Revista Innovación y Software
Vol. 4, No. 1, Mes Marzo-Agosto, 2023
ISSN: 2708-0935
Pág. 52-62
https://revistas.ulasalle.edu.pe/innosoft



Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 22/10/2022 | Aceptado: 10/12/2022 | Publicado: 30/03/2023

Identificadores persistentes:
ARK: ark:/42411/s11/a63
PURL: 42411/s11/a63

# Identificador de sentimientos de comentarios de hoteles utilizando BERT

# *Hotel feedback sentiment identifier using BERT*

**Walther Medina Pauca [1*], Camila Huamani Tito [2]**

[1] Universidad La Salle . wmedinap@ulasalle.edu.pe
[2] Universidad La Salle. chuamanit@ulasalle.edu.pe

[*] Autor para correspondencia: wmedinap@ulasalle.edu.pe

**Resumen**

La forma de escribir del ser humano fue cambiando con el tiempo siendo reducidas/abreviadas por las nuevas generaciones. El proyecto investigará estas formas de escribir de la personas a través de comentarios de hoteles, para poder identificar y realizar su clasificación de acuerdo si este es un comentario formal o informal; a la vez se tratará de identificar cada uno de estos si cuenta con información positiva o negativa. Todos los procesos para identificar textos serán usados con el Procesamiento de lenguaje natural (NLP), así lograremos identificar diferentes oraciones de acuerdo al contexto que se encontrará en el comentario de la base de datos, la cual será sacada de TripAdvisor.

**Palabras clave:** Clasificación, comentarios, dataset, exactitud, promedio, procesamiento de lenguaje natural, NLP.

***Abstract***

*The way of writing of the human being was changing over time being reduced/abbreviated by the new generations. The project will investigate these forms of writing of people through hotel comments, in order to identify and classify them according to whether it is a formal or informal comment; at the same time we will try to identify whether each of these has positive or negative information. All the processes to identify texts will be used with Natural Language Processing (NLP), so we will be able to identify different sentences according to the context that will be found in the comment database, which will be taken from TripAdvisor.*

***Keywords:*** *Accuracy, averaging, classification, comments, dataset, natural language processing, NLP.*

Revista Innovación y Software
Vol. 4, No. 1, Mes Marzo-Agosto, 2023
ISSN: 2708-0935
Pág. 52-62
https://revistas.ulasalle.edu.pe/innosoft

## Introducción

Cuando se requiere buscar un hotel y ver si es adecuado con las características que se requieren, las personas pocas veces van a los comentarios, donde les brindan una idea externa a la experiencia o la crítica que dan por su estancia en un particular hotel, con esto en mente el proyecto consiste en clasificar los comentarios de acuerdo a los ejemplos vistos de otros sitios de reserva de hoteles, entres esto destaca el tamaño del comentario, la forma de describir el entorno y la experiencia en dicho establecimiento, usando algoritmos de clasificación de textos y obteniendo archivos CSV para los comentarios de los hoteles.

Para la prueba del buen funcionamiento del sistema a implementar es adquirir datos de varias páginas de reserva de hoteles que cuenten con comentarios tanto positivos y negativos para sacar la forma de escritura de algunos clientes o críticos para así realizar las debidas comparaciones, de acuerdo a la base de datos lingüística que tendrá el proyecto. Por eso se buscará en kaggle.com, que tiene más de 515,000 comentarios de clientes en inglés, además el documento nos mostrará diferentes atributos, pero sólo tendremos en cuenta: Negative_Review, Positive_Review y Reviewer_Score. Dichos datos serán obtenidos por el CSV que dará la página. Además, su obtención del archivo CSV se usará un árbol para ir buscando las palabras reservadas Como solo 3 atributos son elegidos su almacenamiento irá a una base de datos para que la máquina aprenda los patrones para futuras consultas. Un buen machine learning, además de almacenar datos para futuras llamadas es identificar escritura informal o formal, está usará un árbol de recorrido y búsqueda usando otro método de machine learning para hacer las debidas comparaciones y mandar en el caso de ser formal o informal, la cantidad exacta que contiene el comentario entre las dos formas de escritura. Lo mismo sucedería con la identificación de ser comentarios críticos o positivos, pasará por un método de Machine Learning realizando las debidas consultas. Para la realización de clasificar los comentarios de acuerdo a experiencias positivas como negativas el análisis de sentimiento es el adecuado para realizar dicha acción. El proyecto será realizado en una página web, mostrando los comentarios de diferentes hoteles de manera “filtrada” como pantalla principal, donde los comentarios buenos y descriptivos de las características del hotel serán los primeros.

