Understanding Customers' Opinion using Web Scraping and Natural Language Processing

Alin-Gabriel Văduva
Simona-Vasilica Oprea
Dragoș-Cătălin Barbu
The Bucharest University of Economic Studies,
Department of Economic Informatics and Cybernetics, Romania
vaduvaalin19@stud.ase.ro
simona.oprea@csie.ase.ro
dragos.barbu@ici.ro

Abstract

The web offers large volumes of data that is unstructured and fails to be further processed if not extracted and organized into local variables or into databases. In this paper, we aim to extract data from the Internet using web scraping and analyse it with Natural Language Processing (NLP). Our purpose is to understand customers' opinions by extracting reviews and investigating them in Python. The positive or negative insight of the reviews, along with the word cloud offer additional tools to understand the customers, predict their behaviour and underpin problems signalled in the reviews. TextBlob and BERTweet are applied to analyse the reviews. To enhance the comprehension of the outcomes, a comparison is drawn between the classifications generated by the BERTweet model and those provided by the TextBlob API, a widely used Python library for performing various NLP tasks. Furthermore, the reviews are pre-processed to clean them from line breaks, punctuation characters etc. and a n-grams analysis is performed to better understand the positive and negative reviews. The frequency of the reviews displays the concrete problems faced by customers visiting the hotel in various seasons. It helps decision makers to take measures and improve the quality of the hotel services.

Key words: web scraping, booking, customers opinions, natural language processing
J.E.L. classification: Z13, C55, C81

1. Introduction

Extracting and analysing data from web along with Google Analytics have become significant to understand customers' opinion and visualize metrics that reveal more insights about customers' requirements. A survey on web scraping techniques and tools were provided in (Saurkar and Gode, 2018) underlying the vast volume of data that has to be extracted and processed to obtain useful information. Simply copying data from the web is not always possible, especially when data is structured so that copying the entire data set is not possible. Or instead of repeating the extraction operation manually, a Python script can be scheduled to run at a certain hour using Node-RED or other techniques and collect the fresh data from web. Web scraping is described into three steps in (Niu et al., 2023): understanding the page source structure, creating regular expression to extract data and design the scraper tool for extracting the news from web, blogs and image data.

Various web scraping libraries do exist (Anon., 2020) in several programming language such as Python – Scrapy, BeautifulSoup, Selenium (Thomas and Mathur, 2019), R (Bradley and James, 2019), PHP, etc. Scraping data for academic purposes is frequently performed to create new data sets, extract information from data and bring more insights from web, especially from the human-computer interactions. The speed of the events is another factor that stimulates web scraping as most of the data repositories do not contain meaningful data and are therefore obsolete. Scraping is at least

equally important in marketing (Boegershausen et al., 2022), (Shankar and Parsana, 2022), banking, electronic markets, e-commerce, politics, elections, etc.

Usually, web data is unstructured and requires selection, filtering to allow further processing (Sirisuriya, 2015). The necessity for web scraping comes from different areas: social, security, academic, educational and financial. Therefore, web scraping transforms unstructured data into structured data that can be organized into a database such as SQLite or MongoDB. One can extract various time series and merge them to create relevant data sets using them furthermore for prediction. Natural Language Processing (NLP) gains more importance as it can analyse and process the human language including reviews and customers' opinion. The state of the art, recent trends and challenges are described in (Khurana et al., 2023), emphasizing on an interesting discussion, models and assessing metrics for NLP. NLP is also useful to analyse speeches and evaluate their impact on the economies, prices, stock markets, etc.

A better interaction and interface between humans and computers are ensured by NLP (Patel and Patel, 2021). It is used in spam detection and machine translation, learning from experience like humans. NLP and deep learning algorithms (such as: Convolutional Neural Networks - CNN, Recurrent Neural Networks - RNN, attention model for NLP) (Alshemali and Kalita, 2020) offer many applications of computational linguistics. The state of the art, current trends and ideas for future research are provided by (Otter, Medina and Kalita, 2021).

