When the best reviews are not placed between extremes

Eva Blanco-Mallo*, João Carneiro†, Goreti Marreiros†, Beatriz Remeseiro‡§ and Verónica Bolón-Canedo*§
*Department of Computer Science and Information Technologies
Universidade da Coruña, CITIC, A Coruña, Spain
Email: eva.blanco@udc.es, veronica.bolon@udc.es
†GECAD - Research Group on Intelligent Engineering and Computing for
Advanced Innovation and Development, Institute of Engineering
Polytechnic of Porto, Porto, Portugal
Email: jrc@isep.ipp.pt, mgt@isep.ipp.pt
‡Department of Computer Science
Universidad de Oviedo, Gijón, Spain
Email: bremeseiro@uniovi.es

***Abstract*—Several research studies have demonstrated the strong influence that online reviews exert on consumers' purchasing decisions. Specifically, those with extreme opinions, both favorable and unfavorable, are often considered more useful. This paper is focused on enhancing the detection of extreme reviews through sentiment analysis. For this purpose, a real scenario is taken into account, using the examples of all classes and dealing with the imbalance between them, which is characteristic in online reviews. The main objective is to carry out the classification with a high certainty and incurring as few errors as possible in relation to the examples belonging to the rest of the classes. Therefore, the emphasis is on the quality of the predictions rather than on the quantity. Using XLNet, we show how the transfer of knowledge extracted from the source domain (i.e., the extreme reviews) improves their detection regarding the overall number of errors made in the target domain (i.e., multi-class classification).**

***Index Terms*—extreme reviews, sentiment analysis, transfer learning, transformer models**

I. Introduction

The vast majority of consumers turn to social networks to seek information about products or services before purchasing them. Due to the multitude of options available, making a choice can become a complicated and overwhelming process. The use of information from other consumers can help reduce the alternatives and thus the complexity of the decision. This phenomenon, known as social proof heuristic, refers to the tendency to adopt the option preferred by the majority when the decision is unclear [1]. In fact, a widely adopted resource in marketing is to use statements such as "the preferred by consumers" or "the best-selling product", trying to influence the audience by alluding to the tastes of the majority.

Numerous studies confirm the strong influence that online reviews have on purchasing decisions. Lee et al. [2] analyzed the effect of negative reviews on consumers' attitudes toward the product. One of their conclusions was that consumer attitudes become more unfavorable as the proportion of negative reviews increases. In fact, Chevalier and Mayzlin [3] demonstrated how extreme negative reviews hurt book sales on Amazon, while positive reviews led to an increase in sales. Participants of a study [4] based on a survey of Internet users stated that they were willing to pay between 20% and 99% more for a product rated 5 stars than for a product rated 4 stars, depending on the product category. Anderson [5] also found, by analyzing TripAdvisor data, that the percentage of customers who consulted reviews before booking a hotel room and the number of reviews consulted increased steadily over time. On the other hand, if a hotel increases its rating by 1 point on a scale of 1 to 5, the hotel can increase its price by 11.2% without noticing a decrease in occupancy. In addition, a 1% increase in a hotel's online reputation score leads to an increase in hotel price and occupancy.

For this reason, a great effort has been made to find out what are the main factors in a review that make it useful in a consumer's purchasing decision. Aspects such as length, readability, and extremity of the reviews are often the subject of study. Extreme reviews refer to evaluations with the highest and lowest rating scores in a given range. For example, on TripAdvisor, these reviews would be the ones with a score of 5 and 1 stars, respectively. There are many studies in the literature that highlight the impact that these types of reviews have on the consumers' decision-making process. Individuals tend to focus on extreme values as reference points because they are perceived as less ambiguous and more diagnostic [6]. Even for an expert it would be challenging on many occasions to distinguish a 5-star review from a 4-star review. Since the

This work has been supported by the National Plan for Scientific and Technical Research and Innovation of the Spanish Government (Grant PID2019-109238GB, subprojects C21 and C22), by the Spanish Ministry of Science and Innovation (Grant FPI PRE2020-092608), and by the Xunta de Galicia (Grant ED431C 2018/34) with the European Union ERDF funds. CITIC, as Research Center accredited by Galician University System, is funded by "Consellería de Cultura, Educación e Universidades from Xunta de Galicia", supported in an 80% through ERDF Funds, ERDF Operational Programme Galicia 2014-2020, and the remaining 20% by "Secretaría Xeral de Universidades" (Grant ED431G 2019/01).

§These authors jointly supervised this work.

boundaries on the rating scale are undefined and subjective, according to Pang and Lee [7], extreme opinions are the only ones that can be considered natural, transparent, and unambiguous positive or negative statements. In fact, several studies demonstrate the high correlation between review helpfulness and extremity. Park and Nicolau [8] reported that users perceive extreme ratings as more useful and pleasant than moderate ratings regarding restaurant services, with negative extremes being more useful than positive ones. Filieri et al. [6] also concluded that extreme reviews are more useful than moderate reviews in the hotel industry. In Fang et al.'s research [9], the authors also reviewed expressing extreme sentiments received more helpfulness votes.

