, you could have a model with very low perplexity, but whose topics are not very informative.
- The topics are predicated on the multinomial distribution, while the words are predicated on a different multinomial distribution formed specifically for this topic. The structure may not be properly adjusted if the real structure is more complex than a multinomial distribution or if the data needed to construct the structure are insufficient.
- The user specifies the total number of subjects in the dataset (or bases it on a certain distribution using sampling), which is subjective and may not always reflect the true distribution of subjects.

2) LSA: Latent Semantic Analysis [20], or LSI (Latent Semantic Index), employs a bag-of-words (BoW) model, creating a term-document matrix (occurrence of terms in a document) [21]. Terms are represented in rows, and documents are represented in columns. By applying singular value decomposition to the term-document matrix, LSA can identify latent subjects. It is typically applied as a technique for noise or dimension reduction. In this method, document analysis is done by machines using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a metric that quantifies the significance of a word for a corpus of documents.

Advantages:

- Easy to understand and implement.
- In comparison to the vector space model, it shows improved outcomes.
- Only involves the decomposition of document term matrix which makes it faster than other available algorithms.

Disadvantages:

- In general, LSI is very slow on large corpora and not very accurate compared to LDA.
- The dimension of the latent subject depends on the rank of the matrix.
- The decomposed LSA matrix is extremely dense, making it challenging to index the individual dimension.
- Polysemy cannot be captured by LSA (multiple meanings of a word).
- It provides less accuracy than LDA.

3) NMF: Since the non-negative matrix factorization [22] is an unsupervised method, the subjects on which the model will be trained are not labeled. NMF factors or decomposes high-dimensional vectors into a representation that has a lower dimension. Since the coefficients of these lower-dimensional vectors are nonnegative, they are likewise nonnegative vectors. Consider the general scenario where there is an input matrix V with the form m x n. This approach divides V into two matrices, W and H, whose dimensions are m x k and n x k, respectively. In this case, V stands for the term document matrix, H stands for an embedded word in each row, and W stands for the weight of each word found in each sentence.

Advantages:

- NMF can handle missing values naturally and this property leads to a new method to determine the rank hyper parameter.

Disadvantages:

- NMF cannot be applied to several real-world issues where the domain limited knowledge of experts is available
- It sometimes provides semantically incorrect results

4) CorEx: Contrary to LDA and NMF, the semi-supervised topic model Correlation Explanation [23] allows to give the model "anchor words" which exemplified potential topics that the model might be looking for. CorEx also allows to provide the model with a confidence score for the anchors. If this choice is less certain, the model may forgo the target recommendations if they don't sufficiently match the data. This new capability is strong in guiding a thematic model with the chosen domain expertise. CorEx provides a flexible framework for learning topics that are maximally informative about a text corpus. The CorEx topic model makes few assumptions about the LDA structure and flexibly incorporates domain knowledge through user-specified

"anchor words". With anchor words, one can guide the topic model to topics of substantial interest, interact with the topics, and refine them in ways not possible with traditional topic models.

Advantages: CorEx competes with LDA in terms of producing semantically consistent topics that aid in document classification. By citing above some advantages of this model [23]:

- The CorEx's modeling algorithm is rapid.
- It searches for topics that are "maximally informative" about a set of documents rather than assuming a specific model of data generation.
- Word-level domain knowledge can be flexibly integrated into the CorEx thematic model.
- It consistently creates document clusters with higher homogeneity than LDA in terms of clustering.
- The CorEx anchor steers the thematic model toward topics that don't naturally arise and frequently results in topics that are more coherent and predictable.

Disadvantages: Despite the number of advantages of the CorEx thematic model over LDA, there are some drawbacks[23]:

- The sparse implementation necessitates that every word appears in just one topic. It is not a matter of fundamental theoretical limitations, but rather of computer efficiency.
- CorEx relies on binary accounting data for its parcel-level optimization rather than the usual accounting data that are input into LDA and other theoretical models.

Despite binary number limitations, CorEx nonetheless discovers a reliable and competitive structure in the data.

B. Machine Learning Model

The determination of the direction of the opinions in a text divided into two or more classes on certain features is known as classification of opinions by aspect. The classification of opinions has been done into several categories, such as binary, ternary, etc.

Typically, the classification task is defined as the task of predicting the label that is to say, assigning each given object to a group based on a classification rule. The primary goal at work entails classifying aspect opinions, thus training a classifier to predict the label for each input text is needed. There are three kinds of polarities (positive, negative, or neutral). In this section, the most common employed algorithm shall be outlined. [24].

1) Logistic regression (LR): This is an analytical technique key in the social and scientific sciences [25]. Logistic regression, which also closely resembles neural networks, is the standard supervised Machine Learning approach for classification in natural language processing. A logistic function is used in logistic regression to create discrete dependent variables from a series of data points.

2) Support vector machines (SVM) [26]: This can be applied to both regression and classification tasks. SVM methods aim to partition linearly separable data into two classes with the maximum distance between them. In the high-dimensional space, SVM identifies an ideal hyperplane that separates the input data with the greatest possible margin between it and the point(s) that are closest to it. The points for which the margin is reached are called support vectors. A kernel function can be used to map the data into a higher dimensional space in order to make them linearly separable if the input data are not linearly separable. The polynomial kernel, Gaussian radial basis function, and sigmoid kernel are the three most widely used nonlinear kernels.

