AutoML and Ensemble Transformers for Sentiment Analysis in Mexican Tourism Texts

Victor Gómez-Espinos[1,*], Victor Muñiz-Sanchez[1,*] and Adrian Pastor López-Monroy[2,*]

[1]*Mathematics Research Center (CIMAT), Monterrey, 66628, Mexico.*

[2]*Mathematics Research Center (CIMAT), Guanajuato, 36023, Mexico.*

Abstract

In this paper, we describe the proposed methodology for participating in the sentiment analysis challenge from the Rest-Mex track at IberLEF 2022. This challenge is a classification problem, and consists on two sub-tasks: polarity classification and the type of opinion. Our proposal consists in generating high level features from texts by learning contextualized representations based on an ensemble of fine-tuned BERT models according to each sub-task. Subsequently, these features are aggregated to obtain an extended dataset, where an optimal ensemble of machine learning models are trained. The proposed strategy obtained the third place on this task.

Keywords

Mexican tourism, Sentiment analysis, BERT ensemble

1. Introduction

Tourism is a key activity for the economy of many countries [1]. Particularly for Mexico, before the COVID-19 pandemic, tourism contributed 8.7% of gross domestic product (GDP)[2], and after this difficult pandemic period, this activity is showing a significant improvement in the last quarter of 2021 in terms of GDP according to official statistics [3, 4]. Nowadays, the internet and social media provide us a way to know the opinions and preferences of people who makes use of services related to tourism, such as hotels, lodging, flights, destinations, amenities, among many others. The data generated by users is in the form of text, sometimes together with photos or videos, which can make complex its analysis and implementation of supervised and non supervised machine learning (ML) techniques. Since last year, Rest-Mex [5] offers an evaluation forum which allows the participation of the academic community in some tasks aimed at understanding the perception of users of tourism services in Mexico, helping in this way, to address some issues related to tourism, and at the same time, the reactivation of the tourism sector. To this end, two taks were proposed: recommendation system and sentiment analysis, promoting the application and development of natural language processing (NLP) methodologies for the spanish language, together with state of the art (SOTA) machine/deep learning models.

IberLEF 2022, September 2022, A Coruña, Spain

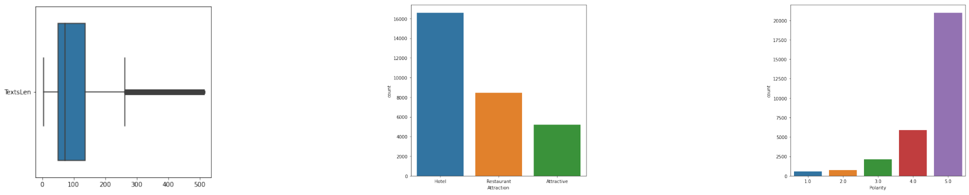
victor.gomez@cimat.mx (V. Gómez-Espinos); victorm@cimat.mx (V. Muñiz-Sanchez); pastor.lopez@cimat.mx (A. P. López-Monroy)

This year, Rest-Mex has focused on three tasks: recommendation system, sentiment analysis and covid semaphore prediction, for whom, a new corpora has been created and provided by the organizers considering Spanish opinions from the TripAdvisor website and news related to covid in Mexico [6][1].

In this paper, we propose a methodology to address the sentiment analysis task, which is described in the oficial website[1] as: "Given an opinion about a Mexican tourist place, the goal is to determine the polarity, between 1 and 5, of the text, and the type of opinion (hotel, restaurant or attraction)". We propose a two-step methodology. In the first step, we obtain high level features from texts based on fine-tuned BERT models, that we use to obtain an extended dataset. In the second step, we obtain an optimal ensemble of classifiers, which give us the final prediction for each subtask. This document is organized as follows. In Section 2, we describe the dataset by doing an exploratory data analysis, which gave us valuable information. In section 3, the proposed methodology is explained in detail. Section 4 shows the results we obtained with our proposal and Section 5 outlines our conclusions.

2. Exploratory data analysis

The dataset of the sentiment analysis task of Rest-Mex, consists of 30212 items, and for each one of them, a title, the opinion, the category (Hotel, Restaurant, Attractive), and the polarity [1-5], are included. From this dataset, we randomly select 60% for training, 20% for validation and the remaining 20% for test. Opinion texts has an average length of 122.21 tokens (after concatenating opinion and title) and 512 tokens as maximum, as can be seen in Figure 1(a). The majority class is Hotel, followed by Restaurant and Attraction (Figure 1(b)), and it can be observed a highly unbalanced dataset respecting to the polarity, where most of the opinions has a very positive polarity (5) and few of them has the most negative polarity (1), as can be seen in Figure 1(c).



