A Proof of Concept Web Application for Sentiment Analysis in Tourism in the Region of Draa-Tafilalet

Loukmane Maada[1][a], Khalid Al Fararni[2][b], Badreddine Aghoutane[1][c],Yousef Farhaoui[3][d], Mohammed Fattah[4][e]

[1]*IA Laboratory, Science Faculty, Moulay Ismail University, Meknes, Morocco*
[2]*LISAC Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco*
[3]*L-STI, T-IDMS, University of Moulay Ismail, Faculty of Science and Technics, Errachidia, Morocco*
[4]*Image laboratory, Moulay Ismail University Meknes, Morocco.*

Keywords: Sentiment Analysis, Machine Learning, Tourism, Big Data

Abstract: Advances in technology have changed how people consume and produce information in the field of tourism. Tourists nowadays tend to share their experiences on social media, forums, *etc*. Those shared experiences became a powerful source of influence in the tourists' community. However, the vast amount of data produced every day on social media makes manual processing an impossible task, making the use of analysis approaches a must. Sentiment Analysis (SA), a sub-field of Natural Language Processing (NLP), is rapidly rising as an automated process of examining semantic relationships and meaning in reviews. This paper provides a proof of concept of a tourism-oriented application that uses Sentiment Analysis to summarize the tourists' impression (positive or negative) vis-à-vis hotels.

1 INTRODUCTION

With the emergence of the new web paradigm in the mid-late 2000s, introduced as Participative Web or the second web generation (Web 2.0 for short), the users now can share a different type of content on platforms for distinct purposes such as microblogging (i.g. Twitter), multimedia sharing (e.g. YouTube), review forums (e.g. TripAdvisor), *etc.*

Genuinely people are influenced by others' opinions in different domains. Pang and Lee (Pang and Lee, 2008) confirmed it and reported that the impact of the online reviews was significant enough to make a person pay 20%-99% for a 5-star item than a 4-star one. This opinion influence is also found in tourism, where tourists tend to look for information on social media or forums. (Assumpció Huertas, 2018) shows in his paper that live videos and stories can be used to make a destination more attractive and enticing. On the other hand, (Narangajavana Kaosiri Y et al., 2019) investigate the impact of social media on tourist satisfaction. The results confirmed that Social Media plays an important part in the pre-traveling period since it influences the whole decision-making process and builds up tourist expectations about the destination, ergo his satisfaction.

The humongous amount of tourism-related user-generated content produced on social media or the internet, in general, complicates the tourist's information gathering process. For instance, in 2015, TripAdvisor received over 200 million feedback and comments, this number gets even bigger when it comes to Facebook; over 800 million active users posted tips and exchanged tourism-related posts. This quantity of data rendered the manual information gathering obsolete and suboptimal, especially that more than 74% of travellers rely on other users'

[a] https://orcid.org/0000-0003-4165-1486
[b] https://orcid.org/0000-0001-5907-6948
[c] https://orcid.org/0000-0002-9555-6786
[d] https://orcid.org/0000-0003-0870-6262
[e] https://orcid.org/0000-0001-6128-9715

Maada, L., Al Fararni, K., Aghoutane, B., Farhaoui, Y. and Fattah, M.
A Proof of Concept Web Application for Sentiment Analysis in Tourism in the Region of Draa-Tafilalet.
DOI: 10.5220/0010735900003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 426-432
ISBN: 978-989-758-559-3

feedback when planning trips (Pelsmacker et al., 2018).

The contribution of this paper is the study of the feasibility of a big data web application (Boulaalam, Oumayma, et al, 2018) that analyses and classifies opinions (negative, positive) based on tourists' feedbacks about hotels, museums, and restaurants in the Draa-Tafilalet (K.AL Fararni, 2021).

The remainder of the paper is structured as follows: Section 2 proposes a short literature review; Section 3 presents the methodology we used in our model; in section 4, we display the results; in section 5, we present our application briefly, and section 6 concludes the paper and displays future work.

2 LITERATURE REVIEW

Sentiment analysis, also referred to as opinion mining or emotion AI, is a set of analytic methods that aims to extract knowledge embedded in subjective text on the internet. Those methods can be categorized into three categories: the lexicon-based methods, the machine learning methods, and the deep learning methods.

The lexicon-based methods are based on the use of a well-rendered sentiment lexicon to determine the text polarity. These methods include the dictionary-based approaches, the corpus-based, and the manual approaches.