3) K-nearest neighbour (K-NN): This is one of the simplest Machine Learning algorithms used for classification and regression problems[27]. The information is subsequently allocated to the class with the closest neighbor based on the nearest measures.

4) Naïve bayes (NB) [28]: This Machine Learning algorithm can be used to divide objects into two or more classes, such as text documents. It is founded on the Bayes theorem, which uses conditional probabilities as its foundation.

5) Decision tree (DT): This is part of the supervised algorithms in the field of Machine Learning [29]. Their principle is to divide learning data into groups whose content becomes increasingly homogeneous until pure data are obtained (belonging to the same class) or a maximum number of partitions is reached. The resulting model is a tree composed of several decision rules and is easily interpretable. As with any supervised learning method, decision trees make use of examples. Building a decision tree by category is necessary if one has to categorize the documents. In order to determine to what extent a category a new document belongs to, the Decision Tree will be used for each category in which the classified document is submitted. Each tree responds with yes or no.

6) Random forest (RF) [30]: This is a prediction method that Ho developed in 1995. In 2001, scientists Leo Breiman and Adele Cutler formally proposed the algorithm. It is made up of various decision trees that each focus on a different aspect of the problem independently. Multiple decision trees are produced by this classifier using a subset of the training data that is randomly chosen. The final class of test objects is then decided by aggregating the votes from various decision trees.

7) ExtraTrees (ET): Extremely Randomized Trees [31] is an ensemble-supervised Machine Learning technique that makes use of decision trees. It builds multiple trees and divides the nodes using random subsets of features, but the sampling for each tree is without replacement. As such, the most important and unique feature of the algorithm is the random selection of a splitting value for a feature, which makes the trees diverse and uncorrelated.

8) AdaBoost (AB) [32]: This is a widely used boosting algorithm. It builds a majority vote iteratively. Over the iterations, it maintains a weight distribution on the training examples so that poorly classified examples see their weight increase and well classified examples see their weight decrease. At each iteration, the weak learning algorithm is trained with the set of weighted examples and the resulting classifier is added to the majority vote.

9) GradientBoost (GB): A meta estimator that fits a series of weak learners is gradient boosting [33]. It is a powerful Machine Learning algorithm used to solve regression and classification problems. It creates a prediction model in the form of a collection of weak prediction models, typically decision trees. It builds the model incrementally, much like other boosting techniques do, and generalizes them by enabling the optimization of an arbitrarily differentiable loss function.

IV. Building the DataFrame

The workflow and the subtasks for each phase are shown in Fig. 2. The learning phase and the testing phase are the two steps that make up this workflow.

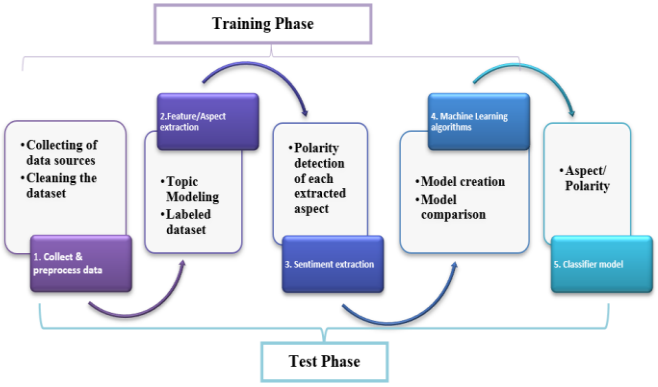

Fig. 2. Aspect-based sentiment analysis workflow

A. Training Phase

Fig. 3 gives an overview of this process namely the collection, the data sources and the pre-processing of the datasets. The main tasks of each step are described in the paragraphs that follow.

1) Data sources: User opinions are the main criterion for enhancing the quality of the services provided and improving the value of the products delivered. These opinions can be found in different data sources namely review sites, blogs and micro-blogs.

a) Review sites: Opinions have the role of decision makers for any user during the purchase phase. User generated reviews of products and services are widely available on the internet. Sentiment ratings or texts use reviewer data collected from websites such as TripAdvisor and Booking (hotel reviews). These sites host millions of visitors' hotel reviews.

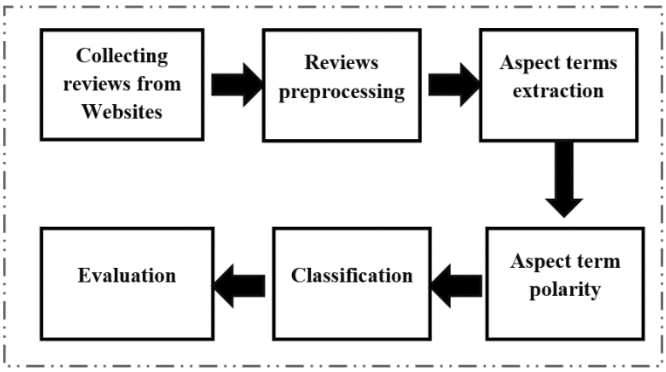

Fig. 3. Sentiment analysis process at aspect level

b) Blogs and micro-blogs: Blogs and micro-blogs are among the most popular communication tools of internet users. Millions of messages are posted every day on well-known microblogging platforms including Twitter, Tumblr, and Facebook. Sometimes Twitter messages express opinions that are used as a source of data to classify sentiments.

