

Aspect-Based Adversarial Attack Against Opinion Analysis Models with an Educational Perspective

Monserrat Vázquez-Hernández¹, Luis Alberto Morales-Rosales²,
Ignacio Algreto-Badillo¹, Luis Villaseñor-Pineda¹

¹Instituto Nacional de Astrofísica, Óptica y Electrónica, INAOE, México,

²Secihti-Universidad Michoacana de San Nicolás de Hidalgo, UMSNH
México

mvazquez@inaoe.mx, lamorales@secihti.mx

Abstract. Analyzing students' opinions about classes, instructors, and the overall learning environment enables institutions to understand students' needs and propose improvements to enhance the learning process. Furthermore, analyzing opinions at aspect-level allows a detailed analysis and helps to identify the relevant aspects that require attention to enrich the students' experience. Currently, opinion analysis systems implement deep learning models to increase precision. However, these models can be misled by strategically modified inputs, known as adversarial examples, producing incorrect classifications. In this work, we analyze adversarial examples and their behavior in aspect-level sentiment analysis models to design aspect-aware modifications. Specifically, we define a model for generating aspect-level adversarial examples that modify inputs by targeting aspect-term relationships. The proposed model reduces accuracy by 20.90%, highlighting vulnerabilities in aspect-level models, where even small and targeted changes can degrade performance, providing a foundation for the development of defense mechanisms to ensure reliable predictions.

Keywords: Adversarial attacks, Vulnerabilities, Aspect-based, Sentiment analysis.

1 Introduction

Student opinion analysis concerns the use of sentiment analysis techniques in the education field to extract and analyze students' opinions from interactive learning environments, institutional digital media, or school administrative systems¹. Through these opinions, students can express difficulties understanding, issues with a professor's teaching style, or other aspects of their educational experience [1].

Several works [2–8] have analyzed student opinions considering digital systems such as response systems, clickers, SMS or forums in learning

¹ <https://educationaldatamining.org/>

management systems. Nowadays, the increasing adoption of digital platforms for collecting student feedback has led to a substantial growth in the volume of available opinions. Consequently, the volume of students' comments has increased considerably, necessitating automated tools for extracting and analyzing opinions.

Sentiment analysis systems are an important tool to analyze opinions and provide summarized information concerning the experiences, positive or negative, that users have had about a product, service, or topic of interest. Sentiment analysis (SA) uses text analysis and machine-learning techniques to automatically extract and process opinions [8]. In the education field, analyzing students' opinions collected from learning environments, assisted collaborative learning, institutional digital media, or school administrative systems enables institutions to track students' learning behavior, progression, and overall experience, thus enhancing the learning process according to students' needs [1]. Figure 1 illustrates the overall functioning of a sentiment analysis system within the educational field.

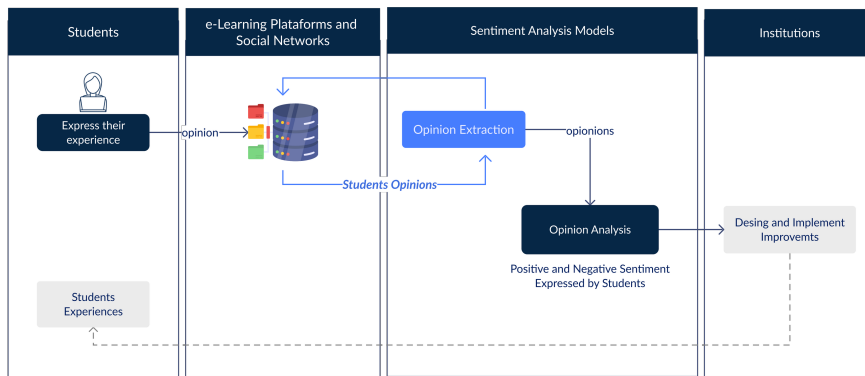


Fig. 1. A general methodology of a sentiment analysis system within the educational field.

According to information needs, sentiment analysis can be performed at different granularity levels: i) document, ii) sentence, or iii) aspect². The analysis at the document-level refers to the positive or negative classification of a full text [9], while at analysis sentence-level the objective is to analyze each sentence in a text to classify them [10]. Finally, aspect-level analysis (or aspect-based sentiment analysis, ABSA) seeks to independently determine the opinion expressed for each aspect mentioned in an opinion [11]. In many cases, the analysis at the document or sentence-level does not provide specific details about aspects. For example, a negative document about an entity does not mean the user has negative opinions about all aspects of that entity [12]; given this, it is necessary to work at a lower granularity. This level of granularity is particularly relevant in educational

² The term "aspect" is used to name components, characteristics or attributes of a product, service or entity.

settings, where different components of the learning experience, such as teaching quality, course content, or evaluation methods, must be evaluated independently. Figure 2 illustrates the differences when performing the sentiment analysis at sentence-level and aspect-level.

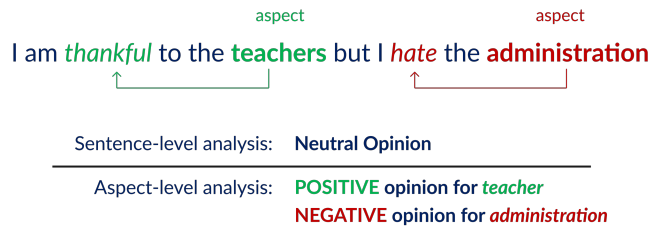


Fig. 2. Aspect-based sentiment analysis. In bold are indicated the evaluated aspects.

Aspect-level analysis, being a detailed task, requires methods that accurately identify the opinion expressed to each evaluated aspect. The use of deep learning (DL) models to address aspect-level analysis has become widespread. These models have demonstrated strong performance across a wide range of tasks. In aspect-level analysis, they enable more accurate analysis by capturing complex semantic relationships within text, supporting more informed decision-making processes.

Despite their success, it has been demonstrated that DL models can be fooled by small modifications to the input data, causing incorrect results [13, 14]. Szegedy *et al.* [13] evaluated state-of-the-art deep neural network (DNN) models for image classification by making minor pixel-level modifications to inputs and showed that these modifications can lead to misclassifications with high probability, even though people could not identify any change. The resulting images, after performing modifications, were denominated as adversarial examples, a term that has since been used to describe intentional modifications applied to input data to mislead deep learning models [46].

Extending the adversarial example concept to text-based tasks, Jia and Liang [15] explored adversarial example design for evaluating DNN models in NLP. In their work, they experimented by inserting text fragments at the end of inputs without changing the original text and showed that DNN-based text models can be misled by such perturbations. Since then, subsequent works on text-based tasks [16–18] have demonstrated that models can be fooled by modifying text at the character, term, or sentence level by adding, deleting, substituting, or swapping text components.

In sentiment analysis, adversarial perturbations can alter correct classifications, leading models to assign incorrect sentiment labels despite minimal changes in the input. Such behavior affects the reliability of sentiment analysis models, particularly when results are used to support decision-making processes. In educational contexts, incorrect sentiment classifications may alter students' interpretations of specific aspects of the learning experience, such as teaching quality or course content. Table 1 illustrates this effect: a minimal

modification at the term level changes the predicted sentiment while preserving the overall structure of the original input.

