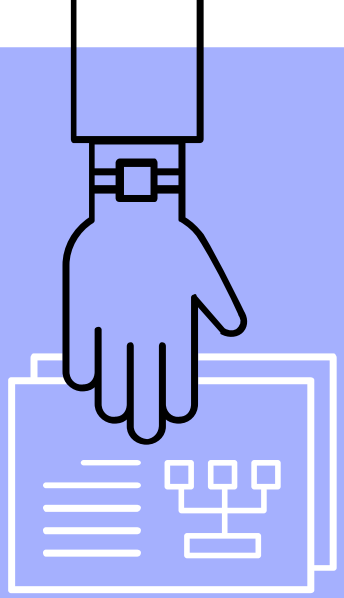
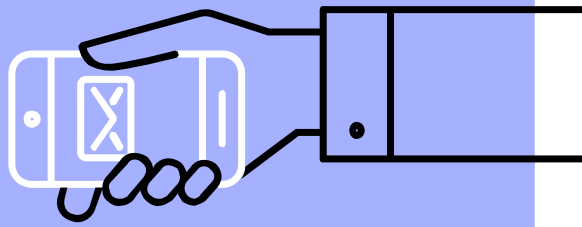
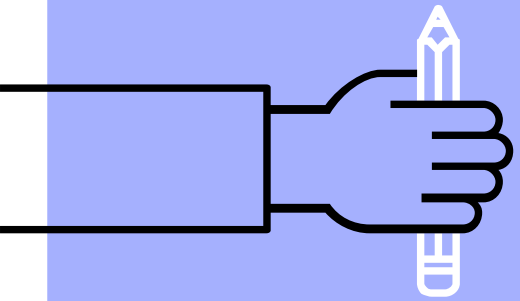
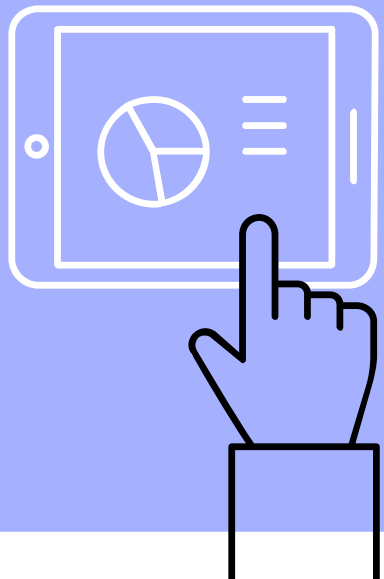




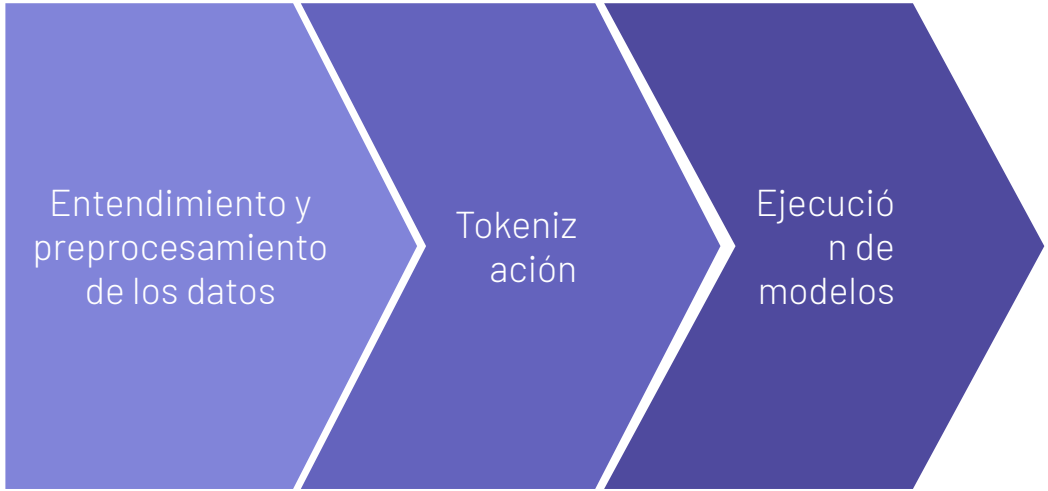
Proyecto 1

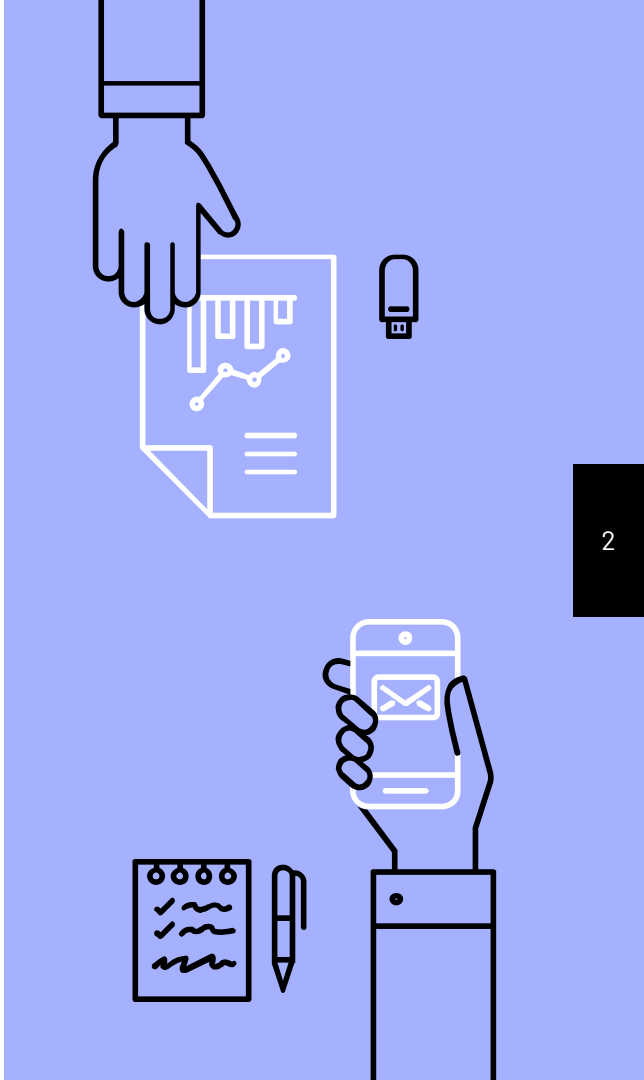
Kevin Babativa
Nicolás Angarita
Nicolás Alvarado