Along with the increased impact of reviews also came a considerable growth in deceptive reviews, which aim to manipulate users' perception of a brand, either favorably (the brand itself) or unfavorably (the competitors). According to Luca and Zervas [10], Yelp's algorithm marks one out of every five reviews as fake. Extreme reviews are closely related to the detection of deceptive opinions, based on the premise that false reviews tend to exaggerate emotions [11]. Therefore, research in this area would contribute to mitigate the effects of unfair competition and distrust that characterize this environment. In this context, extreme review identification could also be used to detect bots on social networks, designed to amplify and spread biased, sometimes false, stories with the aim of influencing public opinion. Furthermore, it would allow companies to analyze the strengths and weaknesses of their products or services from the consumer's point of view. Thus, it would be possible to discover those factors that can be highlighted to increase sales or those that need to be corrected to prevent them from decreasing. In addition, identifying extreme negative opinions offers managers the possibility to act and try to alleviate the negative effect [6], [12], as well as to spread the positive ones in order to expand their impact.

As can be seen, extreme opinions are the subject of study and a determining factor in numerous research questions related to online reviews, and its impact on consumers' decision-making has been demonstrated on several publications. However, to the best of our knowledge, very few studies were specifically dedicated to its detection. The use of sentiment analysis through machine learning techniques to predict ratings in online reviews was successfully applied in the literature [13]–[15]. Our main goal is to improve the detection of extreme reviews by addressing the classification of all classes through deep learning. Specifically, by conducting sentiment analysis using a transformer model [16], [17] and applying transfer learning extracted from the source samples (i.e., the extreme reviews). Proposing a two-stage training, we aim at increasing the certainty of classifying a review as extreme, either positive or negative, minimizing the error with respect to the rest of the classes. Our main contributions are:

- Tackling the classification of extreme reviews from a real scenario, i.e., considering all the available samples and deal with the imbalance that characterizes them.
- Contrasting the effectiveness of transfer learning by starting from the knowledge achieved using only extreme reviews and applying it to the final multi-class classification to improve the certainty in its subsequent detection.

The rest of the paper is organized as follows: Section II provides a review of the literature related to the detection of extreme reviews, plus a brief overview of different techniques currently used to perform sentiment analysis in this context. Section III details the datasets, the proposed methods to detect extreme reviews, and the evaluation procedure. Section IV presents the experiments carried out, as well as the results achieved. Finally, Section V presents the main conclusions drawn from the present study and future lines of work.

II. Related work

To the best of our knowledge, only two publications specifically address the classification of extreme reviews. In the first one, Almatareneh et al. [18] implemented a binary classification task to identify the most negative opinions versus the rest. They tried different strategies to obtain the feature vector representing each review, comparing the performance obtained with unigram features, part of speech (PoS) features, syntactic patterns, and sentiment Lexicons. Then, a support vector machine (SVM) was used to perform the classification. In the experiments conducted, on movie review datasets, low results were obtained in the identification of the most negative class. Authors believe that this was due to the fact that the boundary between very negative and not very negative is more blurred than between positive and negative. In the second one, the same authors [19] reassessed the detection of extreme reviews by testing different feature representations, namely, those based on the frequency of occurrence of terms (TF-IDF and CountVectoricer), neural-based representations (Doc2Vec), and using set of textual features (SOFT), and sentiment lexicons. In the experiments, hotel reviews were used and a balanced set of 1, 2, 4 and 5 star reviews are selected. Ratings of 1 and 2 stars are considered very negative and not very negative examples respectively. Similarly, 5-star ratings are treated as very positive and 4-star ratings as not very positive. Their approach consists in implementing two binary classification tasks using SVMs: one to distinguish the very negative examples from the not very negative ones and the other to distinguish the very positive ones from the not very positive ones. They found that the two feature representations that performed best in extreme classification are neural-based embeddings and textual features. In particular, the best performance in the detection of negative extremes was achieved by Doc2Vec and by SOFT regarding positive extremes. The main difference of the aforementioned research with respect to our approach is that we address the task from a more real and complex scenario, addressing the classification of all classes and tackling the class imbalance. In addition, we focus on improving the certainty of the predictions of extreme reviews and minimizing errors with respect to the other classes.

Another work related to the detection of extreme reviews is the one by Moon et al. [11], who consider that fake reviews are characterized by the use of extreme words, as people tend

to exaggerate when they lie. In order to differentiate a fake review from a genuine one, they rely on the frequency of use of strongly positive or strongly negative words. Using hotel reviews from different platforms, the extreme terms present in the dataset were first identified. Next, a cleaning process was carried out and the remaining terms were reviewed by three researchers to make the final selection. Finally, a trust measure based on the frequency of use of these terms was used to obtain the probability of a review to be genuine.