Table 1. Adversarial examples with a synonym term modification. In the opinion column, the *original* represents the input without modifications and *changes* the generated adversarial example. The modified term is indicated in bold within the student opinion.

Opinion	Student opinion	Sentiment
Original	I like that my teacher pay attention to when we are not emotionally and ready to learn	Positive
Change	I dislike that my teacher pay attention to when we are not emotionally and ready to learn	Negative

Previous adversarial example works in sentiment analysis [16, 17, 19, 20] have proven to effectively fool models via adversarial examples. However, these works have mainly focused on document-level sentiment analysis and have not explicitly addressed aspect-level behavior. Recently, Ekbal *et al.* [21] proposed a method to generate adversarial examples for aspect-level sentiment analysis, integrating specific aspect-level characteristics to preserve semantics. However, this approach assumes that only one aspect is evaluated in an opinion, which is not necessarily true (see Figure 2). This assumption limits its applicability in scenarios where multiple aspects are present within a single opinion, particularly in educational contexts, where different components of the learning experience are expressed simultaneously and must be interpreted independently.

To generate effective adversarial examples, modifications must consider the characteristics of the task, focusing on the terms that determine how the input is interpreted. In our work, by studying aspect-level sentiment analysis, we consider that modifications must be oriented according to the evaluated aspects. As shown in Table 1, each aspect within an opinion is directly related to specific opinion terms, and identifying this aspect-term relationship is necessary to correctly determine the sentiment. Therefore, to perform input modifications and generate adversarial examples at the aspect-level, this relationship must be explicitly considered, and modifications should target these terms to alter the aspect's sentiment classification.

In this work, we propose a model for generating adversarial examples oriented to aspect-level analysis that considers task characteristics to perform input modifications. Based on this model, we define an adversarial attack specifically oriented toward aspect-based sentiment analysis that identifies and modifies terms directly related to the evaluated aspects. We evaluate our adversarial attack against document-level strategies previously applied to sentiment analysis, which have proven effective at misleading DL models. The obtained results show that the proposed attack achieves a 12.8% greater reduction in model accuracy compared to document-level adversarial attacks, while maintaining a semantic similarity of 99% between the original inputs and their adversarial examples. Additionally, the

proposed attack demonstrates transferability across contexts, as it is evaluated on different domain datasets while maintaining its effect on model results.

Since student opinions often involve sensitive information, access to real educational datasets is limited. Therefore, in this work, we evaluate using publicly available datasets that provide aspect-level annotations and sentiment labels, such as SemEval 2016. The results show that aligned perturbations can alter the interpretation of opinions by the model. In educational contexts, such modifications may lead to changes in how students' opinions about their learning environment. Although the dataset used is not in the educational domain, the proposed model captures task-specific behavior, enabling its application to other domains.

Summarizing, the main contributions of this work are:

- A model for generating aspect-based adversarial examples that enables the analysis of sentiment analysis systems by considering aspect-term relations.
- A strategy for deploying an adversarial attack suited to aspect-based sentiment analysis to correctly handle aspect-term relations.
- An adversarial attack that preserves semantic similarity and input readability while applying a reduced number of modifications.
- An adversarial attack that demonstrates transferability across different datasets, maintaining its impact on model performance while preserving semantic consistency.

Together, these contributions enable evaluation of aspect-level sentiment analysis models under controlled input modifications, thereby assessing their robustness to task-specific perturbations. Furthermore, they provide a basis for identifying vulnerabilities and support the development of defenses to mitigate the effects of adversarial examples and improve model reliability.

The remainder of this paper is organized as follows. Section 2 reviews related work, including sentiment analysis in education, adversarial attacks in NLP and sentiment analysis, and discusses the limitations of existing approaches for aspect-based sentiment analysis. Section 3 presents the proposed model for aspect-based adversarial examples. Section 4 describes the experimental setup to evaluate the proposed adversarial model and attack. Section 5 presents the experimental results and discusses their implications for real-world deployments. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Sentiment Analysis in Educational Contexts

Sentiment analysis has been adopted in educational settings to analyze student feedback and support decision-making processes. Educational institutions increasingly rely on digital platforms, such as learning management systems,

forums, and social media, where students express their opinions regarding teaching practices, course content, and overall learning experiences [31–33]. These opinions are no longer merely descriptive but are operationalized to inform pedagogical improvement or institutional evaluation, making their accurate interpretation critical for effective decision-making.

Prior work has shown that analyzing student feedback enables the identification of strengths and weaknesses in teaching quality and course design [31, 32]. In particular, open-ended responses provide richer insights compared to structured questionnaires, allowing students to express nuanced perceptions of their educational experience [32]. As a result, automated sentiment analysis systems have become essential tools for processing large volumes of student-generated content, reducing human bias and enabling scalable analysis [12].

Recent studies have explored the use of sentiment analysis across diverse educational environments, including MOOCs, e-learning platforms, and social media [37–45]. These approaches demonstrate the potential of automated systems to extract meaningful insights from student feedback. However, they implicitly assume that the input data accurately reflects genuine student opinions, overlooking potential vulnerabilities that may distort this information, such as adversarial examples, which introduce subtle modifications that can alter model predictions without changing the perceived meaning of the input.

2.2 Adversarial Attacks in NLP Sentiment Analysis

Deep learning models have significantly improved sentiment analysis performance by capturing complex linguistic patterns and contextual dependencies [26–28]. These models are now commonly used in educational applications due to their ability to process large-scale textual data and provide fine-grained analysis. Despite their effectiveness, these models have been shown to be vulnerable to adversarial examples: carefully crafted inputs that elicit incorrect predictions while preserving semantic plausibility [13, 14].

Formally, adversarial examples are modified inputs designed to alter a model's prediction while preserving the original input's semantics and meaning. Given an input x with a true label y , an adversarial example x' is constructed by applying a minimal perturbation n , such that $x' = x + n$, to change the model classification, i.e., $f(x') \neq y$:

$$\begin{aligned} f(x) &= y, x \in X, \\ x' &= x + n, f(x') \neq y \\ f(x') &= y', y' \neq y \end{aligned} \tag{1}$$

where n represents the worst-case constrained perturbation applied to the input.

In the context of text-based tasks, adversarial attacks operate under the constraint of discrete input spaces, making the generation of imperceptible modifications particularly challenging. Existing approaches address this

limitation through strategies such as synonym substitution, character-level perturbations, paraphrasing, and insertion of distractor text [15–18]. These methods aim to alter model predictions while maintaining semantic similarity and syntactic correctness.

The principal objective of sentiment analysis models is to identify the set of terms that most strongly determine sentiment (positive, negative, or neutral), enabling the classification of an opinion. These terms, often referred to as valuable words, play a crucial role in the final prediction [53, 54]. Recent research focuses on accurately identifying such terms and leveraging them to generate adversarial examples [55]. For instance, TextFool [54] identifies influential terms based on gradient information and applies targeted modifications, while DeepWordBug [51] introduces character-level perturbations in a black-box setting. TextBugger [17] and related methods incorporate multiple similarity constraints to preserve semantic consistency during adversarial example generation. Additionally, population-based optimization approaches generate semantically coherent adversarial examples by selecting synonyms and applying contextual filtering [18].

