Naive Features for Sentiment Analysis on Mexican Touristic Opinions Texts

Gabriela Carmona-Sánchez[1], Ángel Carmona[1], and Miguel Á. Álvarez-Carmona[2,3][0000−0003−4421−5575]

[1] Benemérita Universidad Autónoma de Puebla (BUAP), 72000, Puebla, Mexico
[2] Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad de Transferencia Tecnológica Tepic (CICESE-UT3), 63173, Nayarit, Mexico
[3] Consejo Nacional de Ciencia y Tecnología (Conacyt), 03940, CDMX, Mexico

Abstract. This paper presents a simple approach to extract naive features to represent and classify tourists' opinions in Mexican places to participate in the Rest-Mex 2021 evaluation forum. The proposed approach consists of extracting 15 simple features. Then, various classification algorithms were used to evaluate the quality of these features, such as SVM, KNN, Decision Tree, Random Forest, and Naive Bayes. A weighting scheme was also proposed to obtain the best combination between algorithms and features, where it turned out that the best algorithm for this set of features was KNN with seven neighbors. Of these features, the best turned out to be what had to do with the length of words and characters and the number of stop words. With this approach, 0.76 of MAE was obtained, obtaining 10th place out of 15 teams, which considering the simplicity of this solution, makes it an acceptable result.

Keywords: Naive features · Sentiment analysis · Mexican tourist texts.

1 Introduction

In recent years, tourist texts have taken on great importance in artificial intelligence investigations. This is due to the advantages that can be obtained from analyzing this type of text. One of them is to analyze the sentiment of tourists who leave, writing through digital platforms such as TripAdvisor. In this way, it is possible to automatically determine the user's experience, determine if their comment is positive or negative and through this information, find possible improvements that can be made to improve the experience of other tourists over time.

This task falls in the area of natural language processing, specifically within sentiment analysis. This task determines if the author of a text expresses himself positively or negatively about a product or service received. There are variations

IberLEF 2021, September 2021, Málaga, Spain.

in the task where it is also about determining if the opinion is neutral; it can even go further and determine a numerical scale between 0 and N where 0 would be the most negative and N the most positive [8].

In this way, the sentiment analysis task can be seen as an automatic classification task where the instances are texts, and the class is the text polarity.

Typically, various textual representations are used for this task, such as n-grams, dictionaries, and embeddings, among others, used to feed classifiers, train them, and test them to observe their performance. However, there are scenarios where it is more critical than few computational resources are used both in time and memory due to limitations of specific tasks, for example, implementing solutions for IoT devices [5].

In this work, we propose to study the scope and effectiveness of features based on describing the text to be analyzed. For their simplicity, we will call these features Naive Features.

To test these types of features, the database that was released for the Rest-Mex 2021 evaluation forum will be used [2]. For this edition, a corpus of texts from tourists who visited Guanajuato in Mexico and its attractions was released. In this way, the effectiveness of these features can be tested in texts in Spanish since one of their advantages is that they are independent of the language.

The rest of the document is organized as follows; In section 2, the methodology followed in this work is described. In section 3, the results and their analysis are presented. Finally, section 4 presents the conclusions of this work.

2 Methodology

This section presents the database with which it experimented and the proposed methodology to represent the tourist texts.

2.1 Data set

The analysis of sentiments task in tourism texts, which this year proposed within the Rest-Mex 2021 evaluation forum, predicts a class for each review provided in the evaluation set. The available classes are integers in the range [1, 5]. Reviews were taken from the TripAdvisor website and were written by a tourist who evaluated some of the emblematic places in Guanajuato, Mexico. It is essential to mention that the whole set was in Spanish, being the first data set with these available features for evaluation.

The forum organizers released two different data sets ; one for training and one for evaluation. The training set consists of 5197 opinions with 9 pieces of information described below:

- Index: the index of each opinion.
- Title: The title that the tourist himself gave to his opinion.
- Opinion: The opinion expressed by the tourist.

https://sites.google.com/cicese.edu.mx/rest-mex-2021

Table 1. Distribution of the class on the Rest-Mex 2021 training data set.

Class	instances
1	80
2	145
3	686
4	1596
5	2690
Total	5197

- Place: The tourist place that the tourist visited and to which the opinion is directed.
- Gender: the gender of the tourist.
- Age: The age of the tourist at the time of issuing the opinion.
- Country: The country of origin of the tourist.
- Date: the date the opinion was issued.
- Label: The label representing the polarity of the opinion: [1, 2, 3, 4, 5].