On the other hand, numerous research studies focused on using deep learning approaches to conduct sentiment analysis for review classification [13], [20], [21]. Recently, the focus is being put on using transformer based models. This type of neural networks, presented by Vaswani et al. [22], are mainly composed of a multi-head self-attention mechanism combined with an encoder-decoder structure. A self-attention head receives a sequence and maps it to a new one of the same length. It learns the dependencies between tokens using trainable weight matrices and produces the output sequence by computing a weighted average of entries of the input. The main difference of these models compared to the neural networks traditionally used for these tasks, such as recurrent neural networks or convolutional neural networks, is that the entire text stream is used as input, thus taking into account the entire context. Zhang et al. [23] analyzed the performance of several transformed-based architectures in sentiment analysis for software engineering, like Stack Overflow or GitHub comments. In their experiments, they concluded that pre-trained transformer models, including BERT [24] and XLNet [17] among others, achieve better results than the best performing tools identified in previous studies. They also report that adjusting the weights of the pre-trained models to the downstream task improves overall performance. Michev et al. [25] present a comparative study of different NLP-based methods for sentiment analysis in finance. They analyzed the performance of transformers models compared to different machine and deep learning tools, such as support vector machines, extreme gradient boosting, and recurrent and convolutional neural networks, including different lexicon-based approaches. According to the experimental results, transformers showed superior performances. Alaparthi and Mishra [26] also discussed the effectiveness of transformer models compared to more traditional techniques in sentiment analysis, using movie reviews. Specifically, the performance of BERT was analyzed against logistic regression and long short-term memory (LSTM) networks, among others. Their results confirm once again the superiority of transformer models.

III. Materials and Methods

The aim of this paper is to improve the quality of extreme review detection through transfer learning. The focus is on the use of knowledge drawn from the source domain, i.e., obtained only from extreme reviews. The intention is that the network first learns the features defining these extremes by facing a simpler task (binary classification), and then apply it in a more complex context, i.e. considering examples of all classes (multi-class classification). For this purpose, the classification is addressed through sentiment analysis using a transformer-based model.

A. Datasets

Three datasets of different sizes and domains were considered in the experimental study. In order to facilitate the comparison of future approaches, a publicly available dataset [27], which consists of hotel reviews extracted from TripAdvisor, was selected. The other datasets used [28] were also extracted from TripAdvisor, but they are composed of restaurant reviews from New York City and New Dheli. In the three datasets, the samples are composed of a text review associated with a rating in the range $[1, 5]$. The extreme reviews are those rated with the highest and lowest scores, i.e., 1 or 5 stars. For experimentation purposes, the datasets were randomly split in train and test sets (80% and 20%, respectively). Due to the experimentation requirements, the initial training set was split again, thus obtaining a new train set and a validation set (90% and 10%, respectively). Due to the network architecture, it is necessary to fix the input size of the sentences. Selecting this value based on the maximum word length of the reviews would lead to sparse vectors. Similarly, selecting the average word value of the reviews would lead to discarding too many samples. Therefore, we relied on the upper fence value, which in statistics represents the upper limit beyond which the values found can be considered outliers. The upper fence was calculated for each class in all the datasets and the mean of their maximum values was computed, obtaining 332 as a result. Rounding, a maximum number of 300 words was selected, so all the reviews that exceeded this value were discarded. The size of the different partitions for each class in each dataset is shown in Table I.

Figure 1 displays the polarity versus subjectivity of the reviews of each class, showing the New York dataset as an example. Due to space constraints, only this dataset, which is the largest one, was selected to show the trend followed by the review. For this purpose, the Textblob library [29] was used, which analyzes the sentiment of the sentences using a lexicon composed of negative and positive words. A negative polarity indicates that the sentiment of the analyzed sentence is negative and a positive polarity indicates that it is positive. Similarly, the library allows to analyze the subjectivity present in each sentence, where 0 means that the review is objective and 1 the opposite. As can be seen, some of the extreme negative reviews have a positive polarity, and this also applies to the extreme positive reviews. There is also a large overlap between intermediate classes. One of the main problems in sentiment analysis is that opinions are subjective and everyone has their own priorities when giving an assessment. Even for a human, it is tremendously complicated to classify reviews according to their rating, especially to discern between immediate classes. For example, in the work of Lin et al. [30] in which sentiment analysis was performed for software engineering, there was 18.6% disagreement in the manual evaluation of Stack Overflow sentences. Similarly, Murgia et al. [31] explored

TABLE I
NUMBER OF SAMPLES PER DATASET FOR THE THREE PARTITIONS: TRAIN, VALIDATION, AND TEST.

