UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Guatemala 06 de mayo 2024

Daniel Gómez 21429

Inteligencia Artificial

Proyecto No. 02

Github

https://github.com/JDgomez2002/proj2_AI.git

Análisis de datos exploratorio (EDA)

Para el Exploratory Data Analysis se determinó que las variables del dataset, son únicamente V1 y V2, las cuales tienen la información vital para poder determinar el modelo. Respecto a la Figura 1 se puede observar que V1 únicamente puede tener dos valores: Spam o Ham. Esto nos permite clasificar y entrenar el modelo para determinar el conjunto de prueba y entrenamiento. También se puede ver en la Figura 2, que casi todas la data de V2 es única, exceptuando algunas respuestas que se repiten, ya que se puede ver que Unique no es el mismo número del total del data set.

También se puede observar en la Figura 3, que para este dataset de ejemplo, hay 4,825 ham y 747 spam. Esto nos puede dar más importancia y puede tener relevancia en el entrenamiento de los modelos.

Limpieza de datos

Para la limpieza de datos, como se puede observar en la Figura 4, se utilizaron diferentes técnicas como lower cases, tokenization, eliminación de marcas de puntuación, eliminación de tokens numéricos, eliminación de Stopwords, lematización y el stemming. Esto ayuda a poder entrenar al modelo y remover variables que pueden interferir con la exactitud de las predicciones. Se recomienda realizarlo para que el modelo no pierda efectividad al evaluar el menor número de factores posibles.

Modelo

Para el modelo se consideró principalmente las probabilidades de S dado W. Esta se obtuvo por medio de una función que calculó individualmente los factores de la fórmula para la probabilidad.

$$P(S|W) = \frac{P(W|S)P(S)}{P(W|S)P + P(W|H)P(H)}$$

Pruebas de rendimiento

En la Figura 5 se puede observar el las pruebas de rendimiento para el modelo de entrenamiento y prueba. Para estas pruebas se iteran cada una de las filas del modelo y se evalúan las predicciones para cada uno de los modelos de prueba y entrenamiento.

Discusión de resultados

Los resultados fueron favorables para las predicciones ya que se pudo tanto en la matriz de confusión en la Figura 6, como en el Reporte de clasificación en la Figura 7 que los resultados son los esperados para cada uno de ellos. Sin embargo, se puede notar que la precisión del Spam es menor a la de ham, por mucho. Esto puede deberse a que el data set tuvo una menor cantidad de spams como se puede ver en la Figura 3. Esto pudo provocar que el modelo tuviera menos alimentación para poder entrenarse de la manera correcta y predecir con más precisión los textos de spam.

En la matriz de confusión de la Figura 6 se ven con mayor claridad estos resultados, teniendo un gran acierto para los Hams predecidos, y nuevamente menos acierto para los falsos spams (los hams que se supone eran spams). Con esto se puede concluir que mi modelo posee cierta dificultad para poder detectar los Spams. Sin embargo, al momento de la presentación el modelo demuestra poder identificar de manera efectiva a los spams, acertando en todos los casos y ejemplos provistos.

Como se puede observar en la Figura 8, los resultados del modelo son favorables, pudiendo identificar de manera correcta los textos, pasando por todos los procesos, como la limpieza de datos, y la tokenización de palabras clave. También se puede observar que puede identificar de manera correcta las Key Words para los spams y para los hams respectivamente.

Finalmente, se puede ver que la probabilidad de spam es acertada para los textos y que la predicción también es correcta para cada uno de ellos.