2) Data collection: The data acquisition or collection phase consists of obtaining the corpus to be analyzed. The "web scraping" method [34] is used to collect the reviews, since the goal is to collect reviews from various hotels in Marrakech. As shown in Fig. 4, each entry in this dataset is structured as follows:

- Hotel_name: designates the name of the establishment (Hotel)
- Title_review: refers to the title written by the client to give a general summary
- Reviews_hotel: contains reviews (text), written in English
- Rating_date: indicates the date of publication of a journal
- Score_rating: the evaluation given by each customer between 10 and 50

The reviews for 10 different hotels in the city of Marrakech are obtained from two websites (booking and TripAdvisor). The dataset consists of 21619 reviews in English, but only 14356 reviews are used for this study. Table I represents a summary of the dataset used:

TABLE I. Summary of the Dataset

Domain	Numbers of reviewers	Words average
Hotels	14356	7.06

	Hotel_Name	Title_review	reviews_hotel	Score_rating	Rating_Date
0	Radisson Blu Hotel, Marrakech Carre Eden	Such a nice experience! My best hotel in marra...	A peaceful hotel that I completely loved & enj...	50	29 April 2021
1	Radisson Blu Hotel, Marrakech Carre Eden	Good experience and good service, nice staff!	I appreciated the service and the warm welcomi...	50	18 April 2021
2	Radisson Blu Hotel, Marrakech Carre Eden	Not as expected	Needs more adjustmentsInlt dosent seems that y...	30	10 April 2021
3	Radisson Blu Hotel, Marrakech Carre Eden	Great Room Service	Have to say my start with the Raddison was iff...	40	30 March 2021
4	Radisson Blu Hotel, Marrakech Carre Eden	Refund Refused despite the Pandemic	I was an expat living in Marrakech. Four of us...	10	19 November 2020

Fig. 4. Example of datasets

3) Review pre-processing: The pre-processing procedure followed in this work aims to clean up the notices and to make them as close as possible to a formal language. First, the notices were filtered considering only those written in English. Because a corpus of different languages is a corpus that contains noise. To do this, a Python library called Langdetect is used and then proceeded to a pre-processing that follows the following steps:

Split the text into several rows: As represented in Fig. 5, this task consists of splitting the text contained in each cell of the "Reviews_hotel" column into several rows by the '.' delimiter, for example:

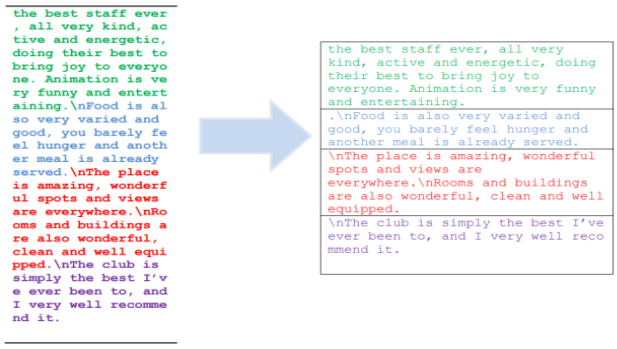

Fig. 5. Example of paragraph splitting into multiple lines

Noise cleaning - spacing, special characters, lowercasing: In a text you can find various characters such as numbers, free white spaces, all kinds of punctuation and some terms are put in random capital letters. This form of noise cleaning takes care of the spacing and special characters. Turning all words into lower case is also a very common pre-processing step. Next, all punctuation and special characters will be removed since they serve no purpose once analyzing the data begins.

Eliminate emoji: One cannot ignore the content of the notices full of emoticons, symbols and pictographs as well as a set of flags. So, this task concerns the elimination of these characters.

Tokenization / eliminating words below three letters: First, separating the corpus into a vocabulary of single terms is essential, which is called tokenization. Individual terms and overwrite all words below three letters can be tokenized.

Delete Stop Words: Some words in English, while necessary, do not contribute much to the meaning of a sentence. These words, such as "when", "had" or "before", are called stop words and should be filtered out.

Stemming and lemmatization of the text: The process of turning a word into its Racine form is known as racinization.

Rooting can create non-real words. Lemmatization, as opposed to racinization, aims to obtain the canonical (grammatically correct) word forms, or lemmas. In terms of calculation, lemmatization is much more difficult and computationally expensive than racinization. In actual practice, the two methods have little impact on the performance of text classification.

4) Aspect terms extraction: To accomplish the Machine Learning task, the aspect or category detection of each review must be first tackled. This can be done according to different tools described in Section III. Topic Modeling has been tested using LDA techniques as well as CorEx and NMF.

For this experimentation, 9 Aspects are predefined: Rooms/Cleanliness, Location, Staff, Food/Restaurant, Experience/Value, Price/Quality, Service, Amenities/Activities, and Hotel/Property.

5) Topic modeling: In this section, the steps as well as the result of the three topic modeling methods have been presented, namely, LDA, NMF and CorEx. The steps followed using LDA are: corpus vectorization which allows to create the term matrix document using Counvectorizer which is an excellent tool provided by Scikit-learn library in Python. It is used to transform a given text into a vector based on the frequency (number) of each word that occurs in the whole text. Then, the LDA model will be built, to evaluate this performance with perplexity and log probability. Gridsearch was used to choose the "right" number of topics for the LDA model, then two hyper parameters (learning decay and number of topics) were tested. And finally, the labeled topics joined the original text.

