

INFORME

PROYECTO SAD

GRUPO F.A.G.B

Integrantes:

Pablo Leclercq
Aitor Gonzalo
Unai De León
Pablo Martínez
Alain Pedrueza
Adrián Fernández

Sistemas de Apoyo a la Decisión

Ingeniería Informática

22 de julio de 2025

Índice general

1. Tableau: Análisis de los Datos iniciales	8
1.1. Decisiones para generar la historia	9
1.1.1. Preproceso de los datos	9
1.1.2. Generación de los gráficos	12
2. Datos para clasificación: Análisis, Preproceso y Experimentación	41
2.1. Datos	41
2.1.1. División entre Train Dev y Test	41
2.1.2. Distribución de las clases en cada conjunto	41
2.1.3. Descripción del preproceso	41
2.1.4. Primeros resultados	41
2.1.5. Descripción del Proceso de Submuestreo o Sobremuestreo	42
2.2. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados	42
2.2.1. Experimentación: Algoritmos empleados y Breve Descripción	42
2.2.2. Resultados sobre el Development	44
2.2.2.1. Optimizando los resultados de la clase negativa	44
2.2.2.2. Optimizando los resultados de la clase positiva	45
2.2.2.3. Sin optimizar ninguna clase en particular	45
2.2.2.3. Discusión sobre el Sentiment Analysis	45
2.2.2.4. Conclusión sobre el Sentiment Analysis	45
2.2.2.5. Como ejecutar	50
3. Datos para el Topic Modeling: Experimentación	51
3.1. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados	51
3.1.1. Experimentación: Algoritmos empleados y Breve Descripción	52
3.1.2. Resultados	53
3.1.3. Discusión sobre los descubrimientos realizados en la tarea de Topic Modeling	85
3.1.4. Conclusión de la tarea de Topic Modeling	92
4. Bibliografía	93
4.1. Tableau	93
4.2. Sentiment Analysis	93
4.3. Topic Modelling	93

Índice de figuras

1.1.	Division Route	9
1.2.	Preprocesado Aerolineas	10
1.3.	SepararAerolinea	11
1.4.	Combinar Aerolinea	11
1.5.	Overall Rating	12
1.6.	Flight Entertainment	13
1.7.	Food and Beverages	14
1.8.	Value For Money	15
1.9.	Seat Comfort	16
1.10.	Staff Service	17
1.11.	Business Class	18
1.12.	Economy Class	19
1.13.	Business	20
1.14.	Couple Leisure	20
1.15.	Family Leisure	21
1.16.	Solo Leisure	21
1.17.	Valoraciones Generales	22
1.18.	Valoraciones Especificas	23
1.19.	Comparacion Traveller And Class	24
1.20.	Mapa Origen Emirates	25
1.21.	Estimacion Origen Emirates	26
1.22.	Verificadas Origen Emirates	26
1.23.	Mapa Escala Emirates	27
1.24.	Estimacion Escala Emirates	27
1.25.	Verificacion Escala Emirates	27
1.26.	Mapa Destino Emirates	28
1.27.	Estimacion Destino Emirates	29
1.28.	Verificacion Destino Emirates	29
1.29.	Origen Emirates/Qatar	30
1.30.	Destino Emirates/Qatar	30
1.31.	Valoraciones Origen Emirates/Qatar	31
1.32.	Valoraciones Destino Emirates/Qatar	31
1.33.	Recuento valoraciones Reales	32
1.34.	Recuento valoraciones Prediccion	32
1.35.	Medidas Train	33
1.36.	Medidas Test	33
1.37.	Malo Nuestro	34
1.38.	Bueno Nuestro	36
2.1.	ValoresDev	42
2.2.	ValoresTrain Naive Bayes	46
2.3.	ValoresTest Naive Bayes	46
2.4.	ValoresTrain LogisticRegression	47
2.5.	ValoresTest LogisticRegression	47
2.6.	ValoresTrain LinearSVM	48
2.7.	ValoresTest LinearSVM	48
2.8.	ValoresTrain XGBoost	49