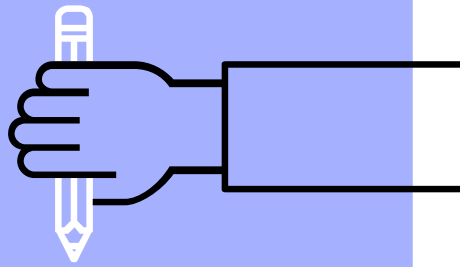


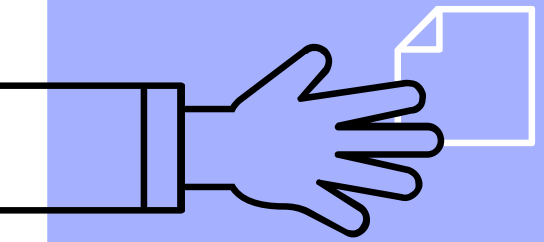
Metodología





Entendimiento y preprocesamiento





Entendimiento

Se tienen 314562 datos en el archivo.

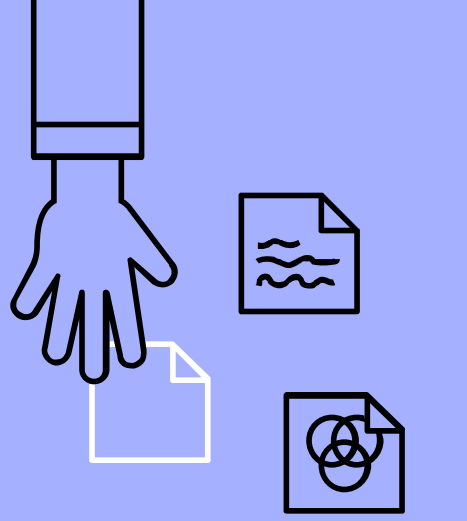
Cada fila consiste en un texto y una calificación de 1 a 5 .

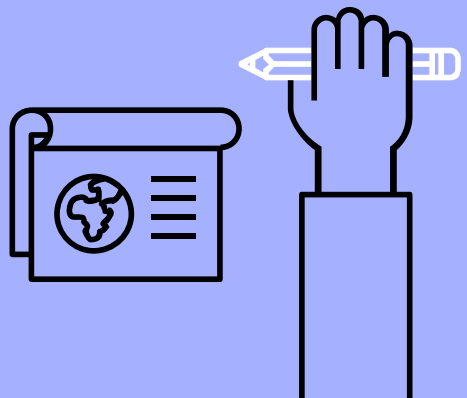
```
[231] data.shape
```

```
(314562, 2)
```

```
[232] data.sample(5)
```

	text	stars
125981	Worst Chinese I have ever eaten. Do yourself a favor and run away!!! Chicken was tough...dry, and flavorless.	1
170239	Good variety of salads, smoothies, and sandwiches. Not the super-upscale, organic, kale & quinoa type of choices; however, chopped fruits and salad ingredients looked fresh and they had many options, including soymilk, whey protein, yogurt, orange sherbet (but not the wildberry listed), & peanut butter. When I asked if I could substitute honey (which was listed as an ingredient in my smoothie), I was told the smoothie would not have had honey, anyway, so I couldn't substitute anything for it. Smoothie was good, nice texture, not too sweet.\n\nPlace is clean, with tables to sit. The older couple that was working when I went were efficient. Refrigerator for bottled drinks. Takes visa/MC/Disc/cash (no sign for AmEx). Nice to see a place like this in town!\n\n(Just an "update" to respond to the owner, I originally found this place on Yelp. There were no reviews, but I was in the city, looking for a smoothie, and intrigued enough to take the risk. Glad I did :)	4
249296	Went here on a Saturday afternoon with my wife. The parking lot across the street was full so we parked several blocks away at a municipal lot which was fine. The place is a monster inside! We got inside and stood in line to ask for a table. We were told it would be a 35 minute wait. At the 35 minute mark we made our way up to the young ladies at the desk and were told another 10 minutes. It was amazing to see alll the people coming in and going out but we saw three parties of 4 that came in after us that were seated before us. I suspect even more groups of four were seated ahead of us because once we were seated (at a 4 top) we could see that they have very few 2 tops and they wanted to fill the tables. I get it. It's a profit thing. However, first come first serve is the proper way or tell people you don't seat 2 tops until you feel like it.They lost a star on that move.\n\nMy wife and I tried the 1905 salad-delicious! Definitely recommend. The bread was warm and fresh. We also ordered house sangria. It wasn't bad at all. I got the eye of round with chorizo and gravy. Very tender and flavorful. My wife got the sampler..an empanada (good) and pork (also good and tender). \n\nIt can be difficult to maintain standards and feed as many people as Columbia does but they do a nice job. Would we return...maybe. The food is better than it should be in a venue like this but it's a zoo with all of the waiters running around, the dropped glasses and plates (3 times during our meal) and the ambient noise just makes for eating to live versus having a nice meal in a pleasant atmosphere. I would tell people that if you go to Ybor City and don't have a clue where you want to eat to go to the Columbia. It seems like they are consistent.	3
104394	Combine the best situational training with top notch instruction and you have firearms training that should be required for ccw permits. MiScenarios provides everything you need to prepare yourself for real world situations, not static paper targets. This is not an arcade. This is a no-nonsense educational environment that you will leave better informed and better prepared.\n\nWhile on our way to Virginia City to enjoy Street Vibration festivities, I booked some time at the facility (you can book online on the website, http://www.miscenarios.com/), My brother-in-law is a firearms aficionado, to say the least! he way blown away at the real life situations and is looking forward to returning. Needless to say, he left grinning ear-to-ear!!\n\nA shout out to John, the best instructor any of us have been around; patient, no ego, just common sense guidance. Thank you!	5
19080	We came here while on vacation in SB. It's on the state streets and not too far away from the pier.\n\nI was super skeptical. \n\nI usually scope a place out before trying it out but we came in here blind. What a pleasant surprise! \n\nIt was really nice outside so we sat outside. We had a total of three people helping us out but one main server.\n\nTheir service was so good! \n\nGranted the place wasn't too busy I was simply really impressed. It was like the perfect service. \n\nFirst timers so we got good recommendations from the server. I hate places that the servers don't seen truthful with suggestions.\n\nAnyways the food was good. I got clam chowder and the killer shrimp Mac and cheese.\n\nThese were both SUPER rich. For the large amount of shrimp in the Mac and cheese I thought the price was right. It was a really good dish. There was spiced and creaminess for sure. \n\nThe clam chowder was very good! It was rich and creamy and clammy hah. It had corn which was unique for me but perfect little sweetness for the rich chowder. \n\nAll of my friends loved their killer shrimp.\nIt's shrimps in their house sauce that you can have in or out of shell. It is served with pasta, bread, or rice.\n\nThey were very generous with the bread that was served with the shrimp.\n\nThey all agreed that hot n juicy has a better sauce though. \n\nOverall we just had a really nice lunch. The food was great, the service team worked together really well, we got checked on the right amount, and the price was worth the entire package. \n\nI honestly can't complain.	4