Conclusiones

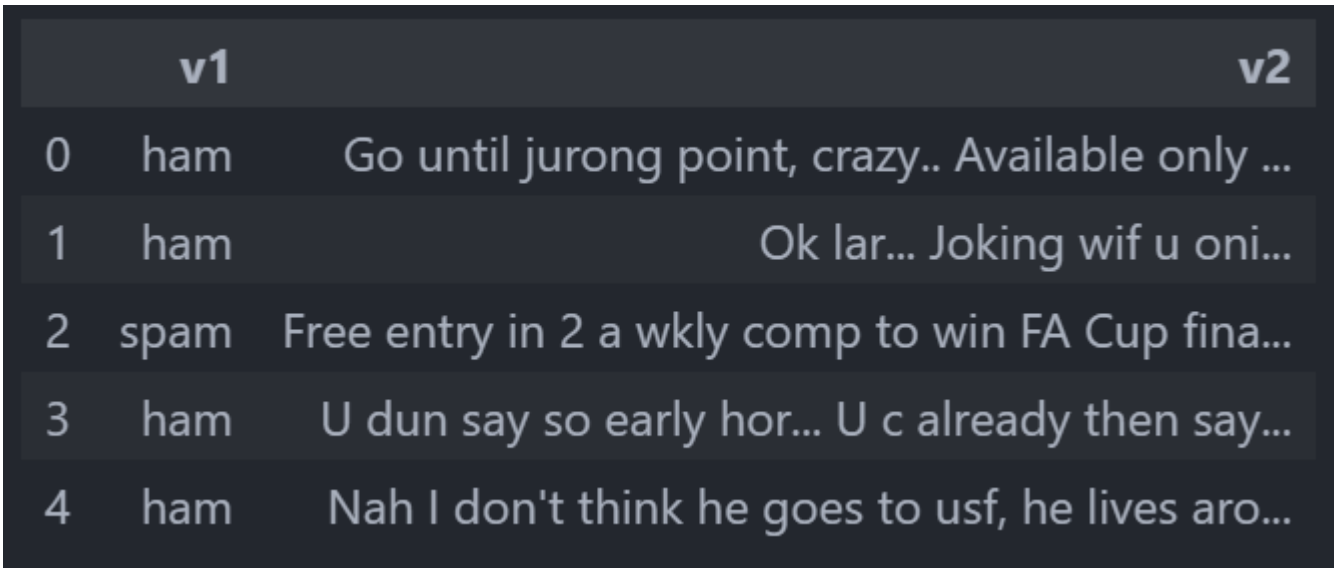
1. Las variables principales en el dataset son V1 y V2, las cuales son críticas para determinar el modelo. La variable V1 solo tiene dos valores posibles: Spam o Ham, lo que permite clasificar y entrenar el modelo de manera efectiva.

2. La variable V2 contiene principalmente datos únicos, pero también hay algunas respuestas repetidas. Esto puede ser relevante para el entrenamiento del modelo y la selección del conjunto de prueba y entrenamiento.

3. El análisis muestra que hay una gran cantidad de datos de tipo "ham" en comparación con los datos de tipo "spam". Esto puede influir en el entrenamiento del modelo y en la precisión de las predicciones, ya que el modelo puede tener menos ejemplos de spams para aprender.
4. Se realizaron diversas técnicas de limpieza de datos, como convertir a minúsculas, tokenización, eliminación de puntuación, eliminación de números, eliminación de palabras irrelevantes, lematización y stemming. Estas técnicas ayudan a entrenar al modelo y eliminar variables que podrían afectar la precisión de las predicciones.

5. Los resultados del modelo fueron favorables en términos de predicciones. Sin embargo, se observa una menor precisión en la clasificación de spams en comparación con los hams. Esto puede atribuirse a la menor cantidad de ejemplos de spams en el dataset utilizado para el entrenamiento del modelo. Aunque el modelo pudo identificar correctamente los ejemplos proporcionados en la evaluación, podría tener dificultades para detectar spams en general.

