Automatic Aspects Extraction in Spanish Reviews Applying Transformers-Model Ensemble

Miguel Ángel Rivero Tapia [1[0009-0003-8289-6567]], Alfredo Simón-Cuevas[1[0000-0002-6776-9434]] and Juan Carlos Calabria-Sarmiento [2[0000-0001-9160-1466]]

[1] Universidad Tecnológica de La Habana "José Antonio Echeverría", CUJAE, Ave. 114, e/Rotonda y Ciclovía, La Habana, Cuba
asimon@ceis.cujae.edu.cu
[2] Universidad Simón Bolívar, Cra. 59 #59-65, Barranquilla, Colombia
Juan.calabria@unisimon.edu.co

Abstract. Aspect-based sentiment analysis is a task aimed at understanding the sentiment expressed in opinion texts or reviews about specific features of an entity, and is currently playing a key role in a variety of scenarios. The automatic extraction of aspects (characteristics) is the most challenging task, as it requires the ability to understand the context and to recognize the relevant and characteristic elements of an entity about which you have an opinion. To increase the quality of the solution to this problem is still a challenge in Spanish, because very few papers have been reported and the reported efficacy rates need to be improved. This paper, an aspect extraction method in which several language models based on Transformer are combined through an ensemble approach is presented. The proposed solution was evaluated using the SemEval2016 review dataset and the results obtained were compared to those reported by other state-of-the-art solutions. The evaluation process developed not only provides a starting point to have a broader perception of the performance of the Transformers models in the solution of this problem, but also highlights the improvement of the quality of the results with the combination of models through ensemble techniques.

Keywords: Aspect-based sentiment analysis, aspect extraction, Transformers models, ensemble models.

1 Introduction

The large volume of textual information in user-generated opinions or reviews on the web (e.g., social networks, blogs, chats, news and e-commerce platforms, etc.) has become a key focus of attention for business, governmental and social organizations. The automatic processing of this information can reveal how users and citizens feel about the products and services provided, public policies, facts, among others, resulting in very valuable knowledge for decision making in these scenarios.

Sentiment analysis (or opinion mining, as it is also known) is the field aimed at the computational study of people's opinions, feelings, ratings, attitudes and emotions towards entities such as products, services, organizations, individuals, problems, events, issues and their characteristics or attributes [1]. Opinions refer to a person's point of view towards a certain issue, while sentiment or polarity refers to the emotion experienced by the person towards that issue in the text, which is usually identified as positive or negative. Aspect-based sentiment mining (ABSA) is one of the categories or lines of work within sentiment analysis, where computational processing is taken to the lowest level of granularity, where entities, aspects and relationships present in opinions are identified and classified separately [2]. Monitoring and analyzing the opinions of users, customers or citizens in general with this approach is essential from a business and governmental perspective, in the current scenario where the content on social networks has such an impact on the decisions of citizens.

ABSA solutions are divided into two fundamental tasks: aspect extraction and aspect-based sentiment classification. The first part aims to identify and extract the aspect about which an opinion is being issued and the second part is responsible for detecting the sentiment polarity associated with that aspect in the opinion [3]. The automatic extraction of aspects is the most challenging task, as it requires the ability to understand the context and recognize the relevant and characteristic elements of an entity about which an opinion or judgment is expressed. Although the use of ABSA solutions has grown in recent years, driven by the need to understand user feedback more deeply, the basic developments and evaluations have been aimed at English text processing, with very few papers reporting results on Spanish language feedback. On the other hand, supervised approaches are the most widely used in the solution of this problem, and specifically those that apply deep neural networks and Transformers models are the ones that have gained the most recognition, due to the good results they have reported in several natural language processing tasks [4][5]. However, this trend is not reflected in the solutions reported for automatic aspect extraction in Spanish opinions, which limits the necessary improvement in the effectiveness of the results. The main objective of this work focuses on improving the quality of the results in this particular task within the ABSA.

This paper presents an aspect extraction method in which several Transformer-based language models are combined through an ensemble approach. As part of the research, 10 models of Transformers were studied to find out their performance in the solution of this problem and to obtain criteria to propose a strategy for combining models through the ensemble technique. In the proposed solution, the Voting Ensemble technique is applied, by majority vote, to combine 5 models (BETO + BERT + ALBERT-base + ALBERT-large + ALBERT-xx-large), given that it was the alternative that achieved the best results. This solution was evaluated using the SemEval2016 reviews dataset, which is highly regarded in the evaluation of this task and one of the few existing datasets with reviews in Spanish. The results obtained were compared with those reported by other state-of-the-art solutions, reflecting the contribution of the proposed solution to increase the effectiveness of the results. The evaluation process developed, not only constitutes a good starting point to have a broader perception of the performances of the Transformers models in the solution of this problem, but also shows the improvement of the quality of the results with the combination of models through ensemble techniques.