The training set classes are unbalanced. as Table 1 shows.

Finally, the test data set contained 2216 rows and the same information as the training set, except the class information.

2.2 Proposed Approaches

To attack the sentiments analysis task in tourist data, it is proposed to use simple features that can capture important information to determine the polarity of an opinion in such a way that it is quick to calculate and represent. Especially to offer an option for restricted applications in time or memory (such as IoT solutions) and that cannot use approaches that, although they have outstanding effectiveness results, can be slow or use much computational power, in addition to having the advantage of being language-independent features.

Given a text in the data set, its representation will be given by the following features proposed:

- F1: Number of capital letters in the opinion
- F2: The length of the longest word in the opinion
- F3: The average words length in the opinion
- F4: Number of words in the opinion
- F5: Number of characters in the opinion
- F6: The ratio between the number of different words and total words in the opinion
- F7: The number of digits in the opinion
- F8: The ratio of the number of stop words and total words in the opinion
- F9: Number of punctuation marks in the opinion
- F10: Number of stop words in the opinion
- F11: Number of characters in the opinion without stop words

– F12: The ratio between the number of different words and total words in the opinion without stop words

The information available for each opinion will also be added as:

– F13: The gender of the person who gave the opinion
– F14: The age of the person who gave the opinion
– F15: The country of the person who gave the opinion

For feature F13 that refers to the opinion author's gender, the value will be 0 if it is a man, 1 if it is a woman, and 2 if the gender of the person is not known. For feature F15, the country will be coded as 0 if the person is from Mexico or 1 if not.

Each option is transformed into a vector representation of dimension 15, which is easy and fast to calculate and independent of the language, which means that data sets in different languages can be evaluated.

The 10 fold cross-validation approach was used to classify the data set [9]. For each partition, the following classifiers were applied:

– Support Vector Machines (SVM) [6]
– k-Nearest Neighbor (KNN) with $k \in \{1, 3, 5, 7\}$ [3]
– Decision Tree (DT) [4]
– Random Forest (RF) [7]
– Naive Bayes (NB) [1]

Accuracy, F-measure, and MAE were used as evaluation measures since it is the measure that the organizers take as official.

3 Results

In this Section, the results obtained for the training partition are presented. Afterward, the chosen model is presented to be evaluated in the training partition together with its obtained result.

3.1 Training data set results

Table 2 shows the results for each classification algorithm for the representation described in the 2 section. This table shows the columns of accuracy (Acc), maco F-measure (F), the F-measure of each class (F CN where N represents each of the 5 classes), and mae (MAE).

In the results, it can be seen that SVM obtains the best results for accuracy and for class 5, which is the majority class; however, it obtains 0 for all other classes. This means that although it performs well for the measure mae, it is only classifying one class. On the other hand, KNN obtains the best results for macro measurement F. Class 1 obtains its best result with KNN-1, class 2 with

Table 2. Training data set results

Algorithm	Acc	F	F C1	F C2	F C3	F C4	F C5	MAE
SVM	**51,74**	0,13	0	0	0	0	**0,68**	0,71
KNN-1	38,21	**0,21**	**0,04**	0,05	0,15	0,29	0,51	0,87
KNN-3	38,59	**0,21**	0,02	0,04	**0,19**	0,25	0,54	0,91
KNN-5	41,96	**0,21**	0	0,01	0,18	0,3	0,55	0,77
KNN-7	44,17	**0,21**	0,02	0,01	0,16	0,28	0,59	0,74
DT	38,57	0,2	0,01	0,04	0,15	**0,3**	0,52	0,87
RF	48,64	0,19	0	0,02	0,07	0,21	0,65	**0,7**
NB	47,46	0,19	0,02	**0,08**	0,01	0,21	0,64	0,79

Table 3. Information gain for the best features

Feature	Description	Information gain
F4	Number of words	0,307
F8	Ratio of stop words and total words	0,299
F7	Number of digits	0,272
F3	Average words length	0,149
F12	Ratio of different words and total words without stop words	0,143
F2	Length of the longest word	0,130

NB, class 3 with KNN-3, and class 4 with DT. The best result of mae is obtained with RF; however, this algorithm cannot capture any instance of class 1.

Table 3 shows the best features under the information gain measure. Only the features that obtained a value greater than 0,1 appear in this table. These results give evidence that the best feature to solve this task is the number of words in the opinion, the second is the number of stop words, while the number of digits, the average of the length of the words, the different words not counting stop words and the length of the longest word complement the list.