Anexos



	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Figura 1

	v1	v2
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

Figura 2

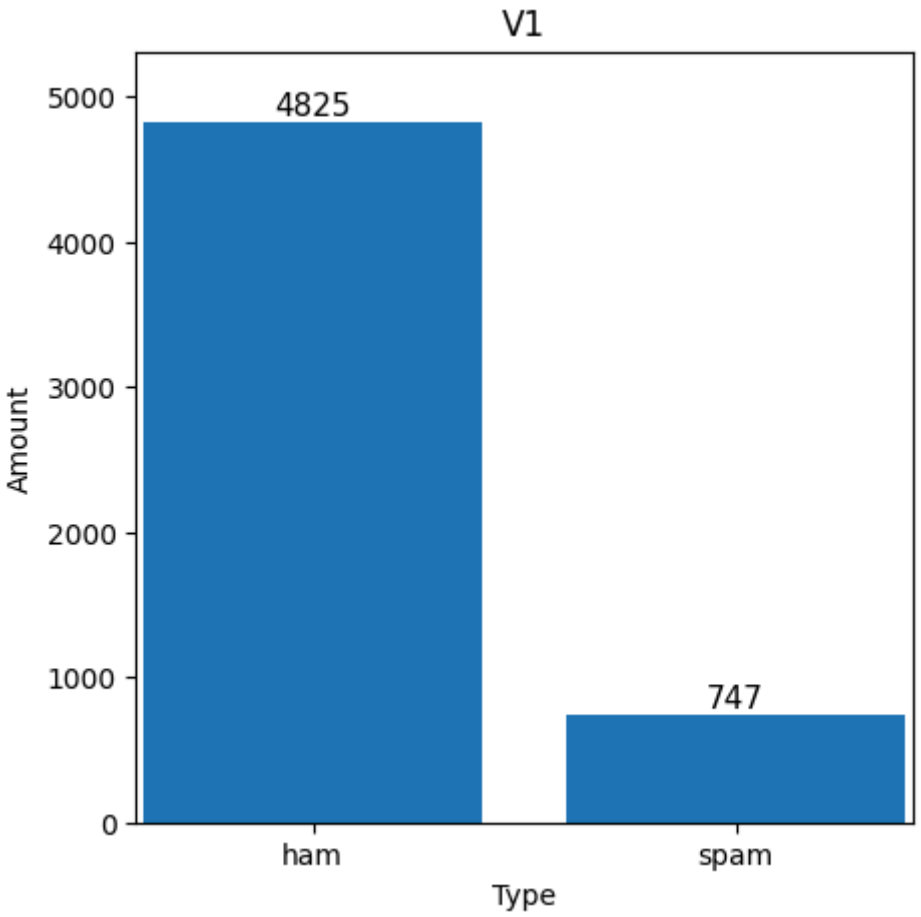


Figura 3

```python
# Lower cases
def toLower(texts):
  return [text.lower() for text in texts]

# Tokenization
def tokenization(texts):
  return [word_tokenize(text) for text in texts]

# Punctuation marks removal
def marksRemoval(texts):
  return [[token for token in text if not re.search(r'\W+', token)] for text in texts]

# number tokens removal
def numTokensRemoval(texts):
  return [[token for token in text if not re.search(r'[0-9]+', token)] for text in texts]

# stopwords removal
def stopwordsRemoval(texts):
  return [[token for token in text if token not in stop_words] for text in texts]

# Lemmatization
def lemmatizater(texts):
  values = [nltk.pos_tag(text) for text in texts]
  return [[lemmatizer.lemmatize(item[0], get_wordnet_pos(item[0], item[1])) for item in value] for value in values]

# stemming
def stemmazer(texts):
  return  [[stemmer.stem(token) for token in text] for text in texts]
```