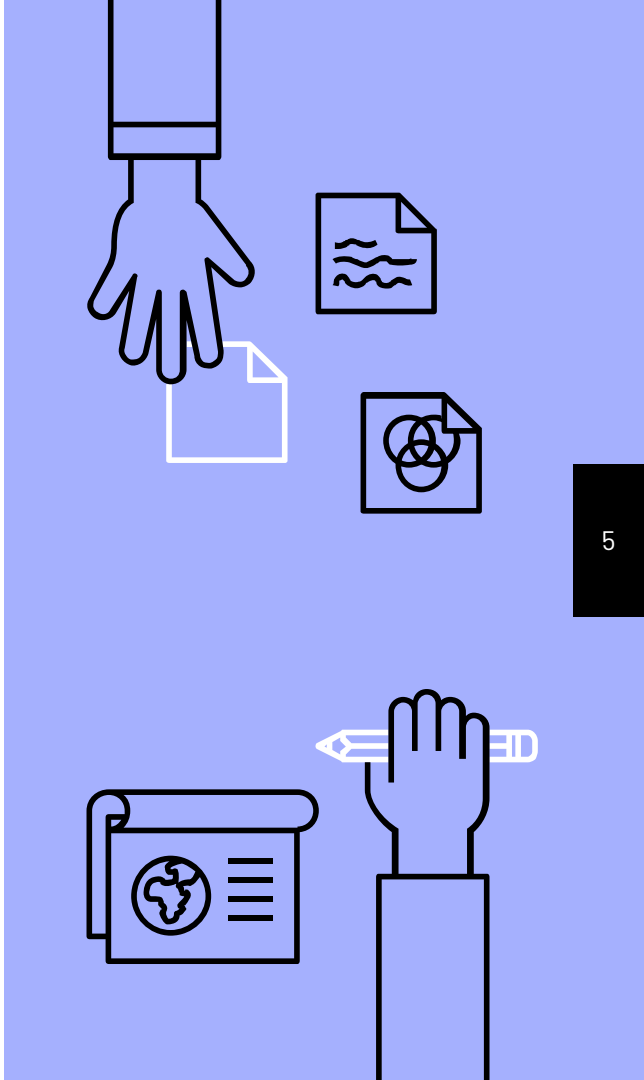




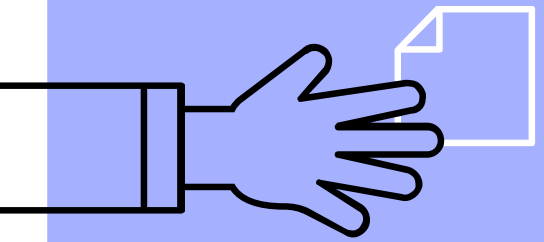
Preprocesamiento

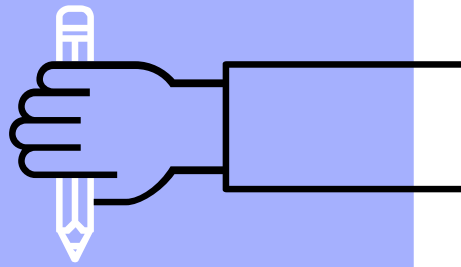
Se divide el dataset en entrenamiento y testeo, se toma el 80% del dataset (251649) como **entrenamiento**.

```
[234] X_train, X_test, y_train, y_test = train_test_split(data['text'], data['stars'], test_size = 0.2, stratify = data['stars'], random_state = 1)
```