Además, como son textos y no simplemente palabras soltadas al azar es un poco más complicado identificar el contexto de estas; no es algo raro de encontrar ya que se considera como el lenguaje neutral siendo ambiguo, pero usando el Part of Speech Tagging, se logrará identificar oraciones para clasificarlas gramaticalmente.

### Materiales y métodos o Metodología computacional

Revista Innovación y Software
Vol. 4, No. 1, Mes Marzo-Agosto, 2023
ISSN: 2708-0935
Pág. 52-62
https://revistas.ulasalle.edu.pe/innosoft

Usando NLP y cado uno de las propuestas que se fueron estableciendo, cuentan con diferentes librerías o la misma librería, en todos los proyectos presentados siempre estará la librería NLTK (Natural Language Toolkit), el cual es la que tiene más información y más material para su desarrollo, el problema que al ser una librería completa es pesada al momento de ser cargada y ejecutada para las aplicaciones. A pesar que el NLTK sea completo, el proyecto usará la librería BeautifulSoup, transformers, torch sklearn, ya que éste simplifica el NLTK de una manera intuitiva sacando lo que es el Part-of-speech tagging, usa el método de Naive Bayes, Decision Tree (métodos más certeros al momento de identificar palabras, pero a la vez los más lentos), también esta librería nos ofrece funciones para una clasificación rápida de las palabras negativas como positivas.

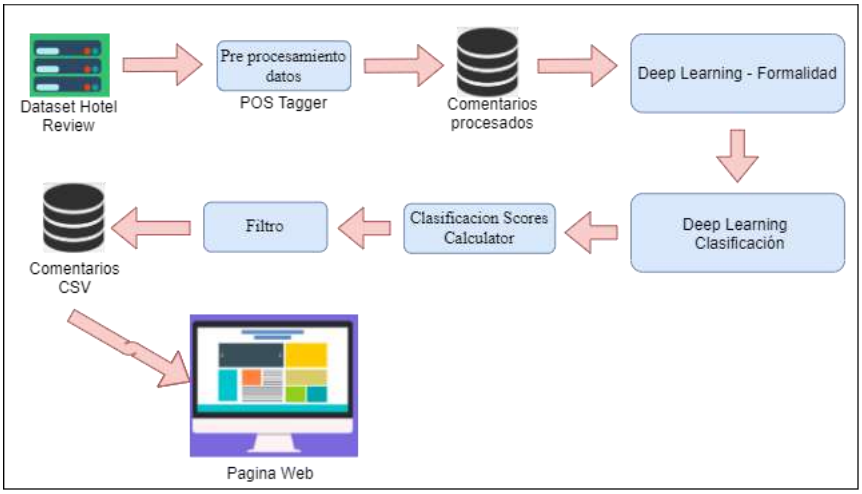


Fig 1. Pipeline de clasificación de comentarios

Para la clasificación de los comentarios/reseñas seguiremos los pasos descritos en la Fig 2.

- Dataset Hotel Review: Estamos obteniendo la información de un dataset que nos servirá de modelo para el software. Este dataset podrá ser cambiado por una base de datos real con reseñas reales.
- Pre-procesamiento datos: Haciendo uso de Part of Speech Taggin para lograr identificar los comentarios reales de comentarios con palabras aleatorias.
- Comentarios procesados: Corresponde al almacenamiento de los comentarios previamente clasificados como reales para utilizarlos posteriormente.

- Deep Learning - Formalidad: Corresponde al identificador de comentarios formales y los comentarios informales que contengan jergas.
- Deep Learning - Clasificación: Corresponde al identificador de comentarios críticos constructivos y de los demás.
- Clasificación Scores Calculator: Permite agregar una clasificación a comentarios que se han considerado como neutros dependiendo de la puntuación que le brindaron en la reseña.
- Filtro: Este filtro está dado por nosotros, pudiendo ser variable permitiendo así obtener los datos que realmente necesitamos.
- Comentarios CSV: Los comentarios ya luego de ser filtrados se almacenarán en formato CSV. Para su creación, utilizamos diferentes librerías que nos permitieron realizar las diferentes operaciones sobre el dataset y este muestre los resultados mejor esperados.

