The Winning Approach for the Recommendation Systems Shared Task @REST_MEX 2022

Cipriano Callejas-Hernández[1], Erika Rivadeneira-Pérez[1], Fernando Sánchez-Vega[1,2], Adrián Pastor López-Monroy[1] and Esaú Villatoro-Tello[3,4]

[1]*Mathematics Research Center, Guanjuato, Mexico.*
[2]*Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico City, México.*
[3]*Idiap Research Institute, Martigny, Switzerland.*
[4]*Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico*

Abstract

This paper presents our approaches for the Recommendation System and Sentiment Analysis shared tasks at Rest-Mex 2022. In the first task, the dataset presented a number of challenges, which we overcome by exploring information organization schemes and traditional data representation. For opinion classification in the case of Sentiment Analysis we found that state-of-the-art pre-trained models by adapting two Bert-based approaches get an acceptable performance. With these two approaches we were able to reach the first place in the recommendation system task while our simple adaptation of state-of-the-art for the sentiment analysis task got a very competitive performance, only 0.58% below the winning approach.

Keywords

Rest-Mex 2022, Recommendation System, Sentiment Analysis, Mexican Tourist Text, Text Information Organization Schemes, BOW, BERT

1. Introduction

Nowadays, recommendation systems and sentiment analysis have gained great relevance in various fields, including the tourism domain. Recommendation systems (RS) are valuable tools in e-tourism platforms (TripAdvisor, Booking, etc) that help users in their decision-making process [1]. For example, state-of-the-art models based on collaborative filtering schemes [2] stands items out that users might like on the basis of past history or reactions by similar users, considering this approach, our system takes advantage of the user history.

On the other hand, Sentiment Analysis (SA) has received notable attention because stakeholders can leverage data from e-tourism platforms in order to perform data-driven decisions. SA could allow to identify the valuation of the products offered in the industry, it helps to identity the flaws and focus the attention to the markets in user's interest. However, modern approaches require a large amount of data to achieve adequate performance, yet these works have been used only in contexts with English text data. For this reason, we focus in generating systems that help to develop intelligent systems for Spanish text data.

cipriano.callejas@cimat.mx (C. Callejas-Hernández); erika.rivadeneiras@cimat.mx (E. Rivadeneira-Pérez); fernando.sanchez@cimat.mx (F. Sánchez-Vega); pastor.lopez@cimat.mx (A. P. López-Monroy); esau.villatoro@idiap.ch (E. Villatoro-Tello)

IberLEF 2022, September 2022, A Coruña, Spain.

IberLEF is an evaluation campaign for Natural Language Processing Systems in Spanish and other Iberian languages. REST-MEX is a task in IberLEF which is focused on recommendation tasks using TripAdvisor as textual source, with texts written in several variants of Spanish (Mexican Spanish being the most common) [3, 4]. In this work, we describe our approaches for the two Rest-Mex 2022 shared tasks:

- *Task 1:* **Recommendation Systems**. For this task we follow a simple, yet effective approach. Given the information about Mexican places, user's profile and their past opinions about places (user history), we structure the data of each user by considering schemes that, first, organize the information in such a way that they gather all the textual information in the same space, and other that separate the information in different spaces. We use a BOW approach to represent all text and a classifier to make a prediction.
- *Task 2:* **Sentiment Analysis**: For this task, we predict the polarity (in a 1-5 scale) and the type of attraction (hotel, restaurant or attraction) using pre-trained and fine tuning BETO and RoBERTuito models over the user's opinions.

In Recommendation Systems, since the dataset was lacking of a homogeneous structure we wanted to pursued an organization scheme approach where we had at least two options, either to use each text data individually or to use it all as a whole, these two will be explained in detail later on. On the other hand, Sentiment Analysis is a more established area and so state-of-the art approaches were an obvious path to follow.

This paper is structured as follows: in sections 2 and 3, we present the methodology used for Recommendation System and Sentiment Analysis tasks respectively; section 4 describes the metrics used for evaluation as well as the results obtained; in section 5 some ethical issues are discussed and in section 6 we state the conclusions.