The remainder of the paper is structured as follows: Section 2 presents a synthesis of fundamentals and characteristic aspects of the related work; Section 3 describes the proposed solution; Section 4 presents and analyses the results of the evaluation process; and the conclusions reached and elements to be addressed in the future are presented in Section 5.

2 Related Works

ABSA plays a crucial role in applications such as customer feedback analysis, market research and product and service improvement, and this new challenge has attracted increasing interest from researchers around the world [4]. This kind of solution stands as a sophisticated tool in the field of sentiment analysis, offering a deeper and more granular understanding of the opinion expressed in a text. In contrast to the simple classification of positive or negative sentiment, the ABSA focuses on to analyzing the polarity of the sentiment towards specific characteristics, or aspects, of a product, service, persons or event. In this context, an 'aspect' is defined as a relevant attribute that contributes to the evaluation of an object.

ABSA solutions are divided into two fundamental tasks: aspect extraction and aspect-based sentiment classification [3][5]. The first is to identify the specific characteristics or attributes, called 'aspects', that are mentioned in a text, for example, in a restaurant review, the aspects could be the quality of the food, the service, the decoration, the price, among others. The second is that, once the aspects have been identified, the polarity of the sentiment expressed towards each of them must be determined. In other words, the objective is to determine whether the opinion on the aspect is positive, negative or neutral. The automatic extraction of aspects is the most challenging task, as it requires the ability to understand the context and identify the relevant and characteristic elements on which an opinion of an entity is expressed. This research focuses precisely on this problem.

The application of ABSA solutions has boomed in recent years, driven by the need to understand users' views in greater depth, leading to the development of a broad diversity of problem-solving approaches. Supervised approaches are the most widely used in the solution of this problem. These include traditional methods such as Support Vector Machine (SVM) and Naive Bayes, where they are used for both polarity detection and aspect extraction, as well as approaches based on deep neural networks and transformer models, which have become more popular in recent years [4][5]. Although there are several solutions reported in the literature, there are still many shortcomings in this field, including the absence of solutions targeted and evaluated in languages other than English. Most of the existing works are designed and evaluated for the English language, so there is an absence of solutions reported and applied to texts in other languages such as Spanish [**Error! Reference source not found.**][9][10]. In this sense, it is of great relevance to carry out studies for the conception of solutions and their evaluation in the field of reviews in Spanish; a solution with good results in English would not necessarily also achieve good results in Spanish. Resolving these linguistic limitations is crucial to expanding the scope and effectiveness of aspect-based sentiment analysis in various linguistic contexts [7]. On the other hand, existing systems are not completely effective in extracting aspects of Spanish-language opinions.

In [9] an unsupervised ontology-based model for explicit and implicit aspect detection is proposed, and a similar approach is also used to classify aspect polarity. The unsupervised approach facilitates the scalability of the system to other languages and domains, because it does not require training processes, but the quality of the results is not high, since the reported F1 results are 0,73. In [**Error! Reference source not found.**] the use of a pre-trained language model is proposed, specifically the BETO model (a model based on BERT, but pre-trained specifically for Spanish language tasks) [11], and report 0,767 of F1-score on the Semeval 2016 dataset. In [10] also considered the use of deep learning models, specifically combining two Convolutional Neural Networks (CNNs) architectures for aspect detection. However, one of the problems with the use of CNNs is that they do not take into account the order of the words in the sentence, so they are not fully effective in capturing the meaning of the words in context [13]. The results they reported using SemEval 2016 are 0,654 as F1-score.

3 Aspect Extraction Using Transformers-Model Ensemble

The proposed method aims to extract the most relevant aspects from user opinion or review texts. The method is comprised of two main phases: (1) Preprocessing and (2) Aspect extraction, where the latter is carried out by applying of an ensemble approach of Transformers models, as shown in Fig. 1.

