Text Analytics Project

PROGETTO REALIZZATO DA **GIUSEPPE MUSCHETTA**

CON LA SUPERVISIONE DELLA PROF.SSA **LAURA POLLACCI**

Sentiment Analysis of Amazon Product Reviews

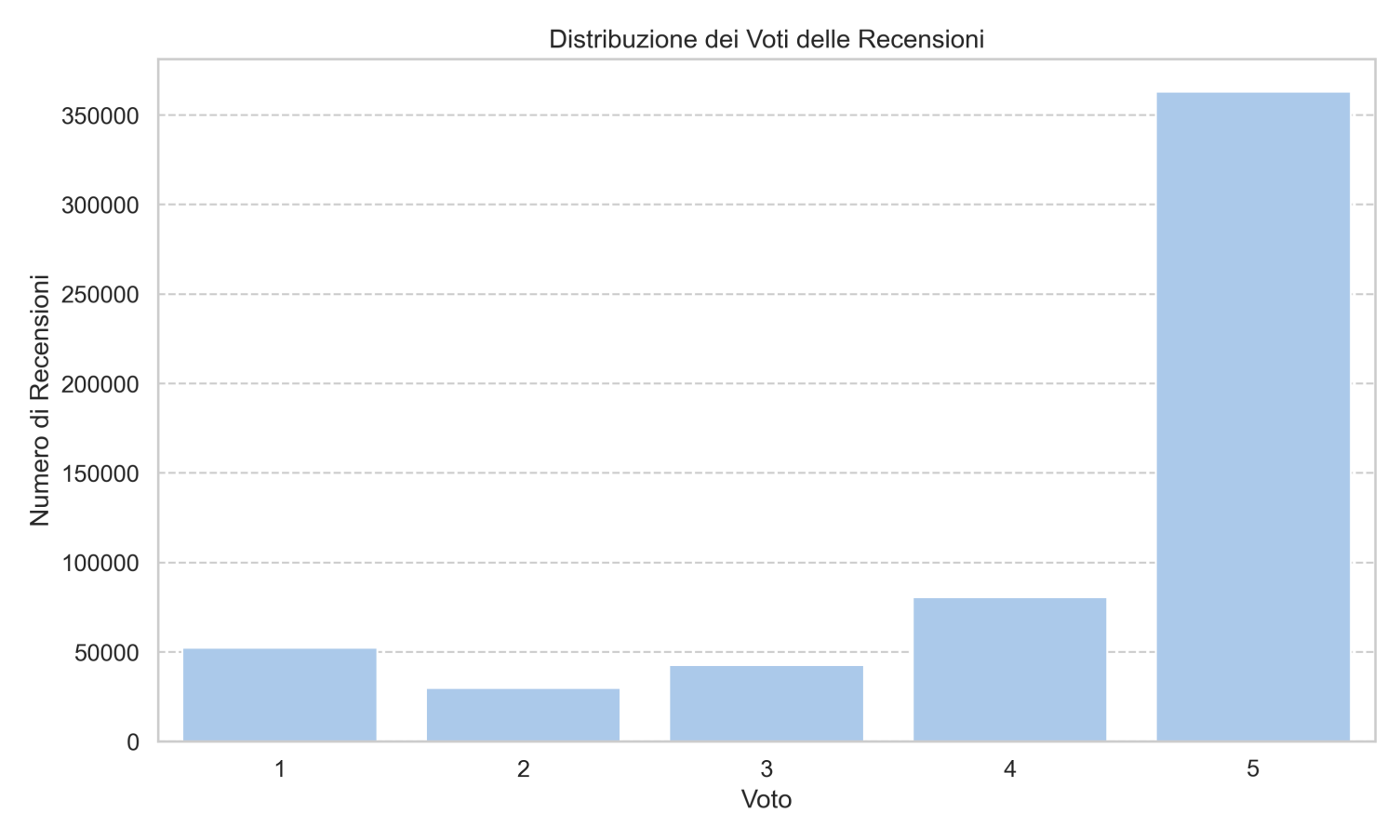
The dataset can be found here:

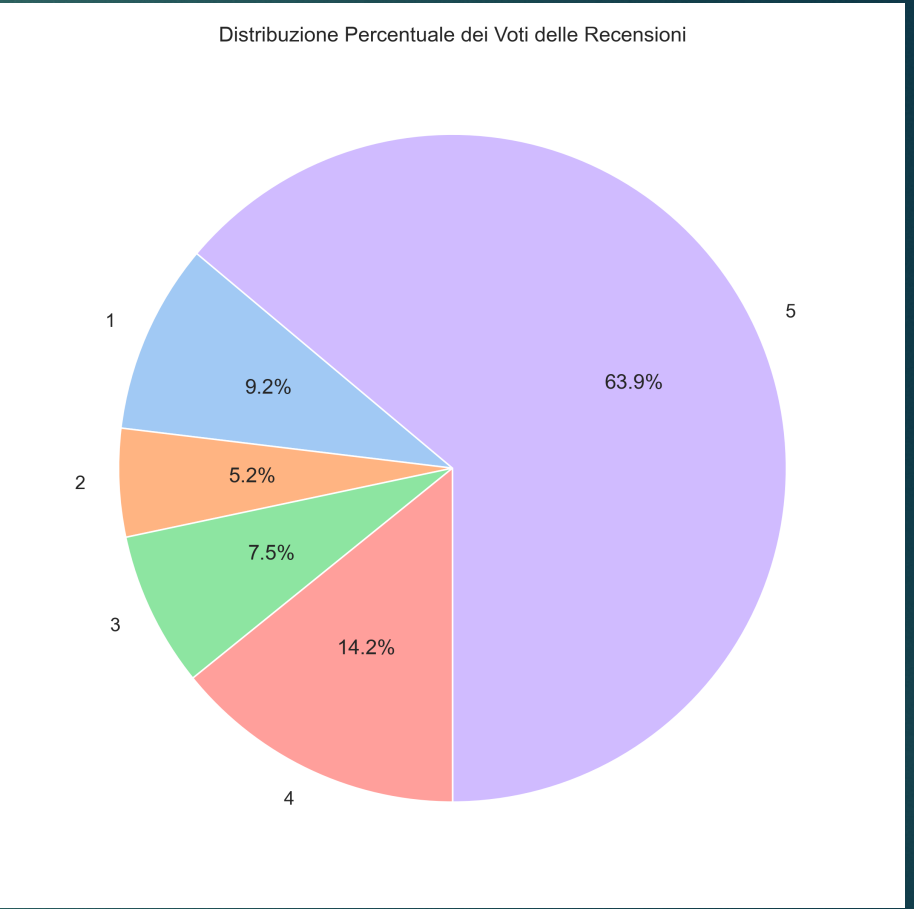
Contains the following attributes:

Total Records: 568454
Total Columns: 10
Domain Name: amazon.com
File Extension: CSV