(a) Text length (b) Type of opinion (c) Polarity

Figure 1: Some basic statistics for the sentiment analysis corpus

We can observe notable differences in the type of opinion according to the distribution of text length, as can be seen in Figure 2, and for the polarity as well, as is shown in Figure 3. For instance, the length of user opinions with polarity 5 are consistently shorter than those with polarity 1.

[1] https://sites.google.com/cicese.edu.mx/rest-mex-2022/home?authuser=0#h.i5o10i3si9z4

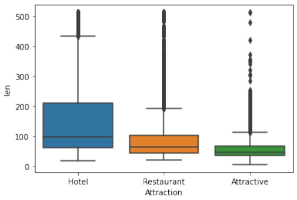


Figure 2: Text length according to the type of opinion

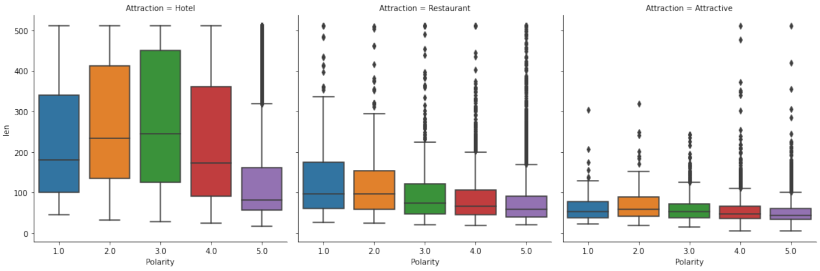


Figure 3: Text length according to the type of opinion and polarity

For the polarity problem, we decided to explore the behavior of a well-known sentiment score based on the Vader lexicon [7], which give us values in $[-1, 1]$, from most negative (-1) to most positive (1). The results of the Vader sentiment scores for each level of polarity is shown in Figure 4. Similarly to the text length, we can observe a notable trend with the Vader score, because the variance of this score becomes smaller when the polarity tends to 5, and the mean of the score tends to 1 for this value of polarity.

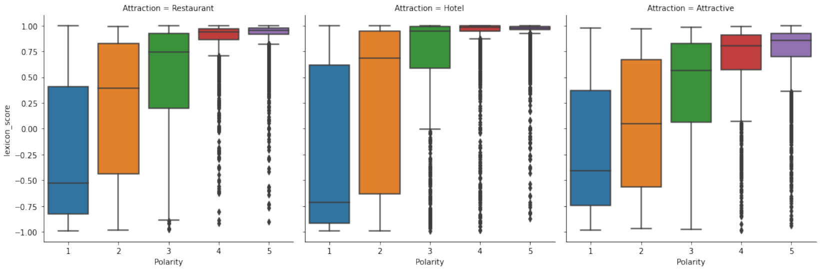
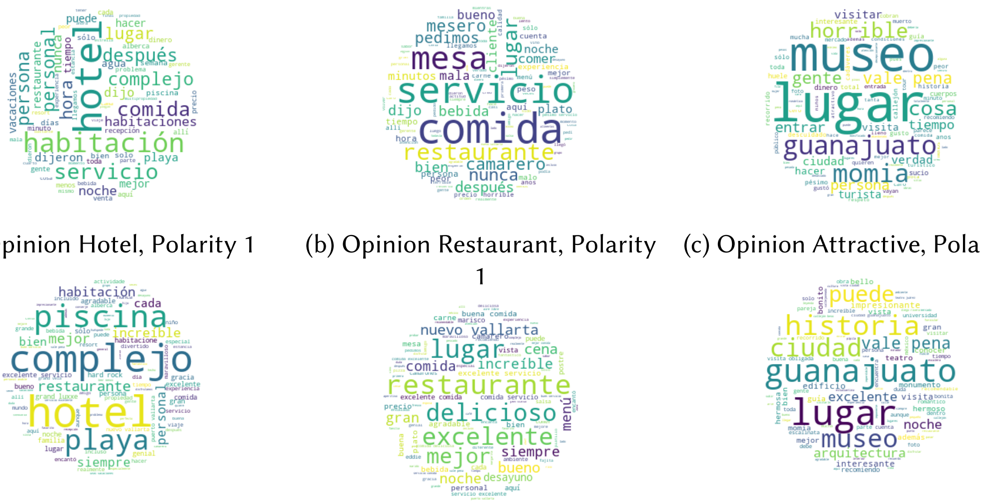