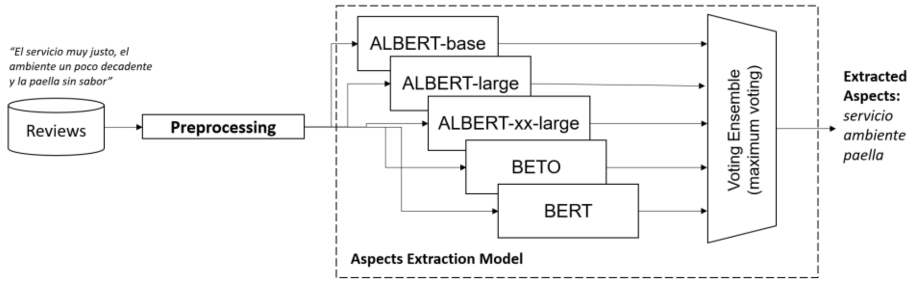


Fig. 1. Pipeline of the Transformers ensemble-based aspect extraction

3.1 Preprocessing

Aspect-based sentiment analysis requires specific pre-processing to prepare the text data and facilitate the extraction of relevant information in the opinions. While preprocessing tasks vary depending on the objective, in the case of Transformer model training, certain steps are crucial and indispensable, while others can be omitted. The Transformers models have been trained with millions of text that preserves information such as capital letters and punctuation marks. Therefore, it is not necessary to convert text to lower case or to remove punctuation marks, as the model has learned to use these features to generate high quality representations. Similarly, the removal of stop words (empty words such as articles and prepositions) is not necessary, since the training corpus already includes them and their tokenizer handles them without problems.

The first step in the pre-processing is to remove special characters, as these are not understood by the models, then for the model training phase it was necessary to bring the text into a specific format that would be understood by the language model. An example of the representation for some reviews is shown in Fig. 2. For example, for the first review: *"La comida estuvo sabrosa"*, the relevant aspect is *"comida"*, therefore, in the target column a 1 is assigned to the word position that constitutes an aspect to be predicted and 0 to the rest of the positions (tokens) that are not considered aspects.

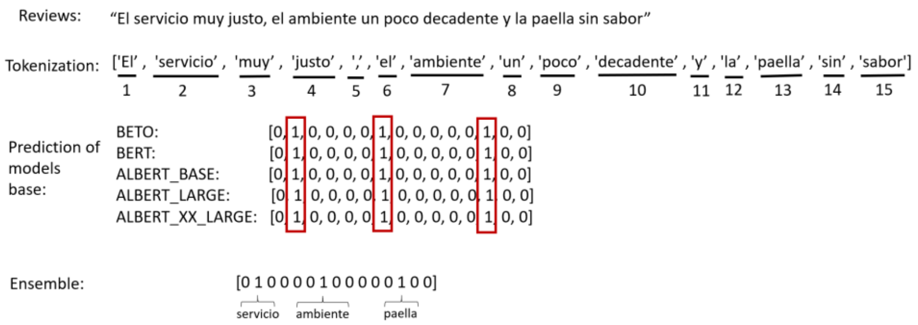
```
text_tokens;tags;polarities
['La', 'comida', 'estuvo', 'sabrosa', '.'];[0, 1, 0, 0, 0];[-1, 1, -1, -1, -1, -1]
['La', 'Carne', ',', 'fenomenal'];[0, 1, 0, 0];[-1, 1, -1, -1]
['Mal', 'Servicio'];[0, 1];[-1, 0]
```

Fig. 2. Example of the result of preprocessing a review.

3.2 Aspect extraction

Aspect extraction was addressed by applying an ensemble technique, which refers to a method in which multiple base models are integrated into the same framework to obtain a more robust classification model. The effectiveness of an ensemble method depends on several factors, including how the base models are trained and how their predictions are combined [14]. There are several methods for applying the ensemble technique, among the most commonly used are: maximum voting method, average voting and weighted average voting, meta-learning methods and bagging methods. In this paper we propose an approach based on the maximum voting method, specifically combining the base models BETO + BERT + ALBERT-base + ALBERT-large + ALBERT-xx-large. The voting-based assembly approach has the advantage of being easy to understand and its implementation does not entail a significant increase in computational cost [15].