Although these methods demonstrate the vulnerability of sentiment analysis models, they primarily operate at the document or sentence level, applying global perturbation strategies without considering task-specific structures. As a result, modifications are often distributed across the input text, without explicitly targeting the components that determine fine-grained sentiment interpretation.

2.3 Limitations on Adversarial Attacks on Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis introduces an additional level of complexity by requiring the identification of sentiment for individual aspects within a text [22–24]. This level of granularity is relevant in educational contexts, where multiple aspects of the learning experience, such as teaching quality, course organization, or assessment methods, are simultaneously expressed within a single opinion.

Most adversarial attack strategies have not been designed to account for aspect-level dependencies. Existing approaches typically apply modifications globally, without explicitly modeling the relationship between aspects and the terms that determine their sentiment polarity. This limitation reduces the effectiveness of adversarial perturbations and may lead to unnecessary modifications that affect text readability and semantic coherence.

Given these limitations, Table 2 presents a comparative analysis of representative adversarial attack approaches for sentiment analysis, organized by key dimensions of the threat model, including model knowledge, perturbation granularity, modification strategy, and evaluation criteria.

Table 2. Comparative analysis of adversarial attack strategies for sentiment analysis models.

Work	Model Knowledge	Granularity	Strategy	Target Model	Evaluation Metric	Aspect-Aware
[49]	White-box	Character, Sentence	Edit	CNN	Model Robustness	No
[17]	White-box	Character, Term	Edit	LSTM, CNN	Success rate	No
[16]	White-box	Term	Edit	CNN	Success rate	No
[18]	Black-box	Term	Edit	LSTM	Success rate	No
[50]	Black-box	Term	Paraphrase	BIDAG, Visual7W, fastText	Success rate	No
[51]	Black-box	Character, Term	Hybrid	CNN, LSTM	Model Robustness	No
[52]	Black-box	Term, Sentence	Hybrid	CNN, LSTM, BERT	Model Robustness	No
[19]	Grey-box	Term	Edit	LSTM	Model Robustness	No
[21]	Black-box	Term	Edit	ABSA Models	Model Robustness	Partial
This work		Term (aspect-aligned)	Edit (semantic-aware)	CNN-ABSA	Model Robustness	Yes

As observed, most existing adversarial approaches do not consider aspect level characteristics when designing perturbations, instead applying general modification strategies that do not explicitly consider the structure of the task. These methods do not model the relationship between aspects and their associated opinion terms, which is essential for fine-grained sentiment analysis. A recent effort by Ekbal *et al.* [21] introduced an aspect-level adversarial attack that preserves aspect terms while modifying the surrounding content. Although this approach improves semantic consistency, it assumes that each opinion has a single aspect, which does not reflect real-world scenarios in which multiple aspects coexist.

The lack of explicit modeling of aspect-term relationships remain a key limitation in current adversarial approaches, highlighting that existing methods do not fully capture the complexity of aspect-level sentiment analysis. This observation reveals a critical gap: current adversarial strategies lack task-aware methods that align perturbations with aspect-level semantic dependencies.

2.4 Task-Aware Adversarial Generation

The limitation identified in prior work suggests that effective adversarial example generation for aspect-based sentiment analysis requires incorporating task-

specific semantic structures, particularly the relationship between aspects and their associated opinion terms. Rather than applying global perturbations, adversarial strategies should focus on modifying the minimal set of terms that directly influence aspect-level classification.

This observation motivates the need for task-aware adversarial attacks that align perturbations with the underlying task structure of the input, enabling a more precise evaluation of model robustness. In this context, adversarial example generation can be interpreted not only as an attack strategy but also as a methodological tool to analyze the reliability of sentiment analysis systems under controlled perturbations. Such a perspective is particularly relevant in educational settings, where evaluating the robustness of sentiment analysis systems is essential to ensure that automated interpretations of student feedback remain consistent and trustworthy under realistic input variations.

3 Aspect-Based Adversarial Examples Modeling

In contrast to prior work, we model the task characteristics to guide the modifications applied to the input, focusing on the terms that determine how aspect-level sentiment is interpreted.

In this work, we focus on aspect-based sentiment analysis, where sentiment is determined at the level of individual aspects rather than at the document or sentence level. For example, as illustrated in Figure 2, each aspect mentioned in an opinion is associated with specific terms that determine the expressed sentiment. Therefore, conventional adversarial formulations do not explicitly capture the relationship between aspects and their associated opinion terms. Based on this observation, we propose a model for aspect-level adversarial examples that incorporates aspect-term relationships into the adversarial setting.

3.1 Formalization

We defined the aspect-based adversarial examples model as follows:

- Given an opinion x composed of a set of terms T , where each $t \in T$ may correspond to a unigram or n-gram, and a set of aspects A present in x , each aspect $a_i \in A$ is associated with a subset of terms that determine its sentiment polarity. Let $t_{a_i} \subseteq T$ denote the term(s) related to aspect a_i , and let y_{a_i} be the sentiment expressed for that aspect. The model prediction can be expressed as:

$$M(x, a_i, t_{a_i}) = y_{a_i} \quad (2)$$

- To identify the terms associated with each aspect, we define a semantic proximity function:

$$SP(a_i, t_i) = [0,1] \quad (3)$$

where values close to 1 indicate a strong semantic relation between term t_i and aspect a_i , while values close to 0 indicate weak or no relation.

- The goal of aspect-based adversarial example generation is to construct a modified input x' by altering the terms associated with a given aspect, such that the predicted sentiment is changed:

$$M(x', a_i, t'_{ai}) \neq y_{ai} \quad (4)$$

The central hypothesis of this work is that adversarial effectiveness in aspect-based sentiment analysis depends on modifying only the minimal set of terms that determine aspect-level sentiment. By explicitly modeling the relationship between aspects and their associated terms, adversarial modifications can be applied more precisely, reducing the number of perturbations while preserving semantic coherence and input readability.

3.2 Model Constraints

Generated adversarial examples x' must satisfy the following conditions:

- Semantic consistency at the term level: each modification applied to t_{ai} must preserve semantic proximity with the original term, i.e.,

$$SP(t_i, t'_i) \approx 1$$
- Global semantic preservation: the modified input x' must remain semantically similar to the original input x .

This task-aware perspective enables the generation of adversarial examples that are both effective and semantically aligned, while also facilitating transferability across aspect-based sentiment analysis models.

3.1 Aspect-Based Adversarial Attack

To evaluate the effectiveness of the proposed model, we design an adversarial attack that implements the aspect-based model introduced in the previous section. The attack is implemented under a white-box scenario, with full access to the target model, to analyze its behavior under controlled perturbations.

The proposed attack, denominated as Aspect-Based Adversarial Attack (ABAA), generates adversarial examples by selectively modifying the terms associated with each evaluated aspect, rather than applying global perturbations to the input. This design ensures that modifications are aligned with the structure that determines aspect-level sentiment. Figure 3 illustrates the overall workflow of ABAA.