Tokenización

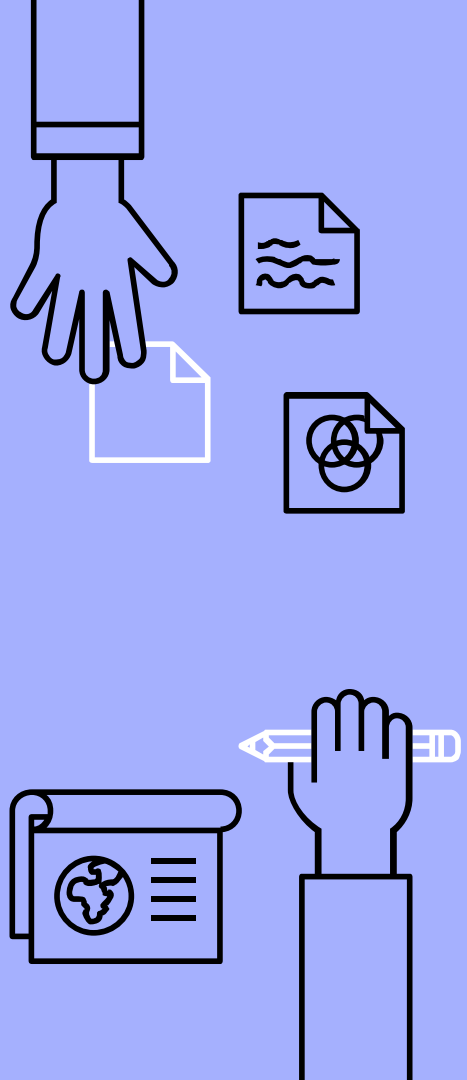




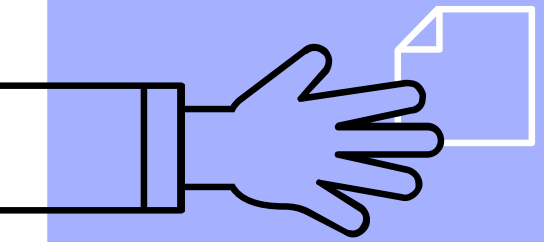
Tokenización

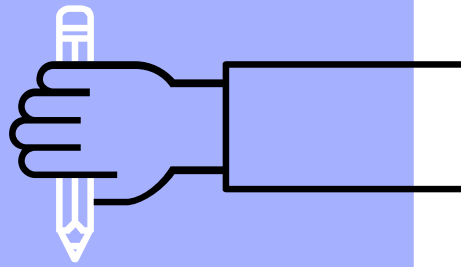
La tarea de tokenización se apoya en la librería NLKT para diversas funciones. De igual manera, se eliminan caracteres especiales y se hace un proceso de *lemmatizing* y *stemming* con el fin de solo conservar la base de las palabras. Todo lo anterior partiendo únicamente del idioma inglés.

```
lemmatizer = WordNetLemmatizer()
ps = PorterStemmer()
def tokenizer(text):
    text = text.replace("\n", " ")
    tokens = word_tokenize(text)
    tokens = [t for t in tokens if t.isalpha()]
    tokens = [lemmatizer.lemmatize(t) for t in tokens]
    tokens = [ps.stem(t) for t in tokens]
    tokens = [t for t in tokens if len(t) > 2]
    return tokens
```



