

Participation of ESCOM's Data Science Group at Rest-Mex 2022: Sentiment Analysis Task^{*}

Julian Alcibar-Zubillaga¹, Yanina De-Luna Ocampo¹, Isaias Pacheco-Castillo¹, Kevin Ramirez-Mendez¹, Juan-Pablo-Minoru Sainz-Takata¹ and Omar Juárez Gambino^{1,*}

¹*Escuela Superior de Cómputo, Instituto Politécnico Nacional, ESCOM-IPN, J.D. Batiz e/ M.O. de Mendizabal s/n, Mexico City, 07738, Mexico*

Abstract

In this paper we describe the participation of the ESCOM Data Science group in Rest-Mex 2022 for the Sentiment Analysis task. For this task, 5 levels of polarity (1-5) as well as the type of attraction (restaurant, hotel and attraction) on which the opinion is given must be predicted. We followed a supervised approach using machine learning methods to train a model and then use it to make predictions. The model was tuned and tested with different text representations and obtained a combined score from both tasks of 0.84, which is only 0.5 points away from the best result.

Keywords

Sentiment Analysis, Machine Learning, Feature Selection,

1. Introduction

Sentiment Analysis (SA) is a challenging task that is defined as the computational treatment of opinion, sentiment and subjectivity[1]. This task is difficult due to the subjectivity that surrounds the analyzed text, the informality in writing and the problems inherent to the sources from which these texts are obtained, such as social networks or web platforms[2]. Sentiment polarity has been addressed as a text classification problem in[3][4]. In order to improve the performance additional resources as sentiment lexicons have been used[5][6].

Rest-Mex is an evaluation forum that proposes several challenges of Natural Language Processing to promote tourism in Mexico [7, 8, 9, 10]. In 2022 edition three task were proposed: recommendation systems, sentiment analysis and covid-19 epidemiological semaphore [11]. This forum is part of IberLEF@sepln 2022 which is an event that promotes Natural Language Processing systems in Spanish.


In this paper we describe our participation in the sentiment analysis task. Despite its simplicity, the trained model obtained competitive results. The rest of this paper is organized as follows. Section 2 describes the task and the corpus. Section 3 describes the method we used. Section 4 explained the performed experiments and the obtained results. Section 5 shows our conclusions and future work.

IberLEF 2022, September 2022, A Coruña, Spain

✉ b150697@sagitario.cic.ipn.mx (O.J. Gambino)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

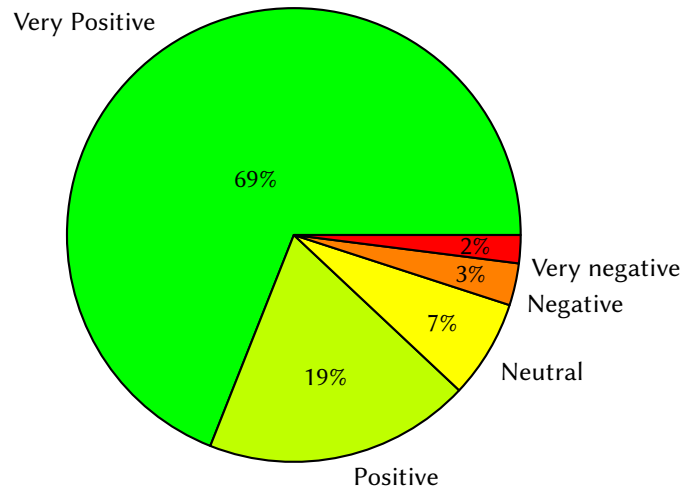


Figure 1: Sentiment polarity class distribution

2. Task and corpus description

The sentiment analysis task for this challenge aims to predict the polarity of an opinion issued by a tourist who traveled to the most representative places, restaurants and hotels in Mexico. The corpus was collected from tourists who shared their opinion on TripAdvisor between 2002 and 2021 and has the following structure:

- Title. Title of the opinion.
- Opinion. Opinion expressed by the user.
- Polarity. Sentiment polarity of the opinion.
- Attraction. Place visited by the user.

In this challenge there are two objectives, to predict the polarity of the users' opinion and the places the users visited. The polarity has the following class labels: Very negative (VN), Negative (N), Neutral (NEU), Positive (P), Very positive (VP). On the other hand, attraction has three class labels: hotel, restaurant, and attractive. For both task the text of title and opinion were concatenated in a single string and was used to train the model (see Section 4).