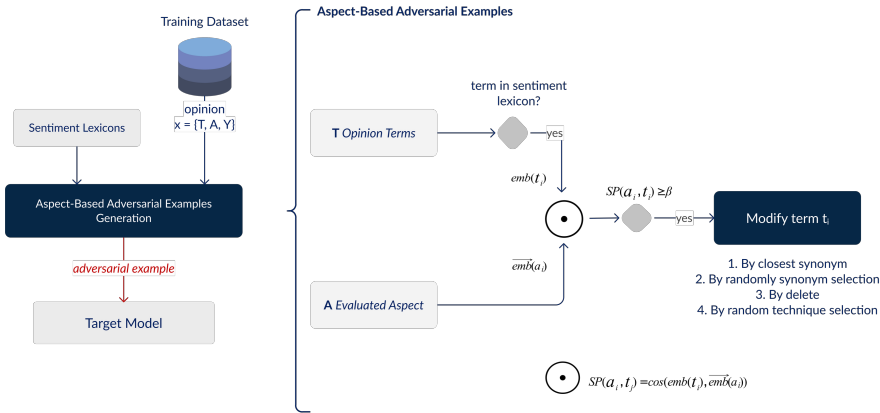


Fig. 3. ABAA: Aspect-Based Adversarial Attack overview.

Unlike existing approaches, which distribute modifications across the input without considering task-specific dependencies, ABAA focuses on aspect-term relationships to guide the modification process. This targeted strategy enables the generation of adversarial examples with fewer modifications, maintaining input readability and semantic consistency.

The following subsections describe the operation of the proposed attack, including the victim model, the adversarial example generation process, and the modification strategies applied.

3.2 Victim Model Knowledge

As a victim model, we selected the existing approach: Sentiment Analysis using Specialized Aspect-Oriented Lexicons [56], which proposes a term weighting scheme for aspect-based sentiment analysis.

The victim model takes as input a set of sentiment-oriented lexicons (one lexicon per sentiment, i.e., positive, neutral, and negative) to represent the common language used to express sentiment. Each sentiment lexicon is modeled as a single vector computed as the average of the term vectors contained in it. Then, each term in an opinion is weighted by its semantic closeness to these vectors. In this way, terms associated with a specific sentiment are emphasized, enabling the final sentiment classification.

To evaluate the weighting scheme, the victim model implements a CNN architecture using the SemEval 2016³ aspect *Restaurant* dataset. The dataset consists of two subsets: (1) a training set with 2,507 reviews and (2) a test set with

³ <https://semeval.github.io/>

889 reviews. In both sets, the customer opinions include annotations identifying the aspects mentioned and their expressed sentiment polarity.

3.3 Aspect-Based Adversarial Examples Generation

Using the sentiment-oriented lexicons from the victim model and the training set, adversarial examples are generated by modifying opinions in the training data prior to model training, without affecting the test set (refer to Fig. 3). This design allows the evaluation of the impact of controlled perturbations on the victim model while preserving an unbiased evaluation setting. To create aspect-based adversarial examples, modifications are performed as follows:

Given an opinion x and one evaluated aspect a_i :

1. Define the set of terms associated with the aspect, denoted as a_terms , which includes the terms composing a_i , and the set of remaining terms in the opinion, denoted as x_terms , which includes all terms in x except those in a_terms . For example, consider the opinion: "*food is tasteless but the support staff was friendly*". If the evaluated aspect is support staff, then: $a_terms = (support, staff)$ and $x_terms = (food, is, tasteless, but, the, was, friendly)$.
2. Each term in a_terms and x_terms is represented with its correspondent embedding⁴.
3. Define a vector representation for the evaluated aspect, denoted as $\overrightarrow{emb}(a_i)$, computed as the average of the embeddings of the terms in a_terms :

$$\overrightarrow{emb}(a_i) = \frac{1}{|a_i|} \sum_{\forall t \in a_i} emb(t) \quad (5)$$

4. To identify candidate terms for modification, compute the semantic proximity between each term $t_i \in x_terms$ and the aspect representation $\overrightarrow{emb}(a_i)$. The semantic proximity is defined using cosine similarity:

$$SP(a_i, t_i) = \cos(\overrightarrow{emb}(a_i), emb(t_i)) \quad (6)$$

$$SP(a_i, t_i) \geq \beta$$

Only terms whose semantic proximity exceeds the threshold β are selected for modification, ensuring that perturbations are aligned with the aspect under evaluation.

⁴ For representing terms and measuring semantic closeness and representation, we use the pre-trained GloVe distributed embeddings on Twitter 200d.

5. Selected terms are modified using the following strategies:

- Replace. The selected term is replaced with a synonym. A list of candidate synonyms is obtained, and their semantic closeness to the original term is evaluated using cosine similarity. The replacement can be selected either: 1) as the most semantically similar candidate or 2) through random selection among valid candidates.
- Delete. The selected term is removed from the opinion.
- Modifications are first applied individually to analyze their isolated effect. Subsequently, a hybrid strategy is evaluated, where the modification type is randomly selected for each candidate term.

4 Experimental Setup

To evaluate the effectiveness of the proposed model, we define a controlled experimental setup comprising the dataset, model configuration, attack parameters, and evaluation metrics.

4.1 Dataset

Since student opinion data is often sensitive and subject to privacy constraints, publicly available educational datasets with aspect-level annotations are limited. For this reason, we employ the SemEval 2016⁵ *Restaurant* dataset to evaluate the proposed model under controlled conditions. This dataset provides fine-grained annotations identifying both the aspects mentioned in each review and their corresponding sentiment polarity. The dataset is divided into training and test sets, as described in Section 3.2.

Importantly, the proposed model is not domain-dependent and can be applied to any dataset that provides aspect annotations and associated sentiment labels, including educational datasets when such data becomes available.

4.2 Model Configuration

The victim model, described in Section 3.2, is implemented using a CNN architecture combined with sentiment-oriented lexicons to weight input terms according to their semantic relevance. Each input opinion is represented through its terms, which are weighted based on their semantic proximity to sentiment lexicon vectors, enabling the model to emphasize sentiment-relevant terms. The

⁵ <https://semeval.github.io/>

model is trained on the original training data and evaluated on both clean inputs and adversarial examples.

4.3 Attack configuration.

The proposed model is used to generate adversarial examples on the training data by applying the modification process described in Section 3.3. In this process, the semantic similarity between each term and the evaluated aspect is computed, and candidate terms are filtered using a threshold $\beta \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$, selecting those whose similarity exceeds it.

The β values are empirically defined to explore different levels of alignment between candidate terms and the aspect representation. Terms with semantic proximity values closer to 1 exhibit a stronger directional similarity with the aspect vector and are therefore more closely associated with the evaluated aspect, thus contributing more significantly to the expressed sentiment. By varying β , the model controls the strictness of the term selection: lower values allow a broader set of candidate terms, while higher values restrict modifications to terms with stronger semantic association. This enables the analysis of how the degree of semantic alignment affects the effectiveness of adversarial perturbations.

4.4 Evaluation metrics.

The effectiveness of the proposed model is assessed using standard evaluation metrics for adversarial attacks. Model performance is measured via Before-Attack Accuracy (BA) and After-Attack Accuracy (AA) to quantify the impact of adversarial examples on classification accuracy. In addition, the attack success rate (SR) is computed as the proportion of inputs for which the model's prediction is altered. To ensure that perturbations preserve input quality, the semantic similarity (SS) between the original and modified inputs is also measured, indicating the degree to which modifications remain imperceptible.

