Analisis NPL sobre Women's E-Commerce Clothing Reviews

Maria Elena Bedolla Zamudio
Tecnologias para la informacion en ciencias
UNAM ENES Morelia
candynena_06@hotmail.com

Figure 1: Women Nation

ABSTRACT

En este reporte se incluye la exploracion de un dataset sobre reseñas en compras online de ropa de mujer, mediante el uso de tecnicas de aprendizaje no supervizado. Este problema de exploracion entra en la categoria NPL, la cual es una rama del Machine Learning en la que se explora las palabras mas relevantes de las opiniones de las personas, en este caso las mujeres que compran en determinada empresa por medio del comercio electronico.

CCS CONCEPTS

• **Data Mining**; • **Aprendizaje no supervisado**; • **NPL**;

KEYWORDS

Exploracion, Reseñas, Compras, Palabras, E-commerce

Mineria de datos, Proyecto final, UNAM

ACM ISBN 978-1-4503-1234-5/17/07.
https://doi.org/10.1145/8888888.7777777

ACM Reference Format:
Maria Elena Bedolla Zamudio. 2021. Analisis NPL sobre Women's E-Commerce Clothing Reviews. In *Proceedings of Mineria de datos.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/8888888.7777777

1 INTRODUCTION

En este reporte se muestra el analisis NPL sobre las reseñas sobre las compras en linea de ropa de mujer. El conjunto de datos que se utiliza se llama Women's E-Commerce Clothing Reviews, su autora es nicapotato [Nicapotato 2018] y este se encuentra en Kaggle. El conjunto de datos a trabajar trata sobre el comercio electrónico de ropa de mujer que gira en torno a las reseñas escritas por los clientes. Este dataset es un excelente entorno para analizar el texto a través de sus múltiples dimensiones. Debido a que se trata de datos comerciales reales, se han anonimizado y las referencias a la empresa en el texto y el cuerpo de la reseña se han reemplazado por "minorista". Este conjunto de datos incluye 23486 filas y 10 variables de características. Cada fila corresponde a una revisión de un cliente e incluye las variables:

- Clothing ID: Variable categórica entera que se refiere a la pieza específica que se está revisando.
- Age: Variable entera positiva de la edad de los revisores.
- Title: Variable de cadena para el título de la reseña.

- Review Text: variable de cadena para el cuerpo de la revisión.
- Rating: Variable entera ordinal positiva para la calificación del producto otorgada por el cliente de 1 Peor a 5 Mejor.
- Recommended IND: Variable binaria que indica dónde el cliente recomienda el producto donde se recomienda 1, no se recomienda 0.
- Positive Feedback Count: número entero positivo que documenta la cantidad de otros clientes que encontraron positiva esta revisión.
- Division Name: Nombre categórico de la división de alto nivel del producto.
- Department Name: nombre categórico del nombre del departamento del producto.
- Class Name: nombre categórico del nombre de clase de producto.

En dicho analisis se aplican algunas tecnicas de aprendizaje no supervisado.

2 EXPOSITION

Primero se extrae las frecuencias de palabras del texto, pues es necesario obtener informacion numerica o cuantitativa para conocer el punto sobre el que se iniciara el analisis.

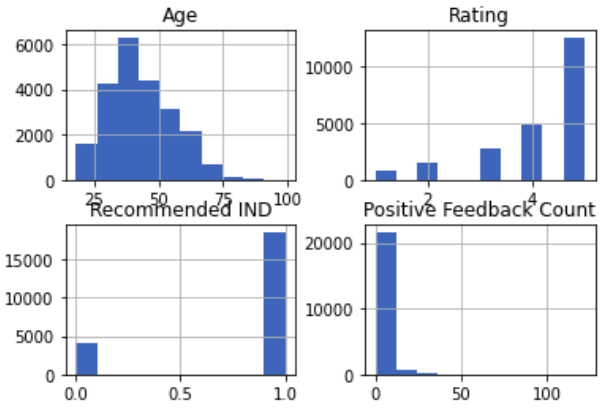


Figure 2: Extraccion de frecuencias de palabras del texto

2.1 Preprocesamiento

El preprocesamiento es una de las principales tareas a realizar, pues de esto depende de las decisiones que vamos a tomar a la hora de procesar los datos. En esta parte conocemos el comportamientos de los datos como tal y sobre esto elegimos las herramientas que se adecuen mas al primer acercamiento de la manipulacion de los datos. Las siguentes tres graficas muestran la distribucion de los tipos de ropa que estan en el dataset.Va de lo mas general a lo mas especifico, pues Division name es lo mas general, sigue Departament name que esta contenido en Division name y finalmente Class name que esta contenido en Departament name.