2. Literature review

Numerous scientific research papers approached web scraping and NLP, analysing opinions and reviews of the customers. (Arbane et al., 2023) applied bidirectional Long Short-Term Memory (LSTM) to create a social media-based COVID-19 sentiment classification model. NLP assisted various fields including management research (Kang et al., 2020) and software testing, providing a systematic mapping of the literature in (Garousi, Bauer and Felderer, 2020). Limited data learning in NLP can be a challenge. (Chen et al., 2023) investigated this challenge, proposing a review for data augmentation considering limited data learning in NLP, providing appropriate augmentations recommendations in different settings. They also discussed the challenges and future directions in this area. Online reviews are investigated (Biswas et al., 2022), offering a critical evaluation of customers' reviews using a hybrid NLP methodology. The effects of various reviews are investigated. Shannon's Entropy Theory, Dual Process Theory and text mining are used to analyse the helpfulness of reviews. The study provides interesting insights on electronic commerce using NLP and its implications.

Usually, customers choose a product or accommodation based on the reviews, but the number and diversity of the reviews can be overwhelming. Thus, the right choice cannot be a facile one for customers. Websites for e-commerce may use algorithms such as the Naïve Bayes classifier, Logistic Regression and SentiWordNet algorithm to classify reviews. COVID-19 pandemic increased the online shopping and interactions of customers with web platforms (Burlacioiu, 2023). Reviews for telecom and utilities services are investigated emphasizing NLP tools to assess the customers' opinions.

After the lockdowns were removed, many people have intensified their traveling activities. NLP has been applied to tourism research (Álvarez-Carmona et al., 2022) providing a systematic survey of 227 most relevant studies and future research directions in this field. Six major topics in applying NLP to tourism issues were identified: Sentiment analysis, Travel, Recommendation systems, Destination branding, Sentiment analysis for hotels, and Destination recommendation. The results showed that countries like China, the United States, Thailand and Spain have similar tourism challenges. The authors answered the following concerns: the NLP techniques applied in tourism industry, NLP algorithms applied for tourism issues, data requirements for NPL in tourism.

In booking, the impact of latent topic extracted from online purchasing reviews for the accommodation industry was analysed in (Sim, Lee and Sutherland, 2021). 400,000 reviews for accommodation in South Korea were investigated using Latent Dirichlet Allocation (LDA) in order to identify the topic of the review content. With CNN, the authors identified the valence of the reviews, whereas with the spatial probit models, they identified the impact of the review content and valence on booking intention. Positive and negative reviews on several variables such as the

ambiance, service, accessibility, surrounding neighbourhood and room space lead to higher/lower booking intentions. LDA and fuzzy C-Means clustering were applied in (Geethalakshmi et al., 2021) to extract the opinion from hotel reviews. An interesting evaluation of sentiment analysis on smart entertainment and devices reviews was provided in (Gamal et al., 2019). Scoring tourist attractions based on sentiment lexicon was furthermore analysed in (Ding et al., 2017).

3. Research methodology

The subsequent section proposes an approach for generating classifications of reviews obtained from the Booking.com website, a case study regarding hotel services located in Dubai.

Initially, the process entails leveraging the web-scraping library BeautifulSoup alongside the Python programming language to retrieve pertinent data from the Booking website. The procedure involves iteratively traversing 30 pages of the hotel's website section and parsing user-specific data, including the reviewer's name, the publication date of the review, the human-annotated rating, the user's positive and negative reviews and its general review based on its experience. Subsequently, the parsed data is structured and stored in a specialized data structure, Pandas DataFrame, in which the columns represent the identified information mentioned above.

Upon incorporating the data into the Pandas DataFrame, the next step involves passing the column comprising the user's general review to a sentiment analysis pipeline that utilizes the "transformers" Python library provided by Hugging Face.

Hugging Face is an AI-centered community which promotes the values of open-source contributions. Its primary aim revolves around the creation and refinement of the models belonging to Natural Language Processing, Computer Vision, and other relevant subfields of Artificial Intelligence. The models developed by the Hugging Face community are presently being employed in various practical applications, including Language Modelling, Sequence Classification, Sentiment Analysis, Question Answering, and several other domains (Wolf et al., 2019). In the present study, the Sentiment Analysis pipeline is being implemented to classify the users' general reviews, as outlined on the hotel's Booking page.