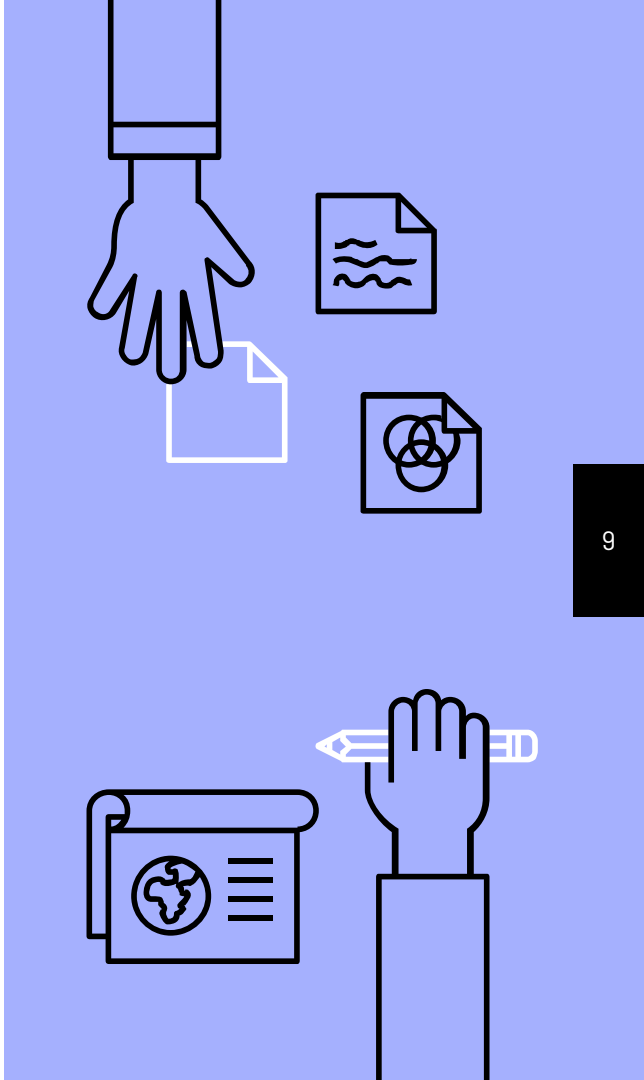
Modelos





Generalidades

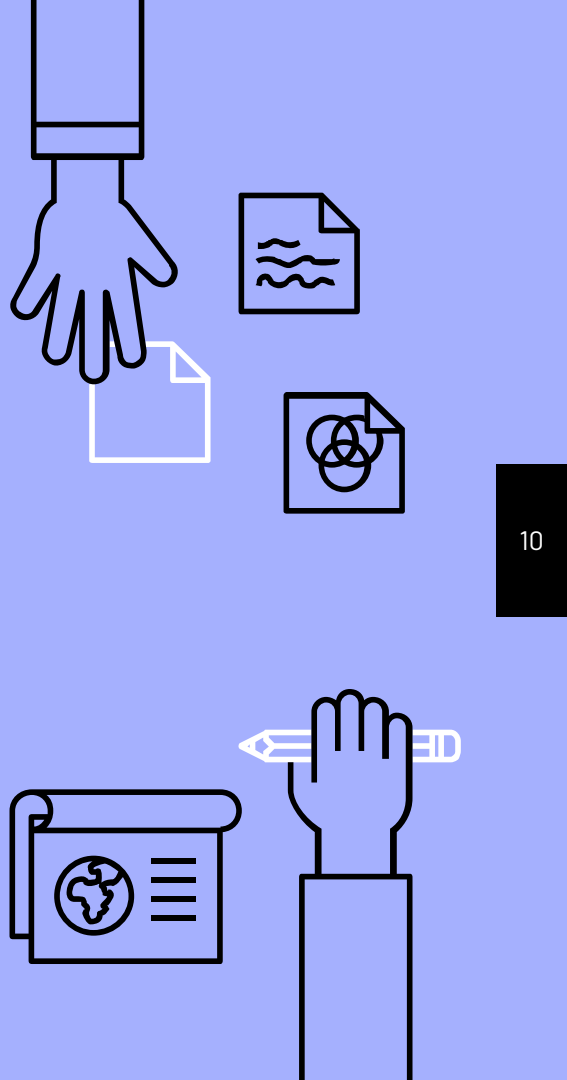
Para cada modelo se crearon dos Pipelines, uno cuya primera tarea era realizar la función CountVectorizer, que nos ayuda a vectorizar los tokens calculados. Como segunda tarea el pipeline realiza el modelo escogido. Y el segundo Pipeline hace lo mismo pero con TfidfVectorizer.



Naive Bayes

Fue escogido por su rapidez dentro de los algoritmos que necesitan de transformación. La transformación se hizo con la estrategia OneVsRest y después se aplicó el modelo de Bayes.

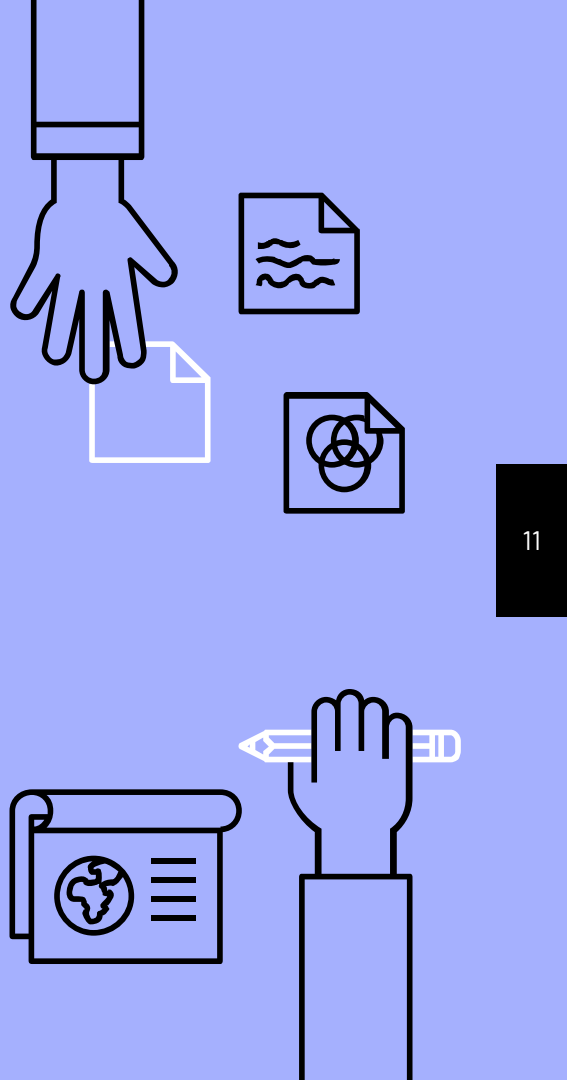
```
pipelineB = Pipeline([
                ('bow', CountVectorizer(tokenizer = tokenizer, stop_words = stop_words, lowercase = True)),
                ('clf', OneVsRestClassifier(MultinomialNB())),
            ])
pipelineB.fit(X_train, y_train)
predictionsBayes = pipelineB.predict(X_test)
```



Decision Tree

Fue escogido por su simplicidad y los insights que fácilmente se pueden determinar a partir de los resultados.

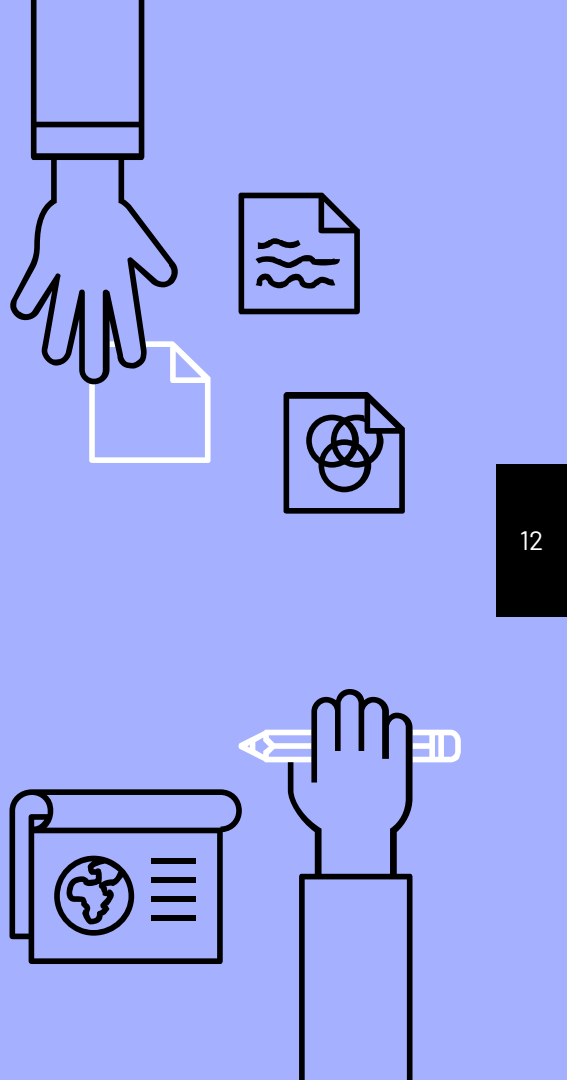
```
pipelineDT = Pipeline([
                ('bow', CountVectorizer(tokenizer = tokenizer, stop_words = stop_words, lowercase = True)),
                ('clf', DecisionTreeClassifier()),
            ])
pipelineDT.fit(X_train, y_train)
```