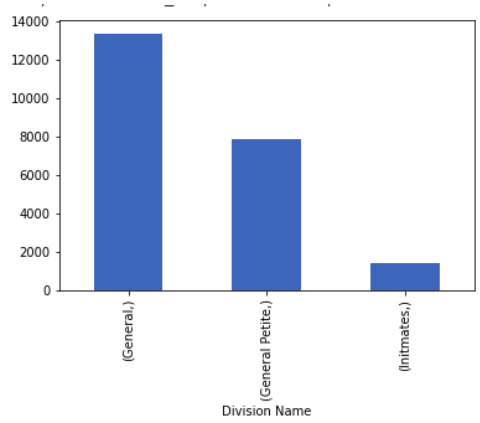


Figure 3: Division Name

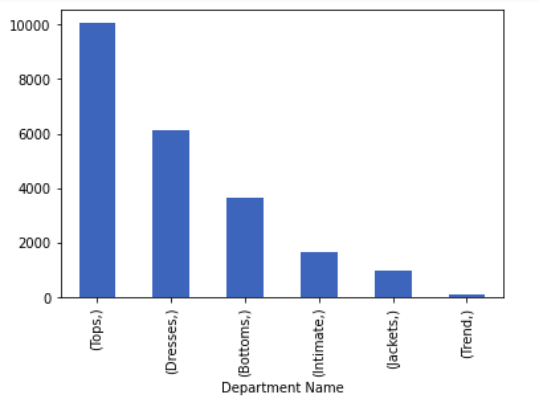


Figure 4: Departament Name

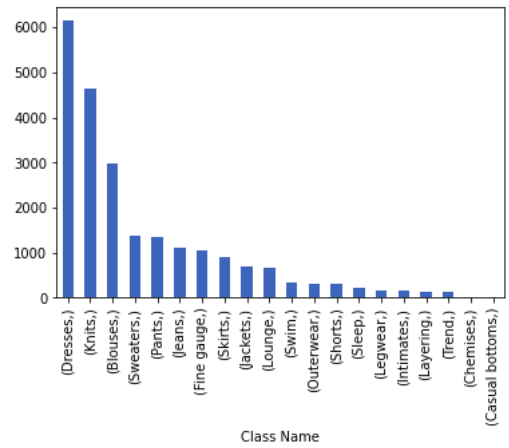


Figure 5: Class name

2.2 Procesamiento

Se usa un método del codo avanzado que usa como parte de su criterio derivadas para medir el cambio experimentado. Se pretende encontrar el primer codo tal que su segunda derivada sea localmente mayor. Dicha derivada indica el grado de cambio de un punto a otro, pudiéndose usar como indicador para encontrar el codo. Y además, se usan los coeficientes silueta para medir qué tan parecidos entre sí son los elementos de un clúster, factor que se quiere maximizar también.

Para maximizar ambos, se ordenan ascendentemente los coeficientes silueta primero y en base a ellos se ordenan las segundas derivadas para encontrar aquella tal que funja de máximo local. En general no se observa mucha distincion entre los clusters usando unicamente informacion de las reseñas. Pareciera ser un cluster unico, y el metodo solo encontro dos clusters que realmente parecen estar muy conectados entre si como se puede ver en la figuras 6 y 7.