## Bibliotecas utilizadas:

**Transformers.-** Proporciona APIs para descargar y entrenar fácilmente modelos pre entrenados de última generación.

**Torch.-** Es un framework de redes neuronales, para diseñar y entrenar redes neuronales.

**BeautifulSoup.-** biblioteca de Python para extraer datos de archivos HTML y XML.

**Sklearn.-** Proporciona algoritmos para machine learning.

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import requests
import numpy as np
import pandas as pd
from bs4 import BeautifulSoup
import re
from sklearn.metrics import *
```

Entrenando el algoritmo con el diccionario de datos de BERT MULTILINGUAL.

```
tokenizer = AutoTokenizer.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')
```

```
model = AutoModelForSequenceClassification.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')
```

Se generan tokens para hacer la prueba con la oración ingresada manualmente.

```
tokens = tokenizer.encode('meh, it was okay', return_tensors='pt')
result = model(tokens)
##print(result.logits)
```

Se carga el dataset para comenzar con la limpieza del mismo.

```
df = pd.read_csv('dataset2.csv')
print( df.head())
```

Reduciendo los datos mostrados:

Filtramos el dataframe.

```
dfreview = df[['Review']]
```

Verificamos datos duplicados

```
print('Cantidad de datos duplicados: ', dfreview.duplicated().sum())
```

Verificamos la existencia de datos NULL

```
print('Cantidad de datos nulos: ', df.isnull().sum())
```

Hacemos un conteo de los datos luego de la limpieza.

```
print('Cantidad de datos en Review: ', df['Review'].value_counts())
```

Encapsulamos el proceso de sentimiento en una función, esto hace que sea más fácil procesar múltiples cadenas.

```
def sentiment_score(review):
    tokens = tokenizer.encode(review, return_tensors='pt')
    result = model(tokens)
    return (int(torch.argmax(result.logits))+1)
```

Se utiliza la función para poder generar la revisión en el marco de datos.

```
df['Sentiment'] = df['Review'].apply(lambda x: sentiment_score(x[:512]))
```

Se convierte el valor de las columnas en listas, para poder ser procesadas.

```
sentiment_column = df.loc[:,'Sentiment']
sentiment_nums = sentiment_column.values
sentiment_nums = sentiment_nums.tolist()
rating_column = df.loc[:,'Rating']
rating_nums = rating_column.values
rating_nums = rating_nums.tolist()
```

Convirtiendo los valores de ambas listas a clases A, B y C.

```
filter_actual = []
for element in rating_nums:
    if element < 3:
        filter_actual.append('A')
    elif element == 3:
        filter_actual.append('B')
    else:
        filter_actual.append('C')
filter_pred = []
for element in sentiment_nums:
    if element < 3:
        filter_pred.append('A')
    elif element == 3:
        filter_pred.append('B')
    else:
        filter_pred.append('C')
```

Se construye una serie con pandas, para que se pueda construir la matriz de confusión

```
filter_actual = pd.Series(filter_actual, name='actual')
```

```
filter_pred = pd.Series(filter_pred, name='pred')
```

Se halla la matriz de confusión, un informe de clasificación y el valor exactitud.

```
cm = confusion_matrix(filter_actual, filter_pred)
matrix = classification_report(filter_actual, filter_pred)
accuracy = accuracy_score(filter_actual, filter_pred)
```

### Macro average metrics:

Macro average precision implementado para casos en los que se tengan más de dos clases.

```
print('La presicion Score de macro es: ', precision_score(filter_actual,filter_pred, average='macro'))
```

### Macro average recall

```
print('El macro promedio recall es: ', recall_score(filter_actual,filter_pred, average='macro'))
```

Informe de clasificación para la precisión, el recall, la f1-score y la accuracy.