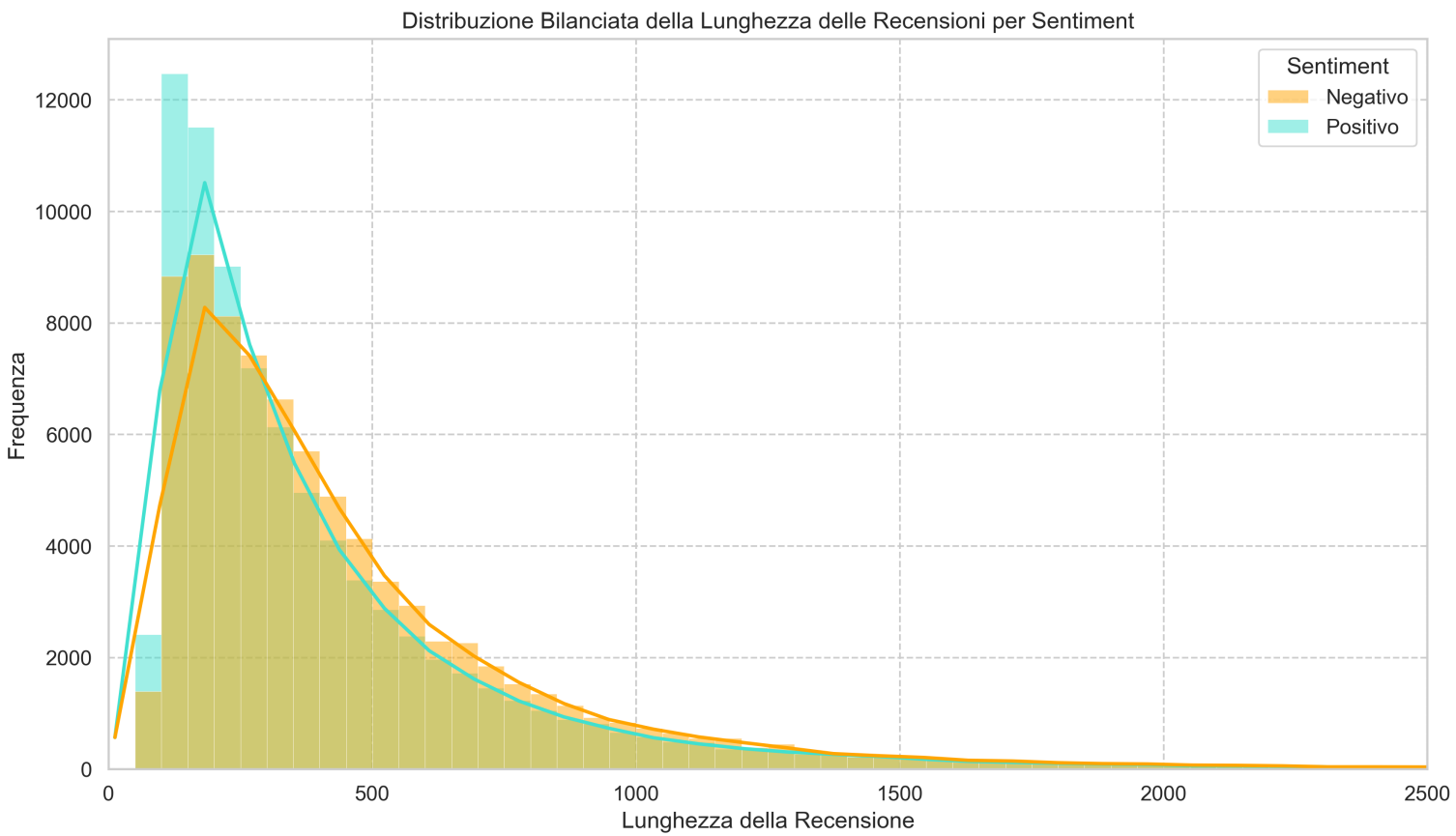
Available Fields: Id, ProductId, UserId, ProfileName, HelpfulnessNumerator, HelpfulnessDenomenator, Score, Time, Summary, Text

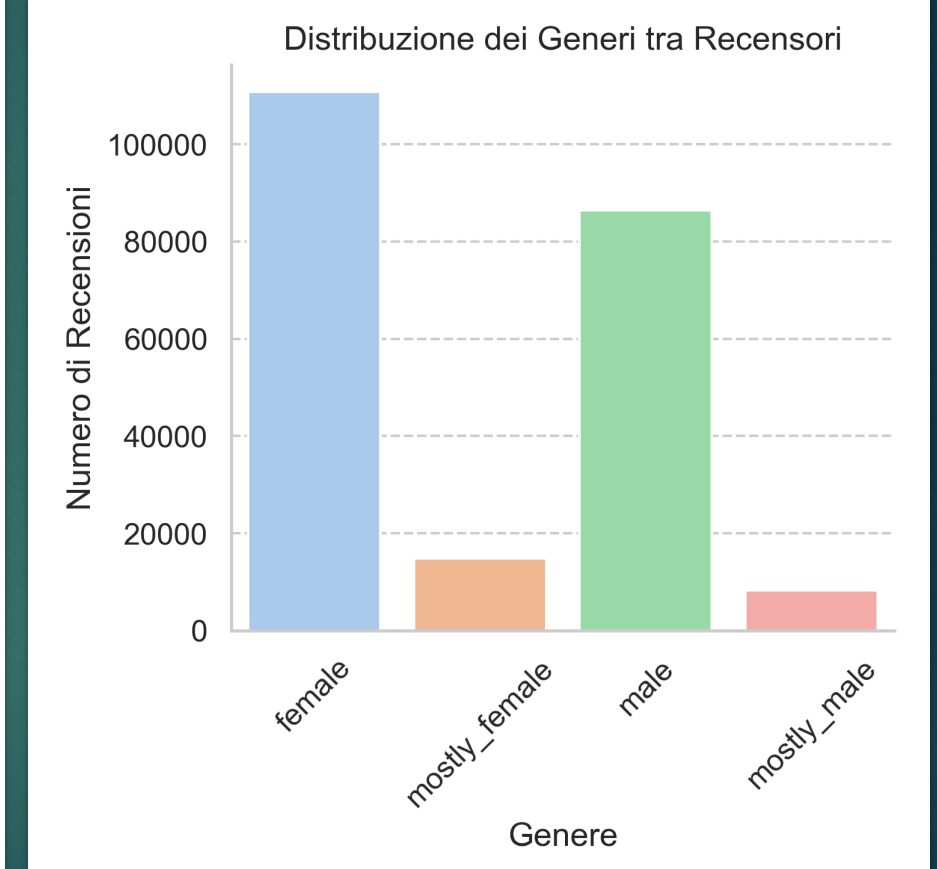
Data **Understanding** and **Exploring**