	Hotel Reviews				Restaurant Reviews New Dheli				Restaurant Reviews New York			
	Train	Val	Test	Total	Train	Val	Test	Total	Train	Val	Test	Total
5 stars	6269	692	1737	8698	25657	2853	7129	35639	63520	7067	17636	88223
4 stars	4089	454	1139	5682	14337	1585	3977	19899	28725	3200	7973	39898
3 stars	1468	157	414	2039	4311	479	1199	5989	8416	939	2340	11695
2 stars	1185	132	336	1653	1226	140	345	1711	2807	316	777	3900
1 star	965	104	270	1339	1198	134	334	1666	2037	236	558	2831
Total	13976	1539	3896	19411	46729	5191	12984	64904	105505	11758	29284	146547

Fig. 1. Polarity versus subjectivity for each class in the Restaurant Reviews New York dataset.

the emotions present in the comments made by a software development work team. In their research, they tried to detect the presence or absence of different emotions, such as joy, love, surprise, or sadness. Only in approximately 46% of the comments experts agreed on their ratings, and they typically agreed on the absence of emotions in simple comments, such as "commited" or "done". Therefore, when detecting extreme reviews, a minimal error rate is to be expected. However, the goal is to reduce this rate to the minimum.

B. Methods

A generalized autoregressive pre-training method was used to perform the sentiment analysis, XLNet [17]. This network is the state-of-the-art in many natural language processing tasks, such as natural language inference, question answering, document ranking, and also sentiment analysis. In addition to the advantages of transformers mentioned in Section II, XLNet incorporates a mechanism called permutation language modeling, learning the dependencies between all combinations of tokens in the input. Thus, it is able to capture bidirectional context, since it maximizes the expected log likelihood over all possible permutations in a sequence.

In these models, transfer learning is a key factor. The intuition behind transfer learning is that the knowledge learned in solving a previous task could help in solving new tasks. Starting from a task T that we have to solve in a given domain D, the knowledge achieved in the source domain D_S for solving the task T_S could help in learning the target domain D_T focused on solving the task T_T. Transformers exhibit tremendously complex architectures, with hundreds of millions of parameters. Therefore, training such a model from scratch with small datasets would result in overfitting. Hence, it is better to use pre-trained models in larger source domains and adjust the parameters later with the target task data, a process called parameter fine-tuning. Thus, XLNet-Base (12 layers of transformer blocks, 768 hidden layers, and 12 self-attention heads) with sentence classification weights available in the Huggingface API [16] is used as a baseline. This network uses the databases BookCorpus, English Wikipedia, Giga5, ClueWeb 2012-B, and Commom Crawl for pre-training.

Our approach is also based on transfer learning, in the sense that we use the knowledge extracted from the extreme reviews (source domain) and transfer it to face their classification against all classes (target domain). For this purpose, a two-stage training is proposed[1] (see Figure 2). The first stage consists in solving a binary classification task (T_S) using only the extreme reviews (D_S). During the second stage, a multi-class classification task (T_T) is carried out, where the goal is to detect the extreme samples among all the classes (D_T). The steps followed in the process are detailed below.

1) Data pre-processing: Prior to the training procedure, data pre-processing is conducted. In the network, each word is identified as a token. In order to unify them and avoid that equal words are represented by different tokens, all

[1]https://github.com/evablanco/extreme_reviews_detection

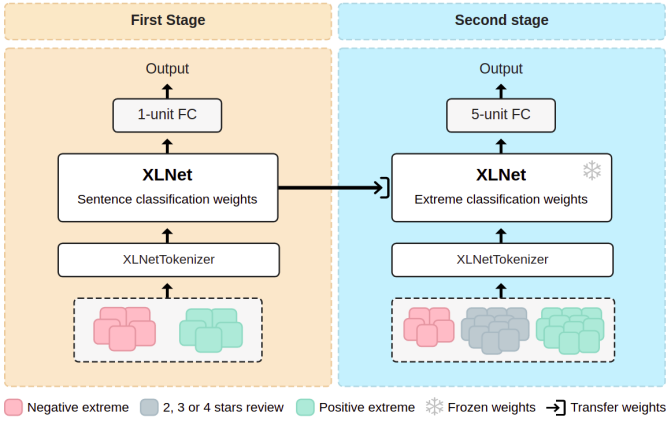

Fig. 2. Training phases of the proposed method.

of them were transformed to lowercase. Next, symbols and special characters were removed. However, periods, commas, and question and exclamation marks were retained, as they contribute to the meaning of the sentences. A common step in this context is to eliminate stop-words and carry out lemmatization. In this case, they were not performed since the goal of the transformer models is to capture the bidirectional context through its different attention and permutation mechanisms. Therefore, the overall structure is preserved. Finally, XLNetTokenizer, available also in the Huggingface API [16], was used to tokenize the input sequences.

2) Learning the features from the source domain: The first training stage is designed to learn the features that define the source domain, i.e., the extreme reviews. Therefore, we define a binary classifier trained only with the 1 and 5 star reviews. To avoid the large imbalance between the two classes (see Table I), an undersampling process is performed on the majority class (5 star). To this end, the same number of extreme positive reviews is randomly selected as the number of extreme negative reviews available.