The **dictionary-based** approach consists of building from a small number of sample sentiments in which polarity was manually set. The number of words iteratively increases using a well-known lexicon, e.g., WordNet (Miller George A et al.,1990). Like the dictionary-based one, the corpus-based approach starts with a small set of manually calibrated sentiment words. The number of words then increases utilizing a large corpus and following a set of predetermined rules and formulas like the LDA and the PMI. Unlike the dictionary-based approach, the Corpus-based ones can link words to context; in other words, the word polarity depends on the context, not on a predefined value. The **manual-based** approach relies on the manual collection and labeling of the lexicon, which requires a significant amount of human effort and time. (Qiu et al., 2010) ameliorated the targeted advertising strategy using a dictionary-based approach. The presented approach extracted topics and opinions from sentences, which helped consumers' attitude identification towards topics, resulting in more accurate advertising. (Rajput and Haider, 2016) performed a lexicon-based sentiment analysis on students' evaluations of professors at the end of a course. The presented approach is based on using a dictionary to compute the sentiment score of each feedback; this metric is then used to determine if the feedback is positive, negative, or mixed.

The machine learning methods capitalize on classification methods, supervised and unsupervised, to determine the textual content polarity. Those methods exploit the bag of words (BOW), part of speech, the n-gram feature, and the TF-IDF model.

Among the various classification strategies for detecting users' emotions from their text: **SVM**, **LDA**, **Linear regression**, **Naïve Bayes**, and **artificial neural networks** are more common and achieve the highest performance. (Nikhil Kumar Singh et al., 2020) compared different algorithms (SVM, LR, NB, RF) with multiple feature extraction techniques (BOW, POS, Hash Tagging) on two databases: the Twitter sentiment corpus data set and the Stanford data set. The results showed that SVM and NB with POS were on top with, respectively, 83.27% and 83.13% on the first data set and 81.34% and 80.12% on the second one. (Fang Luo et al., 2016) proposed a feature selection algorithm, CHIsquare Difference between the Positive and Negative Categories (CDPNC), that uses both Document Frequency (DF) and Chi-Squared (CHI). This method was tested with three algorithms: SVM, KNN, and NB. The experimental results show that the classification efficiency of the proposed system outperforms the state-of-the-art, especially when used with SVM.

The deep learning methods use deep network architectures to output text polarity. The strength of this approach comes from the use of advanced word embedding tools like word2vec and GloVe (Pennington Jeffrey et al., 2014). The RNN class of neural networks was the obvious choice in the context of sequential data like text. However, recent research showed that combining RNN and CNN classes results in better performance (Behera Ranjan Kumar et al., 2021). (Zhou et al.,2016) have proposed two bi-directional LSTM models for sentiment analysis. The first one combines biLSTM and a two-dimensional pooling (BLSTM-2DPool), the second one combines biLSTM and a two-dimensional Convolution Layer (BLSTM-2DCNN). The BLSTM-2DCNN did not only outperform RecNN, RNN, and CNN models but also the BLSTM-2DPool on multiple databases, namely Stanford Sentiment Treebank (SST), TREC, and 20Newsgroups (20Ng). In 2018, (Weihang Huang et al.,2018) proposed a document-level sentiment analysis model SSR-LSTM. This model first removes sentences with weak emotions and then proceeds with two layers of LSTM. The first layer employs the word embedding technique to generate sentence vectors, which are then fed into the second layer to generate document representations. The SSR-LSTM outperforms the state-of-the-art, according to the empirical result.

3 METHODOLOGY

This section displays the sentiment analysis approach; we choose our application, the system architecture, the preprocessing steps we are using, and the feature extraction technique we have chosen.

3.1 Models

We adapted the machine learning approach since it does not need as much computational power as the deep learning approach and still gets good results compared to the state-of-the-art. The models we choose to work with are the Support vector machines (SVM), Naive Bayes (NB), and Linear regression (LR).

3.1.1 Support Vector Machines

It is a binary classification method based on maximizing the margin between two classes using a hyper plan.

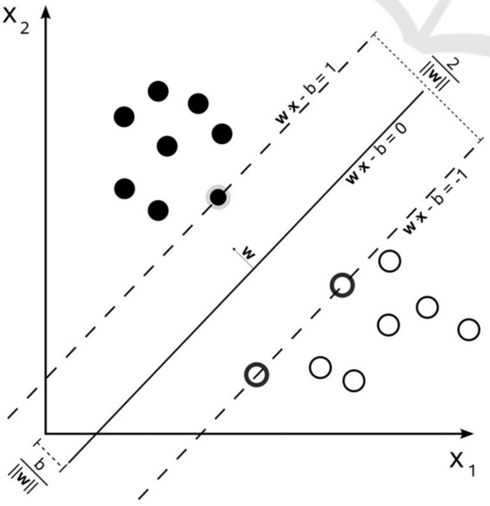


Figure 1: SVM algorithm illustration.