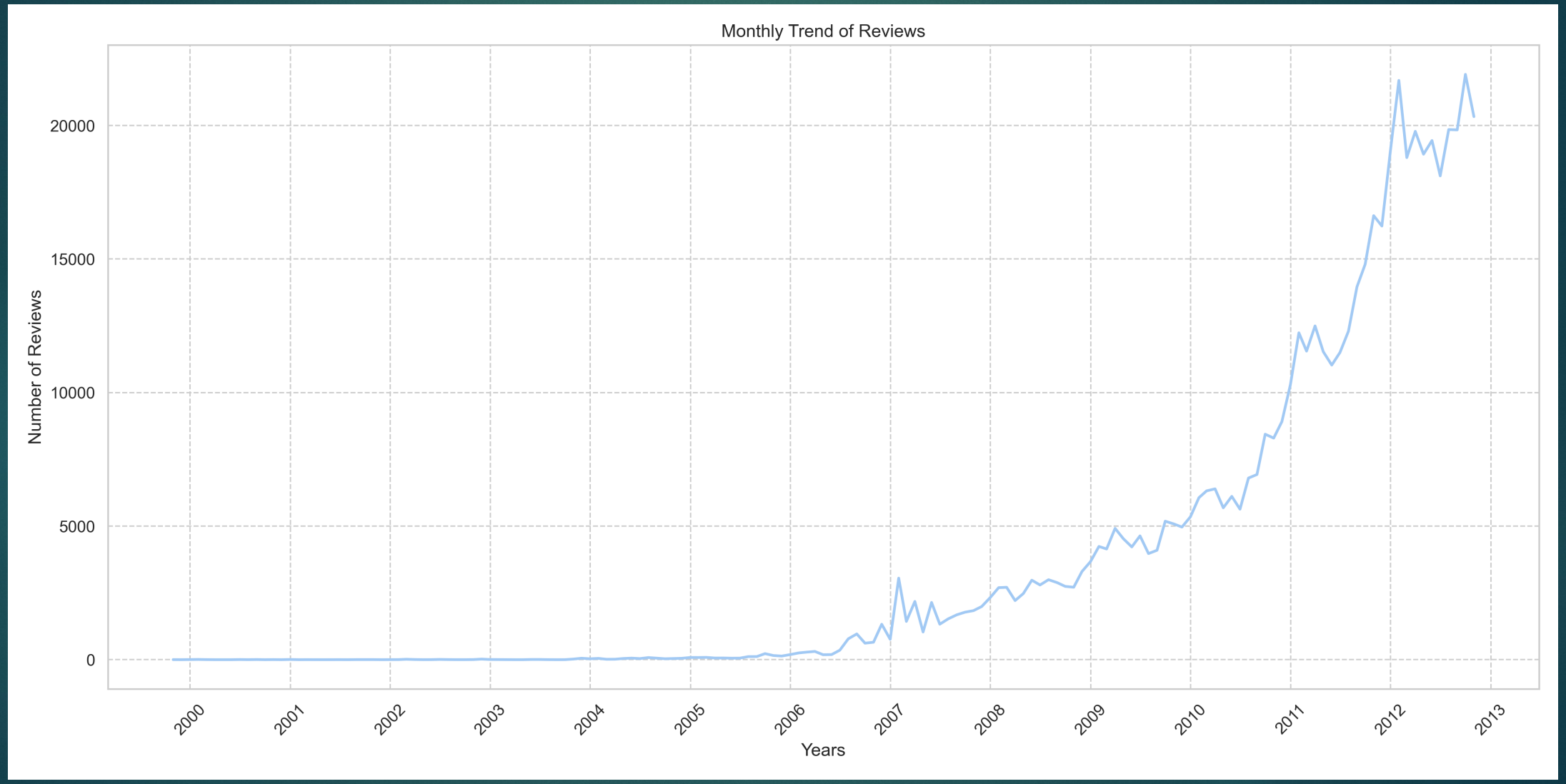


Data **Understanding** and **Exploring**

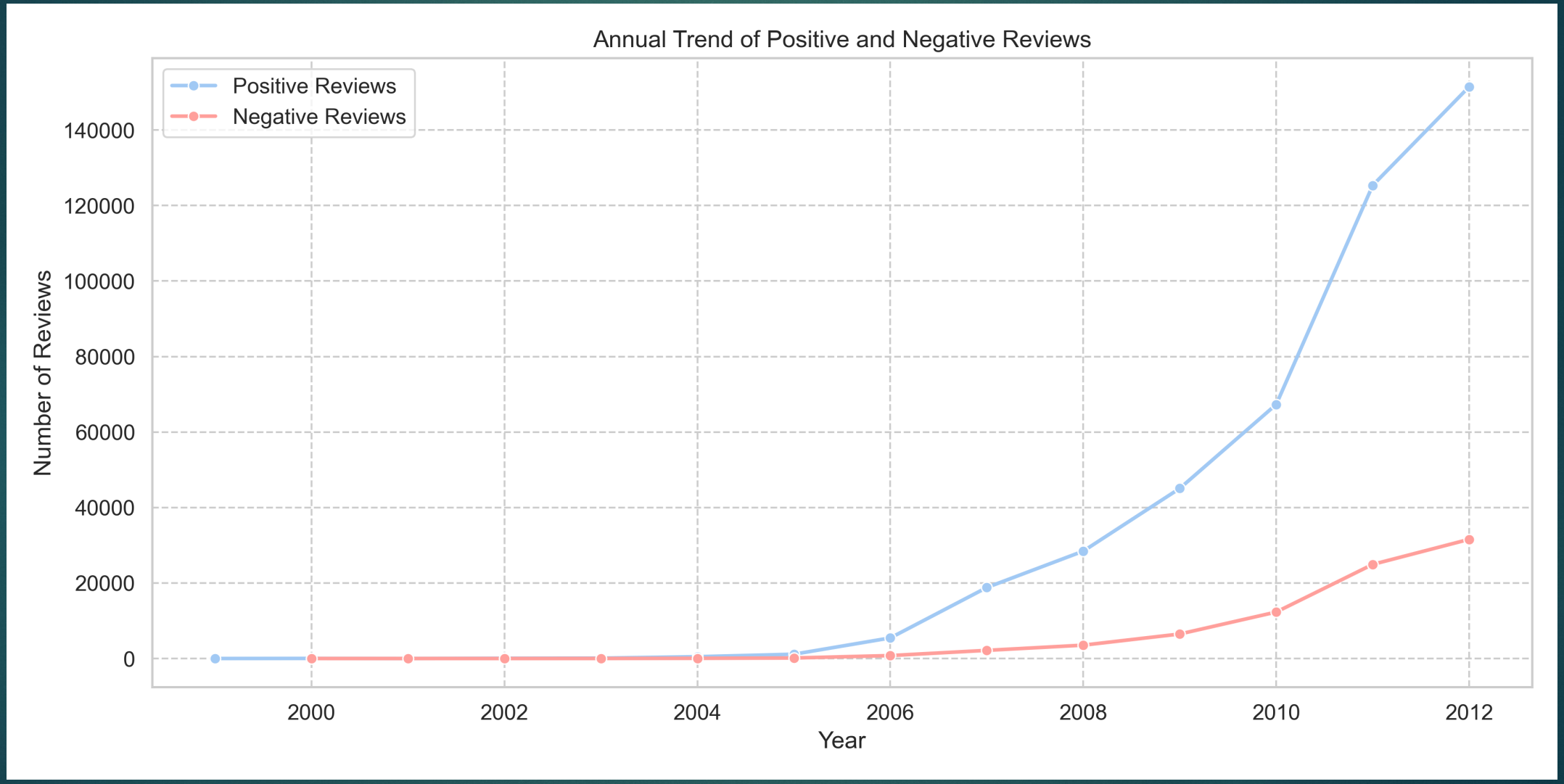