4.5 Reference Adversarial Attack

To establish a comparative reference attack, we define a set of non-aspect-aware adversarial modifications that simulate document-level perturbation strategies. These attacks are designed using the same sentiment-oriented lexicons and training dataset employed by the victim model.

In contrast to our proposed model, this reference attack applies modifications without considering the relationship between aspects and their associated terms. Specifically, opinion terms are modified if they are present in the sentiment-oriented lexicons, regardless of their relevance to the evaluated aspect. This design reflects conventional strategies that rely on global perturbations rather than task-specific dependencies. The purpose of this attack is to provide a reference point for evaluating the proposed model's effectiveness. By comparing

both approaches, it is possible to observe how incorporating aspect-term relationships improves adversarial examples' impact while preserving the meaning of the input.

In this reference attack, modifications are performed as follows:

- Replace. Terms included in the sentiment-oriented lexicons are replaced using semantically similar alternatives. A list of candidate synonyms is obtained for each term, and their semantic closeness is measured using cosine similarity. The replacement is selected either as the most semantically similar candidate or through random selection.
- Delete. Terms included in the sentiment-oriented lexicons are removed from the opinion.
- Each modification strategy is evaluated independently to analyze its individual effect. Subsequently, a hybrid scenario is considered, where the modification type is randomly selected for each candidate term.

5 Experimental Results

This section presents the experimental evaluation of the proposed model and attack. The objective is to assess the effectiveness of aspect-based adversarial perturbations in comparison to non-aspect-aware strategies. The evaluation focuses on three main aspects: 1) the impact of adversarial perturbations on model accuracy, 2) the success rate of the generated adversarial examples, and 3) the preservation of semantic similarity between original and modified inputs. All experiments are conducted under the setup described in Section 4.

First, Table 3 presents the performance of the target model without input modifications and the results obtained under the reference adversarial attack (non-aspect-aware) described in Section 4.5. The results were calculated by executing the target model 10 times; mean and standard deviation ($\pm$ std) are reported. Semantic similarity is computed using cosine similarity between the original input x and the modified input x' .

Table 3. Reference adversarial attack results by the modification technique. BA denotes Before-Attack Accuracy, AA denotes After-Attack Accuracy, and SS denotes Semantic Similarity. Bold values indicate the best results obtained.

Technique	BA	AA	SS
Target model	82.60 $\pm$ 0.46	-	-
Replace	-	74.48 $\pm$ 0.67	0.84
Random replace	-	79.28 $\pm$ 0.37	0.81
Delete*	-	74.50 $\pm$ 0.60	0.65
Hybrid	-	78.86 $\pm$ 0.47	0.73

According to the results, the delete technique is selected as the reference attack, as it produces the largest degradation in model accuracy, reducing it from 82.60% to 74.50%. Among the evaluated modification strategies, this technique consistently leads to the lowest accuracy values, making it suitable as a reference

for comparison. In terms of attack success rate, the target model resisted 607 modified instances, resulting in a success rate of 9.80% (66/673) and an accuracy after attack of 74.47% (607/815). These results reflect the proportion of inputs that remain correctly classified after the application of adversarial modifications, indicating a limited ability of global perturbations to alter model predictions. Although delete introduces stronger perturbations, it significantly degrades semantic quality, leading to a semantic similarity of 0.65%. This reduction in similarity indicates that the modified inputs differ substantially from the original text, suggesting a trade-off between perturbation strength and semantic preservation. Overall, these results indicate that global modifications are insufficient to effectively mislead the model.

Table 4 presents the results of the proposed ABAA attack under different modification strategies and values of β . Each row corresponds to a specific threshold value, while columns report the accuracy after attack and semantic similarity for each modification strategy. The best results for each technique are marked in bold; results marked with * indicate the best values per technique, and ** indicates the overall best performance achieved by ABAA. This organization allows a direct comparison of the impact of different configurations on model performance and input similarity.

Table 4. Aspect-based adversarial attack results under different modification strategies and threshold values β . AA denotes After-Attack Accuracy, and SS denotes Semantic Similarity. Bold values indicate the best results for each modification technique. Results marked with * indicate the best result within each technique, while ** indicates the overall best result achieved across all configurations.

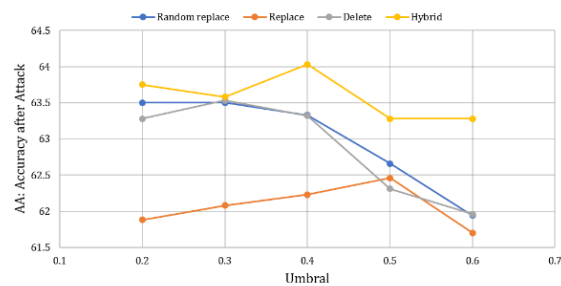
β	ABAA Random replace		ABAA replace		ABAA delete		ABAA Hybrid	
	AA	SS	AA	SS	AA	SS	AA	SS
0.2	63.50 $\pm$ 0.74	0.86	61.88 $\pm$ 0.99*	0.93	63.28 $\pm$ 0.96	0.87	63.75 $\pm$ 0.85	0.90
0.3	63.50 $\pm$ 0.56	0.88	62.08 $\pm$ 0.90*	0.94	63.53 $\pm$ 0.69	0.89	63.58 $\pm$ 0.51	0.99
0.4	63.33 $\pm$ 0.68	0.91	62.23 $\pm$ 0.93*	0.95	63.32 $\pm$ 0.74	0.91	64.03 $\pm$ 0.59	0.91
0.5	62.66 $\pm$ 0.66	0.95	62.46 $\pm$ 0.47	0.97	62.31 $\pm$ 0.70*	0.96	63.28 $\pm$ 0.65	0.92
0.6	61.94 $\pm$ 0.69	0.99	61.70 $\pm$ 0.37**	0.99	61.96 $\pm$ 0.59	0.99	63.56 $\pm$ 0.65	0.93

Target model 82.60 $\pm$ 0.47

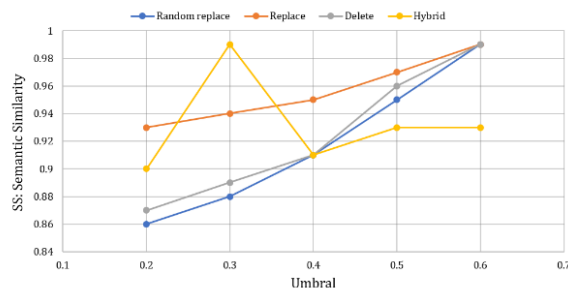
Reference attack 74.50 $\pm$ 0.60

The results show that ABAA reduces model accuracy by up to 20.90% compared to the reference attack. This reduction is observed across different configurations, depending on the selected modification strategy and the value of β . In terms of attack success rate, the model resisted 503 modified instances, corresponding to a success rate of 25.26% and an accuracy after attack of 61.70%. These values indicate the proportion of modified inputs that lead to a change in the model prediction under the evaluated conditions, reflecting a higher effectiveness compared to the reference attack.

Figure 4a illustrates the effect of each modification strategy on model accuracy. The replace strategy produces the strongest degradation in performance, particularly when combined with higher values of β , which restrict modifications to terms more strongly associated with the evaluated aspect. Figure 4b illustrates the effect of ABAA on semantic preservation. In contrast to the reference attack, aspect-aware modifications maintain higher semantic similarity, demonstrating that targeted perturbations can effectively alter model predictions while preserving input readability, indicating a more controlled modification process.



a) Accuracy by technique according to β .



b) Semantic similarity by technique according to β .