In Non-negative Matrix Factorization (NMF), the first step is to convert the document into a term-document matrix which is a collection of all the words in the given document using the TfidVectorizer. Then, build the NMF model with Scikit-learn and view the original topics. And as far as CorEx is concerned, the first step is the corpus vectorization which converts the document into a document-Terms-matrix using the TfidVectorizer by creating a vocabulary containing the topics. The second step is the creation of the model, starting with the identification of the Anchors (anchored words) and then creating the model that allows to generate the set of topics. In each of these models, topics related to four distinct themes for LDA and NMF can clearly be seen. But it is also clear that these topics contain words that can apply to multiple contexts and cause problems in certain circumstances. The following table (Table II) illustrates the comparison between these techniques:

TABLE II. COMPARISON OF THE THREE EXTRACTION METHODS

	LDA	NMF	CorEx
Topic 0	'room','leave', 'shower', 'bathroom', 'table', 'change', 'towel','work', 'wasnt','tv', 'dirty', 'open','service', 'expect','door', 'drink','provide', 'toilet','available', 'reception'	'room', 'clean', 'comfortable', 'spacious','bed', 'bathroom','beautiful', 'shower', 'big', 'small','towel', 'need', 'view','large','size', 'balcony','wifi', 'daily', 'work', 'floor'	room, clean, bed, comfortable, bath room,shower, towel, spacious,table, door, balcony, room clean,bedroom,room spacious, room pool, petal, clean room,hotel clean, size, toilet
Topic 1	'staff','friendly','helpful', 'hotel','spa','nice', 'reception', 'ok', 'massage','ask','tour', 'manager','extremely', 'speak','french', 'stay', 'english', 'owner', 'polite'	'staff', 'friendly', 'really','attentive', 'polite',' welcome ','reception','amaze', 'extremely','service', 'professional','amazing','member','animation','nice', 'make', 'team','kind','really'	staff,friendly,helpful,team,animation team,manager, waiter, professional, staff friendly, receptionist, reception staff, hotel staff,attentive, friendly helpful, staff helpful, member,polite, lifeguard,restaurant staff, helpful staff, staff polite
Topic 2	'room', 'clean','pool', 'nice','hotel','bed', 'view','lovely', 'comfortable','beautiful', 'small', 'spacious', 'sun','garden','large', 'great','big','wifi', 'terrace','ground'	'great','value','location','experience','time' ,'service','breakfast', 'atmosphere','staff', 'family','view','overall','team','animation', 'trip', 'kid','money', 'visit', 'people','spa'	food, restaurant, breakfast,drink,dinner,meal, fresh, lunch, menu, buffet, delicious,cook,coffee,fruit,snack,omelette,salad, eat, ate, tea, bread,juice,order, mint
Topic 3	'team','staff','make', 'animation', 'work', 'bar','entertainment', 'really','help','amaze', 'brilliant','hard', 'special','aqua','attentive','great','time','especially','best', 'kid'	'hotel', 'best','beautiful', 'amazing', 'lovely','time','book', 'nice','fantastic','like','city, 'locate','boutique','wonderful', 'shuttle','medina','visit','really','experience', 'return'	hotel, stay, back, recommend, experience return, enjoy, come back, good, highly recommend, recommend hotel, stay hotel,highly, enjoyed, place stay, beautiful hotel, nice hotel,boutique hotel, lovely hotel

While topic models can be rapidly run, they are not necessarily as accurate in their classification decisions as the more complicated supervised learning models, and occasionally their outputs can even be outright false. Semi supervised topic modeling will be utilized to determine the main topics of these documents in order to prevent ambiguities between topics. This most recent development gave a middle ground between supervised classification modeling and unsupervised topic modeling.

From the comparison table, CorEx provides more specification of aspects than the others, CorEx is chosen as the best, as the grouping of each aspect seems better. Fig. 6 shows what the dataset looks like at this point.

	review_text_clean	Score_rating	dominant_topic	Categorie_Aspect
0	nice experience hotel amaze view	50	4.0	Experience/Value
1	netflix amaze	50	7.0	Amenities/Activities
2	peaceful hotel completely love enjoyed beginning	50	4.0	Experience/Value
3	staff professional welcoming	50	2.0	Staff
4	shoutout rachid lansari manager team	50	2.0	Staff
...	...	...	...	...
14351	hotel close everywhere	30	1.0	Location
14352	recommend little cafe door delicious fresh juice	30	4.0	Experience/Value
14353	pool lovely	30	7.0	Amenities/Activities
14354	food	30	3.0	Food/Restaurant
14355	price pay overall happy	30	5.0	Price/Quality

14356 rows × 4 columns

Fig. 6. Datasets after aspect extraction

6) Annotation of journals: For review annotation, reviews were labeled using two tools VADER (Valence aware Dictionary and Sentiment Reasoner) and TextBlob. There are three types of sentiments in this dataset: positive, negative, and neutral. To pursue the supervised learning approach, the type of sentiment (polarity) of each review should be known.