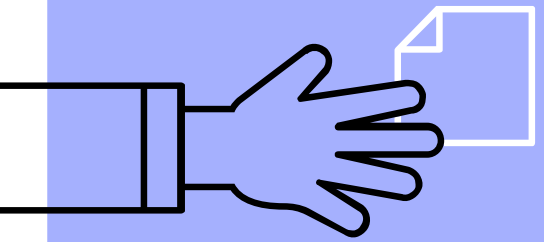
Random Forest

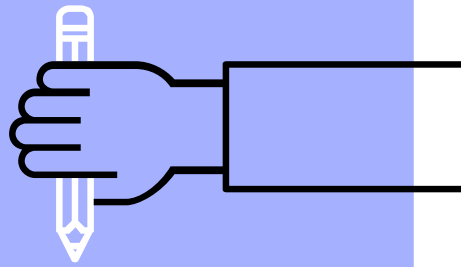
Fue escogido por ser la sucesión de Decision Tree, ambos algoritmos en conjunto aportan grandes insights para el proyecto.

```
pipelineRF = Pipeline([
                ('bow', CountVectorizer(tokenizer = tokenizer, stop_words = stop_words, lowercase = True)),
                ('clf', RandomForestClassifier(random_state = 2)),
            ])
pipelineRF.fit(X_train, y_train)
```



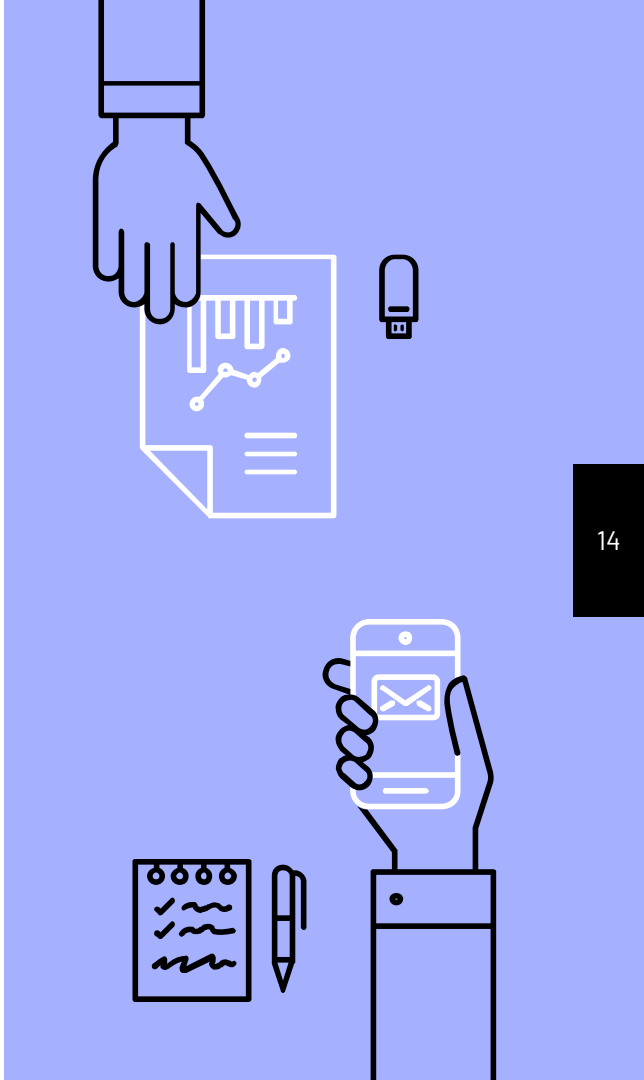
Resultados





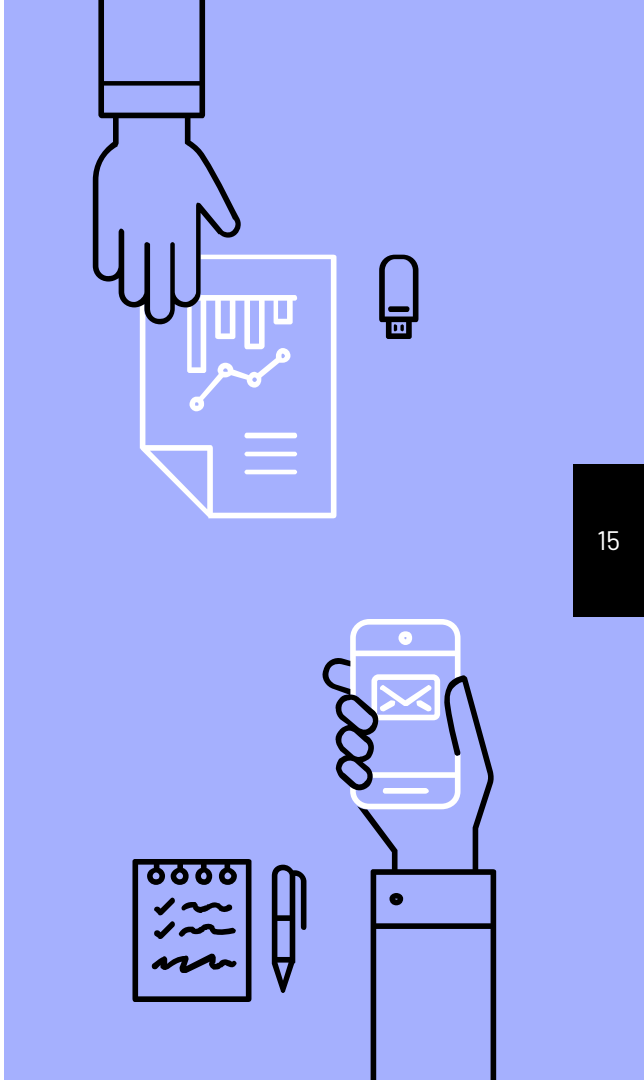
Comparación (CountVectorizer)