The default model used in the pipeline is the pre-trained bertweet-sentiment-analysis model, which has been finetuned on the SemEval 2017 dataset, comprising approximately 40.000 scraped tweets. This model was initially developed for Twitter sentiment analysis, with its base model being BERTweet - a RoBERTa model trained on English Tweets.

Upon processing the data through the Sentiment Analysis pipeline, the method generates a Python list of dictionaries, where each dictionary contains two primary elements: the assigned label (positive, negative, or neutral) for each general commentary, along with the corresponding classification score. To enhance the comprehension of the outcomes, a comparison is drawn between the classifications generated by the BERTweet model and those provided by the TextBlob API, a widely used Python library for performing various NLP tasks.

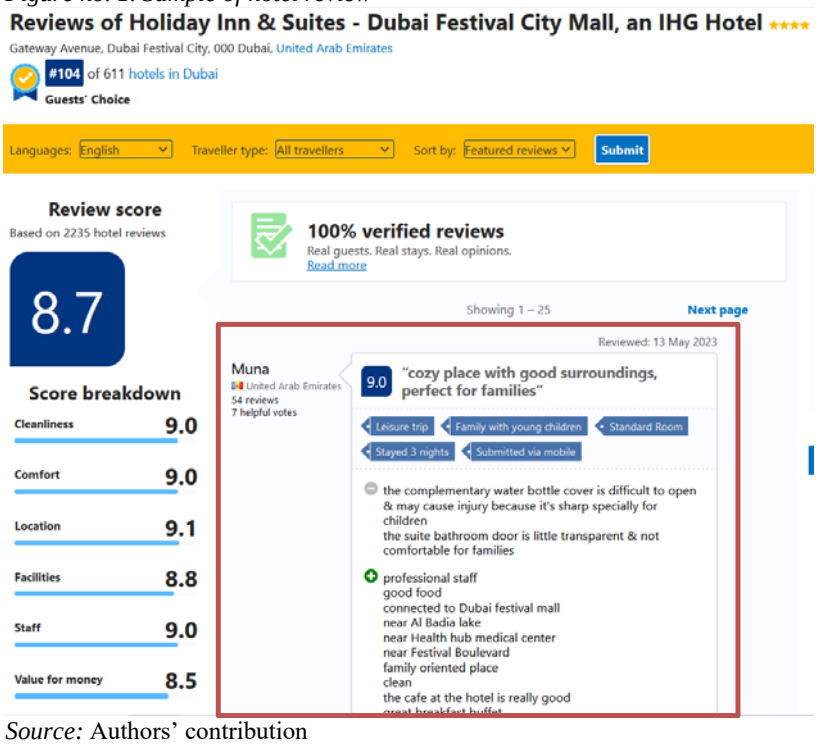
In the domain of sentiment analysis concerning text data, the key performance metrics are typically characterized by the semantic orientation and intensity of each word contained in a given phrase. These metrics include polarity, which ranges between -1 (indicating negative sentiment) and 1 (indicating positive sentiment), and subjectivity, which gauges the degree of personal opinion expressed within a sentence. In certain cases, polarity may also provide insights into neutral commentaries left by users (Montoyo, Martínez-Barco and Balahur, 2012). The TextBlob library provides both polarity and subjectivity indicators, which are particularly useful for simpler cases, such as general user reviews.

To know the general customers' opinion regarding hotel services is a useful goal, but equally important is to identify the issues and concerns that customers may have. Such issues are voluminous in terms of data and requires special NLP techniques such as n-grams analysis of text. Furthermore, to better understand the reviews of customers, a n-grams analysis is performed to identify the issues encounters during their stay. The n-gram analysis is usually performed after the text is cleared from line breaks, punctuation etc. and consists in extracting n consecutive words and counting their frequency. The script is run weekly to extract the sentiment and provide fresh and useful information to the hotel manager.

4. Findings

Following the extraction of data from the Booking hotel's website, 748 instances of user reviews were retrieved. 30 pages were investigated, one page may have 24-25 reviews. One sample of such reviews is displayed in Figure 1.