VADER has been chosen since it provides a better classification and more negative feelings than the other (Fig. 7).

	review_text_clean	Score_rating	dominant_topic	Categorie_Aspect	scores	compound	Sentiment_Vader
0	nice experience hotel amaze view	50	4.0	Experience/Value	{'neg': 0.0, 'neu': 0.323, 'pos': 0.677, 'comp...	0.7430	Positive
1	netflix amaze	50	7.0	Amenities/Activities	{'neg': 0.0, 'neu': 0.222, 'pos': 0.778, 'comp...	0.5423	Positive
2	peaceful hotel completely love enjoyed beginning	50	4.0	Experience/Value	{'neg': 0.0, 'neu': 0.21, 'pos': 0.79, 'compou...	0.9056	Positive
3	staff professional welcoming	50	2.0	Staff	{'neg': 0.0, 'neu': 0.408, 'pos': 0.592, 'comp...	0.4404	Neutral
4	shoutout rachid lansari manager team	50	2.0	Staff	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	Neutral

Fig. 7. Datasets after polarity detection

7) Multi-target classification (aspect/sentiment): In order to use machine learning algorithms in the textual data, there is a need to represent the text in the document as a vector of fixed size and this in order to plunge the data in a metric space. Among the vectorization techniques the TF-IDF and CountVectorizer are two ways to convert text into numbers.

a) Count vectorizer: Count Vectorizer offers a straightforward method for tokenizing a group of text documents, creating a vocabulary of recognized words, and encoding new documents using that vocabulary.

b) TF-IDF vectorizer: TF-IDF, which stands for Term Frequency - Inverse Document Frequency, is a statistic which is based on a word's frequency in the corpus. It also gives a numerical representation of a word's importance for statistical analysis.

B. Testing and Evaluation Phase

In this experiment, the most commonly used classifiers in the sentiment analysis literature are applied. "Documents x Terms" vectorization method will be evaluated using nine supervised classification models: Bayesian Naive, SVM, Logistic Regression, K-nearest Neighbor, Decision Trees, Random Forests, Extratrees, Adaboost and Gardient Boost. The performance of the selected models will be compared using their Accuracy, Precision, Recall and F1-scores to determine the best decision model.

1) Description datasets: The sentiment analysis as well as the aspect analysis were performed on a dataset that contains 14356 English dialect reviews from TripAdvisor and Booking websites and labeled as follows: 4337 positive texts, 397 negative texts and 8647 neutral, still 2822 texts labeled with the aspect Room, 386 with Service, 1952 with Food/Restaurant, 2118 with Staff, 1352 labeled with Location, 2833 with Experience/Value, as well as 992 are labeled with the aspect Amenities/Activities and 420 reviews are with Price/Quality.

2) Performance measure: The choice of classifier for the current data is based on the performance measures [35]. The evaluation of the optimal solution in classification training can be defined based on the confusion matrix. From the given confusion matrix one can determine, the number of positive and negative that are correctly classified. Meanwhile, the number of negative and positive cases are misclassified respectively. The performance measure of the various classifiers is evaluated using the accuracy, precision, recall and F1-scores.

The Accuracy metric: indicates the percentage of correct predictions. It refers to the ratio of the number of correct predictions to the total number of input samples or observations, which is shown in eq. (1).

$$Accuray = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: is the number of correct positive results divided by the number of positive results predicted by the classifier. The result is a value between 0.0 for no precision and 1.0 for total or perfect precision.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall: In order to complete the accuracy, the recall is also calculated, which is the fraction of true positives to real positives, which is shown in eq. (3), i.e., the proportion of positives that were correctly identified.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score: The calculated average harmony of precision and recall is used to assess how well these two metrics—rappel and precision—compromise. This unique score ranges from 0 to 1, with 0 being the worst possible outcome and 1 being the best possible outcome. It can be calculated as follows.

$$F1 = \frac{2*(Precision*Recall)}{Precision+Recall} \quad (4)$$

3) Results: The learning phase will be followed by the testing phase in order to evaluate the classifier. For performance validation, 80/20% rules is used to check the model, the corpus is divided into two parts, 80% for the training phase and 20% for the testing phase. Several tests, whose results are presented in the following tables, were made:

a) CountVectorizer embeddings classification

- Aspect Classification

TABLE III. RESULTS OBTAINED FROM ALL CLASSIFIERS IN COUNTVECTORIZER WEIGHTING METHOD FOR ASPECT CLASSIFICATION

Metrics / Classifiers	Accuracy %	Recall%	Precision %	F1-Score %
LR	83.25	73.01	75.56	73.42
RF	**84.26**	**74**	**77.14**	**75**
NB	76	55.78	67.68	56.29
DT	48.70	28.12	43.75	26.53
KNN	71.59	63.71	67.22	64.95
SVM	83.51	70.91	86.84	72.88
ET	81.67	71.41	74.66	72.49
AB	69.83	59.33	65	60.44
GB	76.13	61.07	69	62.72

The results of aspect extraction are reported in Table III. The latter shows that the best performances are obtained in: Accuracy (84.26%), Recall (74%), Precision (77.14%) and F1-Score (75%) with the RandomForest + CountVectorizer configuration.

