Técnicas para análisis de sentimientos en reviews de Amazon

Sobre el dataset

Se utiliza el Amazon Review Data (2018)[1]. Este dataset contiene un total de 233.1 million reviews y tiene la ventaja de estar subdividido por categorías lo que lo vuelve interesante para experimentar con diversos productos.

“This Dataset is an updated version of the Amazon review dataset released in 2014. As in the previous version, this dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).”

Citation:
Justifying recommendations using distantly-labeled reviews and fined-grained aspects
Jianmo Ni, Jiacheng Li, Julian McAuley
Empirical Methods in Natural Language Processing (EMNLP), 2019

Técnicas de NLP

Metodología

Pre-procesamiento de los datos:

Selección del dataset

Se probó con las subcategorías *Musical Instruments* y *Software* por tener tamaños reducidos. Dado que los tiempos de procesamiento eran elevados se terminó aplicando el código con la última.

Estructura del dataset:

---

[1] **Justifying recommendations using distantly-labeled reviews and fined-grained aspects**
Jianmo Ni, Jiacheng Li, Julian McAuley
Empirical Methods in Natural Language Processing (EMNLP), 2019

	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
7995	1.0	False	12 4, 2013	A2QFUMSDNMZY1H	B00EOI2SR2	{'Format:': ' Software'}	Patriot	I could not get X6 to install properly and Cor...	Compatibility Issues with Microsoft KB2670838 ...	1386115200	19	NaN
8775	5.0	True	05 2, 2015	A25NDDJD0UUALR	B00FZ0FK0U	{'Format:': ' Software Download'}	Christopher	Look no further for security software for mac ...	Excellent	1430524800	NaN	NaN
3364	5.0	False	09 20, 2008	A3JJNSW1XHXX9B	B0018EI3I8	NaN	HealthIT	Kaspersky used to be more geared to admins, co...	Great for the PC beginner, or PC GOD! This is ...	1221868800	11	NaN
5379	4.0	False	05 11, 2012	A2I3XS9TO93Q0F	B004PBG8EC	{'Format:': ' Software'}	Michael Meredith	I know, you can get directions for free to get...	Great planner for a road trip!	1336694400	NaN	NaN
9563	4.0	True	02 11, 2015	A2IOGHTCJ3QAHU	B00M9GTEPA	{'Platform:': ' PC Disc'}	Tall Timbers	I've been using Quicken since 1992 when I got ...	Quicken is a Powerful Tool if You're Dedicated...	1423612800	NaN	NaN

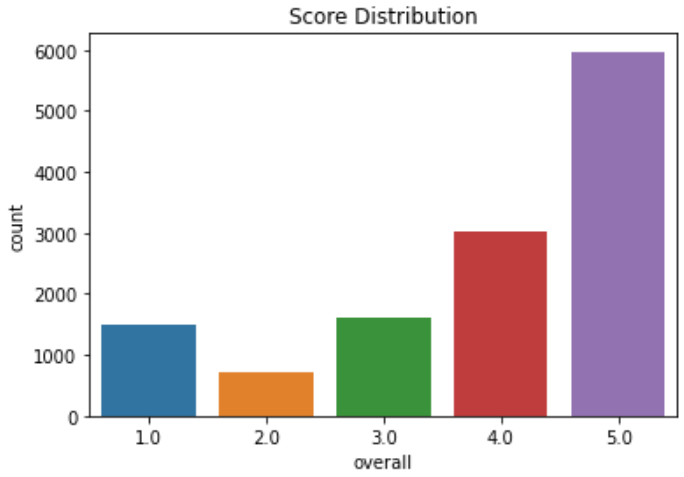
Para aplicar sentiment analysis se utilizaron únicamente las columnas *reviewText*, que contiene el texto del review y *overall* que contiene el rating asignado por los usuarios. Este último valor va de 1.0 a 5.0 (1 a 5 estrellas)

Dado que se buscó aplicar una clasificación binaria del tipo positivo/negativo se realizó una conversión de la escala anterior de la siguiente manera:
1.0 y 2.0 - Negativo
4.0 y 5.0 - Positivo

Para reducir sesgos se decidió descartar los valores neutrales 3.0, ya que podría presentarse la ambigüedad de que usuarios consideren el valor 3 como positivo o negativo. Otra posibilidad era asignar aleatoriamente positivo/negativo a la valoración 3, como se plantea en este trabajo[2], pero se buscó reducir al mínimo los datos por la escasa posibilidad de procesamiento.