2.9. ValoresTest XGBoost	49
3.1. Perplexity y coherence c v en malo nosotros	53
3.2. Perplexity y coherence c v en malo nosotros - Bigramas	53
3.3. Perplexity y coherence c v en malo nosotros - Trigramas	53
3.4. Perplexity y coherence c v en malo nosotros - Unigramas y Bigramas	54
3.5. Perplexity y coherence c v en malo nosotros - Unigramas y Trigramas	54
3.6. Perplexity y coherence c v en malo nosotros - Bigramas y Trigramas	54
3.7. Perplexity y coherence c v en malo nosotros - Unigramas, Bigramas y Trigramas	55
3.8. Perplexity y coherence u-mass en malo nosotros	55
3.9. Perplexity y coherence u-mass en malo nosotros - Bigramas	56
3.10. Perplexity y coherence u-mass en malo nosotros - Trigramas	56
3.11. Perplexity y coherence u-mass en malo nosotros - Unigramas y Bigramas	56
3.12. Perplexity y coherence u-mass en malo nosotros - Unigramas y Trigramas	57
3.13. Perplexity y coherence u-mass en malo nosotros - Bigramas y Trigramas	57
3.14. Perplexity y coherence u-mass en malo nosotros - Unigramas, Bigramas y Trigramas	57
3.15. Perplexity y coherence C-NMPI en malo nosotros - Bigramas y Trigramas	58
3.16. Perplexity y coherence C-UCI en malo nosotros - Bigramas y Trigramas	58
3.17. c-v de 30 a 40 tópicos en malo nosotros	59
3.18. Barrido de 10 a 30 tópicos	59
3.19. Mejor visualización conseguida	60
3.20. Perplexity y coherence c-v en bueno nosotros - Unigramas	61
3.21. Perplexity y coherence c-v en bueno nosotros - Bigramas	61
3.22. Perplexity y coherence c-v en bueno nosotros - Trigramas	62
3.23. Perplexity y coherence c-v en bueno nosotros - Unigramas y Bigramas	62
3.24. Perplexity y coherence c-v en bueno nosotros - Unigramas y Trigramas	62
3.25. Perplexity y coherence c-v en bueno nosotros - Bigramas y Trigramas	63
3.26. Perplexity y coherence c-v en bueno nosotros - Unigramas, Bigramas y Trigramas	63
3.27. Perplexity y coherence c-uci en bueno nosotros - Unigramas	64
3.28. Perplexity y coherence c-uci en bueno nosotros - Bigramas	64
3.29. Perplexity y coherence c-uci en bueno nosotros - Trigramas	64
3.30. Perplexity y coherence c-uci en bueno nosotros - Unigramas y Bigramas	65
3.31. Perplexity y coherence c-uci en bueno nosotros - Unigramas y Trigramas	65
3.32. Perplexity y coherence c-uci en bueno nosotros - Bigramas y Trigramas	65
3.33. Perplexity y coherence c-uci en bueno nosotros - Unigramas, Bigramas y Trigramas	66
3.34. Perplexity y coherence C-NMPI en bueno nosotros - Bigramas y Trigramas	66
3.35. Perplexity y coherence U-MASS en bueno nosotros - Bigramas y Trigramas	67
3.36. c-v de 15 a 30 tópicos en malo nosotros	67
3.37. Mejor visualización conseguida Bueno Nosotros	68
3.38. Perplexity y coherence c v en malo competidor	69
3.39. Perplexity y coherence c v en malo competidor - Bigramas	69
3.40. Perplexity y coherence c v en malo competidor - Trigramas	70
3.41. Perplexity y coherence c v en malo competidor - Unigramas y Bigramas	70
3.42. Perplexity y coherence c v en malo competidor - Unigramas y Trigramas	70
3.43. Perplexity y coherence c v en malo competidor - Bigramas y Trigramas	71
3.44. Perplexity y coherence c v en malo competidor - Unigramas, Bigramas y Trigramas	71
3.45. Perplexity y coherence c-npmi en malo competidor	72
3.46. Perplexity y coherence c-npmi en malo competidor - Bigramas	72
3.47. Perplexity y coherence c-npmi en malo competidor - Trigramas	72
3.48. Perplexity y coherence c-npmi en malo competidor - Unigramas y Bigramas	73
3.49. Perplexity y coherence c-npmi en malo competidor - Unigramas y Trigramas	73
3.50. Perplexity y coherence c-npmi en malo competidor - Bigramas y Trigramas	73
3.51. Perplexity y coherence c-npmi en malo competidor - Unigramas, Bigramas y Trigramas	74
3.52. Perplexity y coherence U-MASS en malo competidor - Bigramas y Trigramas	74
3.53. Perplexity y coherence C-UCI en malo competidor - Bigramas y Trigramas	75
3.54. c-v de 8 a 20 tópicos en malo competidor	75
3.55. Mejor visualización conseguida para malo competidor	76
3.56. Perplexity y coherence c-v en bueno competidor - Unigramas	77

3.57. Perplexity y coherence c-v en bueno competidor - Bigramas	77
3.58. Perplexity y coherence c-v en bueno competidor - Trigramas	78
3.59. Perplexity y coherence c-v en bueno competidor - Unigramas y Bigramas	78
3.60. Perplexity y coherence c-v en bueno competidor - Unigramas y Trigramas	78
3.61. Perplexity y coherence c-v en bueno competidor - Bigramas y Trigramas	79
3.62. Perplexity y coherence c-v en bueno competidor - Unigramas, Bigramas y Trigramas . . .	79
3.63. Perplexity y coherence u-mass en bueno competidor	80
3.64. Perplexity y coherence u-mass en bueno competidor - Bigramas	80
3.65. Perplexity y coherence u-mass en bueno competidor - Trigramas	80
3.66. Perplexity y coherence u-mass en bueno competidor - Unigramas y Bigramas	81
3.67. Perplexity y coherence u-mass en bueno competidor - Unigramas y Trigramas	81
3.68. Perplexity y coherence u-mass en bueno competidor - Bigramas y Trigramas	81
3.69. Perplexity y coherence u-mass en bueno competidor - Unigramas, Bigramas y Trigramas .	82
3.70. Perplexity y coherence C-NPMI en bueno competidor - Bigramas y Trigramas	82
3.71. Perplexity y coherence C-UCI en bueno competidor - Bigramas y Trigramas	83
3.72. c-v de 35 a 45 tópicos en bueno competidor	83
3.73. Mejor visualización conseguida para bueno competidor	84
3.74. Mejor visualización conseguida separada por subtemas malo nosotros	85
3.75. Mejor visualización conseguida separada por subtemas bueno nosotros	87
3.76. Mejor visualización conseguida separada por subtemas malo competidor	89
3.77. Mejor visualización conseguida separada por subtemas bueno nosotros	91

Índice de cuadros

2.1. División Train, Dev y Test	41
2.2. Distribución Train, Dev y Test	41
2.3. Resultados Clase Negativa	44
2.4. Resultados Clase Positiva	45
2.5. Resultados generales	45

Acrónimos

- **LR**: Logistic Regression
- **XGB**: XGBoost
- **MNB**: Multinomial Naive Bayes
- **BoW**: Bag of Words
- **Tf-Idf**: Term frequency – Inverse document frequency
- **LDA**: Latent Dirichlet Allocation
- **NMF**: Non-negative Matrix Factorization

1. Tableau: Análisis de los Datos iniciales

Para comenzar con la parte de Tableau, comenzaremos respondiendo las preguntas iniciales para ponernos en contexto y posteriormente mostraremos los gráficos elegidos, así como el comienzo de la historia.