The corpus is divided in a training set that has 30,212 opinions while the testing set has 12,938 opinions. The distribution of class labels in the training set is shown in Figure 1 and Figure 2.

As can be seen, both tasks show an imbalance in class labels, especially in sentiment polarity. Machine learning methods have problems when classes are imbalanced and this problem affects performance, as described in Section 4.

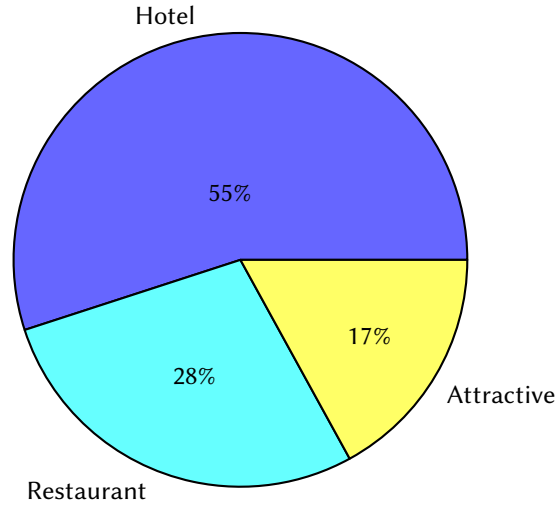


Figure 2: Attraction class distribution

3. Method

3.1. Preprocessing

The challenge corpus is provided as an XML file. The first step in our method was to extract the data and preprocess it. The following preprocessing phases were applied:

1. Cleaning data. There are misspelled words, words in another language and symbols that can affect the performance of the model. These words and characters were removed.
2. Tokenization. Text was separated into pieces called words or tokens.
3. Lemmatization. Inflectional forms of words were reduced to a common base form.
4. Stop words. Some common words are not useful for the proposed task, therefore these words were removed.

Tokenization and lemmatization were performed using the Freeling language suite[12]. Words such as articles, pronouns, and prepositions were considered stop words.

3.2. Text representation

The next step is to represent the text in a form appropriate for the machine learning model. The vector space model helps to transform the text into a vector with numerical values considering a vocabulary generated from all the words. These values can be the presence or absence of the word (binarized), the frequency of the word, or the relevance of the word in documents (TF-IDF)[13]. For our experiments binarized and frequency text representation were selected and we used scikit-learn [14] *CountVectorizer* function for this transformations.

3.3. Machine learning algorithm

Once the text has been vectorized, a machine learning algorithm can be used to train a model to tackle the tasks. We propose to consider both task—polarity prediction and attractive prediction—as a multiclass classification problem. Several supervised learning methods have been used for sentiment analysis and text classification[15][16] obtaining promising results. For our experiments we have selected the Logistic Regression classifier based on the performance reported in previous works[17][18].

4. Experiments and results

We carried out experiments using different text representations models and adjusting some parameters of the classifier. Two runs were performed with different configurations. These runs are explained in the following subsections.

4.1. Run 1

In this experiment we wanted to reduce the number of attributes used for training the classifier. The vector of frequencies had 32,339 different features and we considered that the size can be reduced using only the 1,000 most frequent words. A previous process that removed stop words (see 3.1) was also performed.

We created two sets with the training corpus. The first set with 80% of data used to adjust the parameters of the classifiers and the second set with 20% of data for testing the model. GridSearch method of Scikit-learn was used to adjust the hyper-parameters of Logistic Regression as penalty, solver, max_iter and multi_class (details of parameters are provided in the official web site <https://scikit-learn.org>). The class labels of sentiment polarity is unbalanced (shown in Figure 1, therefore the parameter class_weight was set to {VP:0.75, P:0.75, NEU:1, N:1, VN:1} with the idea that the minority class has a greater weight than the more frequent classes. In the case of the attraction class labels, the imbalance was smaller, so no adjustment was necessary.

The best parameters for the model that predicts the polarity and the attraction of the training dataset are presented in Tables 1 and 2.

Parameter	Value
solver	saga
C	2.5
multi_class	ovr
penalty	l2
class_weight	{5:0.75, 4:0.75, 3:1, 2:1, 1:1}

Table 1

Best parameters for sentiment polarity prediction in run 1

Tables 3 and 4 shows the results of the models. As can be seen, the model for predicting attraction obtained higher accuracy than the model for predicting sentiment polarity. This difference could be due to the high imbalance in the polarity classes and that the number of classes to be predicted are only 3.