Figura 4

{'Category': 'ham', 'Text': 'We left already we at orchard now.', 'Recognized words': 'leav → 0.0223,alreadi → 0.0082', 'Spam probability': '0.0004', 'Prediction': 'ham'}
{'Category': 'spam', 'Text': 'Hi babe its Jordan, how r u? Im home from abroad and lonely, text me back if u wanna chat xxSP visionsms.com Text stop to stopCost 150p 087124006
{'Category': 'ham', 'Text': 'Nothing. I meant that once the money enters your account here, the bank will remove its flat rate. Someone transfered <#> to my account an
{'Category': 'ham', 'Text': 'The word \\Checkmate\\" in chess comes from the Persian phrase \\"Shah Maat\\" which means; \\"the king is dead..\\" Goodmorning.. Have a good day
{'Category': 'ham', 'Text': 'Lol, oh you got a friend for the dog ?', 'Recognized words': 'oh → 0.0076,get → 0.0829,friend → 0.0957,dog → 0.4773', 'Spam probability': '0.0
{'Category': 'ham', 'Text': 'If you ask her or she say any please message.', 'Recognized words': 'ask → 0.0327,pleas → 0.3134,messag → 0.3093', 'Spam probability': '0.0138'
{'Category': 'ham', 'Text': 'Hello- thanx for taking that call. I got a job! Starts on monday!', 'Recognized words': 'take → 0.08,call → 0.421,get → 0.0829,start → 0.1544'
{'Category': 'ham', 'Text': 'I think you should go the honesty road. Call the bank tomorrow. Its the tough decisions that make us great people.', 'Recognized words': 'think →
{'Category': 'spam', 'Text': 'Reply to win â£100 weekly! What professional sport does Tiger Woods play? Send STOP to 87239 to end service', 'Recognized words': 'repli → 0.614
{'Category': 'ham', 'Text': "K come to nordstrom when you're done", 'Recognized words': 'k → 0.0077,come → 0.0154,do → 0.08', 'Spam probability': '0.0', 'Prediction': 'ham'
{'Category': 'ham', 'Text': 'Haha yeah, 2 oz is kind of a shitload', 'Recognized words': 'yeah → 0.009,oz → 0.549', 'Spam probability': '0.0218', 'Prediction': 'ham'}
{'Category': 'ham', 'Text': 'We are both fine. Thanks', 'Recognized words': 'fine → 0.0166,thank → 0.1071', 'Spam probability': '0.004', 'Prediction': 'ham'}
{'Category': 'spam', 'Text': "I want some cock! My hubby's away, I need a real man 2 satisfy me. Txt WIFE to 89938 for no strings action. (Txt STOP 2 end, txt rec â£1.50ea. OT
{'Category': 'ham', 'Text': 'Good! No, don\x89Ûªt need any receipts\x89ÛÓwell done! (\x89Û_) Yes, please tell . What\x89Ûªs her number, i could ring her', 'Recognized words':
{'Category': 'ham', 'Text': 'Just sleeping..and surfing', 'Recognized words': 'Words not recognized :('}
{'Category': 'spam', 'Text': 'TheMob>Hit the link to get a premium Pink Panther game, the new no. 1 from Sugababes, a crazy Zebra animation or a badass Hoody wallpaper-all 4 F
{'Category': 'spam', 'Text': "Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 084528100
{'Category': 'ham', 'Text': 'Ok.ok ok..then..whats ur todays plan', 'Recognized words': 'ok → 0.0133,what → 0.0447,ur → 0.2756,today → 0.1687', 'Spam probability': '0.0',
{'Category': 'ham', 'Text': 'Reason is if the team budget is available at last they buy the unsold players for at base rate..', 'Recognized words': 'team → 0.3784,avail → 0.
{'Category': 'ham', 'Text': 'Hi:)cts employee how are you?', 'Recognized words': 'hi → 0.0678', 'Spam probability': '0.1356', 'Prediction': 'ham'}
{'Category': 'ham', 'Text': "Solve d Case : A Man Was Found Murdered On <DECIMAL> . <#> AfterNoon. 1,His wife called Police. 2,Police questioned everyone. 3,Wife
{'Category': 'spam', 'Text': 'Congrats! 2 mobile 3G Videophones R yours. call 09061744553 now! videochat wid ur mates, play java games, Dload polyH music, noline rentl. bx420.
{'Category': 'ham', 'Text': 'No no:)this is kallis home ground.amla home town is durban:)', 'Recognized words': 'home → 0.005,home → 0.005,town → 0.0633', 'Spam probability
{'Category': 'spam', 'Text': "Sorry I missed your call let's talk when you have the time. I'm on 07090201529", 'Recognized words': 'sorri → 0.0101,miss → 0.0556,call → 0.42
{'Category': 'ham', 'Text': 'This is wishing you a great day. Moji told me about your offer and as always i was speechless. You offer so easily to go to great lengths on my be
...
{'Category': 'ham', 'Text': 'The Xmas story is peace.. The Xmas msg is love.. The Xmas miracle is jesus.. Hav a blessed month ahead & wish U Merry Xmas ... ', 'Recognized wo
{'Category': 'spam', 'Text': 'You are a winner U have been specially selected 2 receive â£1000 cash or a 4* holiday (flights inc) speak to a live operator 2 claim 087127781081
{'Category': 'ham', 'Text': "Sorry, in meeting I'll call you later", 'Recognized words': 'sorri → 0.0101,meet → 0.0261,call → 0.421', 'Spam probability': '0.0004', 'Predic
{'Category': 'ham', 'Text': "I'm a guy, browsin is compulsory", 'Recognized words': 'guy → 0.0282', 'Spam probability': '0.0564', 'Prediction': 'ham'}