Visualización

Se visualiza la distribución del rating en los reviews del subconjunto Software del dataset.



Se utilizó WordCloud para visualizar el dataset. Se experimentó con shape, aplicando una máscara con el logotipo de Amazon para hacer más interesante la visualización.

[2] http://www.narimanfarsad.com/cps803/docs/samples/CPS803-SampleReport-SentimentAnalysis.pdf

Luego se realizó limpieza de los datos, normalización y eliminación de stop-words.

Método 1: Clasificador Naive Bayes

A partir de los datos pre-procesados se aplicó uno de los métodos vistos en el curso: Naïve Bayes.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Para esto se utilizó el módulo *NaiveBayesClassifier* incluído en el toolkit NLTK[3].

A partir del pre-procesamiento del dataset descrito anteriormente, se procede con:

- Limpieza, normalization, stop-words
- Clasificador bayesiano ingenuo
- Resultados

[Enlace al notebook]

Resultados obtenidos:

```
# Accuracy
print(nltk.classify.accuracy(classifier, training_set[0:500]))
```

0.912

```
print(nltk.classify.accuracy(classifier, test_set[0:500]))
```

[3] NLTK: Learning to Classify Text. https://www.nltk.org/book/ch06.html

0.866

Al igual que en el caso trabajado en clase, la menor accuracy en el test respecto al training está indicando un posible overfit.

Método 2: LSTM (Long Short Term Memory)

A partir de los datos pre-procesados se aplicó una red neuronal tipo LSTM

A partir del pre-procesamiento del dataset descrito anteriormente, se procede con:

- Limpieza, normalization, stop-words
- Tokenization
- Armado de la RNN (Se utilizan las capas vistas en clase):
    - Embedding: Utilizada para bajar la dimensionalidad
    - LSTM
    - Capa Densa con función de activación softmax y la cantidad de nodos necesarios para el problema de clasificación
- Resultados

Algunas dificultades encontradas

Tokenización:
Al realizar la tokenización se encontró la dificultad de las dimensiones en la secuencia de salida, se redujeron las *max_features* para obtener un resultado que corriera en la plataforma Colab.

```
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences

# Se limita a 2000 features por eficiencia
max_features = 250
#Inicializo el Tokenizer
tokenizer = Tokenizer(num_words=max_features, split=' ')
#Entrenamiento
tokenizer.fit_on_texts(df['reviewText'].values)
#texts_to_sequences transforma el input en arrays numéricos
X = tokenizer.texts_to_sequences(df['reviewText'].values)
#pads_sequences transforma las secuencias al mismo largo
X = pad_sequences(X)
```

```
[ ] X.shape

    (11206, 2671)
```

Entrenamiento del modelo:

Debido a lo anterior, el entrenamiento del modelo fué extremadamente lento por lo que se ajustó a batch_size=16 y epochs=2. Esto arroja una accuracy muy baja.

