Neural Processing Letters manuscript No.
(will be inserted by the editor)

Cross-domain Polarity Models to evaluate User eXperience in E-learning

Rosario Sanchis-Font · Maria Jose Castro-Bleda · José-Ángel González · Ferran Pla · Lluís-F. Hurtado

Received: date / Accepted: date

Abstract Virtual Learning Environments are growing in importance as fast as e-learning is becoming highly demanded by universities and students all over the world. This paper investigates how to automatically evaluate User eXperience in this domain using Sentiment Analysis techniques. For this purpose, a corpus with the opinions given by a total of 583 users (107 English speakers and 476 Spanish speakers) about three Learning Management Systems in different courses has been built. All the collected opinions were manually labeled with polarity information (positive, negative or neutral) by three human annotators, both at the whole opinion and sentence levels. We have applied our state-of-the-art sentiment analysis models, trained with a corpus of a different semantic domain (a Twitter corpus), to study the use of cross-domain models for this task. Cross-domain models based on Deep Neural Networks (Convolutional Neural Networks, Transformer Encoders and Attentional BLSTM models) have been tested. In order to contrast our results, three commercial systems for the same task (MeaningCloud, Microsoft Text Analytics and Google Cloud) were also tested. The obtained results are very promising and they give an insight to keep going the research of applying sentiment analysis tools on User eXperience evaluation. This is a pioneering idea to provide a better and accurate understanding on human needs in the interaction with Virtual Learning Environments and a step towards the development of automatic tools that capture the feed-back of user perception for designing Virtual Learning Environments centered in user's emotions, beliefs, preferences, perceptions, responses, behaviors and accomplishments that occur before, during and after the interaction.

Keywords Machine Learning · Artificial Neural Networks · Sentiment Analysis · User Experience · Virtual Learning Environments · Learning Management Systems

Partially supported by the Spanish MINECO and FEDER founds under project TIN2017-85854-C4-2-R. Work of J.A. González is financed under grant PAID-01-17.

VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Valencia, Spain
E-mail: rosanfon@doctor.upv.es, mcastro@dsic.upv.es, jogonba2@dsic.upv.es, fpla@dsic.upv.es, lhurtado@dsic.upv.es

1 Introduction

Human Computer Interaction (HCI) tools developers, agents and industry require to focus their interactive systems on end-users in order to design and provide quality systems upon the international standards requirements ISO. These interactive systems are the "combination of hardware, software and/or services that receives input from, and communicates output to, users" (ISO 9241-210:2019) [18]. This international standard is related to ergonomics of human system-interaction and human-centered design for interactive systems. It provides requirements and recommendations for human-centered design principles and activities throughout the life cycle of computer-based interactive systems. It is intended to be used by those managing design processes, and is concerned with ways in which both hardware and software components of interactive systems can enhance human–system interaction.

Therefore "User eXperience" (UX) enhances human interaction within the hardware or software components, being the UX concept multidimensional and centered in human needs. This UX concept goes beyond usability, interaction experience and design by involving two main qualities: traditional HCI usability and accessibility balanced with hedonic and affective design [42]. In this perspective, in [14], UX is described as a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.), and the context (or the environment) within which the interaction occurs (e.g. organizational/social setting, meaningfulness of the activity, voluntariness of use, etc.). Therefore, these authors conclude that UX is considering three perspectives: emotion and affect of the user, technology and the hedonic instrument and the experiential aspect. As a result, UX includes a multidimensional concept and focuses in human needs and the aspects of beauty, fun, pleasure, and personal growth rather than the value of the product or instrument used [14], which improves or worsens along the time of use [21].

UX has to be considered when designing and redesigning hardware and software applications. In this way, in the last years, UX has been taken into account when designing Virtual Learning Environments (VLEs) [42]. VLEs includes a wide range of technology-enabled learning environments, such as Learning Management Systems (LMSs), computer games or Virtual Worlds.

Traditionally, evaluation of UX in VLEs (or in any other product or service), has always been addressed by conventional questionnaires. In this regard, we used the validated User Experience Questionnaire (UEQ) [32], conducted on students of biomedical postgraduate studies and Massive Open Online Courses (MOOCs) students. Three LMSs have been evaluated using this adapted UEQ: an ad-hoc system called "Conecto" (in Spanish and English languages), an open-source Moodle personalized system (in Spanish), and an edX platform (both in Spanish and English languages).

In this paper, instead of evaluating UX in this traditional way, we have addressed the problem in a novel way applying machine learning tools to the users' opinions, expressed freely in natural language. Preliminary work was done in [35]. Deep learning cross-domain models (Convolutional Neural Networks, Transformer Encoders and Attentional BLSTM models) trained with tweets and different general systems for text analytics (such as MeaningCloud [27, 28], Google Cloud [13],

and Microsoft Text Analytics [29]) are used to this end. The application of sentiment analysis tools on UX opinions will provide a better and accurate understanding on human needs in the interaction with VLEs. The ultimate goal of this work is to develop further tools of automatic feed-back of user perception for designing user-centered VLEs valued by users for its usability, quality and pleasure of use.

This paper is organized as follows. Next Section gives a brief overview about the state of the art of the work presented here. Section 3 describes the data collection from the questionnaires and the labeling process. The evaluation metrics are introduced in Section 4. The used models for sentiment analysis are described in Section 5 and Section 6 presents our proposal. Section 7 presents the experimental results and their analysis. Finally, the conclusions and future directions are drawn in the last Section.