The model architecture is composed of a XLNet followed by a 1-unit fully connected (FC) layer with the sigmoid activation function. Regarding the training process, a fine-tuning approach is followed, using a XLNet pre-trained for sentence classification and adjusting its weights to the downstream class. Learning is optimized according to the accuracy and monitored using the validation set, applying an early stopping strategy and restoring the state with the highest performance. The convergence of pre-trained models is usually fast, so we used a maximum of 10 epochs and a patience of 3, which is the number of training epochs set by the authors of BERT [24] and XLNet [17].

3) Transferring the features to the target domain: The aim of the second training stage is to identify the extreme reviews using the features learned in the previous step. For this purpose, a multi-class classification task is conducted, using the reviews belonging to all classes. The architecture used is similar to the previous one, using a XLNet followed by a 5-unit FC layer with the softmax activation function.

In this case, a feature-based approach is followed, which consists in using the pre-trained model as feature extractor. That is, we used the XLNet weights obtained in the first step, which were frozen during the second training stage; and only the weights of the 5-unit FC layer were adjusted. As a result, one part of the model performs the feature extraction, specially trained for the detection of the extreme classes; and the other part the multi-class classification. As shown in Figure 1, reviews with intermediate ratings are also characterized by positive or negative polarities. The intention is that the more extreme features a sample presents, negative or positive, the higher the probability of belonging to that class, extreme negative or extreme positive, respectively. In this case, to cope with the complexity present in a real scenario, undersampling was not performed. Instead, we followed the strategy of using class weights in the loss function. As in the previous stage, early stopping was used with the same hyper-parameters, a patience of 3 and a maximum of 10 epochs. Due to the high class imbalance, the training was monitored with the balanced accuracy, which is the average of the recall on each class.

C. Evaluation

In this paper we rely on certainty, considering certainty as the probability with which an example is assigned to the correct class [32]. This practice is common in other contexts, such as sports betting or financial investments. The idea is that avoiding mistakes is fundamental, even though significantly fewer correct examples are detected. The amount of online reviews available is often overwhelming, so it would be more desirable to get a lower number of extreme reviews as long as this number is more accurate. Consequently, we rely on the quality of the predictions (i.e., with the lowest possible error) and not on the quantity obtained.

Thus, to evaluate the models, only the accuracy obtained in the classification of the extreme examples was taken into account. Regarding certainty, only hits obtained with a certainty ≥ 0.9 were evaluated. In the case of binary models, a threshold was established on the probability obtained in the last layer. Therefore, all examples classified with probability ≥ 0.9 are considered positive extremes and those with probability ≤ 0.10 as negative extremes. Similarly, in the case of multi-class models, hits in extreme examples are only counted if the probability of the corresponding unit is ≥ 0.9. In addition to the accuracy of the extreme classes, errors committed in classifying the rest of the classes as extreme reviews were also taken into account, categorized as follows:

- Undesired errors: 4-star reviews classified as positive extremes and 2-star reviews classified as negative extremes.
- Critical errors: 3-star reviews categorized as positive or negative extreme reviews.
- Very critical errors: 1- or 2-star reviews classified as positive extremes, and 4- and 5-star reviews classified as negative extremes.

Finally, to evaluate the performance of the models tested in the experimentation, a multi-criteria decision analysis (MCDA) was carried out. For this purpose, the weighted sum

model (WSM) method was chosen, available in the Scikit-Criteria library [33]. This method computes the importance of an alternative as the sum of the different weighted criteria. The evaluation criteria considered are the hits and the different errors mentioned above, maximization and minimization criteria, respectively. For the analysis, equal weights were given to all criteria. Prior to applying the method to the results, the minimization objectives were inverted by calculating the inverse value of each criterion. Next, the values were normalized by dividing each criterion by the total sum of all criteria.

IV. Experimental results

In order to contrast the effectiveness of the transfer of learning extracted from the extreme revisions, as well as the methods applied in the proposed model, different experiments were carried out as detailed below. Finally, the multi-criteria analysis is conducted among all the approaches. For the training process, a batch size of 8 was set and Adam [34] was used as optimizer, with a learning rate of $2e-5$ as recommended by the XLNet authors. All experiments were performed using an NVIDIA RTX 3080.

A. Feature-based vs fine-tuning

The first experiment consists in analyzing whether the feature-based approach followed in the second phase of training achieves better performance than the fine-tuning approach (first training phase). Therefore, the intention is to analyze if, once the knowledge extracted from the extreme classes has been obtained, it is necessary to adjust again the weights of the XLNet layers to face the multi-class classification. The results obtained with both methods are presented in Table II, where PM stands for our proposal with the feature-based strategy and FTA refers to the fine-tuning strategy. As can be seen, although FTA increases the number of hits in the classification of extreme classes, it also increases the number of errors incurred. Although both approaches show an error rate of less than 1% for critical and very critical errors in all cases, with the immediate classes the performance differs considerably. In the New Dheli restaurant dataset a similar performance is observed, whereas in the other two datasets there are significant differences. Following the fine-tuning approach (FTA), the number of undesired errors committed is elevated, ranging from 2% to 11% of the reviews. Taking into account that a large number of examples are handled and that the highest possible precision is pursued when detecting an extreme review, this value is too high. The feature-based approach (PM) obtains an error rate $< 1\%$ for the undesired errors. Therefore, the readjustment of weights in the second stage leads to an increased confusion in discerning between the extreme reviews and their immediate classes, with the feature-based approach preserving these extremes better defined.