The maximum voting ensemble model is composed of three main stages: training of base models, generation of predictions and voting of predictions. In the first step, the base models are trained separately, for example, in the case of this research, using the dataset of SemEval 2016 reviews. Previously, a fine-tuning process of each pre-trained model was carried out, from which the model was adapted to solve the aspect extraction problem. As part of this fine-tuning the output classes were defined, which in this case would be Aspect / Non-Aspect, were set, and through the binary coding 0 or 1 it is specified whether it is aspect or not. Once the base models are trained, they are used to generate aspect predictions on a test dataset, e. g. restaurant reviews, as in the case of SemEval 2016. For an opinion, each of the base models has as output a vector with a dimension corresponding to the number of tokens, where the binary encoding represents which token is an aspect and which is not. The predictions of each base model are combined using the majority voting ensemble approach, resulting in a new vector representing the majority outcome for each of the tokens. Fig. 1 illustrates an example of the aspects extracted from the review: *“El servicio muy justo, el ambiente un poco decadente y la paella sin sabor”*, which is detailed as follows. In this example, the majority vote of the models can be seen when predicting as an Aspect class the tokens: *“servicio”*, *“ambiente”*, and *“paella”*.



4 Evaluation and Discussion

The experimental evaluation of the solution was carried out about the collection of reviews in Spanish that was offered at SemEval-2016 for Task 5: Aspect Based Sentiment Analysis [12]. This collection is made up of restaurant reviews, labeled with the aspects that should be identified in an automatic process (for example: service, food, atmosphere, attention, among others). It is important to note that this collection is one of the few available in Spanish to evaluate this type of solution, and it is also the most widely used. The results were computed using the Precision, Recall and F1-score metrics. Table 1 shows a characterization of the test collection used.

Table 1. Characterization of the Test Collection.

Dataset SemEval 2016 Spanish ABSA	Reviews	Test	Train	Aspects
4^{th}-level heading	2697	627	2070	247

The experimental evaluation was carried out on a total of 10 pre-trained Transformers models: BETO (BERT-based model, but pre-trained specifically for Spanish-language tasks), BERT multilingual-cased, ELECTRA two versions (Spanish version), BERTIN (and its variants), GPT-2-Spanish, and AlBERT (in three different sizes). Unlike BETO, specific for the Spanish language, BERT multilingual was prepared in different languages, including Spanish, Chinese, Arabic, Hindi, Portuguese, French, among others. A fine-tuning process was applied to each of the models, to adapt them as a solution to the problem of extracting aspects from reviews or opinions. Besides, each model was trained with the Dataset reviews during 5 epochs and a batch size or batch of 8. The choice of batch size was motivated by several factors: large batch sizes can lead to fluctuations in gradient updates; Larger batch sizes can lead to faster convergence, but there is a risk that the model will be overly tight to the training data and not generalize well into new data. Table 2 shows the configuration of the different hyperparameters defined for the execution of the experiments with each of the models.

Table 2. Hyperparameters configuration.

Hyperparameters	Values
Seed	50
Learning rate	$1e^{-5}$
Epochs	5
Batch size	8
Maximum word sequence size	512

Two evaluation tasks were carried out:

1. To evaluating each of the models independently, in order to have a perception of the quality of the results of each model in the extraction of aspects;
2. To evaluating different combinations of models under an ensemble approach, considering the individual results obtained in the formation of the combinations.

The results obtained from the first task are shown in Table 3, highlighting in bold the highest values of each of the metrics. As can be seen in this table, the ALBERT-large and ALBERT-xx-large models obtain better results based on the F1 value. The F1-Score is a metric that balances Precision and Recall, so it's useful for measuring the overall performance of models. The F1-Score is a metric that balances precision and Recall, so it's useful for measuring the overall performance of models. Therefore, ALBERT-large and ALBERT-xx-large are the strongest in terms of balance between Precision and Recall, which makes them the best and most effective models. However, in terms of Precision, the following models stand out with the highest values: BETO, BERT, ALBERT-large. Some models such as BERTIN-large and ELECTRA-base show consistent results across all metrics, given that both have Precision and Recall around 0.83 and 0.81, making them reliable, but they don't stand out as much as ALBERT's models. GPT-2 is the worst performer, particularly in Recall, suggesting that it struggles to correctly identify the actual positives.

The combination of the models was carried out through the Voting Ensemble technique, by maximum vote, and the results of the different variants evaluated are shown in Table 4.