Figure 4: Vader sentiment score according to the type of opinion and polarity

For the rest of polarity levels, this difference is less significant, however, we find it interesting to explore the most frequent words of the opinion texts for the polarity values of 1 and 5. This analysis can be seen in Figure 5. Respecting to the subject of opinion (Hotel, Restaurant or Attractive), it can be seen that it is easy to distinguish the categories, even with the simple frequency of words, however, it is not the case of the polarity of opinions, because it is not easy to identify this category based solely on this feature, even with the extreme categories 1 and 5.

Based on this exploratory data analysis, we expect that the polarity classification will be



(a) Opinion Hotel, Polarity 1

(b) Opinion Restaurant, Polarity 1

(c) Opinion Attractive, Polarity 1

(d) Opinion Hotel, Polarity 5

(e) Opinion Restaurant, Polarity 5

(f) Opinion Attractive, Polarity 5

Figure 5: Wordclouds for the type of opinions and polarities 1 (most negative) and 5 (most positive).

considerably more difficult than the opinion classification subtask.

3. Methodology

The methodology we proposed consists on two steps. In the first step, we obtain high level features from texts by using an ensemble of BERT models trained in both subtasks (type of opinion and polarity). In the second step, we obtain an optimal ensemble of classifiers trained in the features obtained in the first step, which give us the final prediction for each subtask. We did minimal preprocessing of the data, which consists of concatenate the title and opinion and convert the concatenated text to lowercase. We explain in detail these two steps in the following sections.

3.1. High level features

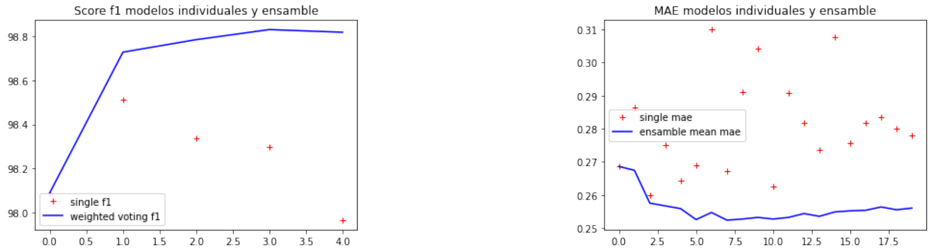
In this step, we obtain high level features by learning contextualized representations of texts based on an ensemble of BERT models [8], particularly, we used the spanish pre-trained model BETO [9] with fine-tunning for each subtask with the following configuration.

- Type of opinion. We address this subtask as a classification problem, and for this case, we did an exhaustive search of the hyper-parameters suggested by Devlin et. al. [8] by maximizing the Macro F1-score in the validation dataset, and using a weighted cross entropy loss function to consider the effect of class imbalance, where the weights were defined according to the class populations (hotel, restaurant, attractive). This results in a sequence length of 128 tokens, batch size of 32 and a learning rate of 5e-05 with ADAM optimizer on 2 epochs.

- Polarity. In this case, we address this subtask as a regression problem to take into account the ordinal nature of the categories. We used the same settings described before but in this case, minimizing the mean absolute error (MAE), and the optimal parameters were a sequence length of 128 tokens, batch size of 16, and a learning rate of 2e-05 with ADAM optimizer on 3 epochs.

Once we obtained the optimal hyper-parameters for each subtask, we used an approach based on an ensemble of BERT models. Recent research has shown that using an ensemble of single BERTs as "weak" models with a weighted voting scheme, provides a more robust model for classification tasks [10, 11].

For the classification problem, we used an ensemble of 4 BERTs with a weighted voting scheme, i.e., by accumulating the softmax layer outputs and selecting the class with the maximum weight. For the regression problem, we used the mean from an ensemble of 9 BERTs, rounding to the nearest integer in the valid range $[1, 5]$. It is worthwhile to say that the size of the ensemble for each subtask, was choosen in such a way that there would no longer be any significant improvement in the final result. This ensemble scheme is shown in Figure 6(a) for the type of opinion (classification) subtask, where we show the Macro F1-score for the ensemble of 4 BERTs with weighted voting scheme and the polarity (regression) subtask in Figure 6(b), were we show the MAE for the ensemble of 9 BERT models by using the mean.