B. Undersampling vs class weights

This experiment involves the training performed during the first phase regarding class imbalance, which results are shown in Table II. The behavior of the proposed method (PM) was analyzed by training with the balanced set of extreme reviews instead of using all of them and applying class weights in the loss function (CW). Based on the results, undersampling achieves better performance. In fact, the model using class weights is not able to classify any review as negative extreme with the required certainty in the restaurant datasets. Consequently, it has a 0% rate in errors related to negative extremes, which should not be considered. Regarding the Hotels reviews dataset, although it gets more hits in the classification of extremes, it commits a 6% error in classifying 2-star reviews as negative extremes. Thus, it can be observed that as the size of the dataset increases, and consequently the imbalance between

TABLE II
Performance comparative of the proposed method versus the different strategies (certainty $>= 0.9$).

Hotel Reviews						
		PM	FTA	CW	1-S	Base
Hits	5 as 5	0.0656	0.5100	0.2199	**0.9102**	0.4352
	1 as 1	0.1000	0.6407	0.5592	**0.9852**	0.4630
Not desired errors	4 as 5	**0.0061**	0.0921	0.0123	0.6866	0.0702
	2 as 1	**0.0089**	0.1101	0.0684	0.9018	0.0505
Critical errors	3 as 5	**0.0000**	0.0024	**0.0000**	0.1763	0.0024
	3 as 1	**0.0000**	0.0010	0.0072	0.4903	0.0048
Very critical errors	1 as 5	0.0000	0.0000	0.0000	0.0000	0.0000
	2 as 5	**0.0000**	**0.0000**	**0.0000**	0.0119	**0.0000**
	4 as 1	0.0009	0.0009	0.0009	0.0685	**0.0000**
	5 as 1	**0.0000**	0.0011	0.0006	0.0127	**0.0000**
Restaurant Reviews New Dheli						
		PM	FTA	CW	1-S	Base
Hits	5 as 5	0.0804	0.1119	0.2027	**0.9815**	0.2361
	1 as 1	0.1198	0.0898	0.0000	**0.9581**	0.0000
Not desired errors	4 as 5	0.0035	**0.0033**	0.0158	0.8926	0.0153
	2 as 1	0.0116	0.0058	**0.0000**	0.9246	**0.0000**
Critical errors	3 as 5	**0.0000**	0.0008	0.0017	0.3119	0.0008
	3 as 1	0.0017	0.0008	**0.0000**	0.5413	**0.0000**
Very critical errors	1 as 5	**0.0000**	**0.0000**	0.0060	0.0269	0.0060
	2 as 5	**0.0000**	**0.0000**	**0.0000**	0.0435	0.0029
	4 as 1	0.0002	0.0002	0.0002	0.0453	**0.0000**
	5 as 1	**0.0000**	**0.0000**	**0.0000**	0.0046	**0.0000**
Restaurant Reviews New York						
		PM	FTA	CW	1-S	Base
Hits	5 as 5	0.1170	0.2526	0.0000	**0.9806**	0.1929
	1 as 1	0.0753	0.3871	0.0000	**0.9767**	0.4480
Not desired errors	4 as 5	0.0061	0.0247	**0.0000**	0.8824	0.0109
	2 as 1	0.0051	0.0759	**0.0000**	0.9202	0.0914
Critical errors	3 as 5	**0.0000**	**0.0000**	**0.0000**	0.2402	0.0004
	3 as 1	0.0008	0.0038	**0.0000**	0.5414	0.0064
Very critical errors	1 as 5	0.0018	0.0018	**0.0000**	0.0125	0.0018
	2 as 5	**0.0000**	0.0013	**0.0000**	0.0180	0.0012
	4 as 1	0.0004	0.0001	**0.0000**	0.0419	0.0005
	5 as 1	**0.0000**	**0.0000**	**0.0000**	0.0044	0.0002

PM: proposed method. FTA: fine-tuning approach (2nd stage).
CW: class weights in loss function (1st stage). 1-S: one-stage training.
Base: XLNet with sentence classification weights.

classes, the model fails to learn how to correctly extract the features of the minority class. Furthermore, in the case of the New York restaurant dataset, it is unable to obtain predictions with the required certainty, so that the features of the extremes are not well defined. The proposed method, using the balanced set, is able to detect about 6% to 12% of the extreme reviews among all classes with a certainty ≥ 0.90 and a rate $< 1\%$ for all error types.