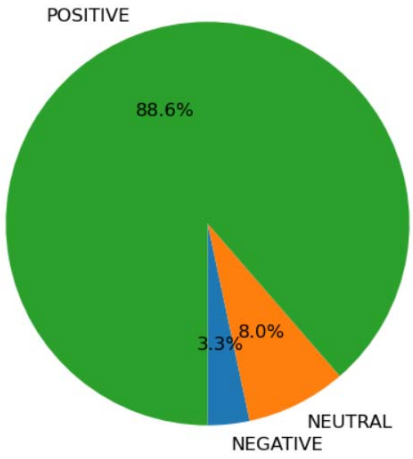
Figure no. 1. Sample of hotel review



Source: Authors' contribution

The distribution of sentiment analysis using BERTweet model revealed that a majority of the reviewers had a positive experience staying at the hotel, as depicted in Figure 2.

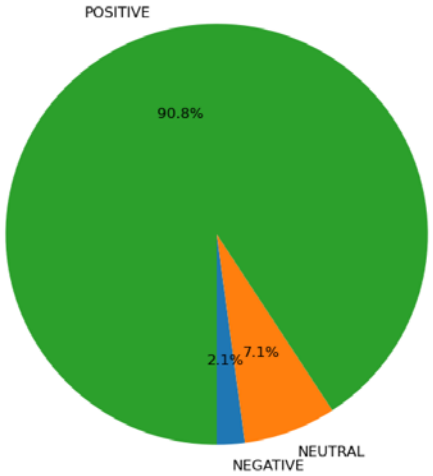
Figure no. 2. Distribution of user reviews - BERTweet model



Source: Authors' contribution

On the other hand, the TextBlob approach produced comparable results (as in Figure 3), categorizing a higher number of reviews as positive than the BERTweet model.

Figure no. 3. Distribution of user reviews - TextBlob



Source: Authors' contribution

The small variations between the TextBlob and BERTweet models originate from the methodology used to determine the polarity of sentences in the TextBlob approach. The polarity is computed as the weighted average sentiment score of all the words in the sentence. It is possible that the presence of certain words, such as "but" and "good", in the sentence can change the overall sentiment of the sentence and may lead to incorrect assumptions about the sentiment of the reviews, due to the way TextBlob calculates polarity as the weighted average sentiment score over all the words in the sentence.

57 instances exhibited differences in sentiment classification results between the TextBlob approach and the BERTweet model. The majority of these instances revealed TextBlob's inclination to classify the reviews as NEUTRAL, while BERTweet correctly assigned them as either POSITIVE or NEGATIVE. In order to illustrate this disparity, a scatter plot was generated and is presented in Figure 4. The BERTweet model's classification was considered to be the true label for each review, and the points were plotted using the Polarity and Subjectivity coordinates of the TextBlob approach.

Figure no. 4. Disparity of points in the Polarity and Subjectivity coordinates

Source: Authors' contribution

The vertical median line in the scatter plot partitions the two side-planes, with the negative reviews mostly falling on the left and the positive reviews on the right. Reviews that are classified as neutral are represented by the points that lie on the median line. The TextBlob approach's method of calculating polarity results in most of the misclassified points being represented as neutral (indicated by red dots) on both sides of the scatter plot.

Using the n-gram analysis, we investigate the content of the review and their frequency in order to identify the complaints that customers may have. For instance, using a 4-gram review analysis, most relevant frequencies are between 2 and 3. In Table 1, a sample of reviews and their frequencies are displayed. Therefore, the manager can see the concrete positive and negative opinions that appeared more than twice. In the first rows of the table, the manager notices some positive aspects – the fact that the hotel is well located from the airport and shopping centres, good quality of breakfast etc. While in the last rows of Table 1, the manager can notice some complaints regarding complementary water bottle and the bathroom door. Furthermore, he/she is able to take measures and correct the issues identified by means of n-gram analysis.