2. Recommendation System Task

This section describes the corpus provided by the REST-MEX organizers, the organization scheme we followed to re-structure the dataset as well as the strategy used for the RS task. Some problems found in the dataset are described and possible ways to fix these problems. This subtask is a classification problem where the system participating should have to predict the degree of satisfaction for a tourist when recommending a destination [3].

For this task, we based our approaches on the collaborative filtering technique [2],that is, we wanted to use as much user data as possible, doing so an organization information scheme was needed and at least two options were available: either to use all textual data as a whole or in a separated way, we called these two as an aaggregated or disaggregated approach, respectively. Because of such magnitude we used a more traditional approach by means of BOW representation.

2.1. Corpus Description

The dataset is given as follows:

Place's description : A set of brief descriptions (most of them with missing values), the type of tourism they have to offer (beach, gastronomy, religious, etc.) and the name of place, 18 places were collected.

User history : The opinion for each of the places the user has visited, as well as their satisfactory degree and a overall label given by TripAdvisor.

Users : Each user is identified with an id, their gender, the last place they visited, location, date of visit, type of visit (Family, Friends, Alone, Couple, Business, etc) and a label corresponding to the satisfaction degree they had.

Notice that is mainly constituted of information about the users and places, 1582 instances were given for training, and 681 for testing. Let us denote the dataset as the pair $(\mathcal{U}, \mathcal{P})$.

For each user (or tourist) $u \in \mathcal{U}$ we have the user's gender, the places the user u has visited and the user's history (e.g., opinions over some other places).

For each place $p \in \mathcal{P}$ we have the place's name and a set of fields mentioning the place's main attractions or a set of categories representing the general features that represent the main offer of the locality (Gastronomy, cultural, beaches, etc).

Our goal is to predict the satisfaction degree (label) that a tourist u may have based on the information of $\mathcal{U}$ and $\mathcal{P}$. The following Entity Relationship Diagram (ERD-like) represents an instance (a tourist) in the dataset described above.

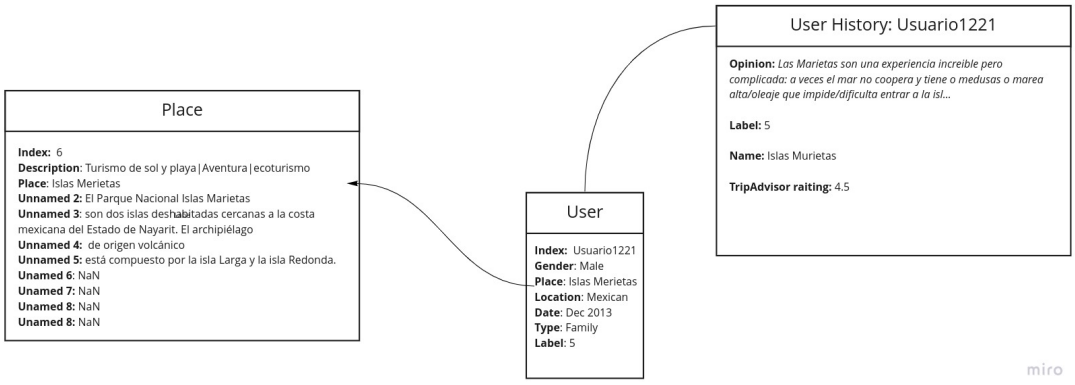

Figure 1: Information given for user 1221.

2.2. Corpus Preprocessing and Representation Selection

The corpus presents several challenges, for example, the target place may not be available in the tourist's history, and the opinions might not be complete as Figure 1 shows, we also face an imbalance classification problem as shown in Figure 2. Furthermore, the following situations were found in the training set:

- Some users didn't have any history, a possible cause is that the user recently joined TripAdvisor.
- Imbalance in the number of opinions per user, between no opinion and 1242, with an average of 7
- Some users had the same opinion history, this implies that they in fact, are the same user.
- Multiple languages were found in the dataset. English being the most predominant after Spanish.