Fig. 4. Comparison of obtained results from ABAA attack according to β .

5.1 Discussion

Figure 5 compares ABAA results against the reference attack. The lowest accuracy is obtained with the replace technique when filtering the terms to be modified with a $\beta = 0.6$ (refer to Fig. 4). Under this configuration, ABAA reduces accuracy by 12.81% compared to the reference attack. Figure 6 shows the effect of the ABAA model in maintaining a reduced number of modifications, with semantic

similarity values of up to 0.99 between the original input and the generated adversarial example.

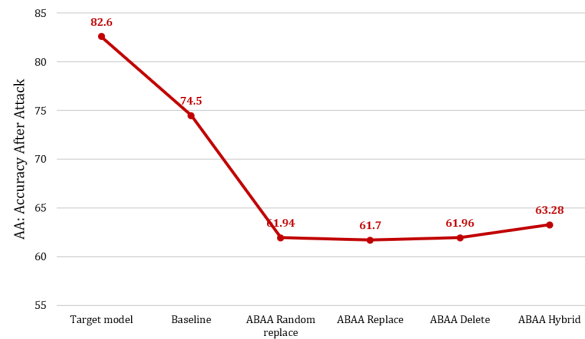


Fig. 5. Comparison ABAA results against target model accuracy and reference attack results.

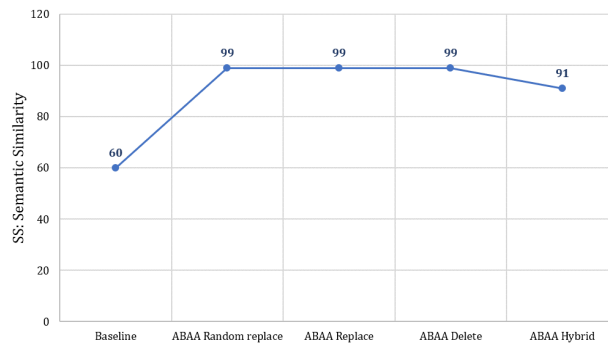


Fig. 6. Comparison achieved semantic similarity by ABAA against attack reference.

Table 5 illustrates the effect of term selection during adversarial example generation by comparing the reference attack and the proposed ABAA model using the delete technique. For each original opinion, the reference attack removes terms based on their presence in sentiment-oriented lexicons, without considering their relevance to the evaluated aspect. As a result, it often deletes words that significantly affect the structure and readability of the sentence, producing incomplete or less coherent inputs (e.g., “*atmosphere*”, “*great*”, “*nice table*”). In contrast, ABAA applies a filtering process based on semantic proximity to the evaluated aspect ($\beta = 0.6$), modifying only terms that are strongly associated with that aspect. This results in fewer modifications, preserving most of the original sentence structure and maintaining readability (e.g., “*everyone raved atmosphere elegant rooms absolutely*”).

These examples show that aspect-based filtering allows more controlled perturbations, where modifications are limited to relevant terms, reducing unnecessary changes while still affecting the model's behavior.

Table 5. Adversarial example generated via reference attack and ABAA attack.

Opinion	Adversarial examples from reference attack by delete technique	Adversarial example from ABAA attack by delete technique with $\beta = 0.6$
everyone raved atmosphere elegant rooms absolutely	atmosphere	everyone raved atmosphere elegant rooms absolutely
great vibe lots people	great	great vibe lots people
very cozy and warm inside	and warm inside	very cozy and warm inside
nice try snag outside table	nice table	nice try snag outside table
like ambience dark original	like ambience	like ambience dark original

To further evaluate the generality and transferability of the ABAA model, an additional experiment is conducted using a different domain dataset. In this case, the English *Laptop* dataset from SemEval 2016⁶ is used, which includes training and test sets with annotated aspects and sentiment polarity. The results are presented in Table 6 and are obtained by executing the target model ten times, reporting the mean and standard deviation ($\pm$ std).

Table 6. Aspect-based adversarial attack results on the *Laptop* dataset. BA, AA, and SS denote Before-Attack Accuracy, After-Attack Accuracy, and Semantic Similarity, respectively.

Technique	BA		AA		SS
Target model	77.48 $\pm$ 0.66		-		-
ABAA Replace	-	0.4	60.248 $\pm$ 3.11		0.786
ABAA Random replace	-	0.5	60.011 $\pm$ 3.16		0.788
ABAA Delete	-	0.4	60.110 $\pm$ 3.25		0.788
ABAA Hybrid	-	0.3	60.110 $\pm$ 3.88		0.776

The results show a reduction in model accuracy under the evaluated configurations, indicating that the generated adversarial examples remain effective across datasets without requiring additional adjustments.

After evaluating the proposed model and compared it against the reference attack, the main observations are:

- Document-level strategies introduce multiple modifications; however, their impact on model accuracy remains limited (refer to Table 5).
- Since the reference attack does not consider aspect-term relationships, modifications are applied to terms that are not directly associated with the evaluated aspect. As a result, some perturbations do not affect the model prediction while still degrading input readability. In contrast, ABAA applies modifications to a smaller set of relevant terms, preserving semantic similarity (up to 0.99) while altering model predictions.

⁶ <https://semeval.github.io/>

- Results obtained on the *Laptop* dataset show similar behavior in terms of accuracy degradation (refer to Table 6), indicating that ABAA model produces similar effects across domains.

From the obtained results, it is observed that designing adversarial examples based on task-specific characteristics influences the effect of modifications on model predictions. The applied modifications are limited in number and preserve semantic similarity while altering the output of the model. These results indicate that models remain sensitive to controlled perturbations applied to relevant input components. Therefore, further analysis of vulnerabilities and the development of defense mechanisms are required to improve the reliability of deep learning models.

Given the benefits that sentiment analysis models provide in the educational field, it is important that such systems incorporate defense mechanisms from their design to mitigate the impact of potential adversarial attacks. This work aims to motivate further research on both adversarial attacks and defense mechanisms for educational sentiment analysis models.

5.2 Implications for Real-World Deployment

Sentiment analysis systems are increasingly integrated into automated pipelines for processing user generated content, including educational feedback platforms and online learning environments. In these settings, aspect-level systems are commonly used to support decision-making processes related to student experience.

In addition, modern NLP pipelines frequently incorporate large language models (LLMs) either as standalone components or as part of architectures for text analysis. These systems often rely on structured inputs, in which aspect-level information plays a significant role in guiding model outputs.

The results obtained in this work indicate that controlled perturbations applied to aspect-related terms can alter model predictions while preserving the input's semantic structure. This suggests that, in practical scenarios, modifications that remain imperceptible at the surface level may still influence the interpretation of specific aspects within user feedback. This observation is particularly relevant for LLM-based systems, where input sensitivity to subtle textual variations can propagate through downstream tasks, potentially affecting generated outputs, summaries, or recommendations derived from user feedback.

These observations highlight the importance of incorporating robustness and validation mechanisms in sentiment analysis systems deployed in real-world environments, particularly in contexts where decisions rely on fine-grained aspect-level analysis.