2 State of the Art

Sentiment analysis is one of the most active areas in Natural Language Processing since the early 2000s. The pioneering works in this field [30, 39] pointed out the importance of "sentiment classification" for a large number of tasks such as *message filtering*, *recommender systems* or *business intelligence applications*. Other sentiment analysis approaches were addressed by manually generating polarity lexicons [23, 41]. However, the efforts required to develop these resources and the good performance of machine learning systems on this task made the research community to move towards data driven approaches. A survey of the most widely used machine learning approaches for the sentiment analysis problem can be found in [22].

Recently, the predominant systems to perform sentiment analysis are neural network based approaches [43]. The most popular models are Convolutional Neural Networks (CNN) [19], Long Short Term Memories (LSTM) [15], and combinations of CNN and LSTM [34]. Moreover, the enrichment of this architectures by using attention mechanisms [2] and Transformers [40] are lately used.

The interest on sentiment analysis has increased along with the popularity of the social networks and the user interactions on them. The most studied social network for sentiment analysis tasks is Twitter, where the users are allowed to broadcast opinions about any topic by using only 280 characters and media content.

Several workshops are organized in order to address the sentiment analysis task in Twitter, providing corpora and resources to the participants for training and evaluating their systems. The most known workshops are the International Workshop on Semantic Evaluation (SemEval) and the Workshop on Semantic Analysis at SEPLN (TASS) for English and Spanish language, respectively.

For the last task of English sentiment analysis presented at SemEval [33], most of the participating teams proposed neural network models mainly based on LSTM and CNN, being the two best systems based on these approaches along with pretrained word embeddings on big collections of tweets. Concretely, the winner team proposed a two layer bidirectional LSTM with attention mechanisms [3], while the second ranked team addressed the task by using a combination of LSTM and CNN [4].

For the Spanish sentiment analysis task of TASS 2019 [7], the predominant presence of deep learning components was also observable, where almost all the systems proposed by the participants made use of them. It is worthly to note the great interest on the Transformer model [40], being used mainly with the aim of fine-tuning pre-trained contextual representations of words [6].

Our team proposed a system focused on encoding pre-trained skipgram word embeddings by using a Transformer encoder to carry out the classification [12]. It turned out to be the best system, being the first ranked system. The system of the second ranked team used a logistic regression classifier on top of different representations, such as word embeddings and bag of characters, by focusing on a novel way of data augmentation [24].

In addition to these kinds of systems, a large number of commercial products and frameworks have also proliferated to facilitate the development and deployment of sentiment analysis systems based on machine learning, such as Google Cloud [13], IBM Watson [17], Microsoft Text Analytics [29], MeaningCloud [28] or Stanford Core NLP [25]. These products allow us to perform text analytics such as sentiment analysis, in a broad variety of domains and languages in an easy way, obtaining also competitive results. For this reason, besides our neural network models, other commercial models will be used in our work as explained in Section 5.2.

But, though the promising results on several tasks of natural language processing and, in particular, on sentiment analysis, generally speaking, UX evaluation is immature in most applications and, especially, in VLEs.

Some work has been done in eCommerce, using natural language processing to improve their UX. For instance, to search products in a more intelligent way, using sentiment analysis to extract insights from the reviews made by the customers on the product or identifying trends and trying to answer best to the customers' concerns. Several new conferences have recently been launched around these ideas, such as the Workshop on Economics and Natural Language Processing[1] or the First International Workshop on e-Commerce and NLP[2].

Another research line covered in this paper is the use of cross-domain polarity classification approach, that is, the texts to be classified belong to a different domain from those used in the training phase. Most work has been done within the classic approach, the so-called single-domain polarity classification, which classifies texts in the same domain to which the texts used in the training phase belong to. Due to the lack of training data (only opinions from 583 users), we used cross-domain models, those trained with another domain (Twitter) and those trained with general data.

3 Experimental Data

The validated User Experience Questionnaire (UEQ) [32] was used in order to automatically evaluate UX in our VLEs. This questionnaire is a list of close-ended questions, but we added questions concerning to sociodemographic data (age, sex, etc.) and an open field "Other comments" (see Figure 1 for a screenshot of the

[1] https://julielab.de/econlp/2019/

[2] https://www.aclweb.org/portal/content/first-international-workshop-e-commerce-and-nlp

32. What kind of user are you?
□ Student
□ Teacher/Tutor
□ Administrator
□ Other

33. Have you previously used other e-learning platforms?
□ Yes, I have used previously other e-learning platforms.
□ No, I have not used previously other e-learning platforms.
□ DK/NA

34. Please, let us know any comments about your experience on the environment of IVI e-learning Master:

By doing this questionnaire you accept the use of your data for scientific purposes. This data is completely confidential and only it will be used for current and future papers, reports and studies that might be produced after processing the information by Fundación IVI and UPV. Please, click on the blue button to send the questionnaire. Many thanks for your contribution and tell us your experience.

Fig. 1 "Other comments" box from UX questionnaire delivered to English speaker users on Conecto LMS of IVI Foundation Biomedical International Master (2017-2018 edition).

questionnaire in one VLE). It is a text entry box to express any opinion or comment related to UX in the course, which is an opportunity to get new and more precise information about their experience, not only by close-ended questions.

Three LMSs have been evaluated using this adapted UEQ: "Conecto"[3], which is an ad-hoc system (for Spanish and English users), an open-source Moodle personalized system (for Spanish users) [4], and an edX platform (both for Spanish and English languages)[5]. We have collected data in different editions of the courses, obtaining an answer to the "Other comments" box from 583 users (107 English speakers and 476 Spanish speakers) .