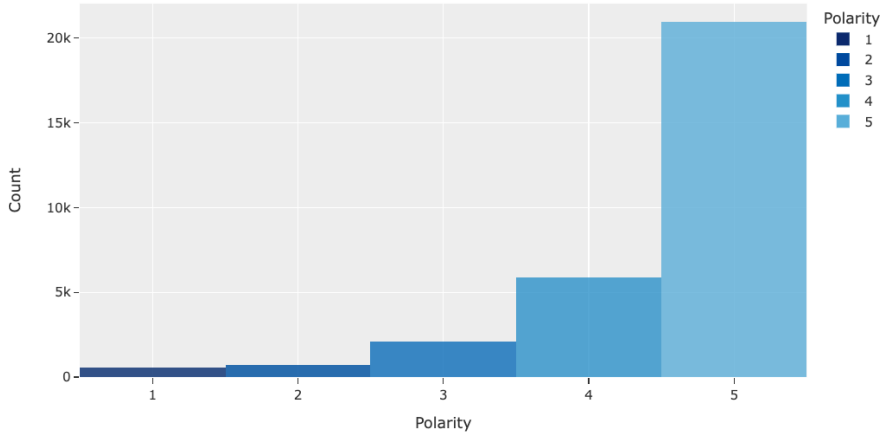

Figure 2: Distribution of the satisfaction degree. 5 is the most frequent satisfaction degree.

In order to solve the above situations we selected only 922 users who had a nonempty history to train the classification method. An effective recommendation system must take advantage of user history information [2], and so a well structured information dataset helps to accomplish this.

At the end, our data, for each user u, the following information was considered:

- All user's opinions were concatenated.
- Gender (*Male, Female*)
- Location (*Mexican*, *Foreign*).
- Type of trip (*Alone, Couple, Business, Family, Alone*).
- The name of the last visited place p, as well as its description.
- Mean satisfaction based on the user's history.
- Mean degree based on a score given by TripAdvisor for each of the places the user visited.

Let us denote this new structured dataset as $\mathbb{X}$. To take advantage of the given information different schemes were considered: one in which the data was structured in such a way that all the textual information is aggregated or unified in the same space and and another in which the information is kept separate through disaggregated representations.

Our goal is to find patters in the triplets $(u, p, label) \in \mathbb{X}$ which represent the relationship between a user u and a target place p given by a *label* using the information stated in the corpus.

For every user u we have a multiple opinions concatenated, this text variable is denoted by T_{opinion}, as well as the place's name they visited, T_{place}, and the corresponding concatenation of the all the brief descriptions available in the place's table, $T_{\text{InfoPlace}}$.

The following transformations were performed on $\mathbb{X}$: the type of trip was recast into the following discrete values: Family (0), Friends (1), Alone (2), Couple (3), Business (4); and the others binarized as described in Table 1, let us denote the set of discrete variables as Z.

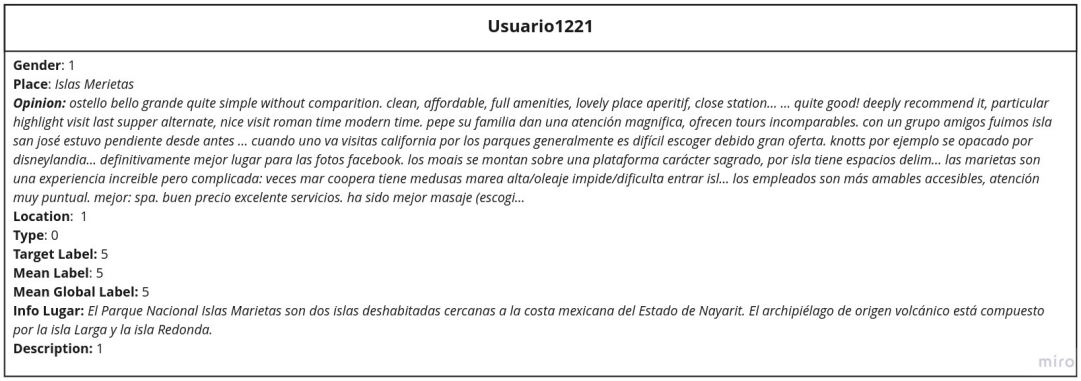

Figure 3: Example of one instance in the resulting dataset.