6 Conclusions

The main contribution of this work is the formalization of aspect-based adversarial examples that considers the aspect-term relationship to determine which terms to modify. Unlike previous approaches, the proposed model uses aspect-level information to guide input modifications that affect model predictions. This allows adversarial examples to preserve readability, semantics, and syntax, achieving up to 0.99 semantic similarity while reducing model accuracy by up to 20.90%.

During the experimental evaluation, the aspect-term relationship is modeled through the semantic proximity between terms and the evaluated aspect, enabling the selection of relevant terms for modification. The results show that aspect-based adversarial examples alter model predictions while preserving the input's semantic structure. In addition, selecting terms based on semantic similarity reduces the perceptibility of modifications. Evaluation across different datasets indicates that the proposed model maintains its effect across domains.

Future work will evaluate the proposed model across different target models, adversary-knowledge settings, and datasets, including transformer-based architectures and large language models (LLMs). This is particularly relevant in educational contexts, where sentiment analysis systems are used to support decision-making, and where ensuring robustness against adversarial perturbations is necessary to preserve the reliability of model outputs.

Acknowledgments. This research was funded by "Secretaría de Ciencia, Humanidades, Tecnología e Innovación (Secihti)" through scholarships 814461.

CRedit author statement. **Montserrat Vázquez-Hernández:** Conceptualization, Methodology, Formal analysis, Software, Writing – original draft preparation. **Luis Alberto Morales-Rosales:** Supervision, Validation, Writing – review and editing. **Ignacio Algreto-Badillo:** Supervision, Validation, Writing – review and editing. **Luis Villaseñor-Pineda:** Supervision, Validation.

Disclaimer/Publisher's Note: The statements, opinions and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. ASLERD and/or the member of the editorial board of IxD&A disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

References

1. Shaik T., Tao X., Li Y., Dann C., McDonald J., Redmond P., Galligan L.: A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis IEEE Access, 10, (2022).

- <https://doi.org/10.1109/ACCESS.2022.3177752>
2. Altrabsheh N., Gaber M.M., Cocea M.: SA-E: Sentiment analysis for education *Frontiers in Artificial Intelligence and Applications*. vol. 255 (2013).
<https://doi.org/10.3233/978-1-61499-264-6-353>
 3. Jimmy, Prasetyo V.R.: Sentiment analysis on feedback of higher education teaching conduct: An empirical evaluation of methods *AIP Conference Proceedings*. vol. 2470 (2022). <https://doi.org/10.1063/5.0080182>
 4. Bal S., Choudhary A.K., Majumdar S., Pal D., Mandal L.: Sentiment Analysis of Online Reviews for Educational Institutions *Applications of Machine Intelligence in Engineering* (2022). <https://doi.org/10.1201/9781003269793-22>
 5. OSMANOĞLU U.Ö., ATAĞ O.N., ÇAĞLAR K., KAYHAN H., CAN T.: Sentiment Analysis for Distance Education Course Materials: A Machine Learning Approach *Journal of Educational Technology and Online Learning*, 3, (2020).
<https://doi.org/10.31681/jetol.663733>
 6. J A., K C., N H., P R., T T., Rokde C.: Sentiment Connect: Opinion Mining on Education System In *India International Journal of Progressive Research in Science and Engineering*, 3, (2022)
 7. Fakhraie N., Hewitt J.: What's in a Note? Sentiment Analysis in Online Educational Forums *Curriculum, Teaching and Learning*, MA, (2011)
<https://utoronto.scholaris.ca/bitstreams/9ab6e236-1f7a-4d2d-ac68-a57eb3f8747a/download>
 8. Weiss S.M., Indurkha N., Zhang T., Damerau F.J.: Text mining: Predictive methods for analyzing unstructured information, (2005). <https://doi.org/10.1007/978-0-387-34555-0>
 9. Pang B., Lee L., Vaithyanathan S.: Thumbs up? Sentiment Classification using Machine Learning Techniques *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002* (2002).
<https://doi.org/10.3115/1118693.1118704>
 10. Riloff E., Wiebe J.: Learning Extraction Patterns for Subjective Expressions *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP 2003* (2003). <https://doi.org/10.3115/1119355.1119369>
 11. Poria S., Hazarika D., Majumder N., Mihalcea R.: Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis *Research IEEE Trans Affect Comput*, 14, (2023). <https://doi.org/10.1109/TAFFC.2020.3038167>
 12. Liu B., Zhang L.: A survey of opinion mining and sentiment analysis. In *Mining text data* Springer, (2012). https://doi.org/10.1007/978-1-4614-3223-4_13
 13. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R.: Intriguing properties of neural networks *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2014).
<https://doi.org/10.48550/arXiv.1312.6199>
 14. Goodfellow I.J., Shlens J., Szegedy C.: Explaining and harnessing adversarial examples *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015). <https://doi.org/10.48550/arXiv.1412.6572>
 15. Jia R., Liang P.: Adversarial examples for evaluating reading comprehension systems *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (2017). <https://doi.org/10.18653/v1/D17-1215>
 16. Tsai Y.-T., Yang M.-C., Chen H.-Y.: Adversarial Attack on Sentiment Classification Presented at the (2019). <https://doi.org/10.18653/v1/W19-4824>
 17. Li J., Ji S., Du T., Li B., Wang T.: TEXTBUGGER: Generating Adversarial Text Against Real-world Applications *26th Annual Network and Distributed System Security Symposium, NDSS 2019* (2019). <https://doi.org/10.14722/ndss.2019.23138>

18. Alzantot M., Sharma Y., Elgohary A., Ho B.J., Srivastava M.B., Chang K.W.: Generating natural language adversarial examples Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018 (2018). <https://doi.org/10.18653/v1/D18-1316>
19. Xu Y., Zhong X., Yepes A.J., Lau J.H.: Grey-box Adversarial Attack And Defence For Sentiment Classification NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference (2021). <https://doi.org/10.18653/v1/2021.naacl-main.321>
20. Zhang Y., Shao K., Yang J., Liu H.: Universal adversarial attack via conditional sampling for text classification Applied Sciences (Switzerland), 11, (2021). <https://doi.org/10.3390/app11209539>
21. Mamta, Ekbal A.: Adversarial Sample Generation for Aspect based Sentiment Classification 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing - Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022 (2022). <https://doi.org/10.18653/v1/2022.findings-aacl.44>
22. Hu M., Liu B.: Mining and summarizing customer reviews KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004). <https://doi.org/10.1145/1014052.1014073>
23. Bing L., Minqing H., Junsheng C.: Opinion observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th international conference on World Wide Web (WWW '05). Association for Computing Machinery, New York, NY, USA, 342-351. <https://doi.org/10.1145/1060745.1060797>
24. Zhu L., Xu M., Bao Y., Xu Y., Kong X.: Deep learning for aspect-based sentiment analysis: a review PeerJ Comput Sci, 8, (2022). <https://doi.org/10.7717/peerj-cs.1044>
25. Gonzalez A.: "Conceptos básicos de Machine learning." Clever Data Big Data Prediction. <https://cleverdata.io/conceptos-basicosmachine-learning> (2014).
26. Ramos D.L., García L.A.: Aprendizaje profundo para la extracción de aspectos en opiniones textuales. Revista Cubana de Ciencias Informáticas, 13(2), 105-145. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992019000200105&lng=es&tng=
27. Bengio Y., Courville A., Vincent P.: Representation learning: A review and new perspectives IEEE Trans Pattern Anal Mach Intell, 35, (2013). <https://doi.org/10.1109/TPAMI.2013.50>
28. Zhang W.E., Sheng Q.Z., Alhazmi A., Li C.: Adversarial Attacks on Deep learning Models in Natural Language Processing: A ACM Trans. Intell. Syst. Technol. 11, 3, Article 24 (June 2020), 41 pages. <https://doi.org/10.1145/3374217>
29. Dolianiti F.S., Iakovakis D., Dias S.B., Hadjileontiadou S., Diniz J.A., Hadjileontiadis L.: Sentiment analysis techniques and applications in education: A survey Communications in Computer and Information Science. vol. 993 (2019). https://doi.org/10.1007/978-3-030-20954-4_31
30. Van Houdt G., Mosquera C., Nápoles G.: A review on the long short-term memory model Artif Intell Rev, 53, (2020). <https://doi.org/10.1007/s10462-020-09838-1>
31. Clarizia F., Colace F., De Santo M., Lombardi M., Pascale F., Pietrosanto A.: E-learning and sentiment analysis: A case study ACM International Conference Proceeding Series (2018). <https://doi.org/10.1145/3178158.3178181>
32. Misuraca M., Scepi G., Spano M.: Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students' feedback Studies in Educational Evaluation, 68, (2021). <https://doi.org/10.1016/j.stueduc.2021.100979>