3.1 Polarity of Observations and Sentences

We have performed experiments at two different semantic levels of decreasing complexity:

1. *Observation.* We measured the polarity of the whole observation. Each entry is composed by one or more sentences. There were 476 Spanish and 107 English observations. An average of 15 words both per Spanish observation and 20 words per English observation is found.
2. *Sentence.* As an observation from one user can be composed by one or more sentences, we automatically split each observation into sentences, being one

[3] `https://postgrado.adeituv.es/es/cursos/salud-7/assisted-reproduction/datos_generales.htm`
[4] `https://medicinagenomica.com/eugmygo/`
[5] `https://www.upvx.es/`

Table 1 Examples of tagged observations and sentences, with their polarity.

Unit	Example	Polarity
Observation	*Overall, this e-learning master environment is very friendly.*	Positive
Sentence	*Overall, this e-learning master environment is very friendly.*	Positive
Observation	*It was good experience to some extent. However, I hope it concentrates more on practical aspect in the future.*	Neutral
Sentence	*It was good experience to some extent.*	Positive
Sentence	*However, I hope it concentrates more on practical aspect in the future.*	Negative
Observation	*Well-organized and structured course. Great study material (articles) but not enough time to read them all. Keep up the good work.*	Neutral
Sentence	*Well-organized and structured course.*	Positive
Sentence	*Great study material (articles) but not enough time to read them all.*	Neutral
Sentence	*Keep up the good work.*	Positive

sentence the text between points. We got 587 Spanish sentences and 184 English sentences. The percentage of observations composed by more than one sentence is 24% for the Spanish observations and 32% for the English ones. An average of 12 words per Spanish and English sentences is found.

As stated before, one observation can be composed of more than one sentence, and it is very usual to mix positive and negative opinions about different concepts in different sentences, so many observations are tagged as neutral (see some examples in Table 1 to illustrate this idea). This fact hides the intention of the user, which is tagged as neutral when she or he is not, that is the reason we automatically split the original observations into sentences and measuring the polarity of each sentence.

3.2 Manual Labeling

The units (whole observations and sentences) had to be labeled according to its polarity (positive, negative, or neutral). In a first step, positive and negative sentences were manually annotated, being tagged as neutral those sentences without presence of any emotion or feelings (e.g., "No applicable.") or when the sentence provided both positive and negative feelings (e.g., "Some of the modules were very interesting and valuable but some of them confusing as too genetic details involved."). Three human annotators (as in [38]) did the annotation of each sentence as positive, negative or neutral.

Secondly, as an observation is a sequence of sentences, and, following Socher's work [38], based on the structure of the discourse of the observations, observation labeling was carried out from the polarity level of the sentences which compose the observation. The core idea is that the polarity of an observation will be automatically set as positive if it is composed of positive sentences; similarly, it will be set as negative if every sentences is negative; and finally, it will be tagged as neutral if it is composed of positive and negative and/or neutral sentences.

In order to evaluate the inter-annotator agreement we used the following measures: Krippendorf's alpha (α) [20], Cohen's kappa (κ) [5] and Scott's pi (π) [37],

Table 2 Observations and sentences extracted from the "Other comments" box.

Unit	Language	Total	Positive	Negative	Neutral
Observation	Spanish	476	338 (71%)	85 (18%)	53 (11%)
	English	107	56 (52%)	30 (28%)	21 (20%)
Sentence	Spanish	587	404 (69%)	142 (24%)	41 (7%)
	English	184	90 (49%)	80 (43%)	14 (8%)

both for the observation and sentence levels of the labeling. The obtained results suggest a high correlation among the labeling work of the three annotators at both levels, concretely, $\alpha = \kappa = \pi = 0.88$ for whole Spanish observations, $\alpha = \kappa = \pi = 0.90$ for Spanish sentences, $\alpha = \kappa = \pi = 0.84$ for whole English observations and $\alpha = \kappa = \pi = 0.90$ for English sentences.

These results seem to suggest that the sentiment is more detectable at sentence level than at observation level, where several opinions with different polarity are more likely to happen, therefore, observations are more difficult to label.

The total number of units and the class distributions can be seen in Table 2. As it can be observed, there are more positive than negative samples. The neutral category decreased from the whole comment (a complex statement) to the sentence (usually, with polarity or, less frequently, with lack of sentiment). All samples were used as test set, and they were automatically labeled by using the proposed models (the neural networks systems developed in this work and other general models from commercial tools) and compared with the ground truth label.

4 Evaluation Metrics

Different evaluation metrics were used in order to test the systems. Concretely, as defined below, accuracy (Acc, Eq. 1) and macro F_1 (MF_1, Eq. 3) were used to reduce the impact of corpus imbalance in the evaluation. Moreover, the F_1 per class c (*positive*, *negative*, or *neutral* class) as defined in Eq. 2 was computed to observe the behavior of our systems at class level.

$$Acc = \frac{\sum_{c \in C} \sum_{x \in \Omega_c} [f(x) = c]}{|\Omega|} \tag{1}$$

$$F_1^c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{2}$$

$$MF_1 = \frac{1}{|C|} \sum_{c \in C} F_1^c \tag{3}$$

Ω is the set of samples, Ω_c are the samples of class c in Ω, $y(x)$ is the prediction of the model f for a given sample x, C is the set of classes, $[\cdot]$ denotes the Iverson bracket, and P_c and R_c are the precision and recall measure of each class, defined as follows:

$$P_c = \frac{\sum_{x \in \Omega_c} [f(x) = c]}{\sum_{x \in \Omega} [y(x) = c]} \qquad R_c = \frac{\sum_{x \in \Omega_c} [f(x) = c]}{|\Omega_c|} \tag{4}$$