Figure 1 visually illustrated the algorithm in the case of points plotted in 2D-space. The SVM separates the two categories (white and black) with a line (hyper plan in 2D) that maximizes the margin.

3.1.2 Naive Bayes

It is a probabilistic model that relies on the Bayes formula Eqs.1 and assumes that the entities are independent (features), which means that for a text (t), out of all sentiments in S, Naive Bayes returns the sentiment s_t, which has the maximum posterior probability Eqs.2.

$$P(s/t) = \frac{P(s)*P(t/s)}{P(t)} \quad (1)$$

$$s_t = argmax_{s\epsilon S} P(s/t) \quad (2)$$

Where S is the set of all the sentiments.
Since a text is a set of words, we can write it as an n-tuple of words **($w_1, w_2, ..., w_n$)**. By replacing t in Eqs.1 and the "naive" assumption of independence, we get Eqs.3.

$$P(s/t) = \frac{P(s)*\prod_1^n P(w_i/s)}{\prod_1^n P(w_i)} \quad (3)$$

We obtain Eqs.4 by substituting P(t/s) in Eqs.2 from the one we got in Eqs.3.

$$s_t = argmax_{s\epsilon S} P(s/t) * \prod_1^n P(w_i/s) \quad (4)$$

3.1.3 Linear Regression

It is a supervised linear model that assumes a linear relationship between the input variables and the output. This regression method creates a linear function by calculating weight values **($w_1, w_2, ..., w_n$)** for each input feature **($x_1, x_2, ..., x_n$)** and outputs a numerical value y Eqs.5. For instance, in Sentiment Analysis, we can associate (-1) to negative and (1) to positive.

$$y = \sum_1^n w_i x_i \quad (5)$$

3.2 Methodology Workflow

The application classifies sentiment polarity into positive and negative categories. The process workflow is illustrated in Figure 2 and discussed in more detail in the following subsections.

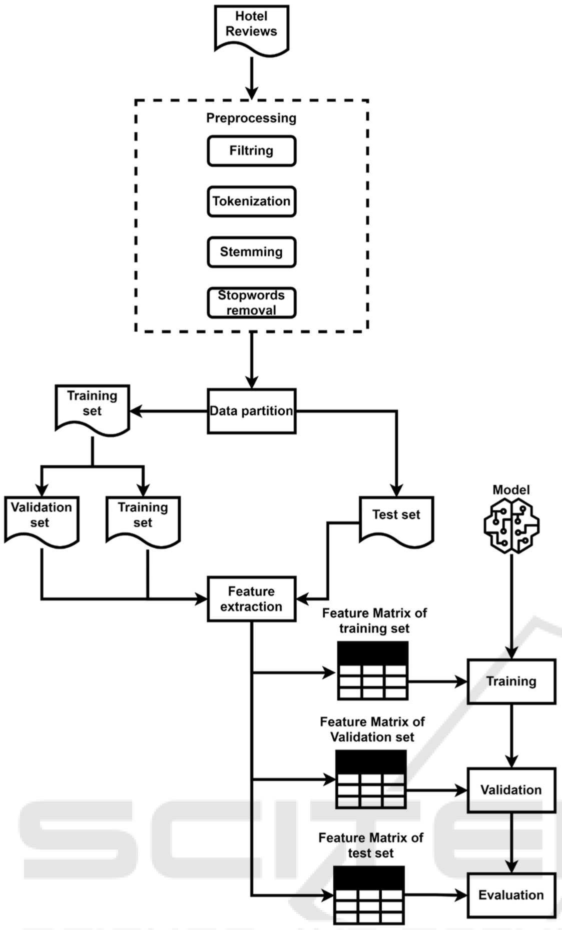


Figure 2: Methodology workflow.

3.3 Preprocessing

Pretreatment is a crucial step in which the data is converted or encoded to bring it into such a state that the machine can now efficiently analyze it. S.Alam and (N. Yao et al., 2019) studied the preprocessing impact on ML algorithm's accuracy. They removed emoticons and applied bi-grams on the dataset, removed stop words, used stemming and word vectorization. The empirical results showed a considerable improvement of accuracy for the Naive Bayes algorithm and a slight improvement of accuracy for the SVM algorithm.