Figura 5

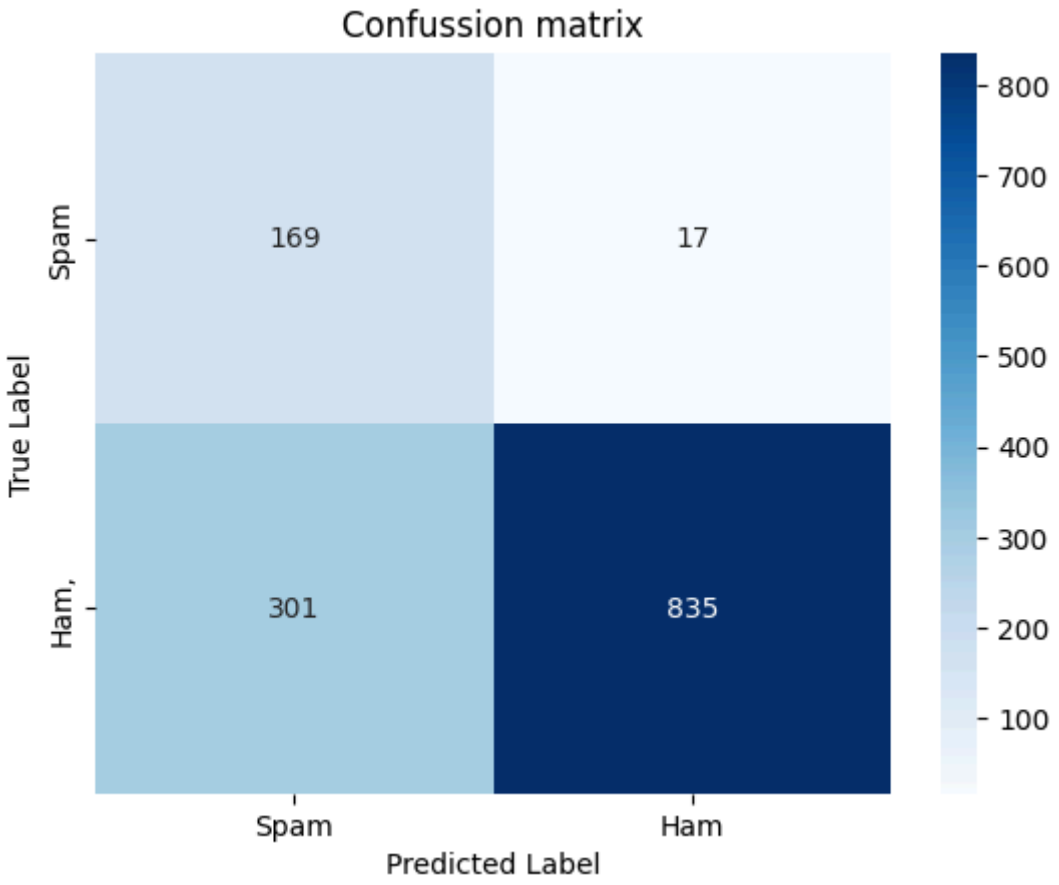
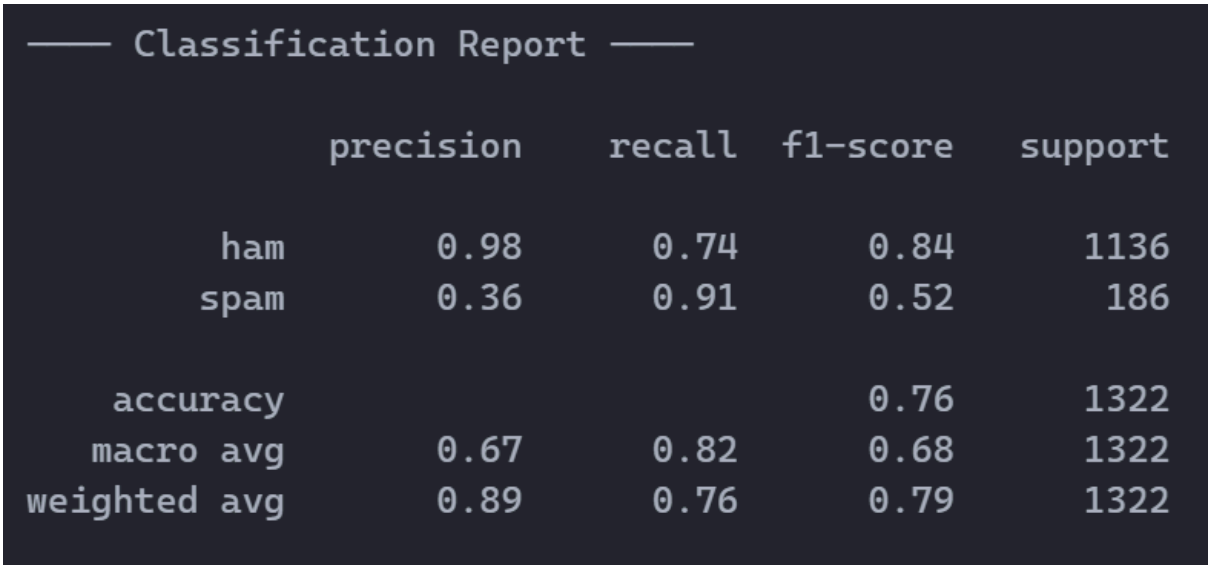