Considering the same analysis guidelines applied, the combination of BETO + BERT + ALBERT_base + ALBERT_large + ALBERT_xx_large by means of a Voting Ensemble is the best performer overall, given that it reached the highest F1 value (0.874). In terms of Precision, the assembly formed by BETO + BERT +

ALBERT_large + BERTIN_large + ELECTRA_base, with 0.869 Precision; A very similar value is achieved by variant 7, but adding another model in the assembly. When looking at the models that appear in the combinations, certain patterns emerge, for example: ALBERT_large and ALBERT_xx_large are present in almost all high-performance combinations; Bert was included in several combinations, and its inclusion seems to aid in Precision; Beto is consistent across all combinations, indicating that he may be providing a solid foundation in terms of overall performance, although he doesn't stand out on his own; and in the case of BERTIN_base and ELECTRA_base they were included in several combinations but do not seem to raise performance significantly, since although their individual performances are not the most remarkable, they do not complement well with other models either.

Table 3. Results of Transformers models in aspect extraction.

Id.	Models	Precision	Recall	F1
1	BETO	**0,85**	0,85	0,85
2	BERT	**0,85**	0,83	0,84
3	ALBERT-base	0,83	0,86	0,84
4	ALBERT-large	**0,85**	0,87	**0,86**
5	ALBERT-xx-large	0,83	**0,90**	**0,86**
6	BERTIN-base	0,78	0,83	0,80
7	BERTIN-large	0,83	0,82	0,82
8	GPT-2	0,78	0,53	0,54
9	ELECTRA-small	0,81	0,78	0,80
10	ELECTRA-base	0,83	0,81	0,82

Table 4. Results of combining Transformers models with an ensemble approach.

Id.	Models	Precision	Recall	F1
1	BETO + ALBERT_base + BERTIN_base	0,829	0,847	0,838
2	BERT + ALBERT_large + BERTIN_large	0,860	0,844	0,852
3	BETO + ALBERT_large + ALBERT_xx_large	0,848	0,888	0,865
4	**BETO + BERT + ALBERT_base + ALBERT_large + ALBERT_xx_large**	0,864	**0,895**	**0,874**
5	BETO + BERT + ALBERT_base + BERTIN_base + ELECTRA_base	0,850	0,838	0,844
6	BETO + BERT + Albert_large + BERTIN_large + ELECTRA_base	**0,869**	0,842	0,855
7	BETO + BERT + ALBERT _base + ALBERT_large + ALBERT_xx_large + BERTIN_large + ELECTRA_base	0,866	0,865	0,866
8	BETO + BERT + ALBERT_xx_large	0.856	0.864	0.860
9	BERT + BETO + ALBERT-large	**0.869**	0.857	0.863

10	BETO + BERT + ALBERT_large + ALBERT_xx_large	0,839	0,894	0,864

These results show an interesting relationship between the number of models combined and performance. On the one hand, more models do not always guarantee the best performance, for example: For example, the combination of 7 models (BETO + BERT + ALBERT_base + ALBERT_large + ALBERT_xx_large + BERTIN_large + ELECTRA_base) achieves an F1 of 0.866, while a combination of only 5 models (BETO + BERT + ALBERT_base + ALBERT_large + ALBERT_xx_large) slightly exceeds that value, indicating that although increasing the number of models might help, The selection of more complementary models is more important. On the other hand, small combinations can also be effective, given that the combination of 3 models BETO + ALBERT_large + ALBERT_xx_large achieves an F1 (0.865) comparable to larger combinations. This suggests that some models have features robust enough to maintain performance, without the need for additional add-ons.

Table 5 shows the comparison of the results of the proposed solution with those obtained by other solutions reported in the literature, which were evaluated using the same dataset. All of these solutions report their results only in terms of F1. As can be seen, the use of the Transformers model assembly approach outperforms the rest of the solutions in the extraction of aspects in Spanish-language opinions. Besides, there is an increase in the value of F1 by approximately 8%, compared to the solutions that participated in SemEval 2016.

Table 5. Comparison of the proposed solution with the state of the art.