Figure 1 shows the decision tree when only the word count feature is used. Ten opinions from class 1 (Bad label) and ten from class 5 (Good label) were used to build this decision tree. These opinions were chosen randomly. For this sample, it is possible to see that negatively valued opinions tend to have more words, which gives evidence that when the tourist is not satisfied, he uses more words to exorcise it, and could be the reason why this feature is important in this task.

It is clear to see that although a good mae result can be considered for all the algorithms, the F-measure results are shallow, which is a consequence of the data imbalance. This makes choosing a classification model not easy.

3.2 Test data set results

To choose the best model from those presented in Table 2 it is proposed to implement a weighting scheme to determine which of all the algorithms presents

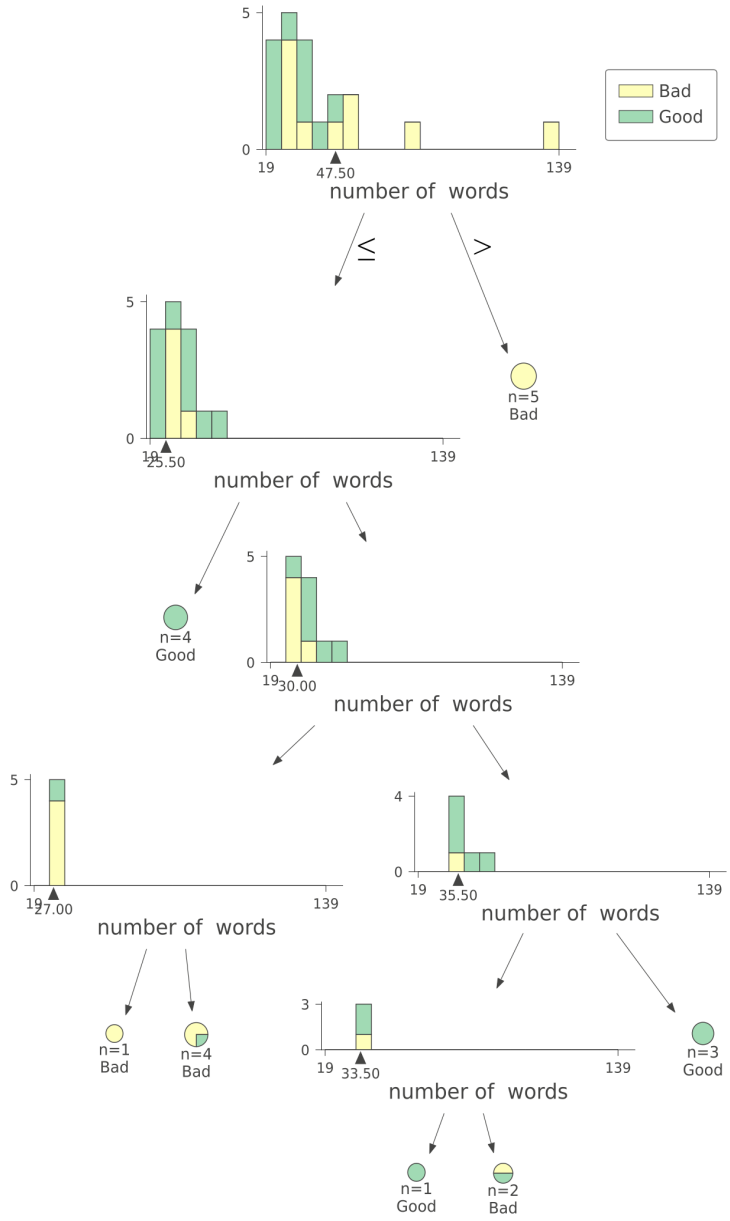

Fig. 1. Decision tree for the number of words feature on 20 random opinions

Table 4. Measure of quality Q

Algorithm	SVM	KNN-1	KNN-3	KNN-5	KNN-7	DT	RF	NB
Q	14,63	19,63	15,20	30,74	**36,73**	17,11	34,22	26,17

the best balance between the different results (accuracy, F-measure of each class, and mae).

It is proposed to generate a linear combination to measure the quality of each of the results obtained as presented in the equation 1.

$$Q = C_1 * Acc + C_2 * F + C_3 * F_1 + C_4 * F_2 + C_5 * F_3 + C_6 * F_4 + C_7 * F_5 + C_8 * MAE \quad (1)$$

Where C_i represents the importance of each variable in the equation. F_j represents the F-measure results for the class j. Acc, F, and MAE represent the valor of accuracy, macro F-measure and mae, respectively.