Table no. 1 Results of n-gram analysis in terms of reviews and their frequency

4-gram Review	**Frequency**
LOCATION CONVENIENT FOR SHOPPING	2
CONVENIENT FOR SHOPPING AT	2
FOR SHOPPING AT FESTIVAL	2
SHOPPING AT FESTIVAL CITY	2
AT FESTIVAL CITY MALL	2
FESTIVAL CITY MALL NOT	2
CITY MALL NOT FAR	2
MALL NOT FAR FROM	2
NOT FAR FROM THE	3
FAR FROM THE AIRPORT	3
OVERALL VERY GOOD.STAYED IN	2
VERY GOOD.STAYED IN MAY	2
GOOD.STAYED IN MAY NOTHING...EVERYTHING...STAYED	2
IN MAY NOTHING...EVERYTHING...STAYED IN	2
MAY NOTHING...EVERYTHING...STAYED IN MAY	2
NOTHING...EVERYTHING...STAYED IN MAY GREAT	2
IN MAY GREAT VALUE	2
MAY GREAT VALUE FOR	2
GREAT VALUE FOR MONEY	3
VALUE FOR MONEY VERY	2
FOR MONEY VERY GOOD	2
MONEY VERY GOOD BREAKFASTSTAYED	2
VERY GOOD BREAKFASTSTAYED IN	3
GOOD BREAKFASTSTAYED IN FEBRUARY	3
THE COMPLEMENTARY WATER BOTTLE	2
COMPLEMENTARY WATER BOTTLE COVER	2
WATER BOTTLE COVER IS	2
BOTTLE COVER IS DIFFICULT	2
COVER IS DIFFICULT TO	2
IS DIFFICULT TO OPEN	2
DIFFICULT TO OPEN MAY	2
TO OPEN MAY CAUSE	2
OPEN MAY CAUSE INJURY	2
MAY CAUSE INJURY BECAUSE	2
CAUSE INJURY BECAUSE ITS	2
INJURY BECAUSE ITS SHARP	2
BECAUSE ITS SHARP SPECIALLY	2
ITS SHARP SPECIALLY FOR	2
SHARP SPECIALLY FOR CHILDREN	2
SPECIALLY FOR CHILDREN THE	2

FOR CHILDREN THE SUITE	2
CHILDREN THE SUITE BATHROOM	2
THE SUITE BATHROOM DOOR	2
SUITE BATHROOM DOOR IS	2
BATHROOM DOOR IS LITTLE	2
DOOR IS LITTLE TRANSPARENT	2
IS LITTLE TRANSPARENT NOT	2
LITTLE TRANSPARENT NOT COMFORTABLE	2
TRANSPARENT NOT COMFORTABLE FOR	2

Source: Authors' contribution

Several combinations of 2 to 7 words were extracted, but the 4 consecutive words revealed the customers' opinion.

5. Conclusions

In this paper, we analysed the text generated by customers of hotel services. It can be applied to other economic activities. Two NLP analysis were performed. First, we classified and analysed the positive and negative reviews. Two approaches were included: BERTweet and TextBlob. Additionally, we cleared the text from line breaks and punctuation and applied n-gram analysis to investigate the content of the reviews and identify the main complaints.

Most of the reviews traversing 30-page website were around 90% positive. 57 instances exhibited differences in sentiment classification results between the TextBlob approach and the BERTweet model. The points were analysed using the Polarity and Subjectivity coordinates of the TextBlob approach. Moreover, applying 4-gram analysis, we identified short combinations of 4 consecutive words that reflected the customers' opinion providing knowledge to correct the problems.

6. Acknowledgement

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project number PN-III-P4-PCE-2021-0334, within PNCDI III.