Id.	**Solutions**	F1-Score
1	GTI/C*	0,685
2	GTI/U*	0,683
3	IIT-T/U*	0,643
4	TGB/C*	0,557
5	*SemEval-2016 – baseline* [12]	0,519
6	Enriquez Miranda & Buelvas [7]	0,730
7	Montanez [**Error! Reference source not found.**]	0,767
8	[9]	0,731
9	[10]	0,654
10	**Voting Ensemble: BETO + BERT + ALBERT_base + ALBERT_large + ALBERT_xx_large**	**0,874**

5 Conclusion and Future Works

In this work, the capabilities of Transformers-based language models as an alternative solution for the task of automatic extraction of aspects in opinions or reviews were explored. An experimental evaluation framework was carried out using the

SemEval2016 dataset in Spanish on 10 Transformers models, which allowed to obtain a new method of extraction of aspects with very promising results. In this method, five Transformers models are combined, specifically BETO, BERT, ALBERT_base, ALBERT_large, ALBERT_xx_large, through a Voting Ensemble, with maximum voting. This solution approach allows us to take advantage of the potentialities of Transformers models in the extraction of aspects, and constitutes a contribution with respect to other solutions reported in the literature for the extraction of aspects of opinions in Spanish. The evaluations carried out and the results obtained showed that the application of the model assembly, as well as the combination of the BETO + BERT + ALBERT_base + ALBERT_large + ALBERT_xx_large models, allowed to reach improvements in the efficiency in the extraction of aspects, surpassing what was reported in the state of the art. In the future, other voting-based assembly techniques will be evaluated, as well as Bagging and Stacking techniques, also extending the evaluation to other datasets. Analyses that are more detailed will also be made about the qualities of each model in order to extend the criteria for the selection of the models to be combined in the assembly scheme, since this is the key to the construction of the assembly models.

References

1. Liu, B., Zhang, L.: A Survey of Opinion Mining and Sentiment Analysis. In Mining Text Data, C. C. Aggarwal and C. Zhai (Eds). Springer US: Boston, MA. 415-463 (2012).
2. Nazir, A., Rao, Y., Wu, L., Sun, L.: Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. IEEE Transactions on Affective Computing 13(2), 845-863 (2022).
3. Geetha, M. P., Karthika R. D.: Improving the performance of aspect-based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks 2, 64-69 (2021).
4. Liu, N., Shen, B., Zhang, Z., Zhang, Z., Mi, K.: Attention-based sentiment reasoner for aspect-based sentiment analysis. Human-centric Computing and Information Sciences 9, 1–17 (2019).
5. López, D., Arcos, L.: Deep learning for aspect extraction in textual opinions. Revista Cubana de Ciencias Informáticas 13(2), 105-145 (2019).
6. Karimi, A., Rossi, L., Prati, A.: Improving BERT performance for aspect-based sentiment analysis. In: Proceedings of the 4th InternationalConference on Natural Language and Speech Processing (ICNLSP 2021), 39–46 (2021).
7. Henriquez, C., Buelvas, E.: AspectSA: Unsupervised system for aspect-based sentiment analysis in Spanish, Prospectiva 17, 87-95 (2019).
8. Montañez, P., Simón, A., Olivas, J. A., Romero, F. P.: Enhancing Spanish Aspect-Based Sentiment Analysis Through Deep Learning Approach. Y. Hernández Heredia et al. (Eds.): IWAIPR 2023, LNCS 14335, 1–10 (2024).
9. HENRÍQUEZ MIRANDA, Carlos; BRICEÑO, F; SALCEDO, Dixon. "Unsupervised model for aspect-based sentiment analysis in Spanish," IAENG International Journal of Computer Science, vol. 46, no. 3, pp. 430-438, 2019
10. Martínez-Seis, Bella, C., Pichardo-Lagunas, O., Miranda, S., Pérez-Cazares, Y., Cazares, I. J., Rodríguez-Gonzalez, J. A.: Deep learning approach for aspect-based sentiment analysis of restaurants reviews in Spanish, Computación y Sistemas, 26(2), 899-908 (2022).

11. Pathan, F., Prakash, Ch.: Cross-domain aspect detection and categorization using machine learning for aspect-based opinion mining. International Journal of Information Management Data Insights 2(2), 100099 (2022)
12. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., AL-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., Eryiğit, G.: Semeval-2016 task 5: aspect-based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 19–30 (2016).
13. Mohammadi, A., Shaverizade, A.: Ensemble deep learning for aspect-based sentiment analysis. Int. J. Nonlinear Anal. Appl. 12(Special Issue), 29–38 (2021).
14. Apostol, Elena-Simona; Pisica, A.-G.: TRUICĂ, Ciprian-Octavian: ATESA-BERT: A Heterogeneous Ensemble Learning Model for Aspect-Based Sentiment Analysis, arXiv preprint arXiv:2307.15920 (2023).
15. Mohammed, A.; KORA, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University-Computer and Information Sciences 35(2), 757-774 (2023).