Moreover, the macro-precision (MP) and macro-recall (MR) are also considered in order to compare the results of our supervised systems with those officially published at SemEval and TASS workshops.

$$MP = \frac{1}{|C|} \sum_{c \in C} P_c \qquad MR = \frac{1}{|C|} \sum_{c \in C} R_c \tag{5}$$

5 Cross-Domain Polarity Models

Cross-domain models for both Spanish and English were used to address the problem of sentiment analysis on VLEs. On the one hand, Deep Neural Networks such as Convolutional Neural Networks (CNN), Attentional Bidirectional Long Short Term Memory, and Transformer Encoders (TE) models were used to train models for sentiment analysis tasks on Twitter, both in Spanish and English, proposed in international competitions [33, 26, 7]. On the other hand, we used the sentiment analysis module provided by several commercial "Software as a service" text analytics products: MeaningCloud [28], Microsoft Text Analytics [29] and Google Cloud [13], which act as general domain polarity classifiers both for the English and the Spanish languages.

5.1 Deep Neural Networks

To determine the polarity of the students' opinions, we used several polarity supervised models based on the use of word embeddings and deep learning. Unfortunately, due to the lack of training data, it was not possible to learn robust models specifically for the task described in this paper.

Instead, we used models trained by our research group, for similar tasks related to the social network Twitter [33, 26, 7] both for English and Spanish. The English corpus (including the partitions for training, development and testing purposes) is provided in the Subtask A of Task 4 from SemEval 2017 [33] intended to detect the overall sentiment of a tweet. The Spanish corpus is a combination of two TASS editions (2017 [26] and 2019 [7]) with the aim of increasing the corpus size and taking into account several Spanish variants (including Spain, Mexico, Costa Rica and Uruguay). Due to the masters are opened to international students, several Spanish variants can be used by them, therefore, it is interesting to consider some of these variants during the training phase of the supervised models. In the Spanish case, partitions for training, development and testing were built following a 80%-10%-10% proportion. Table 3 shows some details of the corpora used to train the models.

As input for CNN, TE and BLSTM models, each opinion is represented as a $N \times d$ matrix where each word of the opinion - up to a maximum of N - is represented as a d-dimensional embedding. Depending on the language, different word embeddins are used. For Spanish, 300-dimensional skip gram word embeddings were learned from 87 millions of tweets [16], whereas for English, 400-dimensional skip gram word embeddings from [8] were used.

Table 3 Characteristics of the corpora used to train the Deep Neural Network models (both for English and Spanish).

Task	Set	Total	Positive	Negative	Neutral	None
SemEval (English)	Train	39656	15705 (40%)	6203 (15%)	17748 (45%)	N/A
	Test	12284	2375 (19%)	3972 (32%)	5937 (49%)	N/A
	Total	51940	18080 (35%)	10175 (19%)	23685 (46%)	N/A
TASS (Spanish)	Train	11755	4211 (36%)	4250 (36%)	1321 (11%)	1973 (17%)
	Test	1507	547 (36%)	543 (36%)	155 (11%)	262 (17%)
	Total	13262	4758 (36%)	4793 (36%)	1476 (11%)	2235 (17%)

5.1.1 Convolutional Neural Networks

This architecture is inspired by the work described in [19], which obtained competitive results in text classification tasks such as sentiment analysis or irony detection.

We applied several one-dimensional (the width of the filter is constant and equal to the dimension of the embeddings) convolutions with different height filters, in order to extract the sequential structure of the text. Concretely, heights from 1 to 3 with 256 filters for each height are used. Subsequently, we applied Global Max Pooling to the feature maps in order to extract the most salient features for each region size.

The final decision is carried out by a softmax fully-connected layer. Table 4 show the performance of the models for the test set of the two tasks [11, 31].

Note that for TASS, the distinction between the classes Neutral (*with both positive and negative feelings*) and None (*lack of sentiment*) is made during training and test. However, when the model trained for TASS is applied to our UX evaluation task, both classes are merged. That is, given a test opinion x, $\text{argmax}_y\, p(y|x) \in \{Neutral, None\} \rightarrow y = Neutral$. This is also true for all the other Deep Learning systems presented in this section.

5.1.2 Transformer Encoder

Our system is based on the Transformer [40] model. Initially proposed for machine translation, the Transformer model dispenses with convolution and recurrences to learn long-range relationships. Instead of this kind of mechanisms, it relies on multi-head self-attention, where multiple attentions among the terms of a sequence are computed in parallel to take into account different relationships among them.

On top of the tweet representations, $Nx = 1$ transformer encoders are applied, which relies on multi-head scaled dot-product attention with $h = 8$ different heads and $d_k = d_q = d_v = 64$ attention dimensionality. To do this we used an architecture similar to the one described in [40], including the layer normalization [1] and the residual connections. Due to a vector representation is required to train classifiers on top of these encoders, a global average pooling mechanism was applied to the output of the encoder, and it is used as input to a feed-forward neural network, with only one hidden layer, whose output layer computes a probability distribution over the classes of the task.

Table 4 Performance of the supervised Deep Learning systems on SemEval 2017 Task 4 (English) and TASS (Spanish) for the test set.