Parameter	Value
solver	lbfgs
C	0.1
multi_class	ovr
penalty	l2
class_weight	auto

Table 2

Best parameters for attraction prediction in run 1

class	precision	recall	f1-score	support
VN	0.74	0.54	0.64	104
N	0.87	0.27	0.41	145
NEU	0.63	0.32	0.43	422
P	0.59	0.29	0.39	1163
VP	0.79	0.97	0.87	4209
accuracy			0.77	6643
macro avg	0.72	0.48	0.54	6043
weighted avg	0.74	0.77	0.73	6043

Table 3

Sentiment polarity results in the train set for run 1

class	precision	recall	f1-score	support
Hotel	0.98	0.98	0.98	3248
Restaurant	0.69	0.96	0.96	1740
Attractive	0.99	0.99	0.99	1055
accuracy			0.98	6643
macro avg	0.98	0.98	0.98	6043
weighted avg	0.98	0.98	0.98	6043

Table 4

Attraction prediction results in the train set for run 1

These two models were used to predict sentiment polarity and attraction in the test set. The same preprocessing and reduction of features were applied to this set. Results of the models are shown in Tables 5 and 6.

Class	F-measure	Recall	Precision
VN	0.10093168	0.0629845	0.25390625
N	0.05353728	0.06730769	0.04444444
NEU	0.1004878	0.08833619	0.11651584
P	0.18380125	0.17998418	0.18778374
VP	0.68073136	0.72563718	0.6410596
Macro	0.223897877	0.224849948	0.248741975
Accuracy	49.81449992		
MAE	0.855928273		
Final rank			
0.59562279			

Table 5
Sentiment polarity results in the test set for run 1

Class	F-measure	Recall	Precision
Hotel	0.74599431	0.8827719	0.64591549
Restaurant	0.59913849	0.56820995	0.63362783
Attractive	0.61216216	0.48920086	0.81768953
Macro	0.652431654	0.646727571	0.699077618
Accuracy	67.18967383		

Table 6
Attraction prediction results in the test set for run 1

Although the results on the training corpus were good, the model performance decreased on the test dataset according to the official Rest-Mex challenge results. We believe this could be due to the large feature reduction (about 97% reduction).

4.2. Run 2

For the second run we used the same partition of the training corpus explained above in order to adjust some parameters. But no feature reduction was done (only stop words were removed). The best parameters found are shown in Tables 7 and 8.

The results during the training phase with these parameters are shown in Tables 9 and 10.

Parameter	Value
solver	lbfgs
C	1.8
multi_class	ovr
penalty	l2
max_iter	10000

Table 7
Best parameters for sentiment polarity prediction in run 2

Parameter	Value
solver	lbfgs
C	1.5
multi_class	ovr
penalty	l2
max_iter	10000

Table 8
Best parameters for attraction prediction in run 2

class	precision	recall	f1-score	support
VN	0.45	0.36	0.40	104
N	0.28	0.13	0.18	145
NEU	0.39	0.32	0.35	422
P	0.43	0.36	0.39	1163
VP	0.83	0.90	0.86	4209
accuracy			0.73	6643
macro avg	0.47	0.41	0.44	6043
weighted avg	0.70	0.73	0.71	6043

Table 9
Sentiment polarity results in the train set for run 2

class	precision	recall	f1-score	support
Hotel	0.98	0.98	0.98	3248
Restaurant	0.95	0.96	0.96	1740
Attractive	0.99	0.98	0.98	1055
accuracy			0.97	6043
macro avg	0.97	0.97	0.97	6043
weighted avg	0.97	0.97	0.97	6043

Table 10
Sentiment polarity results in the train set for run 2

As we can see, the accuracy obtained by the model using more features suffered a small decrease during the training phase. The imbalance in the polarity classes also affects the performance of the classifier in this experiment. Both models were used in the test set and the results are shown in the tables 11 and 12.