Preguntas Iniciales

- ¿Cuál es mi empresa y cuál mi competidor?
 - Nuestra empresa es Emirates y nuestro principal competidor es Qatar Airways.
- ¿Obtengo mejores o peores valoraciones que mi competidor?
 - En general, tenemos peores resultados que nuestro competidor a lo largo de los últimos años en todos los aspectos valorados de nuestro servicio.
- ¿Hay algún destino, escala u origen en el que las opiniones son más relevantes?
 - Las opiniones mas relevantes son en Dubai ya que es la central de Emirates.
- ¿Hay diferencias relevantes con respecto a la fecha?
 - Hasta la cuarentena, nuestros servicios estaban más igualados con los de Qatar Airways, sin embargo, a partir de 2020 Qatar Airways se mantiene y nuestros servicios empiezan a decaer, como veremos en la historia.
- ¿Por qué hemos elegido esos gráficos?
 - Hemos implementado gráficos de barras y lineares para la representación de la comparación. Además hemos implementado mapas para mostrar los aeropuertos más conflictivos y valorar posibles soluciones.

1.1. Decisiones para generar la historia

1.1.1. Preproceso de los datos

Inicialmente los líderes de clasificación nos administraron un CSV con una columna rating” determinando mediante los rangos establecidos, las valoraciones positivas, negativas y neutras.

Una vez esto realizamos los líderes de Tableau realizamos un preprocesado de dicho CSV para dividir la columna Route en 3 columnas: Origen, Destino y Escala y así poder representar mediante mapas en Tableau los aeropuertos conflictivos. Para ello, realizamos un script de python (DivisionRoute.py):

```
import pandas as pd
# usage
def split_route(route):
    parts = route.split(' via ')
    origin_dest = parts[0].split(' to ')
    origin = origin_dest[0].strip() if origin_dest else None
    destination = origin_dest[1].strip() if len(origin_dest) > 1 else None
    stopover = parts[1].strip() if len(parts) > 1 else None
    return origin, destination, stopover

df = pd.read_csv('Modified_AirlinesReviews.csv')

# Obtenemos la posición de la columna 'Route'
route_index = df.columns.get_loc('Route')

# Aplicamos la función para dividir las rutas
new_columns = df['Route'].apply(lambda x: pd.Series(split_route(x), index=['Origen', 'Destino', 'Escala']))

# Insertamos cada una de las nuevas columnas en la posición original de 'Route'
for col in reversed(new_columns.columns):
    df.insert(route_index, col, new_columns[col])

# Eliminamos la columna 'Route' original
df.drop( labels='Route', axis=1, inplace=True)

print(df)

df.to_csv( path_or_buf= 'aerolineas routes split.csv'. index=False)
```

1.1. Figura: Division Route

Una vez realizada la separación, procedimos a realizar un preprocesado de las 3 columnas, generando un diccionario "Siglas IATA.txt". Esto se debe a que la columna Route inicial estaba compuesta por ciudades, países y códigos IATA de los aeropuertos, por este motivo mediante el preprocesado decidimos transformar todas las ciudades y países a los códigos IATA de los aeropuertos más comunes de la ciudad o país. Cabe resaltar, que si en la reseña se especificaba concretamente el aeropuerto se tendrá en cuenta ese. Esto lo realizamos con el siguiente script (PreprocesadoAerolineas.py):

```
import pandas as pd
1 usage
def cargar_abreviaturas(archivo_txt):
    abreviaturas = {}
    with open(archivo_txt, 'r') as file:
        for line in file:
            parts = line.strip().split(': ')
            if len(parts) == 2:
                codigo, ciudad = parts
                abreviaturas[ciudad.strip()] = codigo.strip()
    return abreviaturas

1 usage
def sustituir_ciudades(archivo_csv, abreviaturas, archivo_salida):
    df = pd.read_csv(archivo_csv)
    for columna in ['Origen', 'Destino', 'Escala']:
        if columna in df.columns:
            df[columna] = df[columna].apply(lambda x: abreviaturas.get(x, x))
    df.to_csv(archivo_salida, index=False)

# Carga las abreviaturas desde el archivo de texto
abreviaturas = cargar_abreviaturas('Siglas_IATA.txt')

# Sustituir las ciudades en el archivo CSV y guardar los resultados
sustituir_ciudades(archivo_csv: 'aerolineas_routes_split.csv', abreviaturas, archivo_salida: 'Aerolineas_Preprocesadas.csv')
```

1.2. Figura: Preprocesado Aerolineas

Posteriormente, se dividió el CSV modificado utilizando el script SepararAerolinea.py, resultando en dos archivos: uno para Emirates y otro para Qatar. Esto permitió realizar un análisis detallado y específico para cada aerolínea:

```
import pandas as pd

# usage
def filtrar_por_aerolinea(archivo_csv, aerolinea, columna_aerolinea):
    # Leer el archivo CSV
    df = pd.read_csv(archivo_csv)

    # Filtrar las filas que tienen el valor especificado en la columna de la aerolinea
    df_filtrado = df[df[columna_aerolinea] == aerolinea]

    # Guardar el DataFrame filtrado en un nuevo archivo CSV
    df_filtrado.to_csv("filas_filtradas_qatar.csv", index=False)

    print("Se han guardado las filas filtradas en 'filas_filtradas_qatar.csv'")

archivo_csv = 'updated_routes_split.csv'
aerolinea = 'Qatar Airways' #Emirates
columna_aerolinea = 'Airline'

filtrar_por_aerolinea(archivo_csv, aerolinea, columna_aerolinea)
```