	CNN				**TE**				**BLSTM**			
	Acc	*MP*	*MR*	MF_1	*Acc*	*MP*	*MR*	MF_1	*Acc*	*MP*	*MR*	MF_1
Spanish	66.29	55.63	55.07	54.05	66.82	58.40	57.93	57.63	65.49	55.40	52.81	52.71
English	63.88	63.27	62.24	62.59	63.35	62.24	63.97	62.74	64.54	63.43	65.30	64.19

5.1.3 Attentional Bidirectional Long Short Term Memory

The system is based on Bidirectional Long Short Term Memory (BLSTM) [15] [36] with attention mechanisms. On top of the tweet representations, one 256-dimensional BLSTM is applied and, a context vector is computed from the outputs of the BLSTM network following [2]. In this way, the context vector is a weighted sum of the BLSTM outputs, where the weight associated to each output is computed by means of a feed-forward neural network which is jointly trained with all the other components of the system. Then, the context vector is used as input to a feed-forward neural network with one 256 dimensional hidden layer, whose output layer computes a probability distribution over the classes of the task. Again, the Neutral and None classes are merged when the model is applied on the Spanish UX task, due to the distinction between them in the TASS corpus.

Table 4 show the performance of the models for the test set of the SemEval 2017 and TASS 2019 tasks [11, 31]. Again, the Neutral and None classes are merged when the model is applied on the Spanish UX task, due to the distinction between them in the TASS corpus.

5.2 Commercial Systems

5.2.1 MeaningCloud

MeaningCloud is a Software as a Service product [28] that provides a large number of tools, easy to use and to deploy, for text processing, analytics and text/audio mining, with the aim of facilitating the resolution of natural language processing problems to developers. It includes tools for summarization, topic extraction, language identification and sentiment analysis and it supports several languages.

We have used the sentiment analysis module in this work. This module allowed us to use a classifier, trained in a general domain with texts in multiple languages, to determine the global polarity of user opinions on VLEs. Concretely, we use the field *score tag* in the response of the MeaningCloud API, that indicates the global polarity of the text in 6 different levels: strong positive, positive, neutral, negative, strong negative and without sentiment (None). To carry out our experiments, we collapsed the strong sentiments, i.e., strong positive and strong negative are considered as Positive and Negative, respectively. Moreover, the Neutral/None classes are merged in only one class (Neutral).

5.2.2 Google Cloud

Natural Language API of Google Cloud [13] allows to perform several kinds of analysis such as syntactic parsing, entity or sentiment analysis, in general domain texts and also for several languages. It is based on machine learning with the aim of analyzing the structure and the meaning of documents. For sentiment analysis, it computes two values for each document, *score* and *magnitude*. The overall sentiment of the document is computed by the $score \in [-1, 1]$, where negative values refer to Negative sentiment, positive values refer to Positive sentiment and values closer to 0 could suggest the absence of sentiment (None) or the neutralization of positive and negative sentiments (Neutral). In order to distinguish between None and Neutral, the system computes the $magnitude \in [0, \inf[$ which quantifies the sentiment content in the document, so that documents with *score* closer to 0 will be Neutral if its *magnitude* indicates the presence of sentiment ($magnitude > 0$) or None if there is not ($magnitude = 0$). In our case, we only used the *score* value due to in the UX corpus, the Neutral class indicates both situations.

In order to carry out the experimentation and a fair comparison with the supervised systems based on Deep Neural Networks, we fixed a threshold ϵ to a reasonable value of $\epsilon = 0.15$, so that if $-\epsilon \leq score \leq \epsilon$, then the polarity is Neutral; if $score < -\epsilon$, then it is Negative; and, otherwise, the polarity is Positive.

5.2.3 Microsoft Text Analytics

Text Analytics API of Microsoft Azure [29] provides automatic tools to evaluate opinions and topics, such as sentence extraction, entity recognition and sentiment analysis. The system is based on machine learning and it uses different features such as n-grams, word embeddings and part-of-speech tags, being compatible with several languages. The sentiment analysis module computes a $score \in [0, 1]$, being 0 the most Negative value and 1 the most Positive, where values closer to 0.5 suggests Neutral polarity, due to the objectivity or the neutralization of positive and negative sentiments.

In this case, the classification rule is the same as the previously commented for the Google system, but with $\epsilon = 0.05$ and moving the origin from 0 to 0.5; i.e., if $0.5 - \epsilon \leq score \leq 0.5 + \epsilon$, then the polarity is Neutral; if $score < 0.5 - \epsilon$, then it is Negative; and, otherwise, the polarity is Positive.

6 Our Proposal

In this paper, we address the UX evaluation problem in a novel way. To do this, we propose to apply machine learning tools to the users' opinions, expressed freely in natural language and additionally, due to the lack of training data in the UX domain, we propose to use general-purpose and cross-domain systems. A scheme of our proposal is illustrated in Figure 2. We used six different systems to analyze the polarity of the students' opinions about the learning platform. On the one hand, we have considered three general-purpose commercial systems (MeaningCloud [28], Google Cloud [13], and Microsoft Text Analytics [29]). On the other hand, we have used three deep learning cross-domain models (Convolutional Neural Networks,

