Universidad Autónoma de Nuevo León
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos
Alder López Cerda
Tarea 2 : Análisis de Sentimiento

Resumen

Se realizará un análisis de sentimiento de un dataset de internet; se aplicará el ejercicio en clase y se modificará para comprar resultados.

En este trabajo se enfocará en:

- Mostrar resultados utilizado al menos 3 librerías.
- Explicar los cambios realizados al código de ejemplo (parámetros, etc.), esta descrito en el código del notebook.
- Utilizar random forest y explicar sus resultados y conclusiones.

En general se va a realizar:

1) Preprocesamiento de la entrada en sus oraciones o palabras componentes.
2) Identificar y etiquete cada token con un componente de parte del discurso (es decir, sustantivo, verbo, determinantes, sujeto de la oración, etc.).
3) Asignar una puntuación de sentimiento de -1 a 1, donde -1 es un sentimiento negativo, 0 es neutral y +1 es un sentimiento positivo
4) Devuelver la puntuación y las puntuaciones opcionales, como la puntuación compuesta, la subjetividad, etc.

Análisis de sentimiento con diferentes métodos:

TextBlob

Este analizador de sentimientos devuelve dos propiedades para una oración de entrada dada:

- La polaridad es un flotante que se encuentra entre [-1,1], -1 indica sentimiento negativo y +1 indica sentimiento positivo.
- La subjetividad también es un flotador que se encuentra en el rango de [0,1]. Las oraciones subjetivas generalmente se refieren a opiniones, emociones o juicios.

Out[302]:

	reviews.text	Lemma	Polarity	Analysis
0	Our experience at Rancho Valencia was absolute...	experience Rancho Valencia absolutely perfec...	0.539286	Positive
1	Amazing place. Everyone was extremely warm and...	Amazing place Everyone extremely warm welcom...	0.475000	Positive
2	We booked a 3 night stay at Rancho Valencia to...	book night stay Rancho Valencia play tennis ...	0.484444	Positive
3	Currently in bed writing this for the past hr ...	Currently bed writing past hr dog bark squea...	-0.125000	Negative
4	I live in Md and the Aloft is my Home away fro...	live Md Aloft Home away home stay night Staf...	0.258701	Positive

```
In [303]: tb_counts = fin_data.Analysis.value_counts()
          tb_counts

Out[303]: Positive    4455
          Negative     468
          Neutral       77
          Name: Analysis, dtype: int64
```

VADER

Utiliza una lista de características léxicas (por ejemplo, palabra) que se etiquetan como positivas o negativas según su orientación semántica para calcular el sentimiento del texto. El sentimiento de Vader devuelve la probabilidad de que una oración de entrada dada sea positiva, negativa y neutral.

Out[307]:

	reviews.text	Lemma	Polarity	Analysis	Vader Sentiment	Vader Analysis
0	Our experience at Rancho Valencia was absolute...	experience Rancho Valencia absolutely perfec...	0.539286	Positive	0.9392	Positive
1	Amazing place. Everyone was extremely warm and...	Amazing place Everyone extremely warm welcom...	0.475000	Positive	0.9656	Positive
2	We booked a 3 night stay at Rancho Valencia to...	book night stay Rancho Valencia play tennis ...	0.484444	Positive	0.9763	Positive
3	Currently in bed writing this for the past hr ...	Currently bed writing past hr dog bark squea...	-0.125000	Negative	0.0000	Neutral
4	I live in Md and the Aloft is my Home away fro...	live Md Aloft Home away home stay night Staf...	0.258701	Positive	0.8344	Positive

```
In [308]: vader_counts = fin_data['Vader Analysis'].value_counts()
          vader_counts

Out[308]: Positive    3883
          Neutral      873
          Negative     244
          Name: Vader Analysis, dtype: int64
```

SentiWordNet

Fue creado para la minería de opiniones. SentiWordNet asigna a cada synset de WordNet tres puntajes de sentimiento: positividad, negatividad, objetividad.