1.3. Figura: SepararAerolinea

Además, se desarrolló el script CombinarAerolinea.py para fusionar las instancias de Emirates y Qatar. Este proceso posibilita una comparación directa entre ambas aerolíneas en Tableau:

```
import pandas as pd

# Rutas de los archivos CSV
archivo_csv1 = 'qatar.csv'
archivo_csv2 = 'emirates.csv'

# Leer los archivos CSV
df1 = pd.read_csv(archivo_csv1)
df2 = pd.read_csv(archivo_csv2)

# Concatenar los dos DataFrames
df_combinado = pd.concat( objs: [df1, df2], ignore_index=True)

# Guardar el DataFrame combinado en un nuevo archivo CSV
df_combinado.to_csv( path_or_buf: 'qataremirates.csv', index=False)

print("Los archivos han sido combinados y guardados.")
```

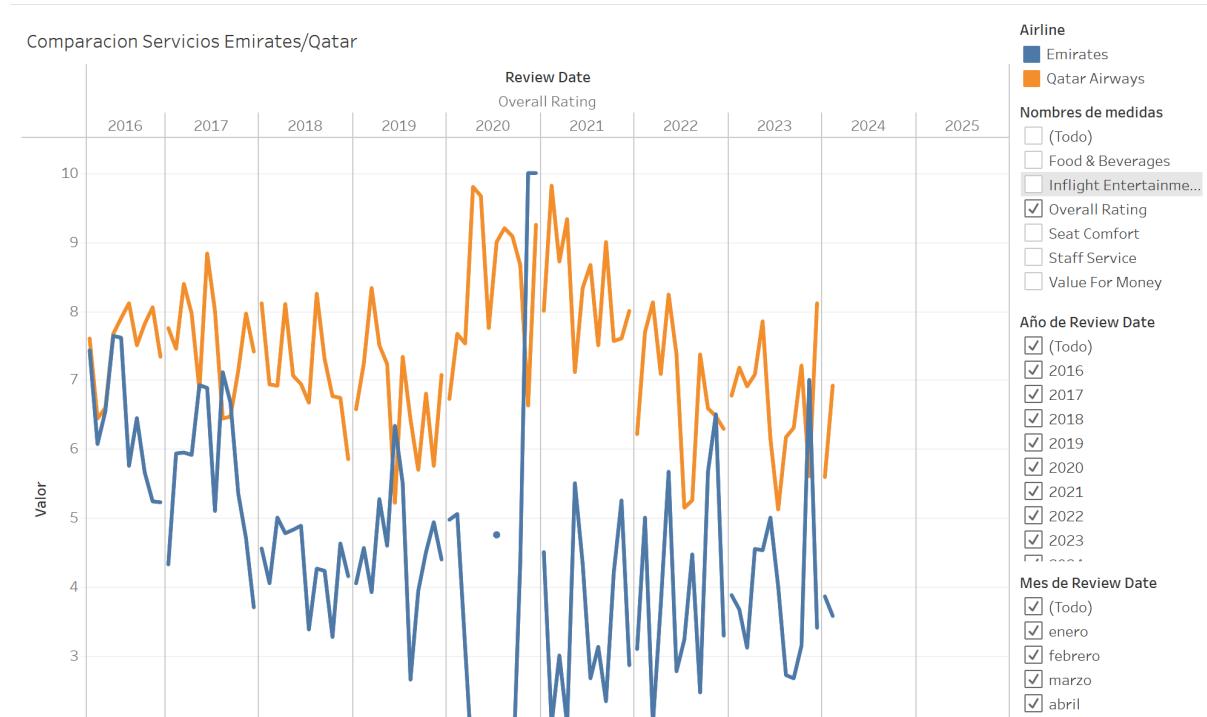
1.4. Figura: Combinar Aerolinea

Es importante destacar que todos estos procedimientos se realizaron utilizando el paquete pandas, el cual es fundamental para el manejo de datos en forma de dataframes durante el preprocesado. Todos los scripts mencionados se encuentran disponibles en la carpeta "Tableau" de la entrega.

1.1.2. Generación de los gráficos

A continuación mostraremos todos los gráficos realizados y implementados en la historia, así como una explicación de cada uno y la importancia de los mismos.

Primero hemos realizado un gráfico dinámico (clickable), en el que podemos observar la comparación de nuestros servicios con los de Qatar, además podemos también seleccionar el año específico, el mes y distinguir entre valoraciones verificadas y no verificadas.