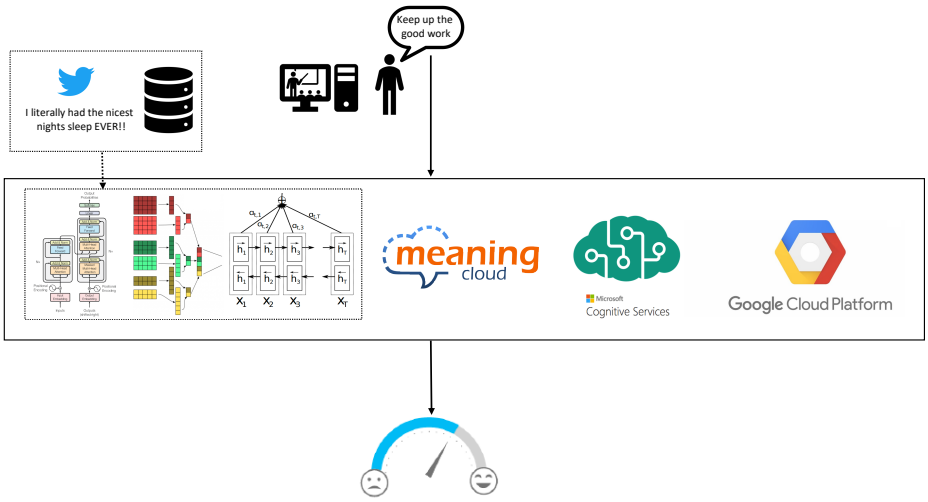

Fig. 2 A scheme of our proposal for evaluating UX about E-learning platforms using general and cross-domain sentiment analysis systems.

Transformer Encoders and Attentional BLSTM models) developed by our team and initially trained for the sentiment analysis problem in Twitter [11, 9, 10].

7 Experimental Results

We applied the cross-domain polarity models presented in Section 5 to the proposed task which consists of determining the polarity (positive, negative, or neutral) of the users' opinions about the learning platform. The opinions, as explained in Section 3, are processed as the whole observation from one user or, if the observation is composed of more than one sentence, each single sentence. The results obtained with each type of polarity models for observations and sentences in the two considered languages are shown in Tables from 5 to 8.

First of all, it is important to highlight the good results obtained by all the systems for the Positive class (F_1^{pos} row in all tables), which is also the highest frequency class, as observed in Table 2. However, the Neutral class has a totally different behavior. All systems obtained much worse results for this class, with differences of almost 50 points for the same system between F_1^{pos} and F_1^{neu}. It should be noted that the *Acc* confidence intervals (excluded from the tables) are wide, ranging from $\pm 3.23\%$ to $\pm 9.38\%$, mainly due to the small size of the corpora. Therefore, there is a considerable amount of uncertainty on the results and the following conclusions should be taken with caution.

Regarding the experiments using the Spanish observations, BLSTM, TE and MeaningCloud systems achieved the best results (see Table 5), being the TE system the best one in terms of MF_1 and the MeaningCloud system the best one considering only *Acc*. In the case of the Spanish sentence level experiments (Table 6), our BSLTM system achieved the best performance. As in general in all

Table 5 Experiments at observation level on the Spanish samples.

	Spanish Observations					
	CNN	**TE**	**BLSTM**	**Mean.C.**	**Google**	**Microsoft**
Acc	76.68	77.94	78.15	78.57	77.73	70.59
MF_1	53.06	61.72	58.55	59.47	55.64	51.98
F_1^{pos}	87.52	88.86	88.24	88.32	87.72	83.21
F_1^{neg}	66.33	66.67	68.82	68.35	65.25	60.87
F_1^{neu}	5.33	29.63	18.60	21.74	13.95	11.86

Table 6 Experiments at sentence level on the Spanish samples.

	Spanish Sentences					
	CNN	**TE**	**BLSTM**	**Mean.C.**	**Google**	**Microsoft**
Acc	78.19	76.49	80.07	76.15	76.49	68.14
MF_1	55.38	58.73	61.62	55.69	51.71	49.35
F_1^{pos}	87.30	87.41	87.65	87.75	87.20	82.38
F_1^{neg}	70.83	67.92	74.13	66.94	60.71	55.74
F_1^{neu}	8.00	20.87	23.08	12.39	7.23	9.93

the experimentation, the Positive class (both in observations and sentences experiments) obtained the highest F_1 results, and the Neutral class the one with the lowest. At sentence level, our three cross-domain models trained with tweets showed a better behavior than those from general-domain commercial systems. We hypothesize that this is due to the greater similarity of sentences and tweets compared to the whole observation, with larger length than one tweet. That is, the test samples at sentence level are more similar to those used to learn our models and therefore these models perform better.

Regarding the experiments using the English samples, the best accuracy and MF_1 values were obtained by Google, Microsoft and our BLSTM model. At observation level, Google system obtained the best results, whereas at sentence level the most competitive system was BLSTM. The worst results were obtained by the CNN and the MeaningCloud system.

Microsoft model was the best detecting Negative samples. The TE system was the worst system detecting the negative samples at observation level, while the CNN model was the worst for detecting this class at the sentence level. As in the case of the Spanish samples, in the English samples the Neutral class was the class with the worst results, being the performance of all the systems for this class much lower than for the other classes. The Neutral samples are more complex in structure due to the neutralization of positive and negative elements in the same unit or the absence of polarity. If we compare the Neutral results for English and Spanish it is possible to see that the commercial system, with the exception of MeaningCloud, achieved results that are slightly higher for the English opinions, suggesting that these systems have been better adjusted for processing English documents.

It is important to highlight the good behavior obtained by our cross-domain models in comparison with the general-domain commercial systems that have been used in the experimentation. They are quite competitive despite having been only trained with tweets. Different from user opinions expressed in VLE, where a for-

Table 7 Experiments at observation level on the English samples.