Variable	0	1
Gender	Female	Male
Location	Foreigner	Mexican
Place's Description	Cultural	Beach

Table 1
Binarization of variables

For each text variable, we decided to use a BOW, where text is represented as an histogram of its words. BOW disregards grammar and word order and just focuses on presence/absence of words. On the other hand, we use n-grams to keep notion about the order, in fact we used bigrams. In short, our representation is a BOW of bigrams.

With the newly structured dataset we had two approaches, an aggregated and a disaggregated one, both described below.

1. (*1st Run*) We first concatenated all text variables

$$T = [T_{\text{opinion}}, T_{\text{place}}, T_{\text{InfoPlace}}]$$

and then BOW(T), with a Term frequency – Inverse document frequency (TF-IDF) weighting, with no normalization and considering the whole text in lowercase. Later, we concatenated the resulting BOW with the remaining features (i.e. gender, location, etc.). See Figure 4 to visualize the general scheme of the proposed model.

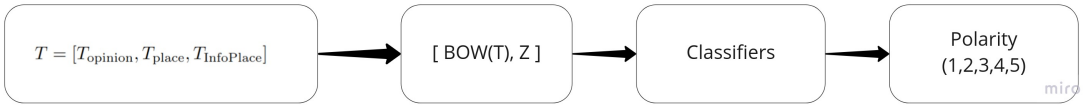

Figure 4: Aggregated approach.

2. (*2nd Run*) For each text variable, a BOW was obtained: $BOW(T_{\text{opinion}})$, $BOW(T_{\text{place}})$ and $BOW(T_{\text{InfoPlace}})$. These BOWs were concatenated with the remaining features (i.e. gender, location, etc.).

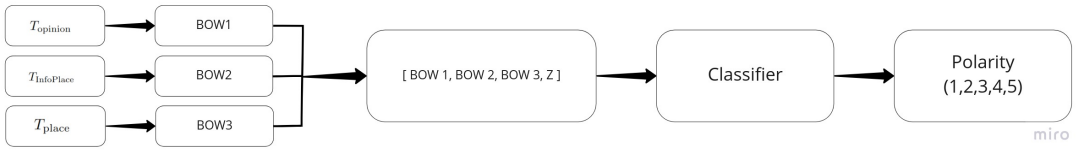

Figure 5: Disaggregated approach.

Different classifiers were used, the ones that got a better performance are: Gaussian Processes for the aggregated approach and XBoost for the disaggregated approach. More traditional, SVM and Multinomial Regression, and even MLP classifiers were tested yet not improvement was found. As a remark, the use of Gaussian processes as a classfier model was motivated because in [5] the results of ordinal regression were equiparable with the ones used with this classifier. Moreover, notice that in both approaches a BOW representation was used yet the big difference between these schemes is how we treated the training set: in the first we split the data and then created the BOWs while in the second, we created the BOWs and then we performed the splitting.

3. Sentiment Analysis Task

The main goal of this task is to predict the polarity and the type of attraction given a user's opinion [3].

3.1. Corpus Description

Training data contains, for each user $u \in \mathcal{U}$, the user opinion, the satisfaction degree, which can take values in $\{1, 2, 3, 4, 5\}$ where 1 represents the worst satisfaction and 5 the best satisfaction, and the type of attraction of the opinion (attractive, hotel, and restaurant). The train dataset contains $30,212$ instances, and the test set contains $12,938$ instances.

3.2. Approaches

Because the reviews in the datasets were written mainly in Spanish we used pre-trained models called **BETO** [6] and **RoBERTuito** [7]. BETO is a BERT-like system pre-trained for user-generated text in Spanish and RoBERTuito is a RoBERTa-like system which was pre-trained using a corpus of tweets in Spanish with a similar size to that of the corpus used to train BertBase [7].