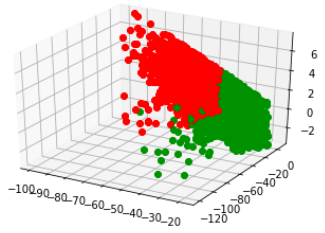


Figure 6: Two clusters projection

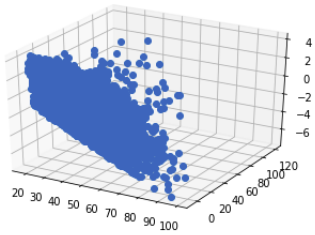


Figure 7: Cluster unique

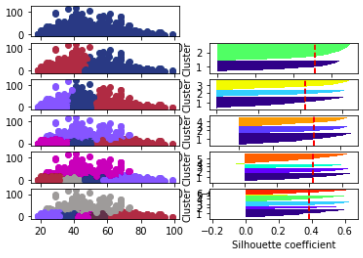


Figure 8: Shilouette Coefficient

El codo optimo encontrado por el algoritmo es 3, como se puede ver la figura 9

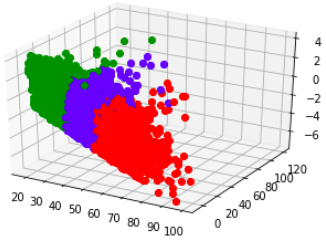


Figure 9: Projection

2.3 Experimentacion y vizualizacion

Ahora se analiza qué palabras suelen acompañarse entre sí, considerando un Lift mayor a 1 para que sea significativa la regla sin que influyan mucho las frecuencias de los elementos. Tras experimentación, parece ser que aquellas reglas de asociación que involucran a más de 2 palabras no suelen ser muy informativas o distinguibles para los conjuntos de calificaciones. Por ello se visualizan aquellas palabras frecuentes por calificacion y su dinamica a lo largo de dichas calificaciones por medio de graficas de barras que van de la figura 10 a la figura 17.