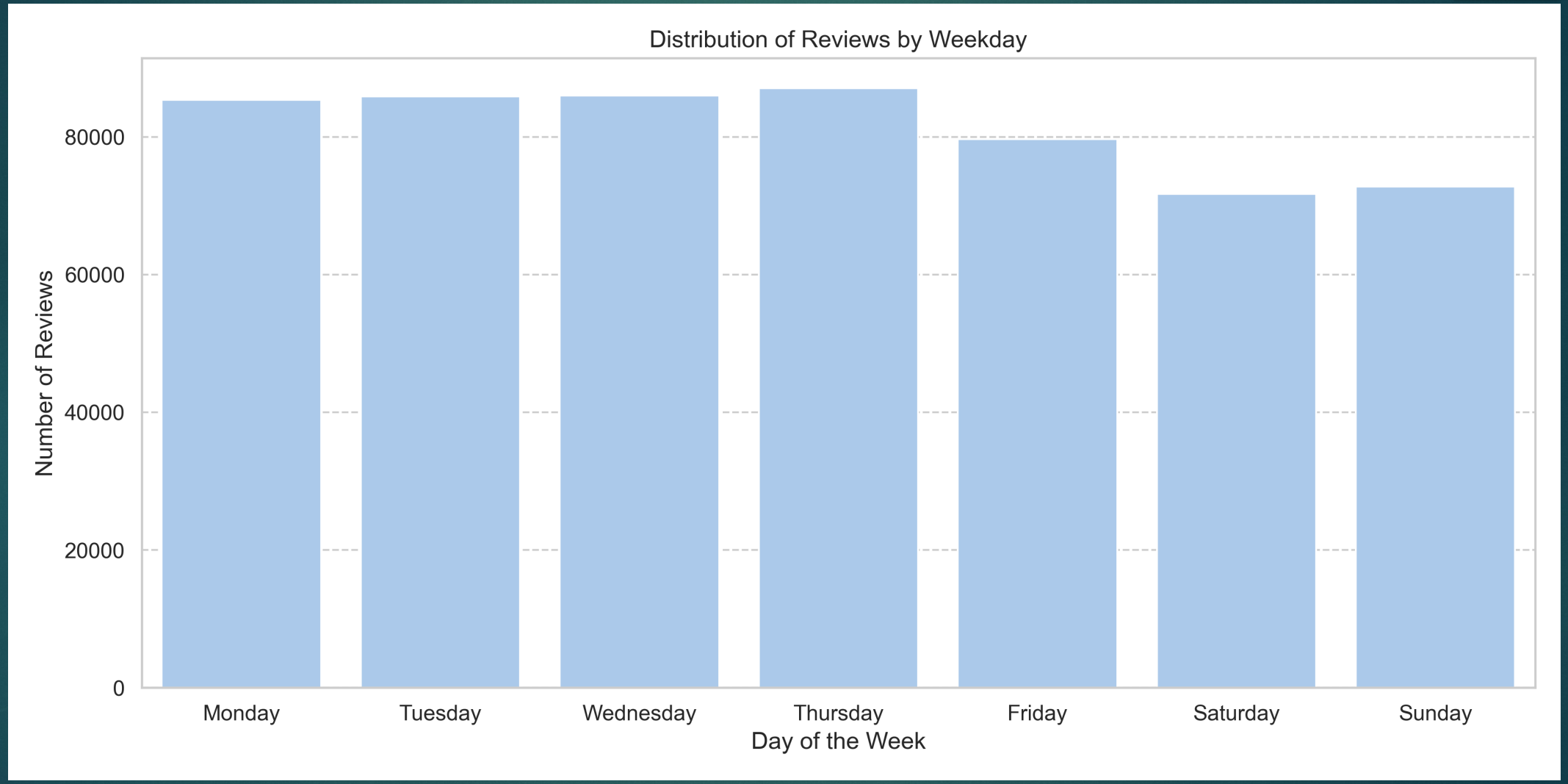
Data **Understanding** and **Exploring**



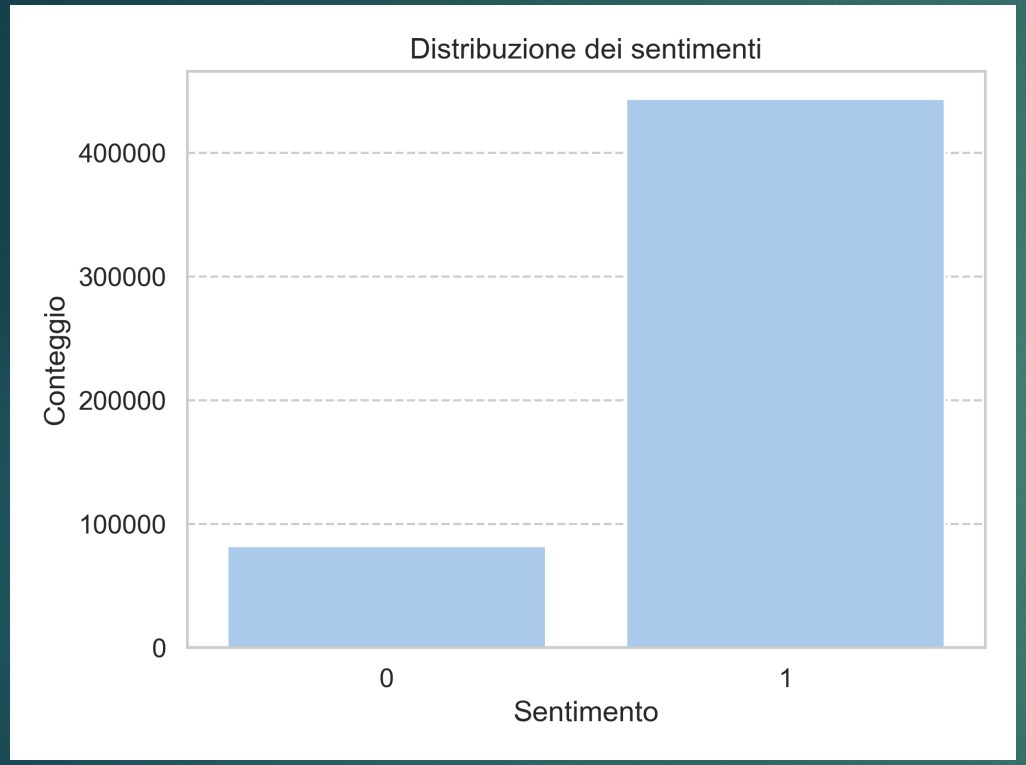
Data **Understanding** and **Exploring**