To choose the value of each constant C_i, the following weights are proposed:

- C_1: Since accuracy is not an important measure because the collection is unbalanced, it will only be given a weight of 1.
- C_2: It seeks to obtain a high F-measure macro result so that it will be given a weight of 10.
- $C_{\{3,4,5,6,7\}}$: The higher the number of opinions in a class, the easier it is to classify, which means that classes with little data are more complicated. In this way, it is sought to reward the well-classified elements of minority classes by putting as a weight $100 - D(i)$ where $D(i)$ is the percentage of the class i.
- C_8. Since MAE is the measure taken into account to order the results by the organizers, it will be given the greatest weight, which must be negative since the ideal is to get as close to zero in this measure. Thus the value of this constant will be -100.

Table 4 shows the results of the equation 1. It is possible to see that the algorithm that presents the best value of Q is KNN-7, which did not present a high individual value of some measure; however, it is the one that obtains the best balance. On the other hand, SVM that obtained good results for accuracy, F-measure for class 5, and mae is the one that obtains the worst value of Q.

For this reason, in order to be evaluated in the Rest-Mex 2021 evaluation forum, it was decided to send the model generated by KNN-7 to the organizers.

3.3 Official results

For the official results, the model proposed obtained the following results:

- Accuracy: 45,71
- Macro F-measure: 0,17

– Mae: 0,76

With these results, the approach proposed obtained 10th place of 15 teams. Also, it is obtained better F-measure results than the baseline, and it was capable of classifying instances in three of the five classes.

4 Conclusions

In this work, a study was presented to measure the performance of naive features to attack the sentiment analysis problem for Mexican tourist texts.

This solution consisted of representing each tourist opinion in 15 simple features that can be extracted very quickly. This simplicity makes this solution ideal for some applications with an extreme limit of space and memory, for example, in IoT devices, and thus they can use some of these features to obtain an acceptable performance in a shorter response time.

When evaluating this solution in the Rest-Mex 2021 corpus, 0,76 of MAE was obtained, where the best result obtained in the competition was 0,47. Considering that the maximum possible error is 4 (when the result can be 5, and the prediction is 1, for example), 0,29 represents 7,25 % of the possible error, which is an acceptable loss considering the simplicity solution.

Evidence is given that the number of words in the opinion gives much information about polarity. Also, the length of the words is an essential source that a classifier can use. Other important features for this task are those that have to do with the stop words. Finally, for this task and in this database, demographic features such as gender, age, and place of origin of the author of the opinion do not seem to provide relevant information for the classification.

As work in the future, it is proposed to apply this solution to a multilingual collection and to be able to exploit its best feature, which is that it is a language-independent solution.

References

1. Ahmed, H.M., Javed Awan, M., Khan, N.S., Yasin, A., Faisal Shehzad, H.M.: Sentiment analysis of online food reviews using big data analytics. Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. Elementary Education Online **20**(2), 827–836 (2021)
2. Álvarez-Carmona, M.Á., Aranda, R., Arce-Cárdenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A.P., Martínez-Miranda, J., Pérez-Espinosa, H., Rodríguez-González, A.: Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism. Procesamiento del Lenguaje Natural **67** (2021)
3. Arslan, H., Arslan, H.: A new covid-19 detection method from human genome sequences using cpg island features and knn classifier. Engineering Science and Technology, an International Journal **24**(4), 839–847 (2021)

4. Charbuty, B., Abdulazeez, A.: Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends **2**(01), 20–28 (2021)
5. Dinculeană, D., Cheng, X.: Vulnerabilities and limitations of mqtt protocol used between iot devices. Applied Sciences **9**(5), 848 (2019)
6. Huang, Y., Zhao, L.: Review on landslide susceptibility mapping using support vector machines. Catena **165**, 520–529 (2018)
7. Mishra, S., Tadesse, Y., Dash, A., Jena, L., Ranjan, P.: Thyroid disorder analysis using random forest classifier. In: Intelligent and cloud computing, pp. 385–390. Springer (2021)
8. Mukherjee, S.: Sentiment analysis. In: ML. NET Revealed, pp. 113–127. Springer (2021)
9. Rohani, A., Taki, M., Abdollahpour, M.: A novel soft computing model (gaussian process regression with k-fold cross validation) for daily and monthly solar radiation forecasting (part: I). Renewable Energy **115**, 411–422 (2018)