This section displays the preprocessing techniques we use in our model.

Filtering: allows us to delete non-alphabetical data like (URL, symbols, emoticons, Html code…)

Tokenization: a technique that consists of converting the text into tokens before converting it into vectors—for example, a document into paragraphs or sentences into words. In our case, we tokenize reviews into words.

Stemming: is a widely used technique in information retrieval to avoid an inadequate lexicon, where the words in the query are not similar to any other words in a document. For instance, "I love this hotel" and "I looooooooooooooove this hotel" are the same, so to improve the accuracy, we transform "looooooooooooooove" into "love."

Stopwords removal: The stopwords in English are: the, is, at, which, on ... The presence of these words can yell bad results (Vicenç Parisi Baradad et al.,2015). We have removed conjunctions (for, and, nor, but, or, yet, so), determinants (a/an, the, this, that, these, those), and preposition (at, in, to).

3.4 Feature Extraction

Feature extraction identifies aspects and relevant attributes to feed to the machine learning algorithm to increase classification accuracy. We choose to use the term frequency-inverse document frequency (TF-IDF) method.

TF-IDF is a widely used algorithm for converting the text into a meaningful numerical representation that can be used to fit machine algorithms for prediction. This simple yet effective approach has proved extraordinarily robust and difficult to beat. The mathematical equations are as follow :

$$\text{TF IDF} = \text{TF}(t,d) * \text{IDF}(t) \qquad (1)$$

$$IDF(t) = log(\frac{1+n}{1+df(d,t)} + 1) \qquad (2)$$

Where t is the term, d the document (the review in our case), n the total number of documents, and df(d,t) document frequency of the term t.

4 RESULTS

This section exposes the results obtained using our three models; we also display the different functionalities we implemented in this proof of concept application.

4.1 Data Set

The data set comes from one of the most popular travel websites and includes hotel reviews sent by customers. This data collection provides ratings for a single hotel, and this data set is available on Kaggle. It contains 38933 comments and two attributes: description containing the comment and is_response representing the two classes (happy, not_happy).

The two classes' distributions are displayed in both Fig. 3 and 4.

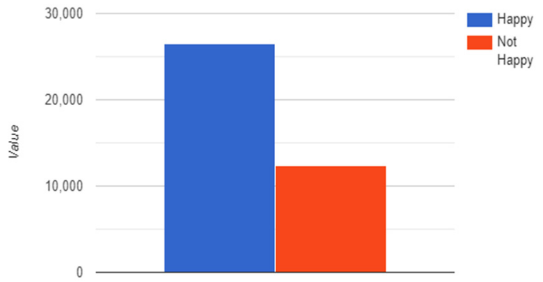


Figure 3: Histogram displaying the number of comments in each category.

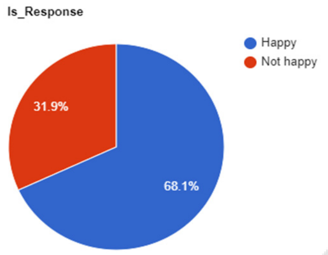


Figure 4: Pie chart displaying the percentage of each category.

4.2 Experiment

First, the data set is imported using the Panda framework. This framework is further used to preprocess the data as mentioned in section 3.3. Once the preprocessing is done, we use the sklearn library to vectorize the data set by applying the TF-IDF function, the vectorized data set is divided later on into a training set (80%) and a test set (20%), the training set was also divided into training (80%) and validation set (20%). Afterward, the training set is used to train the models (SVM, NB, LR) imported from the sklearn library.

4.3 Results and Discussion

We used the accuracy measure to compare and determine which one of the three algorithms performed better. The results were summarized in Fig.5

The SVM came on top with an accuracy of 89% on the test set; the Linear regression algorithm achieved almost the same results with an accuracy of 87%. The Naive Bayes achieved the lowest accuracy of all the classifiers with only 67% accuracy on the test set.

Our method as simple as it looks got the state of art in the tourist review sentiment analysis, (Muhammad Afzaal et al., 2019) reached the same performance (90% accuracy) using a more complex aspect-based method, same goes for (C. A. Martín et al., 2018) that compared a few deep learning approaches (CNN, LSTM…) the highest accuracy they reached was a little bit above 89%.

Based on this result, we choose to keep the SVM classifier in our application.