We use all opinions of the training data on the RoBERTuito pre-trained model since there are evidence that such model performs well when the instances include some English text, see [7]. Considering that BETO and RoBERTuito follow the same design principles as Bert, we proceeded to execute a fine-tuning process for the two substasks: a five classes classification problem (polarity) and a three classes classification problem (type of attraction) for both models.

- **Pre-processing:** First, we preprocess the users opinion, removing the quotation marks in the reviews. Then, we used the BERT tokenizer (loaded with the weights from BETO and RoBERTuito respectively).
- **Fine-Tuning:** For the five classes classification problem we chose a classification layer with a Softmax activation function applied to the output values, the same was done for the three classes classification problem. The hyperparameters chosen for both problems depending on the model for fine-tuning are listed below.

The hyperparameters used for BETO model were:

- Batch size: 16
- Max length: 512
- Learning rate: 5×10^{-6}

On the other hand, the hyperparameters considered for RoBERTuito model were:

- Batch size: 8
- Max length: 120
- Learning rate: 2×10^{-5}

Figure 6 shows the scheme of the proposed approach of this task. Let us remark that in the

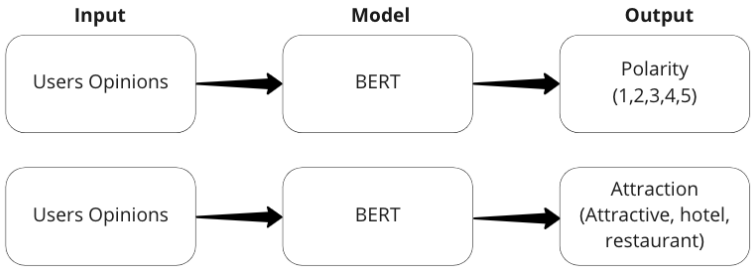

Figure 6: Approach scheme of Sentiment Analysis task

mixed classification problem: the polarity of the opinion and the type of attraction was splitted into two independent classification problems since the second sub-problem perfomed very well on the training set.

4. Results

4.1. Evaluation Metrics

t For the Recommendation System task, the ranking was determined by measuring the systems with the Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|.$$

For the Sentiment Analysis task, the organizers proposed two sub-tasks: the polarity classifications and the type prediction: The polarity is an integer in the interval $[1, 5]$, for which the MAE (Mean Absolut Error) metric was used, while for the type prediction (a three class classification) the Macro F-measure was used [3]. In the overall task of Sentiment Analysis the following metric was used, this is an average of both metrics mentioned before

$$Sentiment = \frac{\frac{1}{1+MAE} + MacroF1}{2},$$

where

$$MacroF1 = \frac{F1_A + F1_H + F1_R}{3}$$

4.2. Results

We have described two approaches for each tasks RS and SA at Rest-Mex 2022: in Recommendation Systems we obtained the first place and in Sentiment Analysis the results were on average obtaining the seventh place in the competition.

The results for *Task 1* of both approaches, the baseline and the average results among participants for the contest are presented in Table 2, where Disaggregated Approach corresponds to the the first run (first row) which got the fourth place in the competence and the greatest accuracy among all participants, while Aggregated Approach represents our second run (second row) this model achieved the lowest MAE in the RS task. We now present the results fot *Task 2*

System	MAE	Accuracy	F-Measure	Recall	Precision
Disaggregated Approach	0.716	**53.663**	0.174	0.181	0.206
Aggregated Approach	**0.693**	52.129	0.196	0.195	0.214
Baseline	0.742	53.304	0.139	0.107	0.2
Average	0.716	50.463	0.187	0.186	0.215

Table 2
Prediction performance for the Recommendation Systems task. One can see that our approaches have better performance than the baseline.

of both systems in Table 3. Both models achieved a similar accuracy, 70%, and MAE of almost 28%, both quite below the MAE and way above the accuracy compared to the baseline and the average metrics among all participants. These results suggest that the use of pre-trained models achieve a good accuracy, nonetheless more work on both the fine tuning and the preprocess of the text needs to be done in future work.