Class	F-measure	Recall	Precision
VN	0.30538922	0.62921348	0.21875
N	0.33962264	0.38095238	0.05079365
NEU	0.32185629	0.40812721	0.26131222
P	0.44867415	0.4184168	0.42756913
VP	0.83167285	0.82908346	0.89359823
Macro	0.4494430291	0.5312243137	0.4465032266
Accuracy	69.24563302		
MAE	0.3533776472		
Final rank			
0.8400073285			

Table 11
Sentiment polarity results in the test set for run 2

Class	F-measure	Recall	Precision
Hotel	0.95936571	0.96441788	0.9543662
Restaurant	0.9364032	0.95258498	0.92076201
Attractive	0.92759888	0.8900871	0.96841155
Macro	0.9411225952	0.935696651	0.9478465865
Accuracy	94.73643531		

Table 12
Sentiment polarity results in the test set for run 2

Although the performance of the model was lower during the training stage for this experiment, better results were obtained with the test set in both task. We believe that the increase in features allows the model to generalize better.

5. Conclusions and future work

Sentiment analysis is a challenging and interesting task. User opinions about places they visited can influence the decision of future travelers, so it is important to have algorithms that analyze these opinions and determine the polarity they express. In this paper we report our participation in the Rest-Mex forum. A Logistic Regression classifier was used to train two models, one for polarity prediction and other for attraction prediction. Two experiments were performed, the first one used only 3% of the features obtained from the vector representation of the text and obtained good results with the training set but the performance was lower in the test set. The second, used the full features and despite obtaining lower results in the training set, showed

a significant improvement in the test set. As future work, we propose to use more advanced feature selection techniques that can help the model to generalize better.

6. Acknowledgments

This research was funded by CONACyT-SNI and Instituto Politécnico Nacional (IPN), through grants SIP-20220620, SIP-2083 and EDI.

References

- [1] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, *Foundations and Trends® in information retrieval* 2 (2008) 1–135. doi:<https://doi.org/10.1561/15000000011>.
- [2] D. M. E.-D. M. Hussein, A survey on sentiment analysis challenges, *Journal of King Saud University - Engineering Sciences* 30 (2018) 330–338. doi:<https://doi.org/10.1016/j.jksues.2016.04.002>.
- [3] H. Calvo, O. Juárez Gambino, Cascading classifiers for twitter sentiment analysis with emotion lexicons, in: *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2016, pp. 270–280. doi:https://doi.org/10.1007/978-3-319-75487-1_21.
- [4] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, End-to-end adversarial memory network for cross-domain sentiment classification., in: *IJCAI*, 2017, pp. 2237–2243.
- [5] O. J. Gambino, H. Calvo, A Comparison Between Two Spanish Sentiment Lexicons in the Twitter Sentiment Analysis Task, in: *Advances in Artificial Intelligence - IBERAMIA 2016*, Springer International Publishing, 2016, p. 127–138.
- [6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, *Computational linguistics* 37 (2011) 267–307. doi:https://doi.org/10.1162/COLI_a_00049.
- [7] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: recommendation system for text mexican tourism 67 (2021). doi:<https://doi.org/10.26342/2021-67-14>.
- [8] R. Guerrero-Rodríguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, *Current Issues in Tourism* (2021) 1–16. doi:<https://doi.org/10.1080/13683500.2021.2007227>.
- [9] M. A. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, *Computación y Sistemas* 26 (2022). doi:<https://doi.org/10.13053/CyS-26-2-4055>.
- [10] 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).

- [11] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, *Journal of Information Science* (2022). doi:<https://doi.org/10.1177/01655515221100952>.
- [12] L. Padró, E. Stanilovsky, FreeLing 3.0: Towards Wider Multilinguality, in: *Proceedings of the Language Resources and Evaluation Conference, ELRA, Istanbul, Turkey*, 2012.
- [13] C. Manning, H. Schutze, *Foundations of statistical natural language processing*, 1999.
- [14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [15] N. F. Da Silva, E. R. Hruschka, E. R. Hruschka Jr, Tweet sentiment analysis with classifier ensembles, *Decision support systems* 66 (2014) 170–179. doi:<https://doi.org/10.1016/j.dss.2014.07.003>.
- [16] K. Shah, H. Patel, D. Sanghvi, M. Shah, A comparative analysis of logistic regression, random forest and knn models for the text classification, *Augmented Human Research* 5 (2020) 1–16. doi:<https://doi.org/10.1007/s41133-020-00032-0>.
- [17] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, *Information* 10 (2019) 150. doi:<https://doi.org/10.3390/info10040150>.
- [18] M. Goswami, P. Sajwan, A comparative analysis of sentiment analysis using rnn-lstm and logistic regression, in: *Trends in Wireless Communication and Information Security*, Springer Singapore, Singapore, 2021, pp. 165–174. doi:https://doi.org/10.1007/978-981-33-6393-9_18.