	Naive Bayes	Decision Tree	Random Forest
Micro F1 Score	0.63	0.52	0.60
Macro F1 Score	0.50	0.40	0.36
Hamming Loss	0.37	0.48	0.40
F1 Score weighted	0.62	0.51	0.51



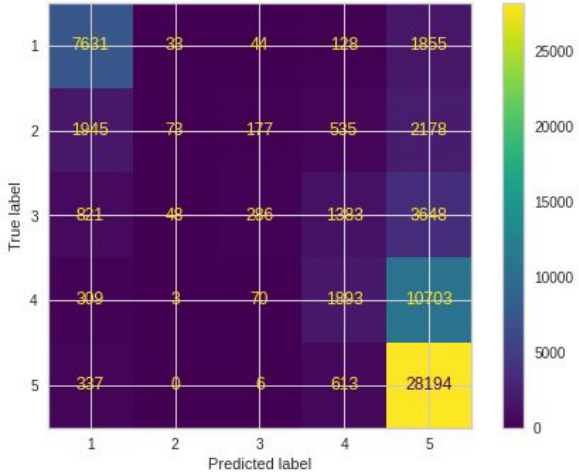
Comparación (TfidfVectorizer)

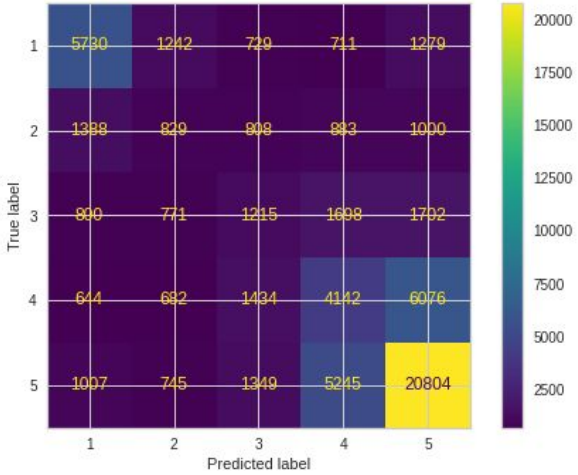
	Naive Bayes	Decision Tree	Random Forest
Micro F1 Score	0.58	0.51	0.60
Macro F1 Score	0.31	0.39	0.35
Hamming Loss	0.42	0.49	0.40
F1 Score weighted	0.47	0.51	0.51