System	MAE	Accuracy	F-Measure	Recall	Precision
BETO	0.267	75.707	0.520	0.571	0.491
RoBERTuito	0.288	75.073	0.479	0.545	0.450
Baseline	0.476	70.026	0.165	0.140	0.2
Average	0.386	70.954	0.433	0.491	0.427

Table 3
Prediction performance for the Sentiment Analysis task. We can see that our approches leads to better performance results than the baseline methods.

5. Ethical Issues

Recommendation Systems collect, curate, and act upon vast amounts of personal data. Inevitably, they shape user preferences and guide choice, both individually and socially. Milano et al. [8] identified two ways in which a recommender system can have ethical impacts: its operations can violate the users rights and second, the risk of imposition, whether the negative impact constitutes an immediate harm or it exposes the relevant party to future risk of harm. As a remark in our work, we notice that the number of opinions was also imbalanced, with one single user who had over hundreds of reviews and some with few, within the same period of time. This might be an indicator of a strong bias present in the data, where collected data might contain fake users and/or reviews. By manually analyzing some of the users we found a few users sharing the same historical information just with different ids.

6. Conclusions

The approach described in section 2.1 obtained the lowest MAE of this year's Recommendation System shared task at Rest-Mex 2022. This suggest that even though deep learning techniques have been used recently in classification problems, a simple BOW approach is still useful when there is not enough data. In particular, we could remark that the effectiveness of such approach relies |on the fact that the BOW model is somewhat independent of the language. This was of great advantage since the dataset had opinions in different languages and BOW is able to easily capture the presence/absence of specific relevant topics and words. Furthermore, it can be concluded that a disaggregated scheme allows a better recommendation by keeping separate information that is not convenient to combine, since we consider that the opinion and the type of place are separate aspects to make a decision.

For Sentiment Analysis task, deep learning techniques outperformed BOW, the use of pre-trained models was of great advantage, particularly, because the opinions were more uniform with respect to the presence of different languages and missing data (compared to the data of RS). The average accuracy for this problem was of 70%, in both of our approaches we were able to reach such baseline with an accuracy of 75%. We are confident that a more selective choice of tokens and using a model with more suitable parameters would greatly improve our results.

In both tasks there is future work to be done, for example, test second order attributes or other deep learning techniques. And if possible, to follow a more text-based approach, that is, to exploit the text diversity.

Acknowledgments

The authors thank *Consejo Nacional de Ciencia y Tecnología* (CONACYT) and *Centro de Investigación en Matemáticas* (CIMAT) for the scholarships assigned with CVUs: 1089020 and 1012686. Author Esaú Villatoro-Tello, was supported partially by Idiap Research Institute, SNI-CONACyT, and UAM-Cuajimalpa Mexico during the elaboration of this work. We also thank the *Instituto Nacional de Astrofísica, Óptica y Electrónica* (INAOE) for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies (*Laboratorio de Supercómputo: Plataforma de Aprendizaje Profundo*) with the project *"Identification of Aggressive and Offensive text through specialized BERT's ensembles"* and CIMAT Bajio Supercomputing Laboratory (#300832). Sanchez-Vega would like to thank CONACYT for its support through the Program *"Investigadoras e Investigadores por México"* by the project *"Desarrollo de Inteligencia Artificial aplicada a la prevención de violencia y salud mental."* (ID. 11989, No. 1311).

References

[1] M. Nilashi, O. Ibrahim, E. Yadegaridehkordi, S. Samad, E. Akbari, A. Alizadeh, Travelers decision making using online review in social network sites: A case on tripadvisor, Journal of computational science 28 (2018) 168–179.

[2] F. Strub, J. Mary, Collaborative filtering with stacked denoising autoencoders and sparse inputs, in: NIPS workshop on machine learning for eCommerce, 2015.

[3] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021).

[4] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[5] J. Cheng, Z. Wang, G. Pollastri, A neural network approach to ordinal regression, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, 2008, pp. 1279–1284.

[6] S. Wu, M. Dredze, Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, arXiv preprint arXiv:1904.09077 (2019).

[7] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, arXiv preprint arXiv:2111.09453 (2021).

[8] S. Milano, M. Taddeo, L. Floridi, Recommender systems and their ethical challenges, Ai & Society 35 (2020) 957–967.

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