- Sentiment classification

TABLE IV. RESULTS OBTAINED FROM ALL CLASSIFIERS IN COUNTVECTORIZER WEIGHTING METHOD FOR SENTIMENT CLASSIFICATION

Metrics / Classifiers	Accuracy %	Recall%	Precision %	F1-Score %
LR	**91**	**69.59**	**82**	**73.27**
RF	**88.87**	67.16	85	**72**
NB	83.28	60.43	67.26	61.58
DT	74.40	43.55	56.09	43.52
KNN	74.03	44.86	77.84	46.24
SVM	87.86	62.48	80.02	65.76
ET	88.53	68.19	82	72.59
AB	84.52	61.07	77.02	65.24
GB	87.03	57.64	58.58	57.85

The sentiment classification results are reported in Table IV. The latter shows that the best performances are obtained in precision (82%), recall (69.59%), accuracy (91%) and F1-score (73.27%) with the LogisticRegression + CountVectorizer configuration.

*b) TF-IDFVectorizer embeddings classification:*Both Tables V and VI show the results of the classifier using the TF-IDF weighting model for sentiment classification.

- Aspect classification

TABLE V. RESULTS OBTAINED FROM ALL CLASSIFIERS IN TF-IDF WEIGHTING METHOD AOR ASPECT CLASSIFICATION

Metrics / Classifiers	Accuracy %	Recall%	Precision %	F1-Score %
LR	83	70.01	78	72
RF	**86**	**74.78**	**81.01**	**76.59**
NB	76.59	72.08	50.55	51
DT	48.70	28.13	44.15	27
KNN	63.73	52.46	59.63	54.58
SVM	82.76	70.69	76.14	71.75
ET	82.80	72	76.04	73
AB	70	59.43	64.01	60.58
GB	71.63	52.04	52.42	51.44

The results of aspect classification with the TF-IDF vectorization method are presented in Table V. This latter shows that using the RandomForest + TF-IDF parameter provides the best performance in terms of precision (81.01%), recall (74.78%), accuracy (86%) and F1-score (76.59%).

- Sentiment classification

TABLE VI. RESULTS OBTAINED FROM ALL CLASSIFIERS IN TF-IDF WEIGHTING METHOD FOR SENTIMENT CLASSIFICATION

Metrics / Classifiers	Accuracy %	Recall%	Precision %	F1-Score %
LR	87.37	60.73	83.84	64.04
RF	**89**	**66.19**	**84.51**	**70.55**
NB	82.87	53.17	57	54.02
DT	74.44	43.57	56.20	43.53
KNN	75.87	50.25	80	54.16
SVM	**87.63**	62.39	82.41	66.54
ET	**87.29**	76.26	78.37	71.10
AB	**84.82**	61.30	76.64	65.58
GB	**87.11**	57.50	59	57.89

The results of the sentiment classification are presented in Table VI. The latter shows that the use of the RandomForest + TF-IDF parameter allows to obtain the best performances in terms of accuracy (89%), Recall (66.19%), precision (84.51%) and F1-Score (70.55%).

V. DISCUSSION

Traditional Sentiment Analysis is done through Sentiment Analysis techniques[36] on documents and sentences which assesses the overall polarity of the feelings of the given opinion target. Nevertheless, if the opinion target contains various aspects with a conflicting sentiment, using a single sentiment label to represent it could be incorrect[37]. The current trend is to move to a deeper level that presents itself as Aspect-Based Sentiment Analysis. ABSA is the sub-field of NLP that essentially breaks the data into aspects and finally extracts the sentiment information[38]. It performs a more advanced and higher quality analysis because it directly examines the sentiment itself. Neither document analysis nor sentence analysis find out what exactly people like and dislike. Specifically, the idea behind this work is to collect customer reviews on tourism sites such as: Booking or TripAdvisor, and assign a sentiment analysis that allows to extract the most relevant characteristics in the review of most customers. Hence the realization of a sentiment analysis as well as an aspect analysis on a dataset that contains customer reviews of hotels located in Marrakech. The results of this research were presented as follows:

The best result in all the tests for the classification of aspects is 86%, it was obtained by the RF with the use of TF-IDF, similarly the classifier GradientBoost, NB and AdaBoost reached their maximum measure (76.13%, 77%, 70% respectively), on the other hand for LR, SVM and KNN, their best results were with CountVectorizer (83.25%, 83.51%, 71.59%) respectively. Moreover, the DT classifier obtained the same result in both tests with TFIDF and CountVectorizer (48.70%). On the other hand, for sentiment classification the best result of Accuracy is 91% in CountVectorizer weighting method, achieved by LR classifier, also NB and SVM classifiers their excellent outcome were with the same method (83.28%, 87.86% respectively), also the DT classifier score is similar for both vectorization techniques with 74%, and both RF, GB, AdaBoost, and KNN classifiers got their excellent score with TF-IDF (89%, 87.11%, 84.82% and 75.87% respectively). Analyzing the results of the confusion matrices the RF+TFIDF that gave as results 2232 True Positives for the classification of the aspects, besides the best classifier for the analysis of the feelings is the LR+CountVectorizer that allows to reach a number of True Positives equal to 3158. These results show that RF is generally considered a better classifier for the aspect extraction task; in return LR is the good classifier for the sentiment classification task. As long as a good measurement is achieved, the model has permitted to perform correct results. These results are listed in Fig. 8 and 9.