(a) Ensemble of BERTs for the classification task

(b) Ensemble of BERTs for the regression task

Figure 6: Results for single BERT models (red crosses) and the ensemble of BERTs (blue line).

With the BERT models trained according to the ensemble scheme described before, we proceed to obtain the embeddings of size 768 given by the special token `[CLS]` for all texts in the corpus and all the single BERT models. These embeddings represents our new features, and they are aggregated to form an *extended* dataset, in a process we called *data augmentation on high level features*.

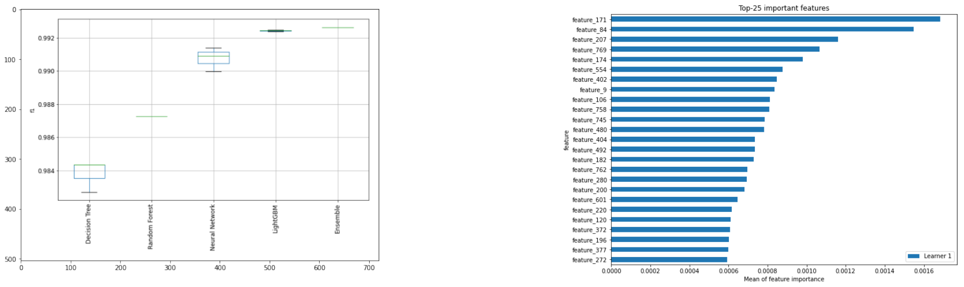
Let $\mathbf{X}^m$ to be the data matrix of size $n \times d$, where n is the size of the training corpus (i.e., the number of text opinions), d the number of features for each text $\mathbf{x}_i$ and $m = 1, \ldots, M$ the number of BERT models in the ensemble. Now, we concatenate each one of the m data matrix (one for each BERT model) to obtain the extended dataset $\mathbf{X}$ of size $(n \times M, d)$.

For the classification problem, the number of features $d = 768 + 1$, where we added the length of the text to the embeddings of the BERT models. For the regression problem, the number of features is $d = 768 + 2$, where we added the length of the text and the sentiment score of Vader lexicon. We decided to add these features because they showed to be relevant for these subtasks in the exploratory data analysis we did in Section 2.

3.2. Ensemble of classifiers

With the extended dataset we obtained in the Section 3.1 above, we search for the optimal prediction algorithm for the classification and regresion subtasks we defined before. To this end, we used the automated machine learning (AutoML) approach given by the `mljar-supervised` Python package [12]. In [12], it is said that AutoML with `mljar` "... abstracts the common way to preprocess the data, construct the machine learning models, and perform hyper-parameters tuning to find the best model" by "trying many different machine learning models (Algorithm Selection and Hyper-Parameters tuning)". This is achieved with the *Compete* mode, that trains highly-tuned ML models with ensembling and stacking [12]. In our case, the metrics to optimize for AutoML were Macro F1-score for the classification subtask and MAE for the regression subtask, with 90% of the extended dataset for training and the remaining 10% for testing.

The optimal ensemble for the classification problem consisted on the models LightGBM [13] and a Neural Network. In Figure 7(a), we can see the performance of different individual ML algorithms compared to the optimal ensemble respecting to Macro F1-score. For the regression problem, the optimal ensemble included LightGBM and XGBoost [14], and we can see in Figure 8(a) the performance of different ML models compared to the optimal ensemble respecting to MAE metric.

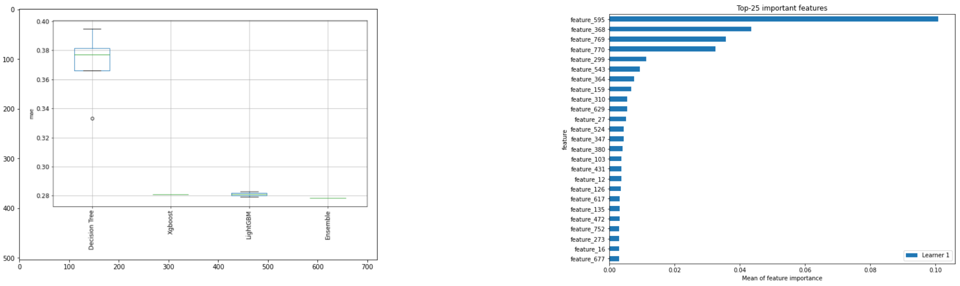


(a) AutoML performance boxplot

(b) AutoML features importance

Figure 7: AutoML results for the classification (type of opinion) subtask.