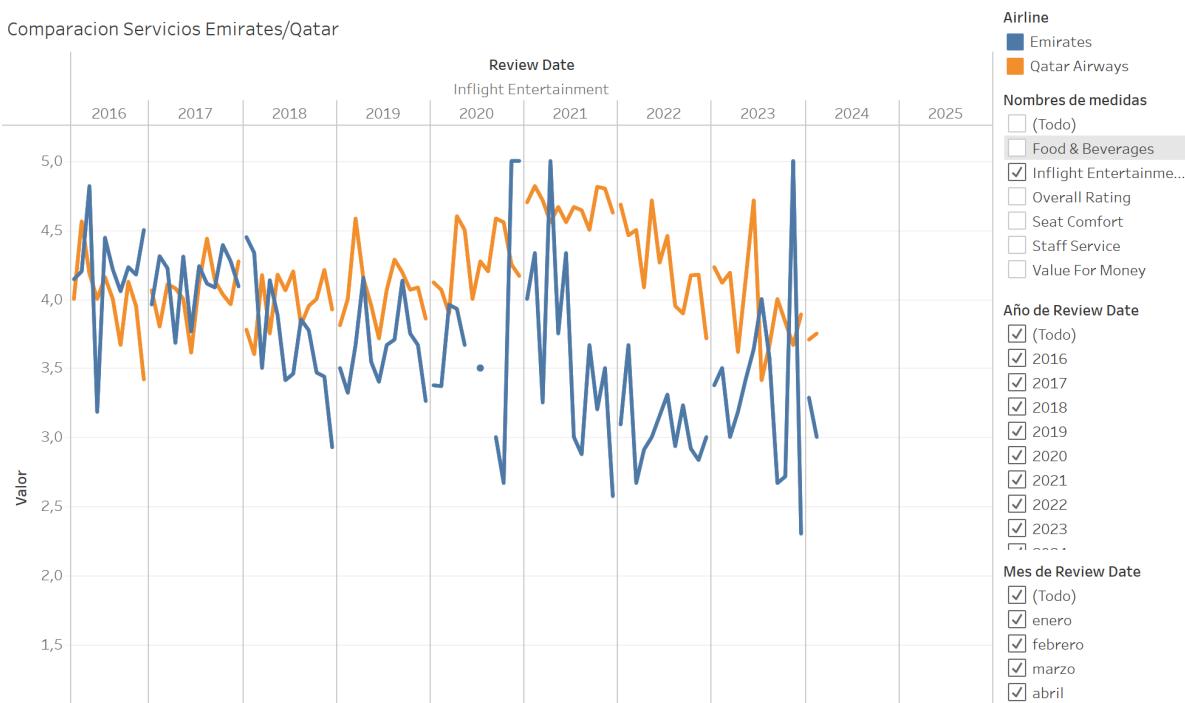
1.5. Figura: Overall Rating

El gráfico muestra la evolución de las valoraciones generales de las aerolíneas Emirates y Qatar Airways desde 2016 hasta 2025. Ambas aerolíneas exhiben fluctuaciones significativas a lo largo del tiempo, lo que indica cambios en la percepción de los servicios por parte de los usuarios.

Inicialmente, en 2016, Qatar Airways comienza con una valoración superior a la de Emirates, sugiriendo un mejor rendimiento en ese momento. A lo largo de los años, ambos muestran una tendencia general de mejora, aunque con bajas notorias en torno a 2020, probablemente debido a la crisis de COVID-19 que afectó globalmente a la industria aérea.

Después de 2020, las valoraciones se recuperan, pero la línea de Emirates muestra más inestabilidad, con picos y caídas más pronunciadas que Qatar Airways. Hacia el final del período, la gráfica muestra que Qatar Airways tiende a tener valoraciones más altas que Emirates, lo que podría indicar una gestión más efectiva de la recuperación pos-pandemia o una mejora constante en sus servicios comparados con Emirates.

Este gráfico es útil para analizar no solo el desempeño individual de cada aerolínea sino también para comparar directamente cómo cada una responde a desafíos externos y evoluciona en términos de calidad de servicio y satisfacción del cliente a lo largo del tiempo.



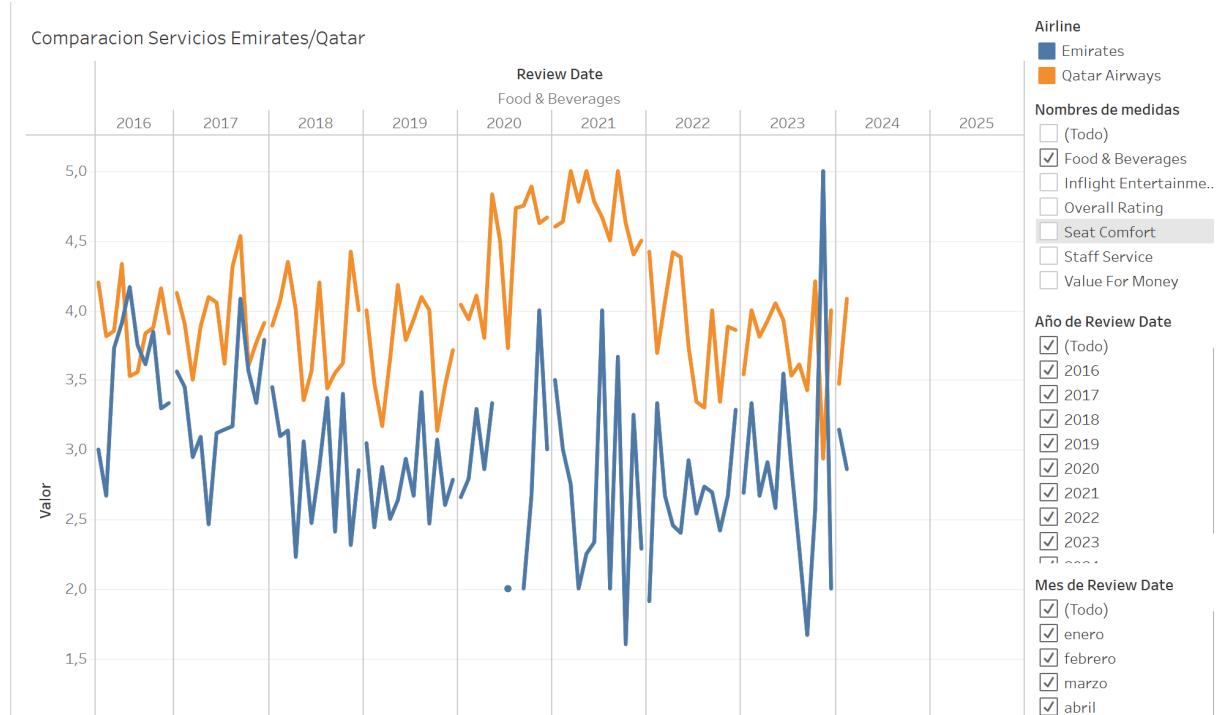
1.6. Figura: Flight Entertainment

En el gráfico para el entretenimiento a bordo desde 2016 hasta 2025, vemos una comparativa detallada entre Emirates y Qatar Airways.