```
matrix = classification_report(filter_actual, filter_pred)
```

*En esta sección se explica cómo se hizo la investigación. Se describe el diseño de la misma y se explica cómo se llevó a la práctica, justificando la elección de métodos y técnicas de forma tal que un lector pueda repetir el estudio.*

#### Subepígrafes en caso de utilizarse

Los párrafos se escribirán en Times New Roman a 11 puntos y con espaciado 1,5 y una línea en blanco como separador.

## Resultados y discusión

Luego de la limpieza y el procesamiento del dataset, podemos presentar los siguientes resultados.

Conforme a la primera obtención de resultados se presentan las siguientes listas.

### Valores de clasificación del dataset:

Revista Innovación y Software
Vol. 4, No. 1, Mes Marzo-Agosto, 2023
ISSN: 2708-0935
Pág. 52-62
https://revistas.ulasalle.edu.pe/innosoft

[4, 2, 3, 5, 5, 5, 5, 4, 5, 5, 2, 4, 4, 3, 4, 1, 2, 5, 5, 3, 5, 5, 4, 5, 2, 3, 4, 3, 4, 4, 4, 4, 1, 2, 4, 4, 4, 5, 4, 4, 1, 4, 2, 4, 2, 2, 3, 3, 5, 5, 5, 4, 5, 5, 3, 5, 3, 5, 5, 5, 4, 5, 5, 5, 4, 1, 5, 3, 3, 1, 4, 2, 5, 5, 4, 5, 1, 1, 4, 2, 2, 5, 5, 2, 4, 4, 5, 4, 1, 4, 4, 5]

**Valores obtenidos luego del procesamiento:**

[4, 2, 3, 5, 1, 5, 4, 5, 5, 5, 2, 4, 4, 3, 5, 1, 2, 4, 4, 3, 4, 5, 5, 5, 4, 4, 5, 3, 4, 4, 4, 4, 2, 3, 4, 4, 4, 5, 4, 3, 1, 5, 2, 5, 2, 1, 2, 3, 5, 4, 5, 5, 5, 5, 2, 4, 3, 5, 5, 5, 3, 5, 5, 5, 5, 1, 5, 2, 3, 3, 5, 2, 5, 5, 4, 4, 1, 1, 4, 1, 1, 5, 4, 2, 4, 4, 4, 4, 1, 2, 3, 5]

En el cual podemos observar que los valores ordenados presentan cierta similitud, ya que al momento de poder comparar con la clasificación dada del dataset, resaltamos casos en los que los valores son exactamente iguales.

**2. Luego de filtrar los valores para crear 3 clases se obtiene el siguiente resultado:**

**Valores de clasificación del dataset:**

['C', 'A', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'C', 'B', 'C', 'A', 'A', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'A', 'B', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'C', 'C', 'C', 'C', 'C', 'B', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'B', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'B', 'B', 'A', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'A', 'A', 'C', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'C', 'C', 'C']

**Valores obtenidos luego del procesamiento:**

['C', 'A', 'B', 'C', 'A', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'C', 'B', 'C', 'A', 'A', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'C', 'C', 'C', 'C', 'C', 'B', 'A', 'C', 'A', 'C', 'A', 'A', 'A', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'B', 'C', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'A', 'C', 'A', 'B', 'B', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'A', 'A', 'C', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'A', 'B', 'C']

Como se puede observar, se presentan más datos resaltados por el rango que se dio al generar las clases.

Confusion matrix:

[[17 2 1]
[ 3 7 1]
[ 2 3 56]]

Accuracy value: 0.8695652173913043

La precisión Score de macro es: 0.7738592824799722

Revista Innovación y Software
Vol. 4, No. 1, Mes Marzo-Agosto, 2023
ISSN: 2708-0935
Pág. 52-62
https://revistas.ulasalle.edu.pe/innosoft

El macro average recall es: 0.8014654744162941

Informe de clasificación:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| A | 0.77 | 0.85 | 0.81 | 20 |
| B | 0.58 | 0.64 | 0.61 | 11 |
| C | 0.97 | 0.92 | 0.94 | 61 |
| accuracy | | | 0.87 | 92 |
| macro avg | 0.77 | 0.80 | 0.79 | 92 |
| weighted avg | 0.88 | 0.87 | 0.87 | 92 |

(1)

## Conclusiones

Terminando de realizar los estudios necesarios. llegamos a la conclusión de que una de las mejores librerías para trabajar NLP es BERT, ya que gracias a su sofisticada tecnología nos proporciona herramientas para poder acercarnos de forma muy clara a las expresiones reales de las personas.