33. Zhou J., Ye J. min: Sentiment analysis in education research: a review of journal publications, (2020). <https://doi.org/10.1080/10494820.2020.1826985>
34. Shapiro H.B., Lee C.H., Wyman Roth N.E., Li K., Çetinkaya-Rundel M., Canelas D.A.: Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers *Comput Educ*, 110, (2017). <https://doi.org/10.1016/j.compedu.2017.03.003>
35. Binali H.H., Wu C., Potdar V.: A new significant area: Emotion detection in E-learning using opinion mining techniques 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, DEST '09 (2009). <https://doi.org/10.1109/DEST.2009.5276726>
36. Ortigosa-Hernández J., Rodríguez J.D., Alzate L., Lucania M., Inza I., Lozano J.A.: Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers *Neurocomputing*, 92, (2012). <https://doi.org/10.1016/j.neucom.2012.01.030>
37. Chauhan G.S., Agrawal P., Meena Y.K.: Aspect-based sentiment analysis of students' feedback to improve teaching-learning process *Smart Innovation, Systems and Technologies*. vol. 107 (2019). https://doi.org/10.1007/978-981-13-1747-7_25
38. Alrajhi L., Alamri A., Pereira F.D., Cristea A.I.: Urgency Analysis of Learners' Comments: An Automated Intervention Priority Model for MOOC Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 12677 LNCS (2021). https://doi.org/10.1007/978-3-030-80421-3_18
39. Chen L., Chen P., Lin Z.: Artificial Intelligence in Education: A Review *IEEE Access*, 8, (2020). <https://doi.org/10.1109/ACCESS.2020.2988510>
40. Leong C.K., Lee Y.H., Mak W.K.: Mining sentiments in SMS texts for teaching evaluation *Expert Syst Appl*, 39, (2012). <https://doi.org/10.1016/j.eswa.2011.08.113>
41. Wen M., Yang D., Rosé C.: Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Proceedings of Educational Data Mining* (2014). <https://api.semanticscholar.org/CorpusID:3243907>
42. Moreno-Marcos P.M., Alario-Hoyos C., Munoz-Merino P.J., Estevez-Ayres I., Kloos C.D.: Sentiment analysis in MOOCs: A case study *IEEE Global Engineering Education Conference, EDUCON*. vol. 2018- April (2018). <https://doi.org/10.1109/EDUCON.2018.8363409>
43. de Oliveira A.B., Alves A.L.F., de Souza Baptista C.: Using opinion mining in student assessments to improve teaching quality in universities *Advances in Intelligent Systems and Computing*. vol. 1181 AISC (2021). https://doi.org/10.1007/978-3-030-49342-4_22
44. Ortigosa A., Martín J.M., Carro R.M.: Sentiment analysis in Facebook and its application to e-learning *Comput Human Behav*, 31, (2014). <https://doi.org/10.1016/j.chb.2013.05.024>
45. Toçoğlu M.A., Onan A.: Sentiment Analysis on Students' Evaluation of Higher Educational Institutions *Advances in Intelligent Systems and Computing*. vol. 1197 AISC (2021). https://doi.org/10.1007/978-3-030-51156-2_197
46. Zhang W.E., Sheng Q.Z., Alhazmi A., Li C.: Adversarial Attacks on Deep-learning Models in Natural Language Processing, (2020). <https://doi.org/10.1145/3374217>
47. Yuan X., He P., Zhu Q., Li X.: Adversarial Examples: Attacks and Defenses for Deep Learning *IEEE Trans Neural Netw Learn Syst*, 30, (2019). <https://doi.org/10.1109/TNNLS.2018.2886017>
48. Zhang J., Li C.: Adversarial Examples: Opportunities and Challenges *IEEE Trans Neural Netw Learn Syst*, 31, (2020). <https://doi.org/10.1109/TNNLS.2019.2933524>

49. Liang B., Li H., Su M., Bian P., Li X., Shi W.: Deep text classification can be fooled IJCAI International Joint Conference on Artificial Intelligence. vol. 2018- July (2018). <https://doi.org/10.24963/ijcai.2018/585>
50. Ribeiro M.T., Singh S., Guestrin C.: Semantically equivalent adversarial rules for debugging NLP models ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). vol. 1 (2018). <https://doi.org/10.18653/v1/P18-1079>
51. Gao J., Lanchantin J., Soffa M. Lou, Qi Y.: Black-box generation of adversarial text sequences to evade deep learning classifiers Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018 (2018). <https://doi.org/10.1109/SPW.2018.00016>
52. Jin D., Jin Z., Zhou J.T., Szolovits P.: Is BERT really robust? A strong baseline for natural language attack on text classification and entailment AAAI 2020 - 34th AAAI Conference on Artificial Intelligence (2020). <https://doi.org/10.1609/aaai.v34i05.6311>
53. Ma Y., Peng H., Cambria E.: Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 (2018). <https://doi.org/10.1609/aaai.v32i1.12048>
54. Xiao Y., Zhou G.: Syntactic edge-enhanced graph convolutional networks for aspect-level sentiment classification with interactive attention IEEE Access, 8, (2020). <https://doi.org/10.1109/ACCESS.2020.3019277>
55. Wang W., Wang R., Ke J., Wang L.: TextFirewall: Omni-Defending against Adversarial Texts in Sentiment Classification IEEE Access, 9, (2021). <https://doi.org/10.1109/ACCESS.2021.3058278>
56. Vázquez-Hernández M., Villaseñor-Pineda L., Montes-Y-Gómez M.: A Semantic-Proximity Term-Weighting Scheme for Aspect Category Detection Procesamiento del Lenguaje Natural, 69, (2022). <https://api.semanticscholar.org/CorpusID:252848418>