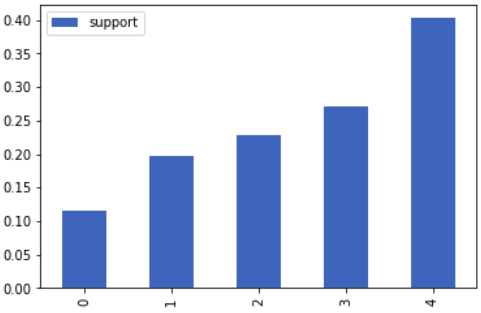


Figure 10: Love

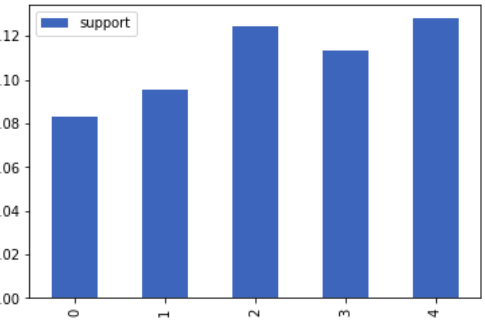


Figure 11: Beautiful

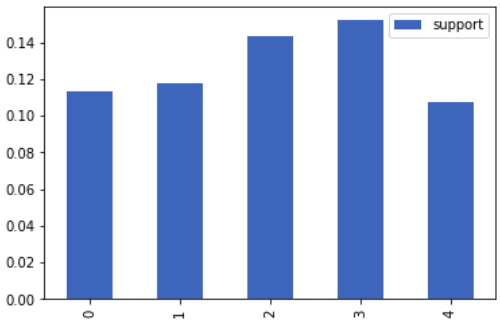


Figure 12: Cute

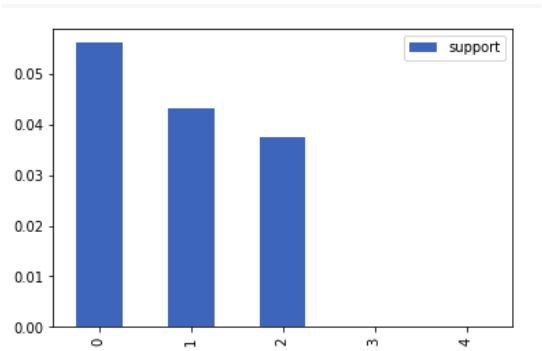


Figure 13: Bad

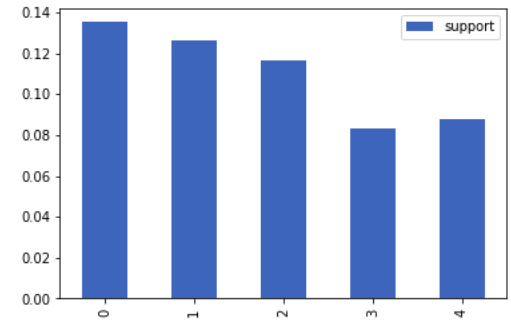


Figure 14: Quality

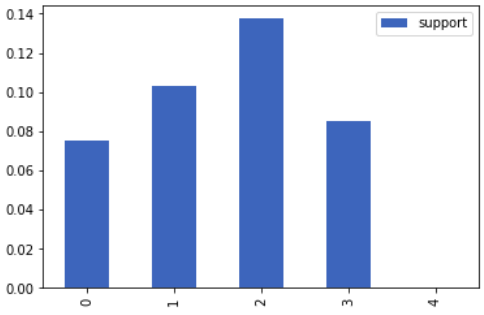


Figure 15: However

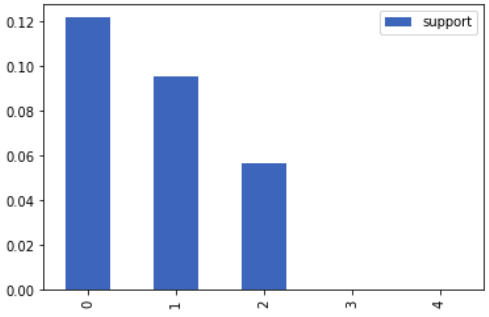


Figure 16: Disappointed

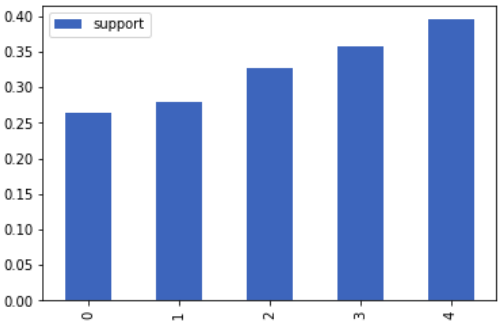
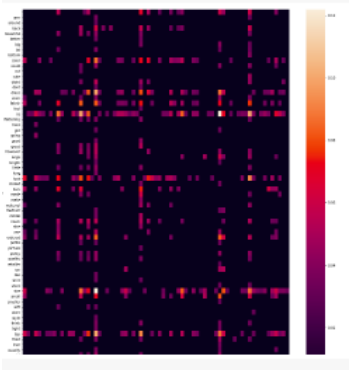


Figure 17: Fit

3 CONCLUSIONS AND FUTURE WORK

Se encontro que la mayoria de las reseñas son positivas. Primero se intento buscar informacion mediante Kmeans, pero no fue lo suficientemente enriquecedor, pues no eran muy distinguibles las reseñas en base a la frecuencia de palabras. Es por ello que se opto implmentar Apriori para descubrir la relacion entre las calificaciones y las palabras usadas. Para la vizualizacion se implemento un mapa de calor para visualizar las distribuciones de palabras para cada calificacion. En el mapa de calor se puede apreciar que las palabras que mas aparecen son aquellas que suelen verse juntas, en especifico aquellas en color más brillante. Lo mas interesante fue descubrir la variacion de la frecuencia de cada palabra dependiendo de la calificacion del articulo. En un trabajo futuro se podria implementar una solucion cuantitativa para medir la relacion que exixte entre los mapas de calor para distintas calificaciones, e incorporar informacion de otras columnas como la retroalimentacion de los usuarios y recomendacion ya que esto solo estuvo enfocado a NLP.

Heapmap



REFERENCES

Nicapotato. 2018. Women's E-Commerce Clothing Reviews. https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-review.