One of the advantages of AutoML, is that we can obtain an interpretable model, meaning that we can obtain a measure of the features importance. It is shown in Figure 7 and 8. We can observe that for the classification task, the length of the text (feature 769) is the fourth most important variable (Figure 7(b)), and for the regression problem, this feature is the third most important variable followed by the Vader score (feature 770), as can be seen in Figure 8(b).



(a) AutoML performance boxplot

(b) AutoML features importance

Figure 8: AutoML results for the regression (polarity) subtask.

BERT model	Single model	Model + AutoML	Improvement (%)
1	98.0900	98.5466	0.47
2	98.5103	98.6028	0.09
3	98.3356	98.5666	0.23
4	98.2963	98.5837	0.29
Ensemble	**98.8168**	**98.9806**	**0.17**

Table 1
Results for the type of opinion (hotel, restaurant or attraction) subtask according to the Macro F1-score evaluated on the test dataset. We shows the results for the single BERT models and the ensemble.

4. Results

The results we obtained with our proposal for both subtasks (classification and regression) are shown in Tables 1 and 2. Those results are evaluated on the test dataset (20% of the data) created previously.

We can observe that in both cases, there is a consistent improvement in the results when we consider the ensemble of BERT models, and furthermore, the best results are obtained when we used the ensemble of classifiers with AutoML, as was described in Section 3.2, even when it is applied to single BERT models. For the type of opinion subtask, we obtain an improvement of 0.17%, and for the polarity subtask, the improvement was 3.4%, when we use the BERT ensemble + AutoML in both subtasks.

5. Conclusions

In this paper, we described our proposal to address the sentiment analysis task from Rest-Mex 2022. Our experiments show the effectiveness of using high level features from texts learned from BERT models, and how they can be used to extend the dataset, in a similar way data augmentation does. By taking our extended dataset as input, we use AutoML techniques, proving to be a useful and efficient technique to obtain good ML models for the classification and regression problems we formulate to tackle this task. Our proposed methodology obtained

BERT model	Single model	Model + AutoML	Improvement (%)
1	0.2685	0.2579	3.95
2	0.2864	0.2624	8.38
3	0.2599	0.2545	2.08
4	0.2750	0.2647	3.75
5	0.2642	0.2498	5.45
6	0.2689	0.2535	5.73
7	0.3099	0.2624	15.33
8	0.2672	0.2525	5.50
9	0.2910	0.2588	11.07
Ensamble	**0.2526**	**0.2440**	**3.40**

Table 2
Results for the polarity $[1-5]$ subtask according to the MAE score evaluated on the test dataset. We shows the results for the single BERT models and the ensemble.

third place in this challenge, where we obtained a MAE score of 0.2642 for polarity and Macro F1-score of 0.9888 for the type of opinion, evaluated in the official test dataset.

References

[1] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodrıguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Sciences (2022). doi:10.1177/01655515221100952.

[2] INEGI, Estadísticas a propósito del día mundial del turismo, September 2021. URL: https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2021/EAP_Turismo21.pdf, comunicado de prensa núm. 539/21.

[3] INEGI, Indicadores trimestrales de la actividad turística. cuarto trimestre de 2021., April 2022. URL: https://www.inegi.org.mx/app/saladeprensa/noticia.html?id=7288, comunicado de prensa núm. 237/22.

[4] INEGI, Indicadores de la actividad turística, ???? URL: https://www.inegi.org.mx/temas/itat/, last accessed May 30, 2022.

[5] 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). doi:https://doi.org/10.26342/2021-67-14.

[6] 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).

[7] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: ICWSM, 2014.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, 2019, pp. 4171–4186. doi:https://doi.org/10.48550/arXiv.1810.04805.

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[10] V. Gómez-Espinosa, V. Muñiz-Sanchez, A. P. López-Monroy, Transformers pipeline for offensiveness detection in mexican spanish social media, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Á. Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 251–258. URL: http://ceur-ws.org/Vol-2943/meoffendes_paper3.pdf.

[11] M. Guzman-Silverio, Á. Balderas-Paredes, A. P. López-Monroy, Transformers and data augmentation for aggressiveness detection in mexican spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 293–302.

[12] A. Płońska, P. Płoński, Mljar: State-of-the-art automated machine learning framework for tabular data. version 0.10.3, 2021. URL: https://github.com/mljar/mljar-supervised.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[14] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. URL: https://doi.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.