C. Two-stage training vs one-stage training

The proposed model performs a second stage of training, freezing the XLNet layers and adjusting a 5-unit fully connected layer to address the multi-class classification. To analyze whether this step is necessary, it was tested if only the knowledge extracted from the extreme classes during the first stage is sufficient to identify them afterwards among all classes. In the results detailed in Table II, column 1-S, it can be seen how a very high number of extreme revisions are detected in this manner, above 90% at both extremes in all the datasets. However, the number of errors committed is exceedingly high as well. Especially in the classification of the immediate classes, where the minimum error rate observed is 68%. Therefore, the model trained for the binary task is able to separate the features characterizing negative and positive extreme examples. However, to discern between the features that define extreme samples and the other classes, further adjustments are definitely needed.

D. Our method vs baseline

Finally, the performance of the proposed model was analyzed in comparison with the base XLNet without using knowledge extracted from the source domain, i.e,. using the weights for sentence classification instead of those obtained from the extreme classes. In the base case, a single training phase was carried out, using the examples belonging to all classes and handling the imbalance with class weights in the loss function. This would be analogous to using the proposed model without performing the first training phase. The performance achieved with both methods are shown in Table II, specifically in columns PM and Base. Concerning the hotel reviews dataset, although the base case detects many more extremes, it also detects many more errors. It presents a minimum error rate in the classification of the examples of the intermediate classes of 5% in contrast to that of the proposed model, which is $< 1\%$. The same applies to the New York restaurant dataset, where the base case classifies 9% of the 2-star reviews as negative extremes, while the proposed model classifies $< 1\%$. Regarding the New Dheli restaurant dataset, the base case is not able to classify any negative extremes with the desired certainty. In fact, it does not detect any example as belonging to this class. Therefore, the error rate is lower in some errors compared to the proposed model, but it should not be taken into account. In the errors concerning the classification of positive extremes, the proposed model presents once again the lowest values. Thus, readjusting the XLNet weights for sentence classification using the knowledge previously extracted from the source classes, achieves better results in relation to the certainty and quality of the classification of the extreme classes.

TABLE III
RESULTS OF THE MULTI-CRITERIA DECISION ANALYSIS IN ALL APPROACHES ACCORDING TO THE WEIGHTED SUM MODEL.

Hotel Reviews					
	PM	FTA	CW	1-S	Base
Rank	1	4	3	5	2
Score	0.3836	0.1019	0.1648	0.0996	0.2500
Restaurant Reviews New Dheli					
	PM	FTA	CW	1-S	Base
Rank	1	4	3	5	2
Score	0.2633	0.1651	0.1797	0.1431	0.2487
Restaurant Reviews New York					
	PM	FTA	CW	1-S	Base
Rank	2	4	1	3	5
Score	0.1282	0.1035	0.6166	0.1153	0.0362

E. Multi-criteria decision analysis

The results obtained from the multi-criteria decision analysis of the different approaches tested using the weighted sum model (WSM) are reported in Table III. As can be seen, our proposal (PM) ranks first in two datasets and second in the other, the New York restaurant reviews dataset. The CW model, which ranks first in this case, has a rate of 0 for all the error types considered because it is not able to classify any example as extreme, with the required certainty. Additionally, it ranks third in the other two datasets. Therefore, our proposal is the first in the ranking that is able to detect extreme reviews, both positive and negative.

V. CONCLUSIONS

Despite the relevance that the detection of extreme reviews has in many contexts, it has not been practically addressed in the literature. This work addresses their classification through sentiment analysis using XLNet, the state-of-the-art in numerous natural language processing tasks. In addition, a real scenario was considered, in which reviews belonging to all classes are taken into account and characterized by a strong imbalance between them. The main objective is to conduct the detection of extreme positive and negative reviews with high certainty and committing as few errors as possible with respect to the rest of the classes. For this purpose, a two-stage training method is proposed using transfer learning. In the first stage, the network learns the features that define both extremes, positive and negative; and in the second stage, this knowledge is transferred to face its subsequent identification in a multi-class classification task.

Through experimentation, it has been demonstrated how the use of the knowledge extracted from the extreme classes improves the performance of the classification to detect extreme reviews while incurring in fewer errors. In this case, using a

balanced set of the source domain to learn its features works better than handling class imbalance with class weights during training. This suggests that finding a quality dataset, rather than trying to fill in the gaps during the training process, is of particular importance in this part of the procedure. Nevertheless, re-adjusting these weights during training with the all-class examples may lead to an increase in the error rate, despite increasing the number of extreme reviews detected. The proposed method leads to a better definition of the features of the extreme classes and diminishes the confusion between immediate classes when dealing with their identification. Thus, performing a first stage of feature learning of the extreme classes and using those features in a second stage of classification, improves the quality of the predictions.

As future research, different classifiers, such as random forest or recurrent neural networks, will be tested in the second training stage, applied to the features extracted by XLNet in the first stage. The intention is to be able to better discern between examples of all classes to obtain a higher number of hits. Other transfer learning techniques using the knowledge of the target classes will also be studied. For example, finding new feature representations that better characterize the extreme classes against the rest of the examples while maintaining a high certainty in their detection. In addition, the effectiveness of the different approaches will be tested using more datasets, not only of larger size, but also in different domains.