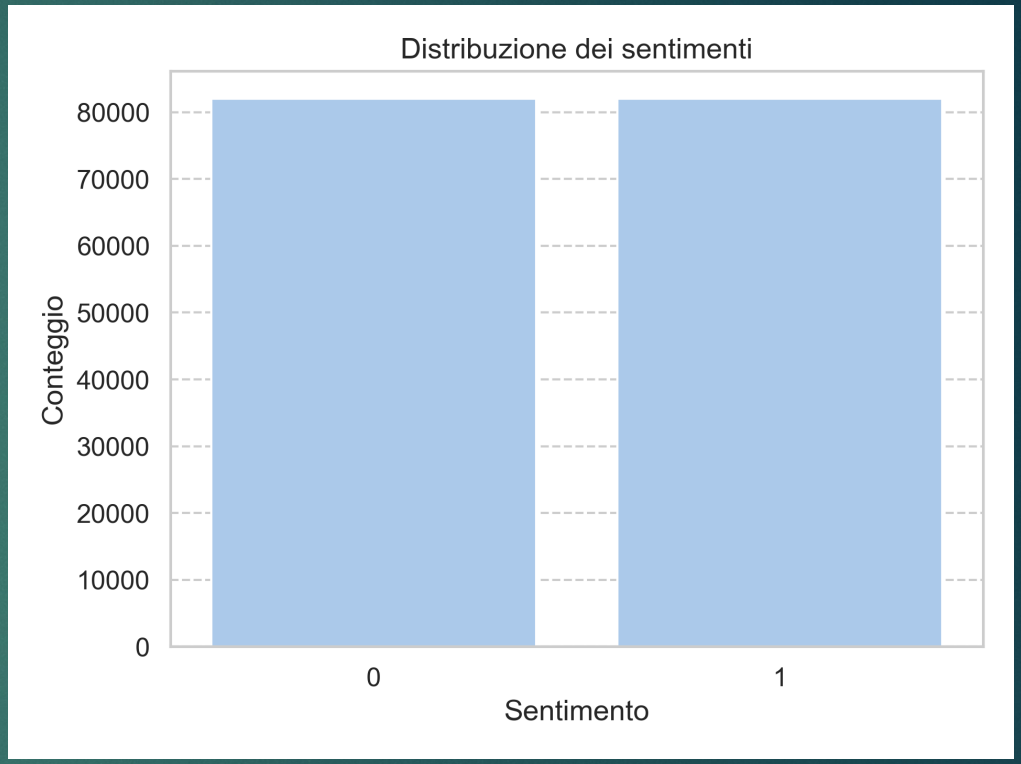


Data **Understanding** and **Exploring**



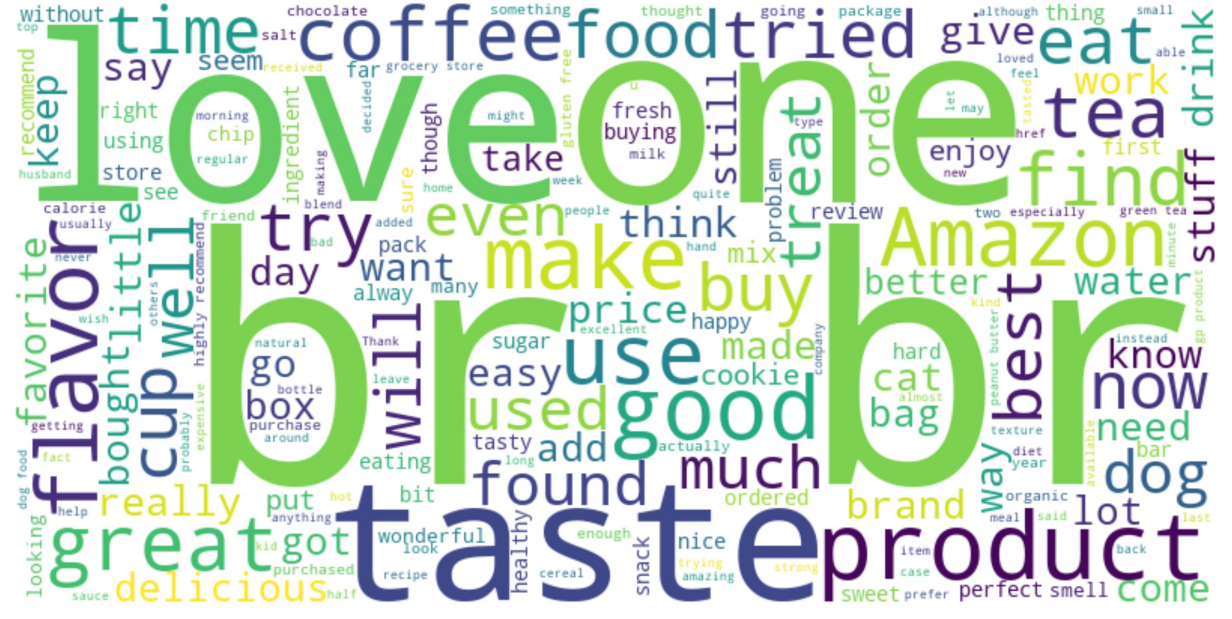
Data **Preparation** and **Cleaning**



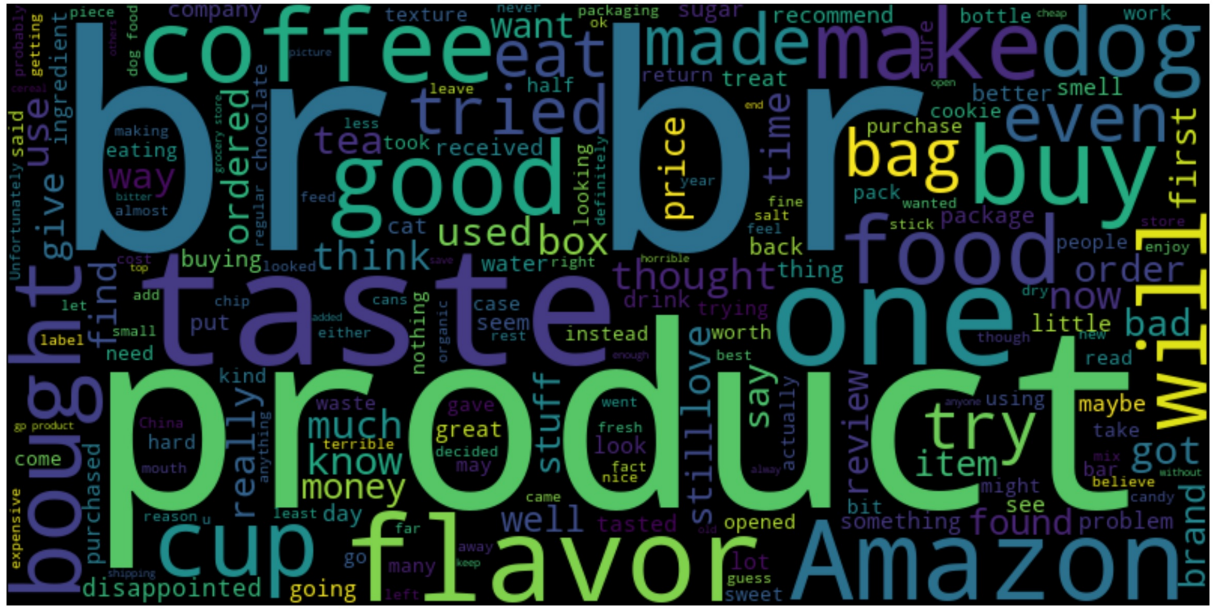


Wordclouds Raw Text

Parole più comuni in recensioni positive



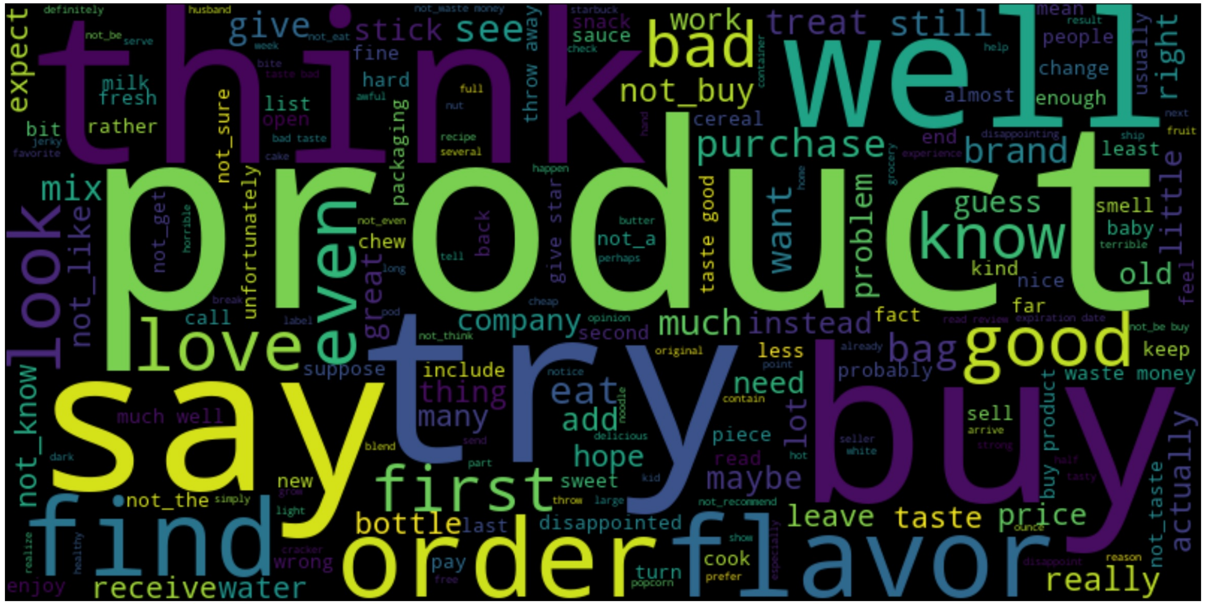
Parole più comuni in recensioni negative