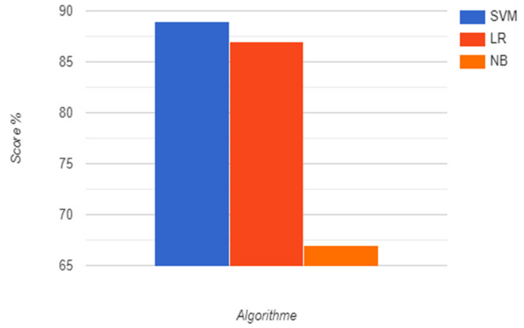


Figure 5: Comparison of accuracy between the three algorithms.

5 APPLICATION

In our proof of concept application, we choose to focus on the Draa-Tafilalet region. First, the user logs into the application; once he is logged in, the user can browse for hotels, review a visited hotel and visualize charts that summarize the other user reviews.

5.1 Use Case

The primary type of system/software specifications for a new software program is the UML use case. It contains the application actors and each one's actions on the system. In our application, we have only one user and three specific actions that require authentication Fig.6.

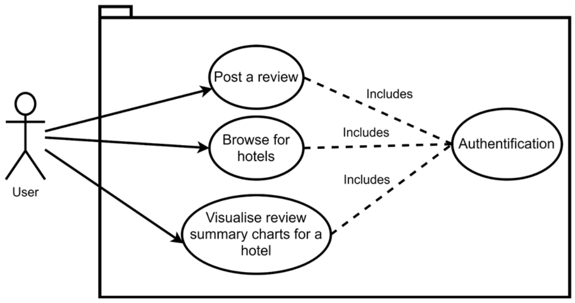
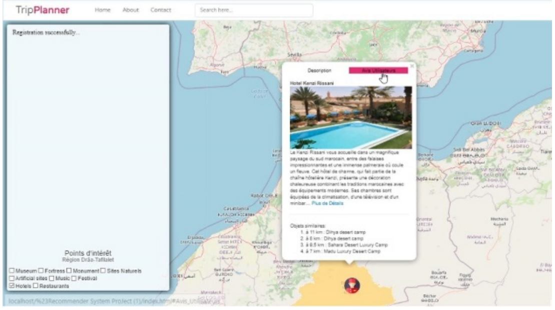


Figure 6: Application's Use case diagram.

5.2 Interfaces



Le Kenzi Rissani vous accueille dans un magnifique paysage du sud marocain, entre des falaises impressionnantes et une immense palmeraie où coule un fleuve. Cet hôtel de charme, qui fait partie de la chaîne hôtelière Kenzi, présente une décoration chaleureuse combinant les traditions marocaines avec des équipements modernes. Ses chambres sont équipées de la climatisation, d'une télévision et d'un minibar.... Plus de Détails

Objets similaires:
1. à 11 km : Dihya desert camp
2. à 8 km : Dihya desert camp
3. à 8.5 km : Sahara Desert Luxury Camp
4. à 7 km : Madu Luxury Desert Camp

localhost/%23Recommender System Project (1)/index.html#Avis_Utilisateurs

Figure 7: Hotel Browsing.

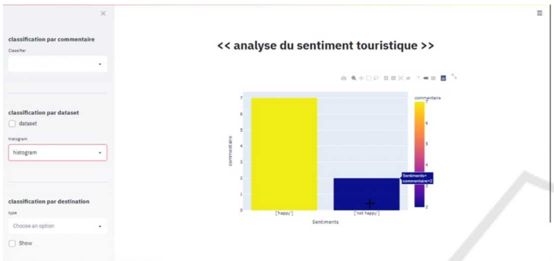


Figure 8: Histogram review analysis

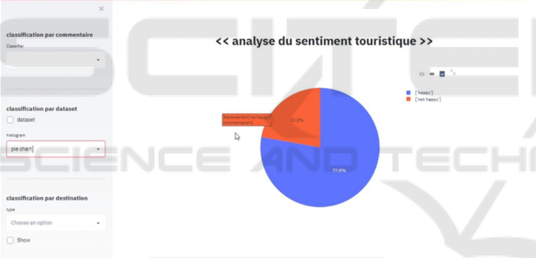


Figure 9: Percentage of each class.

6 CONCLUSION

This paper displayed the different steps we followed to code our proof of concept (POC) application, starting from preprocessing the data, choosing the model, then the training, finally the validation, and test phases. The algorithm with higher accuracy (SVM) was used as the core component in our POC. This application offered the user the possibility to browse for hotels in the Draa-Tafilalet region, give reviews and visualize charts summarizing the polarity of the sentiment in the reviews given by other users, which we considered as the main functionality of our application.