Este gráfico muestra la valoración del entretenimiento en vuelo para Emirates y Qatar Airways. Las calificaciones fluctúan entre 3.0 y cerca de 5.0 para ambas aerolíneas. Inicialmente, las calificaciones para Emirates son ligeramente inferiores a las de Qatar Airways, indicando una preferencia inicial por el entretenimiento ofrecido por QATAR.

A lo largo del tiempo, las dos líneas demuestran una competencia cercana, con múltiples cruces donde una aerolínea supera a la otra en términos de valoración. Esto sugiere que ambos han estado mejorando y ajustando su oferta de entretenimiento para competir más efectivamente.

Desde aproximadamente 2020, se observa una disminución general en la valoración del entretenimiento en vuelo para Emirates, mientras que Qatar mantiene una consistencia relativa. Aunque en 2023, Emirates supera a Qatar lo que se corresponde ya que en este año Emirates consiguió el premio a mejor entretenimiento a bordo.



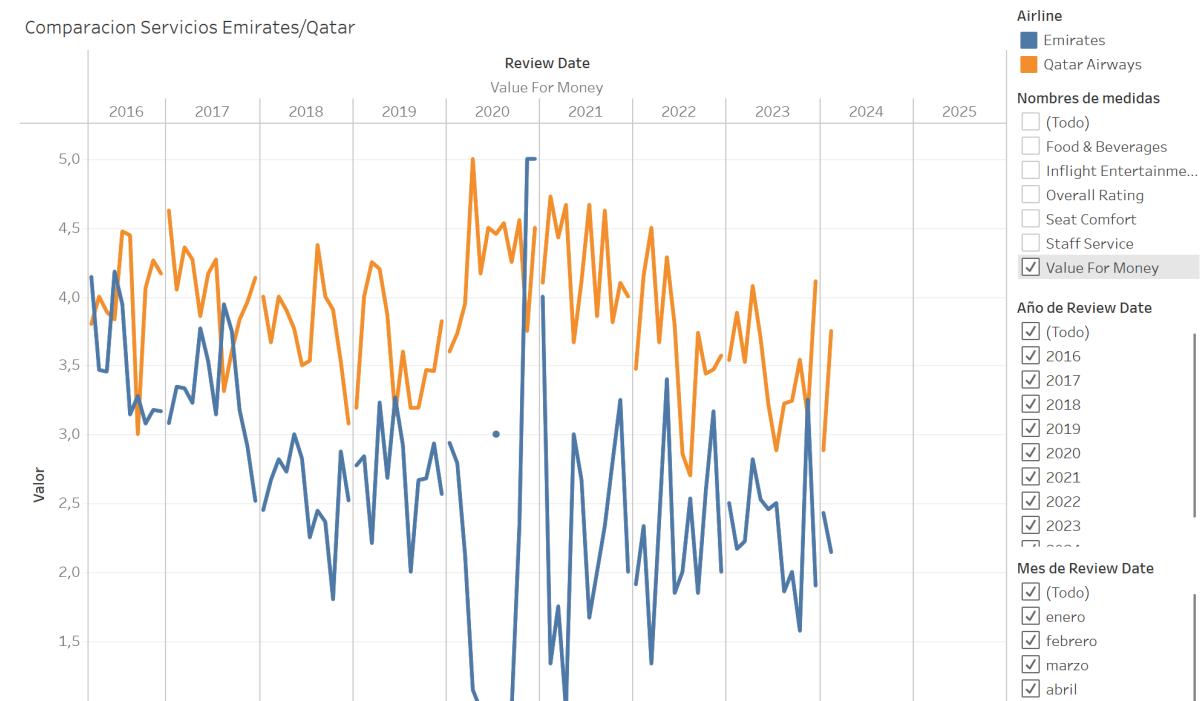
1.7. Figura: Food and Beverages

Este gráfico ofrece una visión clara de cómo los pasajeros han valorado la oferta de comida y bebidas de ambas aerolíneas a lo largo del tiempo. Emirates, muestra una variabilidad significativa en las valoraciones, iniciando en un valor cercano a 3.0 en 2016 y experimentando fluctuaciones constantes, alcanzando picos y valles a lo largo de los años.

Por otro lado, Qatar Airways, inicia también cerca de 3.5 en 2016 y tiende a mantenerse por encima de Emirates en la mayor parte del período observado. Las valoraciones de Qatar Airways exhiben menos variabilidad que las de Emirates, sugiriendo una posible consistencia mayor en la calidad de su comida y bebidas.

Emirates experimenta una caída en las valoraciones alrededor de 2020, lo que puede estar relacionado con los desafíos operacionales durante la pandemia de COVID-19, mientras que Qatar consigue mejorar durante la pandemia. Posteriormente, ambas líneas muestran un acercamiento, aunque la de Qatar Airways se mantiene por encima.