Wordclouds cleaned Text

Parole più comuni in recensioni positive

Parole più comuni in recensioni negative



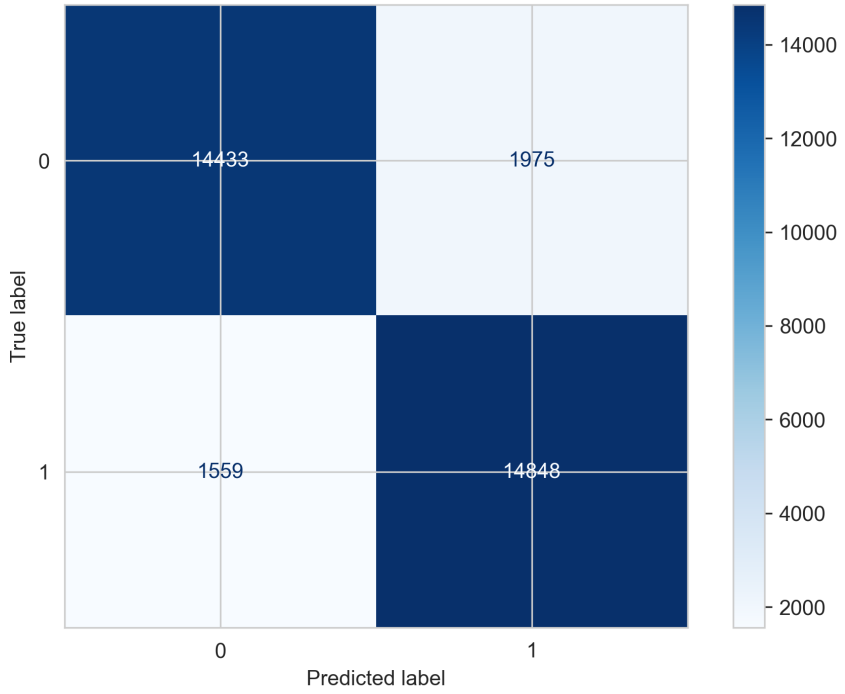
good: 39260 taste: 30479 love: 29221 great: 26968 flavor: 26778
product: 22169 get: 22011 try: 21606 buy: 19391 well: 18924
find: 17971 like: 15256 eat: 15115 really: 12619 also: 12203
price: 12183 much: 11899 order: 11782 little: 11781 bag: 11378
give: 10501 treat: 9734 add: 9407 recommend: 8730
think: 8710 say: 8394 even: 8272 first: 7846 brand: 7695 mix: 7397
favorite: 7347 look: 7342 water: 7335

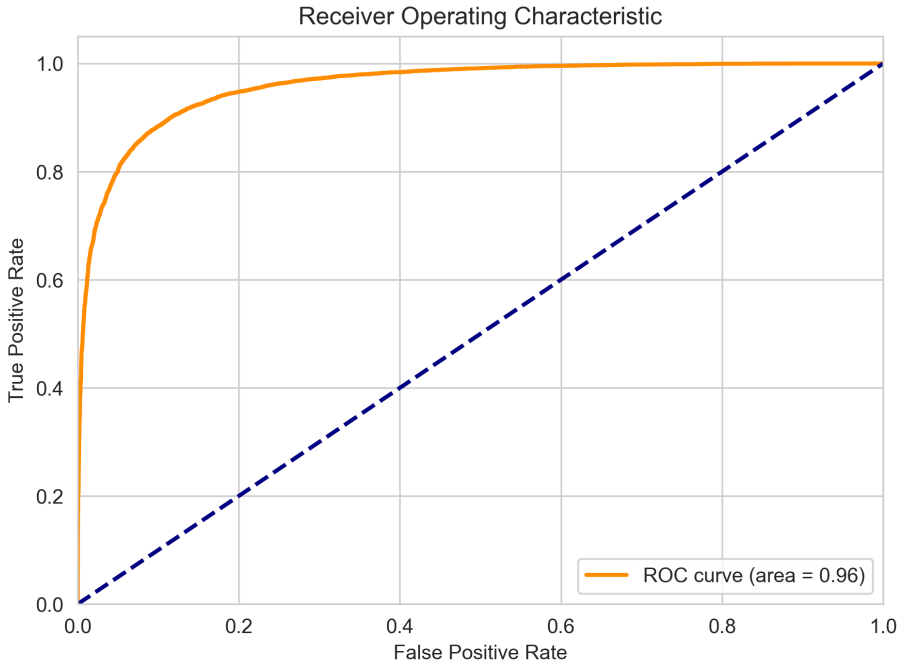
taste: 42293 product: 37509 flavor: 25482 get: 25464 try: 24789
buy: 24697 good: 23178 eat: 16480 think: 16303 order: 16071
well: 15228 bag: 14463 bad: 13672 give: 13471 find: 13065
say: 13064 much: 12907 love: 12446 really: 12412 look: 12055
first: 11341 even: 11285 like: 11198 purchase: 10148 review: 9707
also: 9679 water: 9214 little: 8944 price: 8914 want: 8705
brand: 8679 great: 8241 thing: 8062 smell: 7928 see: 7847

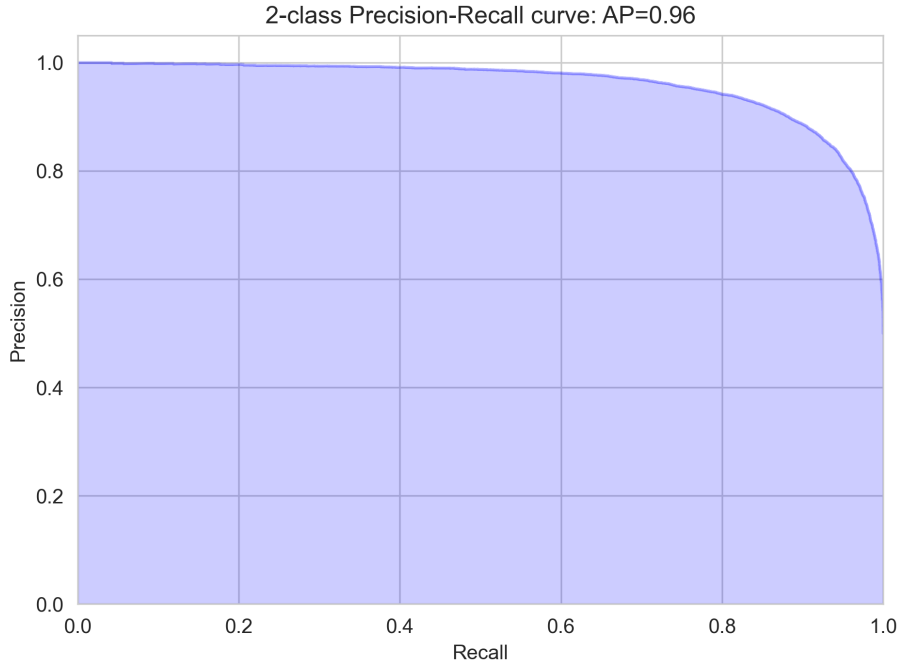
Naïve Bayes Model

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.88	0.89	16408
1	0.88	0.90	0.89	16407
accuracy			0.89	32815
macro avg	0.89	0.89	0.89	32815
weighted avg	0.89	0.89	0.89	32815