Figura 6



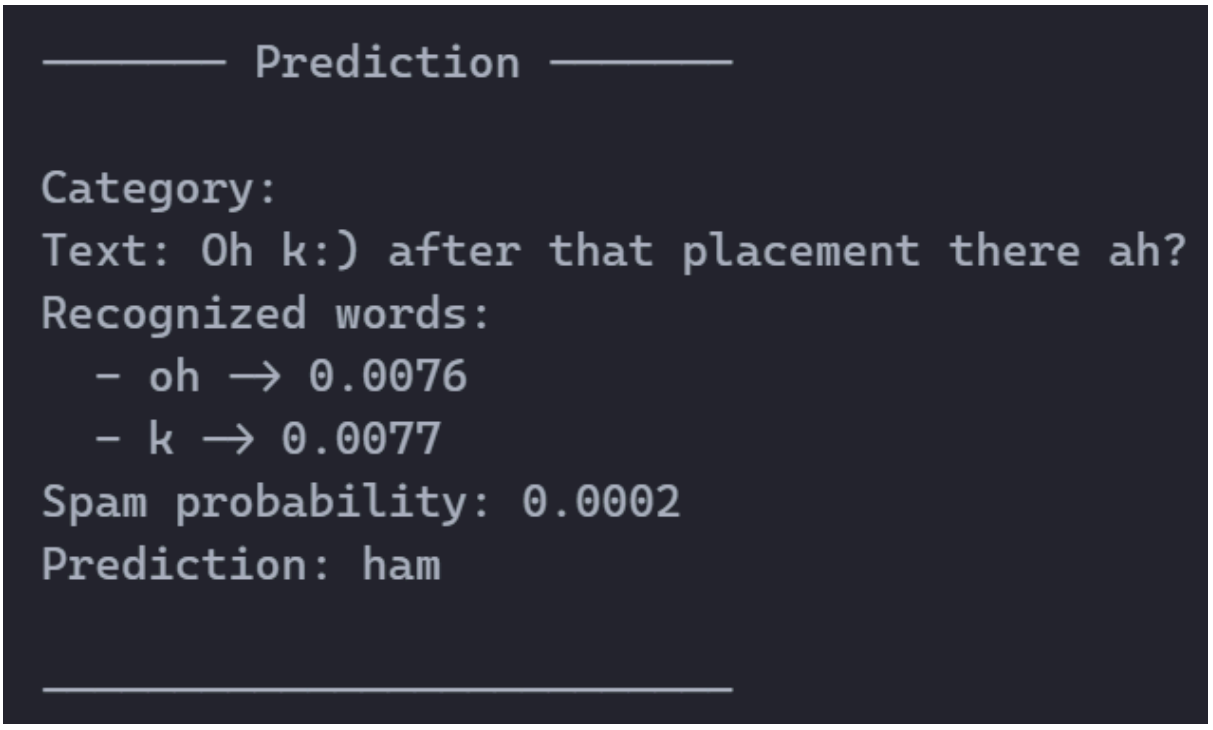
```
─── Classification Report ───

              precision    recall  f1-score   support

         ham       0.98      0.74      0.84      1136
        spam       0.36      0.91      0.52       186

    accuracy                           0.76      1322
   macro avg       0.67      0.82      0.68      1322
weighted avg       0.89      0.76      0.79      1322
```

Figura 7



```
─────── Prediction ───────

Category:
Text: Oh k:) after that placement there ah?
Recognized words:
  - oh → 0.0076
  - k → 0.0077
Spam probability: 0.0002
Prediction: ham

──────────────────────────────
```

Figura 8