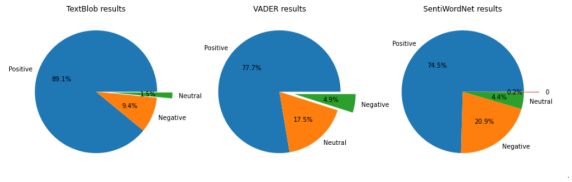
Out[310]:

	reviews.text	Lemma	Polarity	Analysis	Vader Sentiment	Vader Analysis	SWN analysis
0	Our experience at Rancho Valencia was absolute...	experience Rancho Valencia absolutely perfec...	0.539286	Positive	0.9392	Positive	Positive
1	Amazing place. Everyone was extremely warm and...	Amazing place Everyone extremely warm welcom...	0.475000	Positive	0.9656	Positive	Positive
2	We booked a 3 night stay at Rancho Valencia to...	book night stay Rancho Valencia play tennis ...	0.484444	Positive	0.9763	Positive	Positive
3	Currently in bed writing this for the past hr ...	Currently bed writing past hr dog bark squea...	-0.125000	Negative	0.0000	Neutral	Neutral
4	I live in Md and the Aloft is my Home away fro...	live Md Aloft Home away home stay night Star...	0.258701	Positive	0.8344	Positive	Positive

```
In [311]: swn_counts= fin_data['SWN analysis'].value_counts()
          swn_counts
```

```
Out[311]: Positive    3727
          Negative    1046
          Neutral      219
          0              8
          Name: SWN analysis, dtype: int64
```

Comparación



Podemos observar que el método Textblob tiene más resultados positivos y SentiWordNEt menos resultados positivos. Adicionalmente se observa que el método SentiWordNet arroja resultados que no pudieron ser catalogados como Positivo, neutral ni negativo.

Los 3 métodos realizan su cálculo de acuerdo con como fueron diseñados, por los resultados nos dice que los 3 son muy acertados. En el caso de SentiWordNet quizá falta realizar mas limpieza para eliminar esos casos donde no se pudo determinar el sentimiento.

Random Forest

```
In [326]: from sklearn.ensemble import RandomForestClassifier

          text_classifier = RandomForestClassifier(n_estimators=200, random_state=0)
          text_classifier.fit(X_train, y_train)
```

```
Out[326]: RandomForestClassifier(n_estimators=200, random_state=0)
```

Usando la librería de sklearn, randomforestclassifier, podemos entrenar un modelo de aprendizaje de maquina que nos ayuda a predecir el sentimiento del tweet.

```
In [327]: predictions = text_classifier.predict(X_test)
```

```
In [328]: from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

          print(confusion_matrix(y_test,predictions))
          print(classification_report(y_test,predictions))
          print(accuracy_score(y_test, predictions))
```

```
[[ 29   0  58]
 [  0   3  10]
 [  3   3 894]]
              precision    recall  f1-score   support

    Negative       0.91      0.33      0.49        87
     Neutral       0.50      0.23      0.32        13
    Positive       0.93      0.99      0.96       900

    accuracy                           0.93      1000
   macro avg       0.78      0.52      0.59      1000
weighted avg       0.92      0.93      0.91      1000

0.926
```

Como sabemos el método de Random Forest (RF) es un algoritmo de aprendizaje supervisado con el cual crea y combina aleatoriamente los múltiples árboles de decisión, además es un meta estimulador que se ajusta a distintos clasificadores y utiliza el promedio para mejorar la precisión de predicción y controlar el sobre ajuste. Ahora bien, con el input de la clasificación de sentimiento se pudo entrenar y obtener las predicciones con el 92% de precisión, el cual es muy bueno considerando que se acerca al 95% típico de significancia.

Github
https://github.com/pacificIT/MCD-Procesamiento-Datos/tree/main/T2