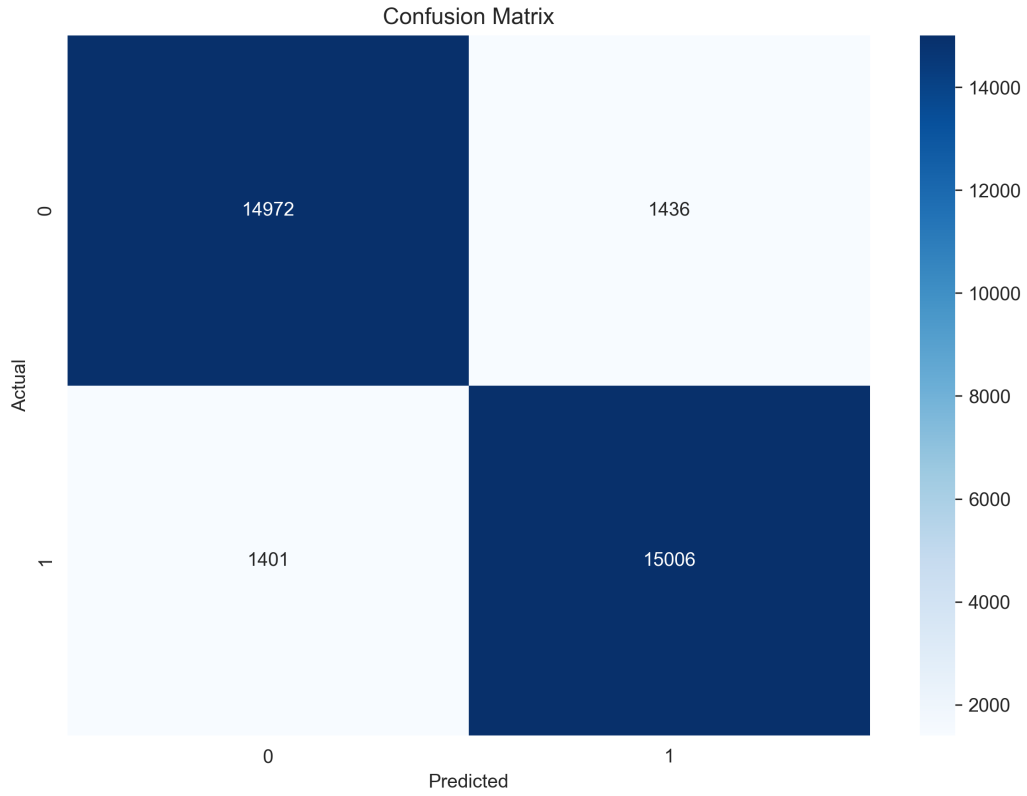




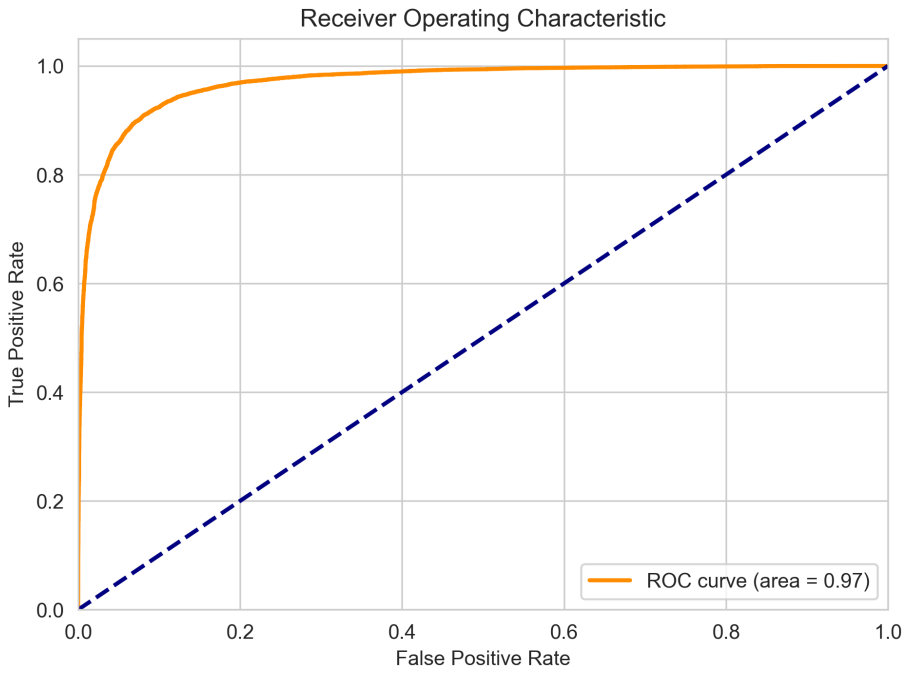
Naïve Bayes Model

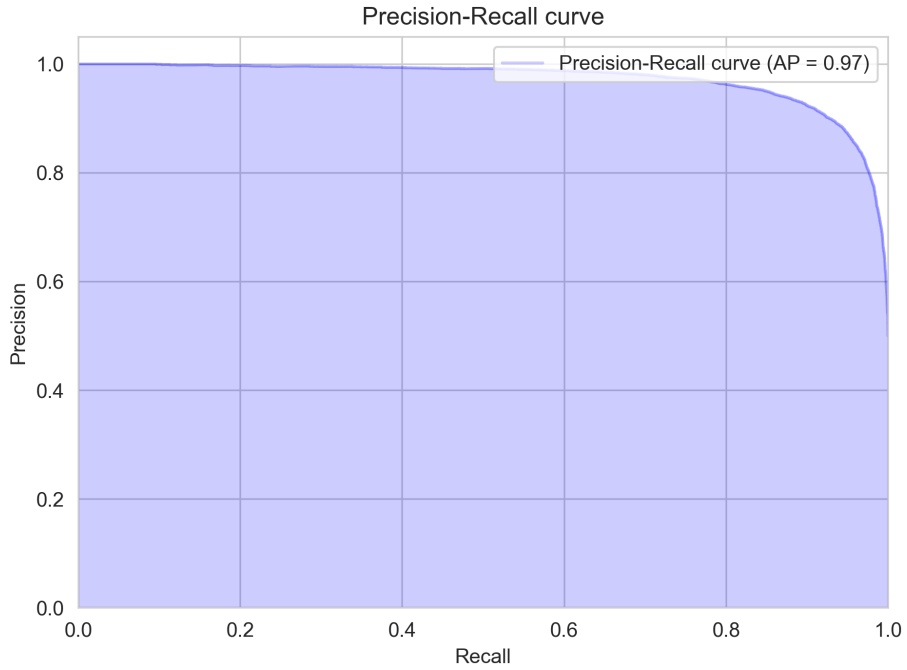
Classification Report:

	precision	recall	f1-score	support
0	0.91	0.91	0.91	16408
1	0.91	0.91	0.91	16407
accuracy			0.91	32815
macro avg	0.91	0.91	0.91	32815
weighted avg	0.91	0.91	0.91	32815



Support Vector Machine Model: **LinearSVC**





Support Vector Machine Model: **LinearSVC**

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.93	0.93	16408
1	0.93	0.93	0.93	16407
accuracy			0.93	32815
macro avg	0.93	0.93	0.93	32815
weighted avg	0.93	0.93	0.93	32815