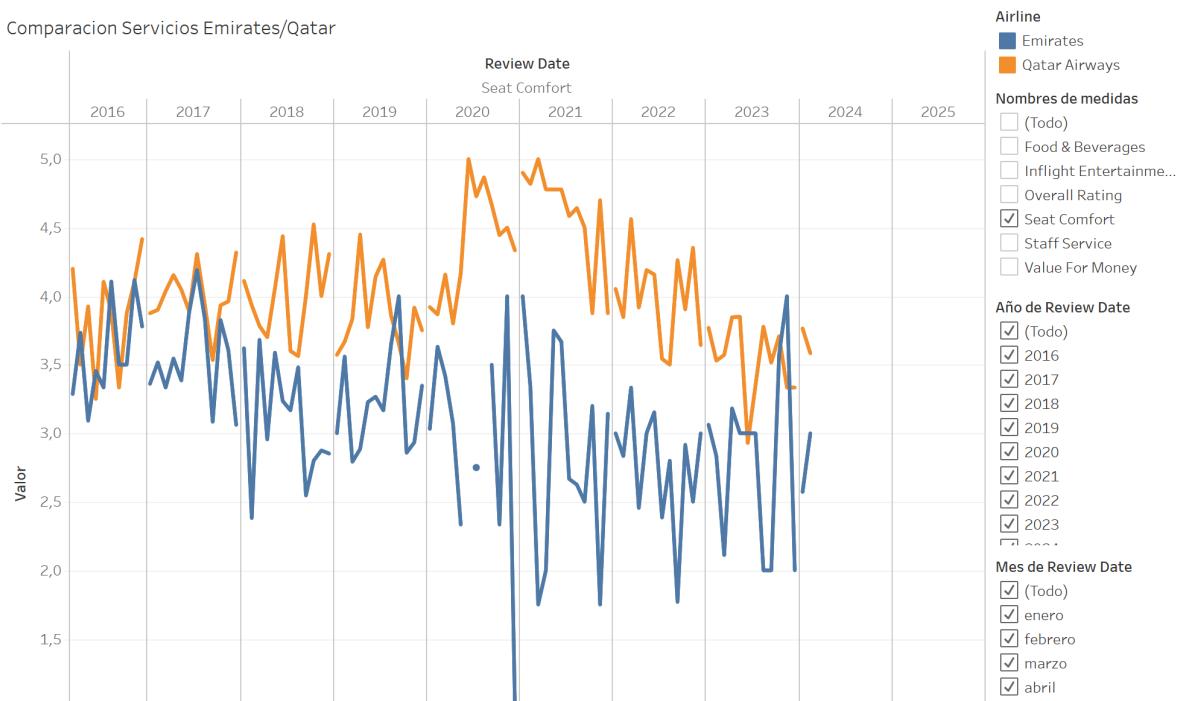
Hacia el final del gráfico, en los años 2023 a 2025, la tendencia de Qatar Airways muestra un declive en la valoración



1.8. Figura: Value For Money

Desde 2016 hasta 2025, las valoraciones sobre la relación calidad-precio de Emirates y Qatar Airways muestran una notable variabilidad, reflejando la percepción cambiante de los consumidores sobre lo que reciben a cambio de su dinero. Inicialmente, Qatar Airways comienza con una ventaja sobre Emirates, sugiriendo una mejor percepción de valor por parte de los pasajeros. A lo largo del tiempo, ambas aerolíneas experimentan fluctuaciones en las valoraciones, pero Qatar generalmente mantiene un ligero liderazgo.

La pandemia de COVID-19 en 2020 afecta negativamente la percepción del valor en ambas aerolíneas pero de manera mucho más notable en Emirates, aunque la recuperación muestra mejoras con el tiempo. Hacia el final del gráfico, ambas aerolíneas enfrentan desafíos para mantener altas valoraciones de relación calidad-precio, posiblemente debido a un mercado más competitivo y cambios en las expectativas de los consumidores. Este análisis ayuda a identificar áreas de mejora potencial en términos de ajuste de precios y mejora de servicios.

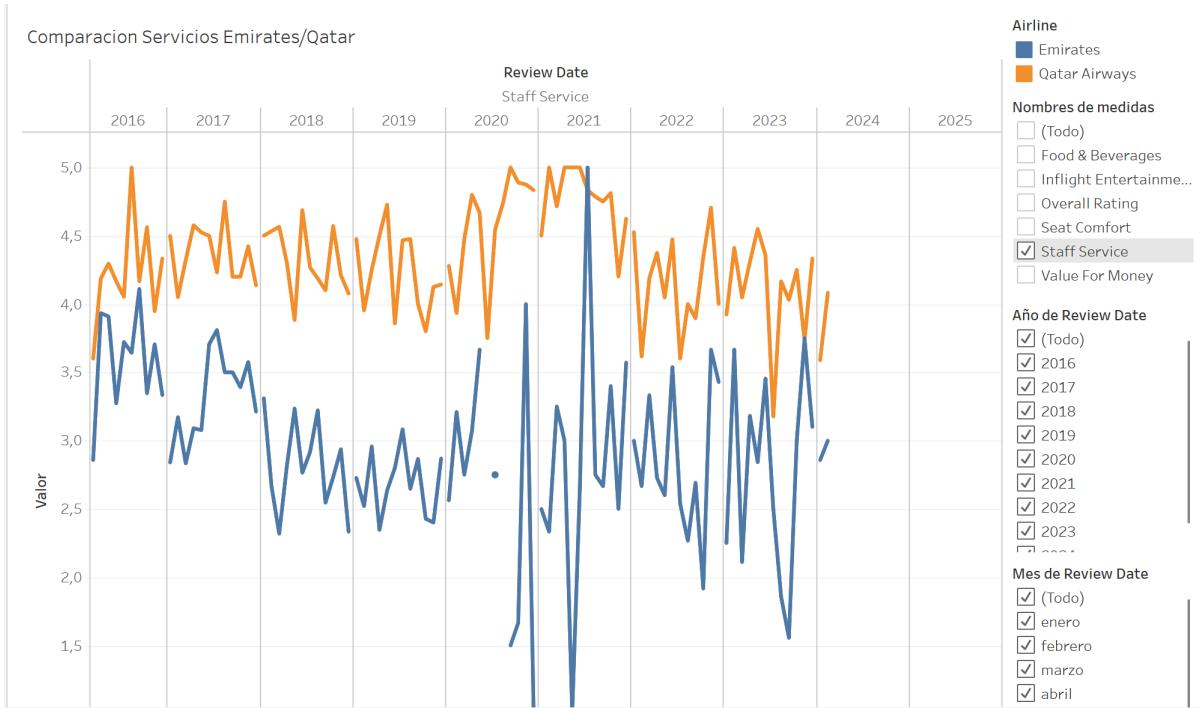


1.9. Figura: Seat Comfort

Desde 2016, Qatar Airways ha mantenido generalmente una valoración superior en comodidad de asientos comparada con Emirates, iniciando con una ventaja y manteniéndola a lo largo de varios períodos. Ambas aerolíneas muestran fluctuaciones significativas en las valoraciones, indicando variaciones en la percepción del confort por parte de los pasajeros.