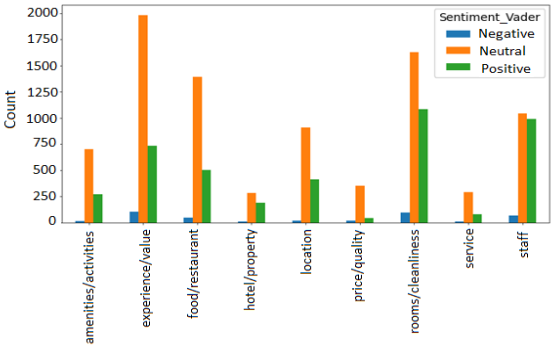

Fig. 8. Sentiment per single aspect

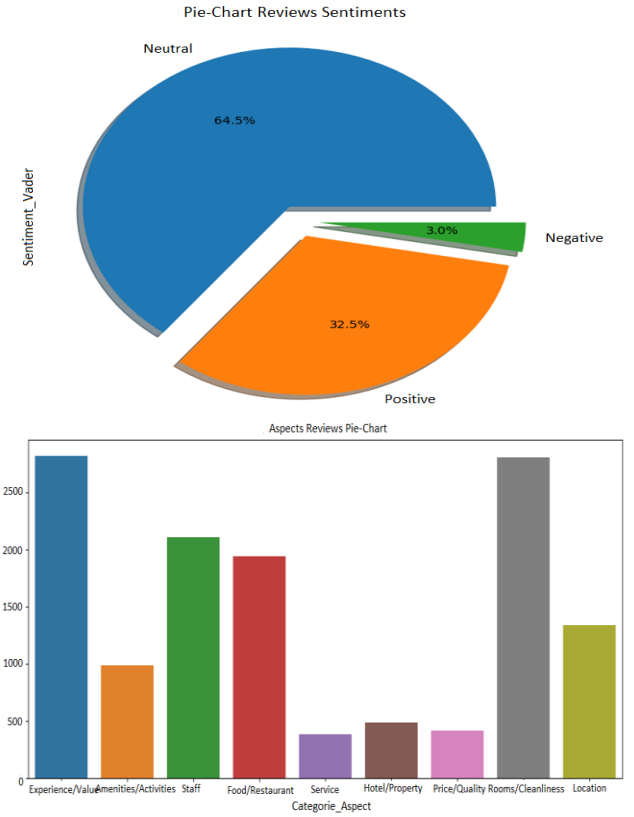

Fig. 9. Data analysis interface

VI. Conclusion

Given the increasing importance placed on online reviews and the evidence that these reviews influence customer behaviors, it is clear that companies are beginning to look at technologies that could automatically analyze what customers are saying about their product or service. In this article, the analysis of the feelings has been performed on the level of aspects in the field of hotels. The hotel reviews that were studied are comments on the service provided, written in English on the rating site TripAdvisor. The preprocessing and vectorization methods are evaluated using nine supervised classification models. The results obtained are very encouraging, and the experimentation conducted on the dataset reveals that a better accuracy score of 92% and 3158 True Positives were achieved when using the LR+ CountVectorizer classifier for sentiment analysis as well as a good Accuracy score of 86% and a number of True Positives equal to 2232 when using RF + TF-IDF Vectorizer for aspect classification. For future work, enriching the dataset with other reviews from other languages such as French and Arabic will be highly recommended and that is by making comparisons between these languages in order to get a broader view on the aspects that are most noticed by travelers from other countries; thus, the use of the mixed class analysis rather than the positive, negative and neutral classes.

References

[1] C. Zong, R. Xia, and J. Zhang, Text Data Mining. Singapore: Springer Singapore, 2021. doi: 10.1007/978-981-16-0100-2.

[2] A. D. Francisco, “Aspect Term Extraction in Aspect-Based Sentiment Analysis Aspect Term Extraction in Aspect-Based Sentiment Analysis,” 2019.

[3] I. E. Vermeulen and D. Seegers, “Tried and Tested : The Impact of Online Hotel Reviews on Consumer Consideration Tried and tested : The impact of online hotel reviews on consumer consideration,” no. February 2009, 2020, doi: 10.1016/j.tourman.2008.04.008.

[4] K. Ravi and V. Ravi, A survey on opinion mining and sentiment analysis : tasks , approaches and applications, no. June. Elsevier B.V., 2015. doi: 10.1016/j.knosys.2015.06.015.

[5] J. Zelenka, T. Azubuike, and P. Martina, “administrative sciences Trust Model for Online Reviews of Tourism Services and Evaluation of Destinations,” 2021.

[6] H. A. Lee, R. Law, J. Murphy, and J. Murphy, “Helpful Reviewers in TripAdvisor , an Online Travel Community,” no. September 2012, pp. 37–41, 2011, doi: 10.1080/10548408.2011.611739.

[7] S. Shayaa, N. I. Jaafar, S. Bahri, A. Sulaiman, and M. A. L. I. Al-garadi, “Sentiment Analysis of Big Data : Methods , Applications , and Open Challenges,” IEEE Access, vol. 6, pp. 37807–37827, 2018, doi: 10.1109/ACCESS.2018.2851311.

[8] V. Pekar and S. Ou, “Discovery of subjective evaluations of product,” vol. 14, no. 2, pp. 145–155, 2008, doi: 10.1177/1356766707087522.

[9] A. Muangon and S. Thammaboosadee, “A Lexiconizing Framework of Feature-based Opinion Mining in Tourism Industry,” pp. 169–173, 2014.