	English Observations					
	CNN	**TE**	**BLSTM**	**Mean.C.**	**Google**	**Microsoft**
Acc	57.01	57.94	68.22	60.75	66.36	71.03
MF_1	49.01	47.55	56.02	50.87	58.54	56.13
F_1^{pos}	74.78	78.33	86.61	76.42	77.69	82.54
F_1^{neg}	60.00	43.48	52.17	57.14	67.93	69.84
F_1^{neu}	12.24	20.83	29.27	19.05	30.00	16.00

Table 8 Experiments at sentence level on the English samples.

	English Sentences					
	CNN	**TE**	**BLSTM**	**Mean.C.**	**Google**	**Microsoft**
Acc	57.61	61.41	70.65	63.04	69.57	73.91
MF_1	50.16	51.44	58.14	49.60	55.55	54.35
F_1^{pos}	80.23	81.97	89.12	79.19	80.42	82.13
F_1^{neg}	53.57	54.70	63.87	57.38	72.59	70.92
F_1^{neu}	16.67	17.65	21.43	12.25	13.64	10.00

mal communication is carried out addressing a set of topics related to the course that they have taken, the tweets are informal and they express opinions of many different topics in a way influenced by the behavior of the Twitter social network (slang, user mentions, hashtags, lexical errors, elongations, etc.). This seems to suggest that there are related properties among the opinions expressed on VLE and those expressed on Twitter.

As stated in [38]: "However, sentiment accuracies even for binary positive/negative classification for single sentences has not exceeded 80% for several years. For more difficult multiclass case including a neutral class, accuracy is often below 60% for short messages on Twitter (Wang et al., 2012)". The accuracies of our models are in the state of the art for sentiment analysis in Twitter, and the evaluation metrics are also very similar when the models are applied to our UX dataset, which is a multiclass problem with 3 classes (Positive, Negative, Neutral).

8 Conclusions and Future Work

In this paper, we have presented a sentiment analysis task to opinions written in natural language extracted from questionnaires of postgraduate biomedical and MOOCs online learning students. As stated in the introduction, the application of sentiment analysis tools on UX comments will provide a better and accurate understanding on human needs in the interaction with VLEs. Three Learning Management Systems have been evaluated, both in Spanish and English, applying cross-domain polarity models trained with a corpus of a different domain (tweets for each language) and general models for the language. The obtained results are very promising and they give an insight to keep going the research of applying sentiment analysis tools on User eXperience evaluation.

The ultimate goal is to develop further tools of automatic feed-back of user perception for designing virtual learning environments valued by users for its usability, quality and pleasure of use. For this, as a future work, we will address

automatic aspect detection (*pleasure of use, pleasure of learning, learning platforms, video, slides, usability, etc.*) and its polarity will be analyzed in order to capture relevant aspects that influence on the students and, possibly, were not considered during the questionnaire, or to analyze which aspects of a course tend to be more negative or positive for the students. Finally, we are now continuing working with questionnaires on MOOCs to collect larger amounts of data in order to train models for the task of sentiment analysis at global and aspect level on VLEs. A transfer learning approach from models trained with data of other domains could also be applied in order to have more robust models for the task.

Acknowledgments

The authors acknowledge the valuable help provided by Carlos Turró-Ribalta and Ignacio Despujol-Zabala in providing the MOOCs samples.

References

1. Ba J, Kiros JR, Hinton GE (2016) Layer normalization. ArXiv abs/1607.06450
2. Bahdanau D, Cho K, Bengio Y (2015) Neural Machine Translation by Jointly Learning to Align and Translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, URL http://arxiv.org/abs/1409.0473
3. Baziotis C, Pelekis N, Doulkeridis C (2017) Datastories at SemEval-2017 Task 4: Deep LSTM with Attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, pp 747–754
4. Cliche M (2017) BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, pp 573–580, DOI 10.18653/v1/S17-2094, URL https://www.aclweb.org/anthology/S17-2094
5. Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1):37
6. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, DOI 10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423
7. Diaz-Galiano MC, et al. (2019) Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR-WS, Bilbao, Spain, CEUR Workshop Proceedings, pp 550–560
8. Godin F, Vandersmissen B, De Neve W, Van de Walle R (2015) Multimedia lab @ ACL WNUT NER shared task: Named entity recognition for

twitter microposts using distributed word representations. In: Proceedings of the Workshop on Noisy User-generated Text, Association for Computational Linguistics, Beijing, China, pp 146–153, DOI 10.18653/v1/W15-4322, URL https://www.aclweb.org/anthology/W15-4322

9. González J, Pla F, Hurtado L (2018) Elirf-upv en TASS 2018: Análisis de sentimientos en twitter basado en aprendizaje profundo (elirf-upv at TASS 2018: Sentiment analysis in twitter based on deep learning). In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018, pp 37–44, URL http://ceur-ws.org/Vol-2172/p2_elirf_tass2018.pdf
10. González J, Hurtado L, Pla F (2019) Elirf-upv at TASS 2019: Transformer encoders for twitter sentiment analysis in spanish. In: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, pp 571–578, URL http://ceur-ws.org/Vol-2421/TASS_paper_2.pdf
11. González JÁ, Pla F, Hurtado LF (2017) ELiRF-UPV at SemEval-2017 Task 4: Sentiment Analysis using Deep Learning. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, pp 723–727, DOI 10.18653/v1/S17-2121, URL https://www.aclweb.org/anthology/S17-2121
12. González JÁ, Hurtado LF, Pla F (2019) ELiRF-UPV at TASS 2019: Transformer Encoders for Twitter Sentiment Analysis in Spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR-WS, Bilbao, Spain, CEUR Workshop Proceedings
13. GoogleCloud (2019) Cloud Natural Language API. URL https://cloud.google.com/natural-language/
14. Hassenzahl M, Tractinsky N (2006) User experience - a research agenda. Behaviour & Information Technology 25(2):91–97, DOI 10.1080/01449290500330331
15. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780, DOI 10.1162/neco.1997.9.8.1735, URL http://dx.doi.org/10.1162/neco.1997.9.8.1735
16. Hurtado Oliver LF, Pla F, González Barba J (2017) ELiRF-UPV at TASS 2017: Sentiment Analysis in Twitter based on Deep Learning. In: TASS 2017: Workshop on Semantic Analysis at SEPLN, pp 29–34
17. IBM (2019) Natural Language Understanding. URL https://www.ibm.com/watson/services/natural-language-understanding/
18. ISO 9241-210:2019 (2019) Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. International Standardization Organization (ISO), URL https://www.iso.org/standard/77520.html
19. Kim Y (2014) Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp 1746–1751, URL http://aclweb.org/anthology/D/D14/D14-1181.pdf
20. Krippendorff K (2004) Reliability in content analysis. Human communication research 30(3):411–433