In future works, we will try to ameliorate the model used in our application and surpass the 89% accuracy; to do so, we will start by collecting our data, then we would use more complex preprocessing steps and features extraction techniques. We will explore the deep-learning approach that has shown a huge potential in the Sentiment Analysis field. In addition we'll try to explore the multilingual models to give tourists more freedom to speak their mind freely. For the application, we will try to give the user more statistics and charts to facilitate the decision-making process; extending our work would be to generalize the application for all tourism-related activities, like restaurants, museums, historical monuments, etc.

REFERENCES

Bo Pang and Lillian Lee, 2008 Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval, volume 2, page 1-135.*

Huertas Assumpció, 2018 How live videos and stories in social media influence tourist opinions and behaviour. In *Information Technology & Tourism volume 19, page 1-28.*

Narangajavana Kaosiri Y, Callarisa Fiol LJ, Moliner Tena MÁ, Rodríguez Artola RM, Sánchez García J., 2019 User-Generated Content Sources in Social Media: A New Approach to Explore Tourist Satisfaction. In *Journal of Travel Research.*

De Pelsmacker, Patrick, van Tilburg Sophie, Holthof Christian, 2018 Digital marketing strategies, online reviews and hotel performance. In *International Journal of Hospitality Management volume 72, page 47-55.*

Miller George A. and Beckwith, Richard and Fellbaum, Christiane and Gross, Derek and Miller, Katherine J., 1990 Introduction to WordNet: An Online Lexical Database*. In *International Journal of Lexicography volume 3, page 235-244.*

Qiu Guang, He Xiaofei, Zhang Feng, Shi Yuan, Bu Jiajun, Chen Chun, 2010 DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis. In *Expert Systems with Applications volume 37, page 6182-6191.*

Rajput Quratulain, Haider Sajjad, Ghani Sayeed, 2016 Lexicon-Based Sentiment Analysis of Teachers' Evaluation. In *Applied Computational Intelligence and Soft Computing volume 2016.*

Singh Nikhil Kumar, Tomar Deepak Singh, Sangaiah Arun Kumar, 2020 Sentiment analysis: a review and comparative analysis over social media. In *Journal of Ambient Intelligence and Humanized Computing volume 11, page 97-117.*

F. Luo, C. Li, Z. Cao , 2016 Affective-feature-based sentiment analysis using SVM classifier. In *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), page 276-281.*

Pennington Jeffrey and Socher Richard and Manning Christopher, 2014 GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference* *on Empirical Methods in Natural Language Processing (EMNLP), page 15320-1543.*

Behera Ranjan Kumar, Jena Monalisa, Rath Santanu Kumar, Misra Sanjay, 2021 Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. In *Information Processing & Management volume 58.*

Peng Zhou and Zhenyu Qi and Suncong Zheng and Jiaming Xu and Hongyun Bao and Bo Xu, 2016 Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. In *arXiv:1611.06639.*

K. AL Fararni, F. Nafis, B. Aghoutane, A. Yahyaouy, J. Riffi, A. Sabri. Hybrid Recommender System for Tourism Based on Big Data and AI: A Conceptual Framework. Big Data Mining and Analytics 2021, 4(1): 47-55.

Rao Guozheng, Huang Weihang, Feng Zhiyong, Cong Qiong, 2018 LSTM with sentence representations for document-level sentiment classification. In *Neurocomputing volume 308, page 49-57.*

Alam Saqib, Yao Nianmin, 2019 The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. In *Computational and Mathematical Organization Theory volume 25, page 319-335.*

Boulaalam, Oumayma, et al. "Proposal of a big data system based on the recommendation and profiling techniques for an intelligent management of moroccan tourism." *Procedia Computer Science 134 (2018): 346-351.*

Vicenç Parisi Baradad and Alexis-Michel Mugabushaka, 2015 Corpus Specific Stop Words to Improve the Textual Analysis in Scientometrics. In *ISSI.*

Martín, C. A., Torres, J. M, Aguilar, R. M.,Diaz, S., 2018 Using Deep Learning to Predict Sentiments: Case Study in Tourism. In *Complexity volume 2018.*

M. Afzaal, M. Usman and A. Fong, Tourism Mobile App With Aspect-Based Sentiment Classification Framework for Tourist Reviews. In *IEEE Transactions on Consumer Electronics, vol. 65, no. 2, pp. 233-242, May 2019, doi: 10.1109/TCE.2019.2908944.*