Resultados

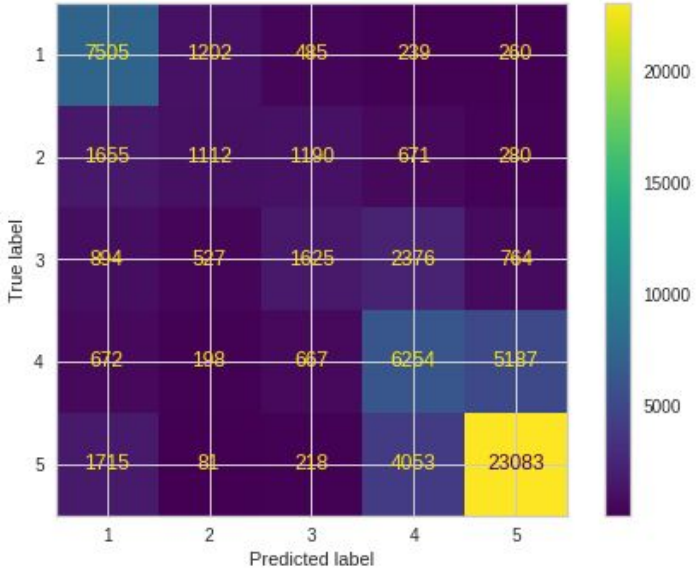
RandomForest y DecisionTree





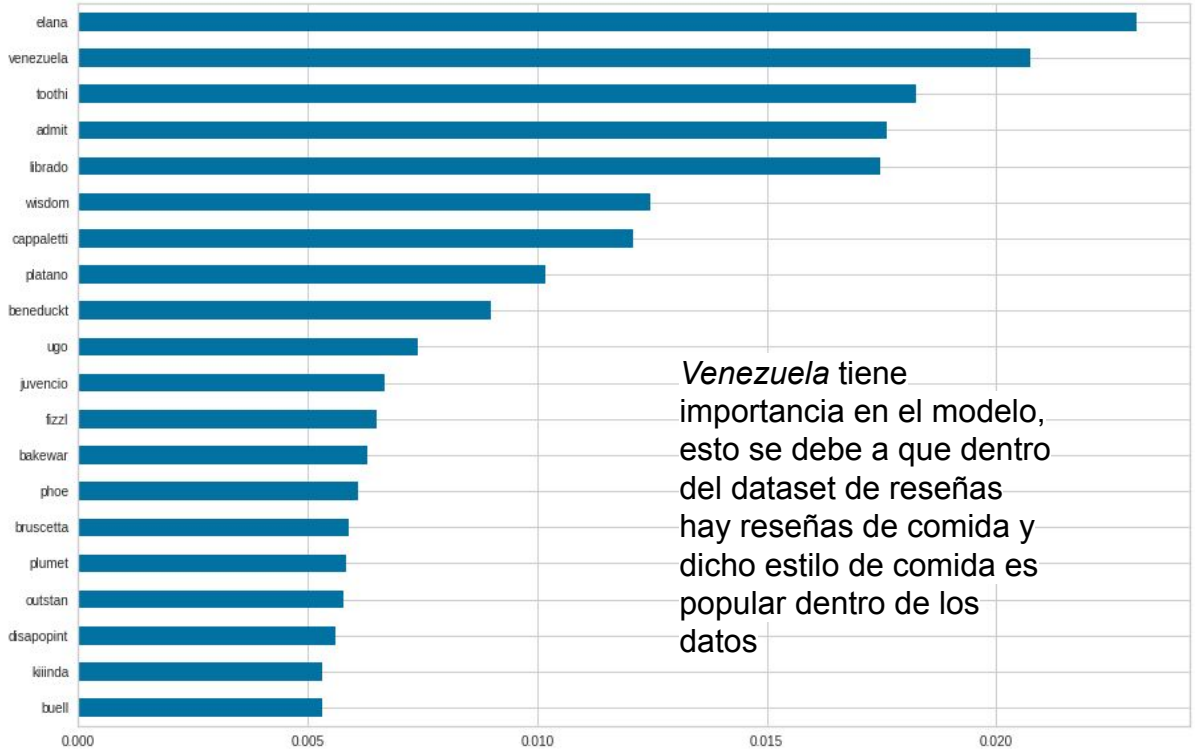
Resultados

Naive Bayes



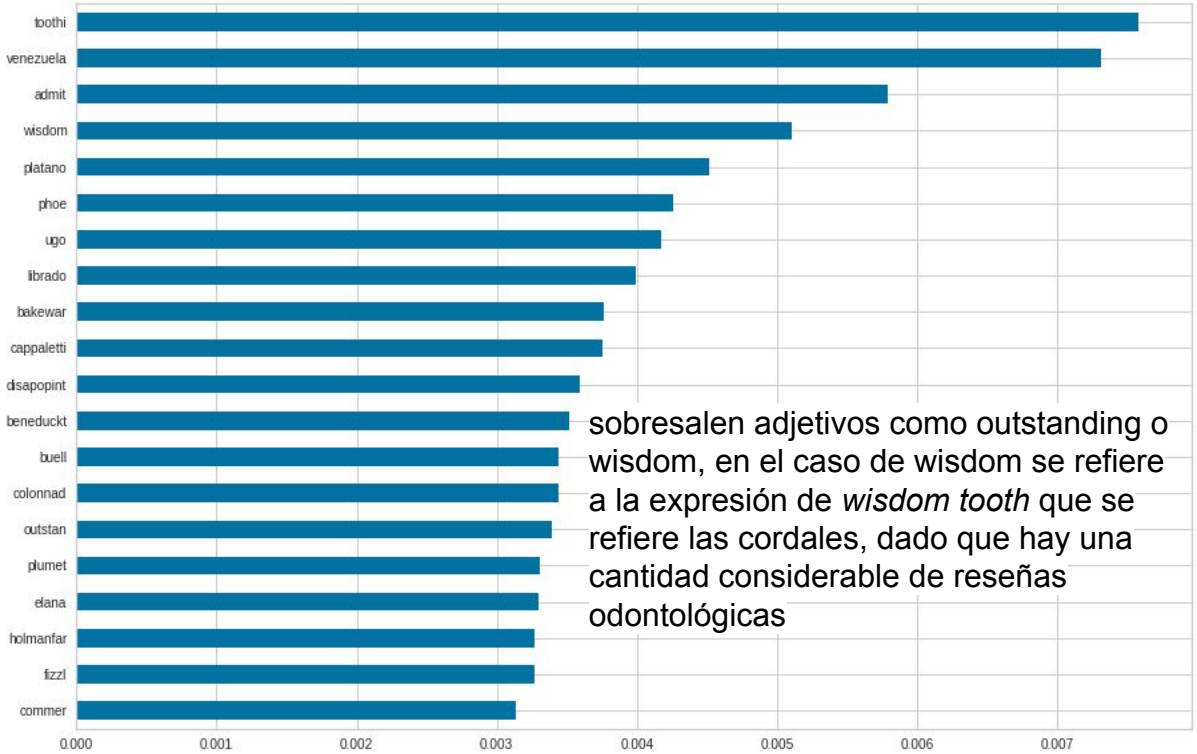
No es lo ideal pero se puede usar para clasificar comentarios entre buenos y malos.

Graficas



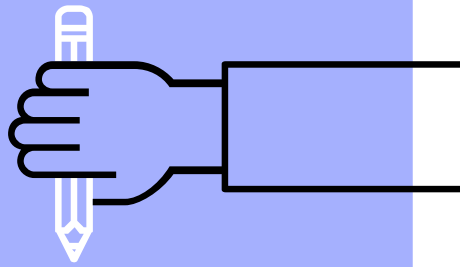
Venezuela tiene importancia en el modelo, esto se debe a que dentro del dataset de reseñas hay reseñas de comida y dicho estilo de comida es popular dentro de los datos

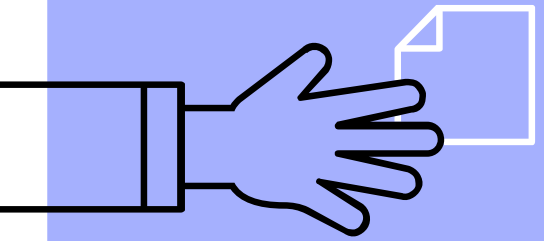
Graficas



sobresalen adjetivos como outstanding o wisdom, en el caso de wisdom se refiere a la expresión de *wisdom tooth* que se refiere las cordales, dado que hay una cantidad considerable de reseñas odontológicas

Conclusiones





Recomendaciones

El algoritmo de Naive Bayes logró la meta propuesta, pero por muy poco por lo que se podría intentar mejorar haciendo cambios en el proceso de entrenamiento. Para los otros dos algoritmos, no están lo suficientemente cerca de la meta por lo que no deben ser tomados en cuenta, más allá de como una fuente de información pero no de toma de decisión.
Se le sugiere a cualquier empresa que quiera usar este modelo que lo use de esa forma, para definir comentarios buenos y malos teniendo en cuenta los comentarios que tienen 1,4 y 5 estrellas. Estos resultados se dieron por la cantidad desproporcionada de comentarios con 5 y 1 estrellas.