Entrenamiento del modelo

```
batch_size = 16
model.fit(X_train, Y_train, epochs = 2, batch_size=batch_size, verbose = 2)
```

```
Epoch 1/2
470/470 - 5011s - loss: 0.4515 - accuracy: 0.8026 - 5011s/epoch - 11s/step
Epoch 2/2
470/470 - 4855s - loss: 0.4017 - accuracy: 0.8235 - 4855s/epoch - 10s/step
<keras.callbacks.History at 0x7f8d03b52610>
```

Resultados obtenidos:

Evaluación de resultados

```
#Tomo 1500 datos para la validación y test
validation_size = 1500

X_validate = X_test[-validation_size:]
Y_validate = Y_test[-validation_size:]
X_test = X_test[:-validation_size]
Y_test = Y_test[:-validation_size]
```

```
score,acc = model.evaluate(X_test, Y_test, verbose = 2, batch_size = batch_size)
print("Score: %.2f" % (score))
print("Acc: %.2f" % (acc))
```

```
138/138 - 77s - loss: 0.3818 - accuracy: 0.8294 - 77s/epoch - 557ms/step
Score: 0.38
Acc: 0.83
```

Finalmente tras model.evaluate se obtiene un accuracy de 83% y un score 38%

Queda pendiente revisar los datos y el procesamiento para mejorar estos resultados.

Por último se aplica el método *predict* para verificar el comportamiento de reviews positivos frente a negativos, buscando posibles sesgos en los datos.

pos_acc 93.69747899159664 %
neg_acc 42.90322580645161 %

Como se visualizó anteriormente en la distribución de ratings en los reviews, los positivos (4 y 5 estrellas) son mucho mayores que los negativos (1 y 2 estrellas). Esta distribución despareja se traslada a los datos de entrenamiento, mejorando el accuracy para detectar reviews positivas.

Anexos

Review del dataset por categoría:

Amazon Fashion	reviews (883,636 reviews)	metadata (186,637 products)
All Beauty	reviews (371,345 reviews)	metadata (32,992 products)
Appliances	reviews (602,777 reviews)	metadata (30,459 products)
Arts, Crafts and Sewing	reviews (2,875,917 reviews)	metadata (303,426 products)
Automotive	reviews (7,990,166 reviews)	metadata (932,019 products)
Books	reviews (51,311,621 reviews)	metadata (2,935,525 products)
CDs and Vinyl	reviews (4,543,369 reviews)	metadata (544,442 products)
Cell Phones and Accessories	reviews (10,063,255 reviews)	metadata (590,269 products)
Clothing Shoes and Jewelry	reviews (32,292,099 reviews)	metadata (2,685,059 products)
Digital Music	reviews (1,584,082 reviews)	metadata (465,392 products)
Electronics	reviews (20,994,353 reviews)	metadata (786,868 products)
Gift Cards	reviews (147,194 reviews)	metadata (1,548 products)
Grocery and Gourmet Food	reviews (5,074,160 reviews)	metadata (287,209 products)
Home and Kitchen	reviews (21,928,568 reviews)	metadata (1,301,225 products)
Industrial and Scientific	reviews (1,758,333 reviews)	metadata (167,524 products)
Kindle Store	reviews (5,722,988 reviews)	metadata (493,859 products)
Luxury Beauty	reviews (574,628 reviews)	metadata (12,308 products)
Magazine Subscriptions	reviews (89,689 reviews)	metadata (3,493 products)
Movies and TV	reviews (8,765,568 reviews)	metadata (203,970 products)
Musical Instruments	reviews (1,512,530 reviews)	metadata (120,400 products)
Office Products	reviews (5,581,313 reviews)	metadata (315,644 products)
Patio, Lawn and Garden	reviews (5,236,058 reviews)	metadata (279,697 products)
Pet Supplies	reviews (6,542,483 reviews)	metadata (206,141 products)
Prime Pantry	reviews (471,614 reviews)	metadata (10,815 products)
Software	reviews (459,436 reviews)	metadata (26,815 products)
Sports and Outdoors	reviews (12,980,837 reviews)	metadata (962,876 products)
Tools and Home Improvement	reviews (9,015,203 reviews)	metadata (571,982 products)
Toys and Games	reviews (8,201,231 reviews)	metadata (634,414 products)
Video Games	reviews (2,565,349 reviews)	metadata (84,893 products)