Al momento de comparar los diferentes resultados obtenidos, vemos que la diferencia entre uno y otro es muy corta, sabiendo que se utilizó una base de datos medianamente corta, el promedio de valores desiguales que se pudieron presentar en este, eran de un nivel muy bajo. Pudiendo observar los valores obtenidos tanto Accuracy y f1-score podemos reconocer que el algoritmo empleado para procesar los sentimientos en comentarios sobre un hotel, es confiable. ya que presenta una exactitud de 0.87.

## Referencias

[1] S. Vidya (2018). “Cross Domain Sentiment Classification Using Natural Language Processing”. Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Institute of Technology, KrishnaSnkoil, Tamil Nadu, India. Recuperado de: https://www.researchgate.net/profile/Vidya-Soundarapandian/publication/339901166_Cross_Domain_Sentiment_Classification_Using_Natural_Language_Processing/links/5e6b74c2a6fdccf321d98d41/Cross-Domain-Sentiment-Classification-Using-Natural-Language-Processing.pdf

[2] Ishaq, A., Umer, M., Mushtaq, M., Medaglia, C., Siddiqui, H., Mehmood, A. and Choi, G., 2020. Extensive hotel reviews classification using long short term memory. Journal of Ambient Intelligence and Humanized Computing, 12(10), pp.9375-9385.

[3] Ghabayen, A. and Ahmed, B., 2019. Polarity Analysis of Customer Reviews Based on Part-of-Speech Subcategory. Journal of Intelligent Systems, 29(1), pp.1535-1544.

[4] Shin, S., Du, Q. and Xiang, Z., 2018. What’s Vs. How’s in Online Hotel Reviews: Comparing Information Value of Content and Writing Style with Machine Learning. Information and Communication Technologies in Tourism 2019, pp.321-332.

[5]Colab.research.google.com. 2022. Google Colaboratory. [online] Available at: <https://colab.research.google.com/github/bentrevett/pytorch-sentiment-analysis/blob/master/1%20-%20Simple%20Sentiment%20Analysis.ipynb?authuser=1#scrollTo=RMDoLMlxaUql> [Accessed 10 July 2022].

[6] Medium. 2022. Churning the Confusion out of the Confusion Matrix. [online] Available at: <https://blog.clairvoyantsoft.com/churning-the-confusion-out-of-the-confusion-matrix-b74fb806e66> [Accessed 10 July 2022].

[7] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014, doi: 10.1109/MCI.2014.2307227.

[8] SALAMI, Salami. IMPLEMENTING NEURO LINGUISTIC PROGRAMMING (NLP) IN CHANGING STUDENTS' BEHAVIOR: RESEARCH DONE AT ISLAMIC UNIVERSITIES IN ACEH. Jurnal Ilmiah Peuradeun, [S.l.], v. 3, n. 2, p. 235-256, may 2015. ISSN 2443-2067. Available at: <http://www.journal.scadindependent.org/index.php/jipeuradeun/article/view/65>. Date accessed: 10 july 2022.

[9] D. Yu, K. Yao and Y. Zhang, "The Computational Network Toolkit [Best of the Web]," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 123-126, Nov. 2015, doi: 10.1109/MSP.2015.2462371.

[10] E. H. Houssein, R. E. Mohamed and A. A. Ali, "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," in IEEE Access, vol. 9, pp. 140628-140653, 2021, doi: 10.1109/ACCESS.2021.3119621.

[11] A. Elnagar, S. M. Yagi, A. B. Nassif, I. Shahin and S. A. Salloum, "Systematic Literature Review of Dialectal Arabic: Identification and Detection," in IEEE Access, vol. 9, pp. 31010-31042, 2021, doi: 10.1109/ACCESS.2021.3059504.

[12] D. Mahendran, C. Luo and B. T. Mcinnes, "Review: Privacy-Preservation in the Context of Natural Language Processing," in IEEE Access, vol. 9, pp. 147600-147612, 2021, doi: 10.1109/ACCESS.2021.3124163.

[13] X. Feng and Y. Zeng, "Neural Collaborative Embedding From Reviews for Recommendation," in IEEE Access, vol. 7, pp. 103263-103274, 2019, doi: 10.1109/ACCESS.2019.2931357.

[14] X. Feng and Y. Zeng, "Multi-Level Fine-Grained Interactions for Collaborative Filtering," in IEEE Access, vol. 7, pp. 143169-143184, 2019, doi: 10.1109/ACCESS.2019.2941236.

[15] S. Salloum, T. Gaber, S. Vadera and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," in IEEE Access, vol. 10, pp. 65703-65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

**Roles de Autoría**

**Walther Mauricio Medina Pauca:** Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Camila Huamani Tito:** Investigación, Metodología, Redacción - borrador original.