7. References

- Alshemali, B. and Kalita, J., 2020. Improving the Reliability of Deep Neural Networks in NLP: A Review. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2019.105210.
- Álvarez-Carmona, M., Aranda, R., Rodríguez-Gonzalez, A.Y., Fajardo-Delgado, D., Sánchez, M.G., Pérez-Espinosa, H., Martínez-Miranda, J., Guerrero-Rodríguez, R., Bustio-Martínez, L. and Díaz-Pacheco, Á., 2022. Natural language processing applied to tourism research: A systematic review and future research directions. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.10.010.
- Anon. 2020. Web Scraping: Applications and Scraping Tools. *International Journal of Advanced Trends in Computer Science and Engineering*. https://doi.org/10.30534/ijatcse/2020/185952020.
- Arbane, M., Benlamri, R., Brik, Y. and Alahmar, A.D., 2023. Social media-based COVID-19 sentiment classification model using Bi-LSTM. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2022.118710.
- Biswas, B., Sengupta, P., Kumar, A., Delen, D. and Gupta, S., 2022. A critical assessment of consumer reviews: A hybrid NLP-based methodology. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2022.113799.
- Boegershausen, J., Datta, H., Borah, A. and Stephen, A.T., 2022. Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing*. https://doi.org/10.1177/00222429221100750.
- Bradley, A. and James, R.J.E., 2019. Web Scraping Using R. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245919859535.
- Burlacioiu, C., 2023. Online Commerce Pattern in European Union Countries between 2019 and 2020. *Societies*. https://doi.org/10.3390/soc13010004.

- Chen, J., Tam, D., Raffel, C., Bansal, M. and Yang, D., 2023. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00542.
- Ding, Y., Li, B., Zhao, Y. and Cheng, C., 2017. Scoring tourist attractions based on sentiment lexicon. In: *Proceedings of 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2017*. https://doi.org/10.1109/IAEAC.2017.8054363.
- Gamal, D., Alfonse, M., Elhorbaty, E.M. and Salem, A.-B.M., 2019. An Evaluation of Sentiment Analysis on Smart Entertainment and Devices Reviews. *Information Theories and Applications*.
- Garousi, V., Bauer, S. and Felderer, M., 2020. *NLP-assisted software testing: A systematic mapping of the literature*. *Information and Software Technology*. https://doi.org/10.1016/j.infsof.2020.106321.
- Geethalakshmi, S.N., Shaambavi, S., Enterpreneur, B., Grey, M. and Com, S., 2021. *Opinion Mining With Hotel Review using Latent Dirichlet Allocation-Fuzzy C-Means Clustering (LDA-FCM)*. *Turkish Journal of Computer and Mathematics Education*.
- Kang, Y., Cai, Z., Tan, C.W., Huang, Q. and Liu, H., 2020. *Natural language processing (NLP) in management research: A literature review*. *Journal of Management Analytics*. https://doi.org/10.1080/23270012.2020.1756939.
- Khurana, D., Koli, A., Khatter, K. and Singh, S., 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13428-4.
- Montoyo, A., Martínez-Barco, P. and Balahur, A., 2012. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. In: *Decision Support Systems*. https://doi.org/10.1016/j.dss.2012.05.022.
- Niu, Q., Kandhro, I.A., Kumar, A., Shah, S., Hasan, M., Ahmed, H.M. and Liang, F., 2023. Web Scraping Tool For Newspapers And Images Data Using Jsonify. *Journal of Applied Science and Engineering (Taiwan)*. https://doi.org/10.6180/jase.202304_26(4).0002.
- Otter, D.W., Medina, J.R. and Kalita, J.K., 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2020.2979670.
- Patel, R. and Patel, S., 2021. Deep Learning for Natural Language Processing. In: *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-981-16-0882-7_45.
- Saurkar, A. V and Gode, S.A., 2018. An Overview On Web Scraping Techniques And Tools. *International Journal on Future Revolution in Computer Science & Communication Engineering*.
- Shankar, V. and Parsana, S., 2022. An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-022-00840-3.
- Sim, Y., Lee, S.K. and Sutherland, I., 2021. The impact of latent topic valence of online reviews on purchase intention for the accommodation industry. *Tourism Management Perspectives*. https://doi.org/10.1016/j.tmp.2021.100903.
- Sirisuriya, S., 2015. A Comparative Study on Web Scraping. *8th International Research Conference, KDU*.
- Thomas, D.M. and Mathur, S., 2019. Data Analysis by Web Scraping using Python. In: *Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019*. https://doi.org/10.1109/ICECA.2019.8822022.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J. and ..., 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv …*.