21. Kujala S, Roto V, Väänänen-Vainio-Mattila K, Karapanos E, Sinnelä A (2011) UX Curve: A method for evaluating long-term user experience. Interacting with Computers 23(5):473–483
22. Liu B (2012) Sentiment Analysis and Opinion Mining. A Comprehensive Introduction and Survey. Morgan & Claypool Publishers
23. Liu B, Hu M, Cheng J (2005) Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '05, pp 342–351, DOI 10.1145/1060745.1060797, URL http://doi.acm.org/10.1145/1060745.1060797
24. Luque FM (2019) Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR-WS, Bilbao, Spain, CEUR Workshop Proceedings
25. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp 55–60, URL http://www.aclweb.org/anthology/P/P14/P14-5010
26. Martínez-Cámara E, Díaz-Galiano M, García-Cumbreras M, García-Vega M, Villena-Román J (2017) Overview of TASS 2017. In: Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017), CEUR-WS, Murcia, Spain, CEUR Workshop Proceedings, vol 1896
27. MeaningCloud (2019) Demo de Analítica de Textos. URL https://www.meaningcloud.com/es/demos/demo-analitica-textos
28. MeaningCloud (2019) MeaningCloud: Servicios web de analítica y minería de textos. URL https://www.meaningcloud.com/
29. MicrosoftAzure (2019) Text Analytics API. URL https://azure.microsoft.com/es-es/services/cognitive-services/text-analytics/
30. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp 79–86
31. Pla F, Hurtado LF (2018) Spanish sentiment analysis in Twitter at the TASS workshop. Language Resources and Evaluation 52(2):645–672, DOI 10.1007/s10579-017-9394-7, URL https://doi.org/10.1007/s10579-017-9394-7
32. Rauschenberger M, Schrepp M, Cota MP, Olschner S, Thomaschewski J (2013) Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ). Example: Spanish Language Version. International Journal of Interactive Multimedia and Artificial Intelligence 2(1):39–45, DOI 10.9781/ijimai.2013.215, URL http://www.ijimai.org/journal/sites/default/files/files/2013/03/ijimai20132_15_pdf_35685.pdf
33. Rosenthal S, Farra N, Nakov P (2017) SemEval-2017 Task 4: Sentiment Analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, pp 502–518, DOI 10.18653/v1/S17-2088, URL https://www.aclweb.org/anthology/S17-2088
34. Sadr H, Pedram MM, Teshnehlab M (2019) A Robust Sentiment Analysis Method Based on Sequential Combination of Convolutional and Re-

cursive Neural Networks. Neural Processing Letters 50:2745–2761, DOI 10.1007/s11063-019-10049-1
35. Sanchis-Font R, Castro-Bleda M, González J (2019) Applying Sentiment Analysis with Cross-Domain Models to Evaluate User eXperience in Virtual Learning Environments. In: Rojas I, Joya G, Catala A (eds) Advances in Computational Intelligence. IWANN 2019, Lecture Notes in Computer Science, vol 11506, Springer, Cham, pp 609–620
36. Schuster M, Paliwal K (1997) Bidirectional recurrent neural networks. Trans Sig Proc 45(11):2673–2681, DOI 10.1109/78.650093, URL http://dx.doi.org/10.1109/78.650093
37. Scott WA (1955) Reliability of Content Analysis: The Case of Nominal Scale Coding. Public Opinion Quarterly 19(3):321–325, DOI 10.1086/266577, URL https://doi.org/10.1086/266577, http://oup.prod.sis.lan/poq/article-pdf/19/3/321/5264776/19-3-321.pdf
38. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, pp 1631–1642, URL https://www.aclweb.org/anthology/D13-1170
39. Turney PD (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: ACL, pp 417–424, DOI http://www.aclweb.org/anthology/P02-1053.pdf
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., USA, NIPS'17, pp 6000–6010, URL http://dl.acm.org/citation.cfm?id=3295222.3295349
41. Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) OpinionFinder: A system for subjectivity analysis. In: Proceedings of HLT/EMNLP on Interactive Demonstrations, Association for Computational Linguistics, pp 34–35
42. Zaharias P, Mehlenbacher B (2012) Editorial: Exploring User Experience (UX) in Virtual Learning Environments. Int J Hum-Comput Stud 70(7):475–477, DOI 10.1016/j.ijhcs.2012.05.001, URL http://dx.doi.org/10.1016/j.ijhcs.2012.05.001
43. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1253