REFERENCES

[1] S. J. Salmon, E. De Vet, M. A. Adriaanse, B. M. Fennis, M. Veltkamp, and D. T. De Ridder, "Social proof in the supermarket: Promoting healthy choices under low self-control conditions," *Food Quality and Preference*, vol. 45, pp. 113–120, 2015.

[2] J. Lee, D.-H. Park, and I. Han, "The effect of negative online consumer reviews on product attitude: An information processing view," *Electronic Commerce Research and Applications*, vol. 7, no. 3, pp. 341–352, 2008.

[3] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, 2006.

[4] I. Comscore and T. Kelsey, "Online consumer-generated reviews have significant impact on offline purchase behavior," https://www.comscore.com/Insights/Press-Releases/2007/11/Online-Consumer-Reviews-Impact-Offline-Purchasing-Behavior, 2007.

[5] C. Anderson, "The impact of social media on lodging performance," *Cornell Hospitality Report*, vol. 12, no. 15, pp. 6–11, 2012.

[6] R. Filieri, E. Raguseo, and C. Vitari, "When are extreme ratings more helpful? empirical evidence on the moderating effects of review characteristics and product type," *Computers in Human Behavior*, vol. 88, pp. 134–142, 2018.

[7] B. Pang and L. Lee, "Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales," in *43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 115–124.

[8] S. Park and J. L. Nicolau, "Asymmetric effects of online consumer reviews," *Annals of Tourism Research*, vol. 50, pp. 67–83, 2015.

[9] B. Fang, Q. Ye, D. Kucukusta, and R. Law, "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics," *Tourism Management*, vol. 52, pp. 498–506, 2016.

[10] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[11] S. Moon, M.-Y. Kim, and P. K. Bergey, "Estimating deception in consumer reviews based on extreme terms: Comparison analysis of open vs. closed hotel reservation platforms," *Journal of Business Research*, vol. 102, pp. 83–96, 2019.

[12] L. Kwok and K. L. Xie, "Factors contributing to the helpfulness of online hotel reviews: does manager response play a role?" *International Journal of Contemporary Hospitality Management*, vol. 28, no. 10, pp. 2156–2177, 2016.

[13] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Computer Science Review*, vol. 41, p. 100413, 2021.

[14] L. Li, T.-T. Goh, and D. Jin, "How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4387–4415, 2020.

[15] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, "Sentiment analysis on product reviews using machine learning techniques," in *Cognitive Informatics and Soft Computing*. Springer, 2019, pp. 639–647.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 1–11.

[18] S. Almatarneh and P. Gamallo, "Searching for the most negative opinions," in *International Conference on Knowledge Engineering and the Semantic Web*, 2017, pp. 14–22.

[19] ——, "Linguistic features to identify extreme opinions: an empirical study," in *International Conference on Intelligent Data Engineering and Automated Learning*, 2018, pp. 215–223.

[20] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4335–4385, 2020.

[21] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] T. Zhang, B. Xu, F. Thung, S. A. Haryono, D. Lo, and L. Jiang, "Sentiment Analysis for Software Engineering: How Far Can Pre-trained Transformer Models Go?" in *IEEE International Conference on Software Maintenance and Evolution*, 2020, pp. 70–80.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: from lexicons to transformers," *IEEE Access*, vol. 8, pp. 131 662–131 682, 2020.

[26] S. Alaparthi and M. Mishra, "Bert: a sentiment analysis odyssey," *Journal of Marketing Analytics*, vol. 9, no. 2, pp. 118–126, 2021.

[27] B. Bansal, "TripAdvisor Hotel Review Dataset," Apr. 2018.

[28] I. L.-R. Botana, V. Bolón-Canedo, B. Guijarro-Berdiñas, and A. Alonso-Betanzos, "Explain and conquer: Personalised text-based reviews to achieve transparency," 2022.

[29] S. Loria, "textblob Documentation," *Release 0.15*, vol. 2, 2018.

[30] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, and R. Oliveto, "Sentiment analysis for software engineering: How far can we go?" in *40th International Conference on Software Engineering*, 2018, pp. 94–104.

[31] A. Murgia, P. Tourani, B. Adams, and M. Ortu, "Do Developers Feel Emotions? An Exploratory Analysis of Emotions in Software Artifacts," in *11th Working Conference on Mining Software Repositories*, 2014, p. 262–271.

[32] O. Hubáček, G. Šourek, and F. Železný, "Exploiting sports-betting market using machine learning," *International Journal of Forecasting*, vol. 35, no. 2, pp. 783–796, 2019.

[33] J. B. Cabral, N. A. Luczywo, and J. L. Zanazzi, "Scikit-criteria: Colección de métodos de análisis multi-criterio integrado al stack científico de Python," in *XLV Jornadas Argentinas de Informática e Investigación Operativa*, 2016, pp. 59–66.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015, pp. 1–15.