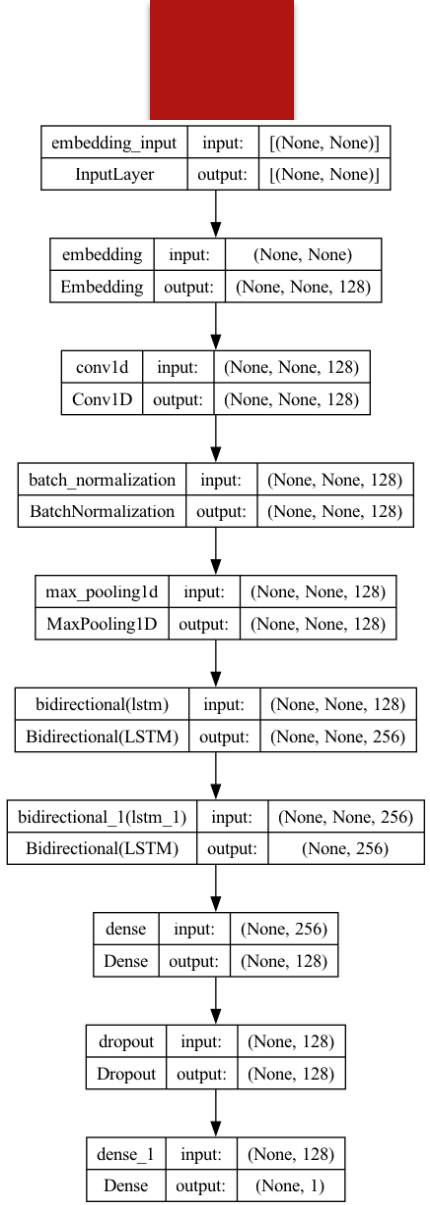
Pre-Trained Model: **DistilBERT**

```
model = Sequential([
        Embedding( input_dim=WORDS_TO_KEEP + 1,
                   output_dim=EMBEDDING_SIZE,
                   weights=[embedding_matrix],
                   trainable=True,
                   mask_zero=True ),

        Conv1D( filters=CONV_FILTERS,
                kernel_size=CONV_KERNEL_SIZE,
                activation='relu',
                kernel_initializer=GlorotUniform(),
                bias_initializer='zeros',
               padding='same', strides=1 ),

        BatchNormalization(),

        MaxPooling1D(pool_size=POOL_SIZE),

        Bidirectional( LSTM( LSTM_UNITS,
                             return_sequences=True,
                             dropout=DROPOUT,
                             kernel_regularizer=kernel_regularizer ) ),

        Bidirectional( LSTM( LSTM_UNITS,
                              dropout=DROPOUT,
                              kernel_regularizer=kernel_regularizer ) ),

        Dense( units=NEURONS,
               activation='relu',
               kernel_regularizer=kernel_regularizer,
               kernel_initializer=HeNormal() ),
        Dropout(DROPOUT), Dense(1, activation='sigmoid')

        Dense(1, activation='sigmoid')
])
```

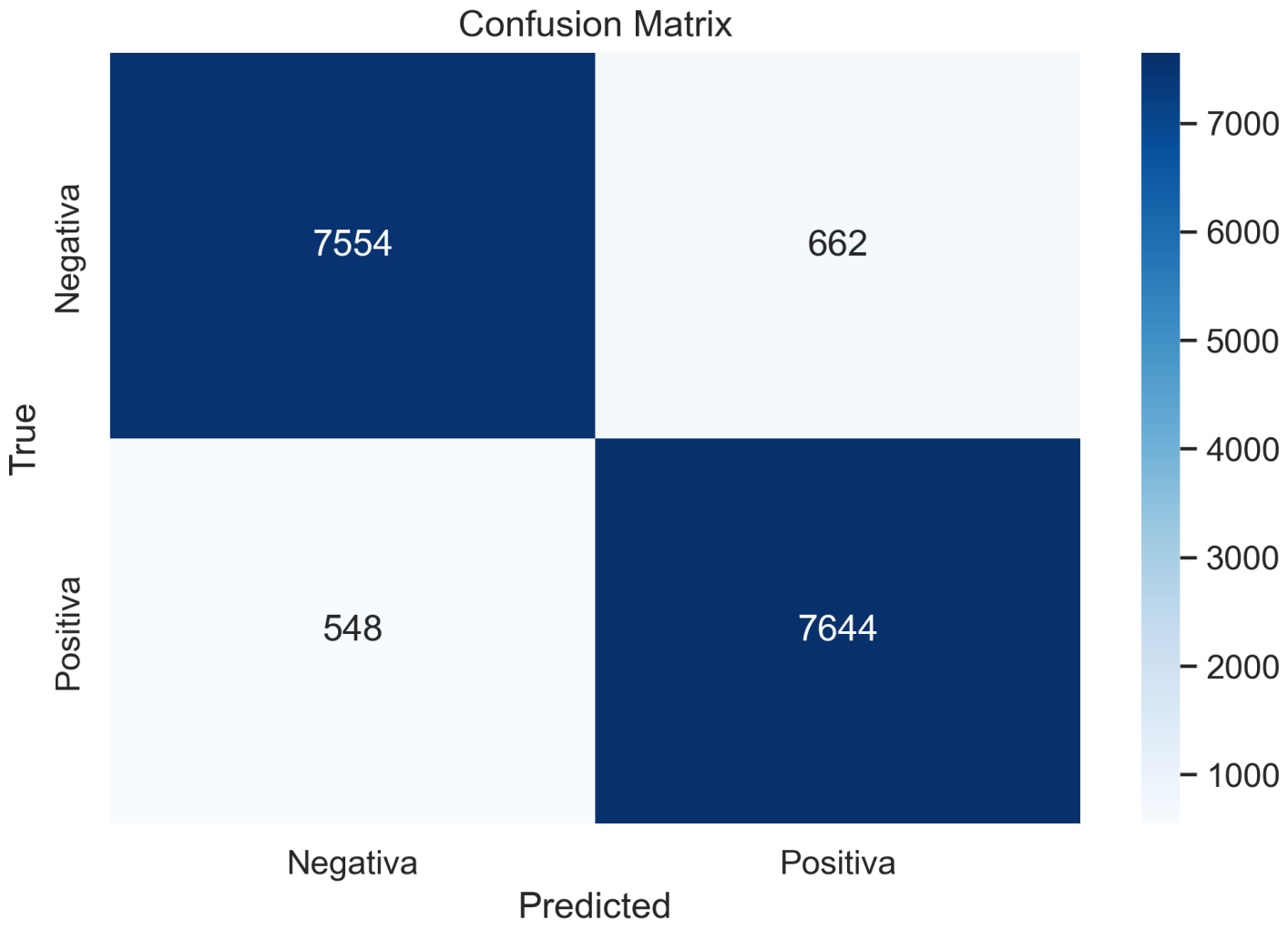
Deep Learning Model

Deep Learning Summary

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 128)         1920128

 conv1d (Conv1D)             (None, None, 128)         49280

 batch_normalization (Batch  (None, None, 128)         512
 Normalization)

 max_pooling1d (MaxPooling1  (None, None, 128)         0
 D)

 bidirectional (Bidirection  (None, None, 256)         263168
 al)

 bidirectional_1 (Bidirecti  (None, 256)               394240
 onal)

 dense (Dense)               (None, 128)               32896

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 1)                 129

=================================================================
Total params: 2660353 (10.15 MB)
Trainable params: 2660097 (10.15 MB)
Non-trainable params: 256 (1.00 KB)
```



Deep Learning Model Performance:



Classification Report:

	precision	recall	f1-score	support
Negativa	0.93	0.92	0.93	8216
Positiva	0.92	0.93	0.93	8192
accuracy			0.93	16408
macro avg	0.93	0.93	0.93	16408
weighted avg	0.93	0.93	0.93	16408