[10] E. Marrese-Taylor, J. D. Velasquez, and F. Bravo-Marquez, “OpinionZoom, a modular tool to explore tourism opinions on the Web,” Proc. - 2013 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IATW 2013, vol. 3, pp. 261–264, 2013, doi: 10.1109/WI-IAT.2013.193.

[11] Z. Hai, K. Chang, J. J. Kim, and C. C. Yang, “Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 623–634, 2014, doi: 10.1109/TKDE.2013.26.

[12] A. Mukherjee and B. Liu, “Aspect Extraction through Semi-Supervised Modeling,” no. July, pp. 339–348, 2012.

[13] W. Kasper and M. Vela, “Sentiment Analysis for Hotel Reviews,” In Computational linguistics-applications conference, vol. 231527, pp. 45–52, 2011.

[14] M. Pontiki et al., “SemEval-2016 Task 5 : Aspect Based Sentiment Analysis,” Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, pp. 19–30, 2016.

[15] A. Nazir, Y. Rao, L. Wu, and L. Sun, “Issues and Challenges of Aspect-based Sentiment Analysis : A Comprehensive Survey,” no. October, 2021, doi: 10.1109/TAFFC.2020.2970399.

[16] C. D. Manning, H. Schütze, and G. Weikurn, “Foundations of Statistical Natural Language Processing,” SIGMOD Rec., vol. 31, no. 3, pp. 37–38, 2002, doi: 10.1145/601858.601867.

[17] K. Mandal,“Topic Modeling: Techniques and AI Models - DZone,” Dec. 15, 2020. https://dzone.com/articles/topic-modelling-techniques-and-ai-models.

[18] B. Dutta, “What is Topic Modelling in NLP? Analytics Steps,” Jan 15, 2022, https://www.analyticssteps.com/blogs/what-topic-modelling-nlp

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” vol. 3, pp. 993–1022, 2003.

[20] T. K. Landauer and P. W. Foltz, “An Introduction to Latent Semantic Analysis,” pp. 259–284, 1998.

[21] Avinash Navlani, “Python LSI_LSA (Latent Semantic Indexing_Analysis) _ DataCamp,” Oct 2018. https://www.datacamp.com/tutorial/discovering-hidden-topics-python

[22] D. D. Lee, M. Hill, and H. S. Seung, “Algorithms for Non-negative Matrix Factorization,” Adv. Neural Inf. Process. Syst., no. 1, pp. 556–562, 2001.

[23] R. J. Gallagher, K. Reing, D. Kale, and G. Ver Steeg, “Anchored Correlation Explanation : Topic Modeling with Minimal Domain Knowledge,” vol. 5, pp. 529–542, 2017.

[24] M. Ounacer, S., Jihal, H., Ardchir, S., & Azzouazi, “Anomaly Detection in Credit Card Transactions,” Adv. Intell. Syst. Sustain. Dev., 2019, doi: 10.1007/978-3-030-36674-2_14.

[25] J. M. Hilbe, Practical Guide to Logistic Regression (1st ed.), Chapman and Hall/CRC, https://doi.org/10.1201/b18678, 2015.

[26] V. Jakkula, “Tutorial on Support Vector Machine (SVM),” School of EECS, Washington State University. pp. 1—13, 2006.

[27] Z. Zhang, “Introduction to machine learning : k-nearest neighbors,” vol. 4, no. 11, pp. 1–7, 2009, doi: 10.21037/atm.2016.03.37.

[28] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 22, pp. 41-46, 2001.

[29] A. Ashari, “Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool,” vol. 4, no. 11, pp. 33–39, 2013.

[30] L. Breiman, “RANDOM FORESTS,” Mach. Learn. 45, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[31] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” no. November 2005, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[32] R. E. Schapire, P. Avenue, and A. Room, “The Boosting Approach to Machine Learning An Overview,” pp. 1–23, 2003.

[33] P. Celio, D. Cellio, S. Experian, M. Forti, M. Witarsa, and S. Experian, “A comparison of Gradient Boosting with Logistic Regression in Practical Cases GRADIENT BOOSTING MACHINES – THEORY,” pp. 1–25, 2018.

[34] V. Carle, “DEGREE PROJECT IN THE FIELD OF TECHNOLOGY Web Scraping using Machine Learning,” 2020.

[35] S. Ounacer, H. Jihal, K. Bayoude, A. Daif, and M. Azzouazi, “Handling Imbalanced Datasets in the Case of Credit Card Fraud,” in Advances in Intelligent Systems and Computing, 2022, pp. 666–678. doi: 10.1007/978-3-030-90633-7_56.

[36] A. R. Alaei, S. Becken, and B. Stantic, “Sentiment Analysis in Tourism: Capitalizing on Big Data,” J. Travel Res., vol. 58, no. 2, pp. 175–191, 2019, doi: 10.1177/0047287517747753.

[37] L. Zhu, M. Xu, Y. Bao, Y. Xu, and X. Kong, “Deep learning for aspect-based sentiment analysis: a review,” PeerJ Comput. Sci., vol. 8, 2022, doi: 10.7717/PEERJ-CS.1044.

[38] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, “A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges,” pp. 1–21, 2022, [Online]. Available: http://arxiv.org/abs/2203.01054