Durante 2020, coincidiendo con la pandemia, ambas líneas registran una caída en las valoraciones, lo que refleja posiblemente un impacto negativo en la percepción del servicio. Posteriormente, ambas aerolíneas experimentan una recuperación, aunque Qatar Airways frecuentemente supera a Emirates, destacando una posible mayor consistencia en la comodidad de sus asientos.

Hacia los años más recientes (2023-2025), las valoraciones de ambas aerolíneas comienzan a converger, sugiriendo esfuerzos por parte de Emirates para mejorar en esta área y reducir la brecha con Qatar Airways. Este gráfico resalta la importancia de la comodidad del asiento como un componente vital de la experiencia del pasajero en vuelo.



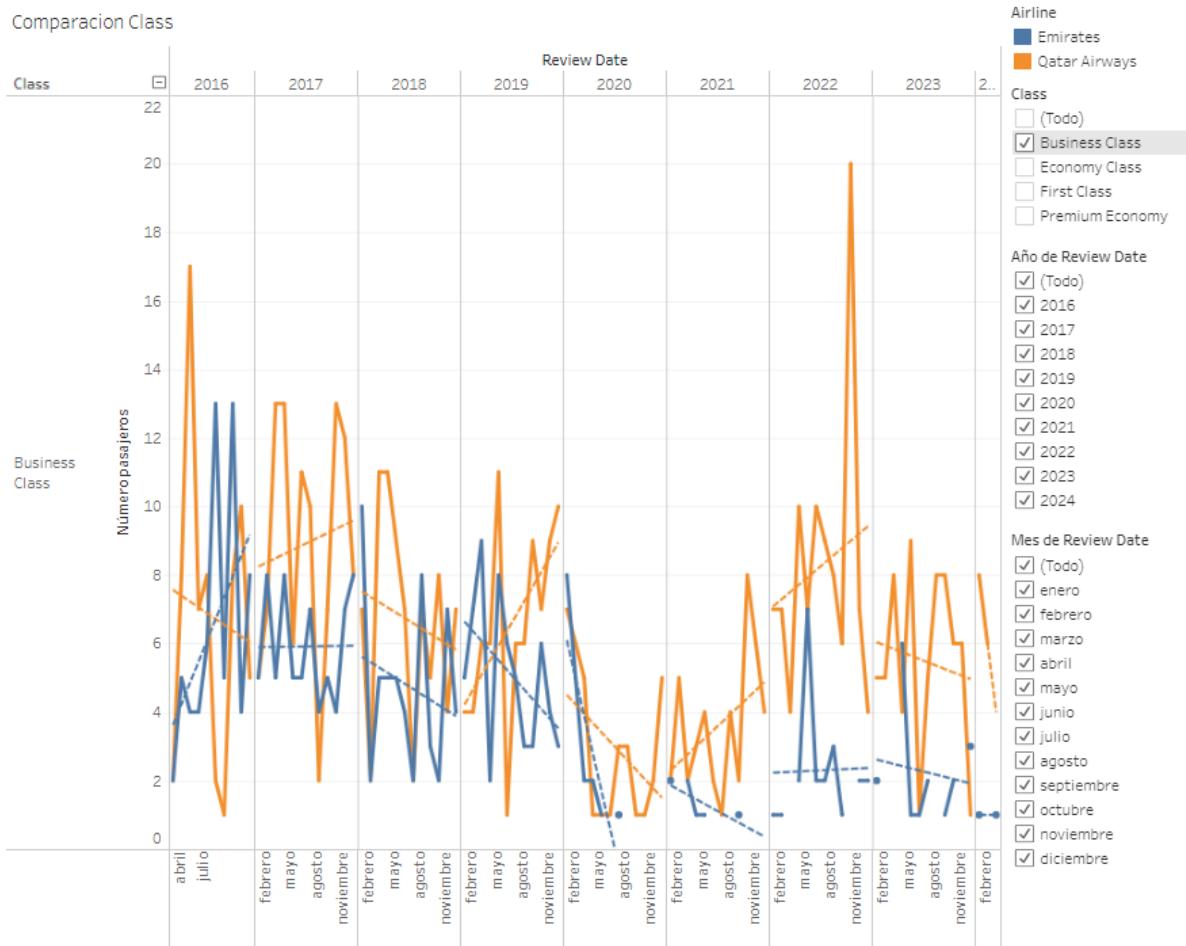
1.10. Figura: Staff Service

En 2016, Qatar Airways ha mantenido consistentemente valoraciones superiores en servicio al cliente en comparación con Emirates, lo que sugiere una mayor satisfacción general de los pasajeros con el personal de Qatar. Qatar muestra menos fluctuación y mantiene un nivel de valoración generalmente alto, por encima de 4,0, destacando un compromiso constante con la calidad en el servicio al cliente.

Emirates, muestra más variabilidad y períodos donde sus valoraciones caen significativamente por debajo de las de Qatar Airways, especialmente notable alrededor de 2020 durante la pandemia. Sin embargo, después de 2021, Emirates muestra una mejora notable, cerrando la brecha con Qatar Airways hacia 2023.

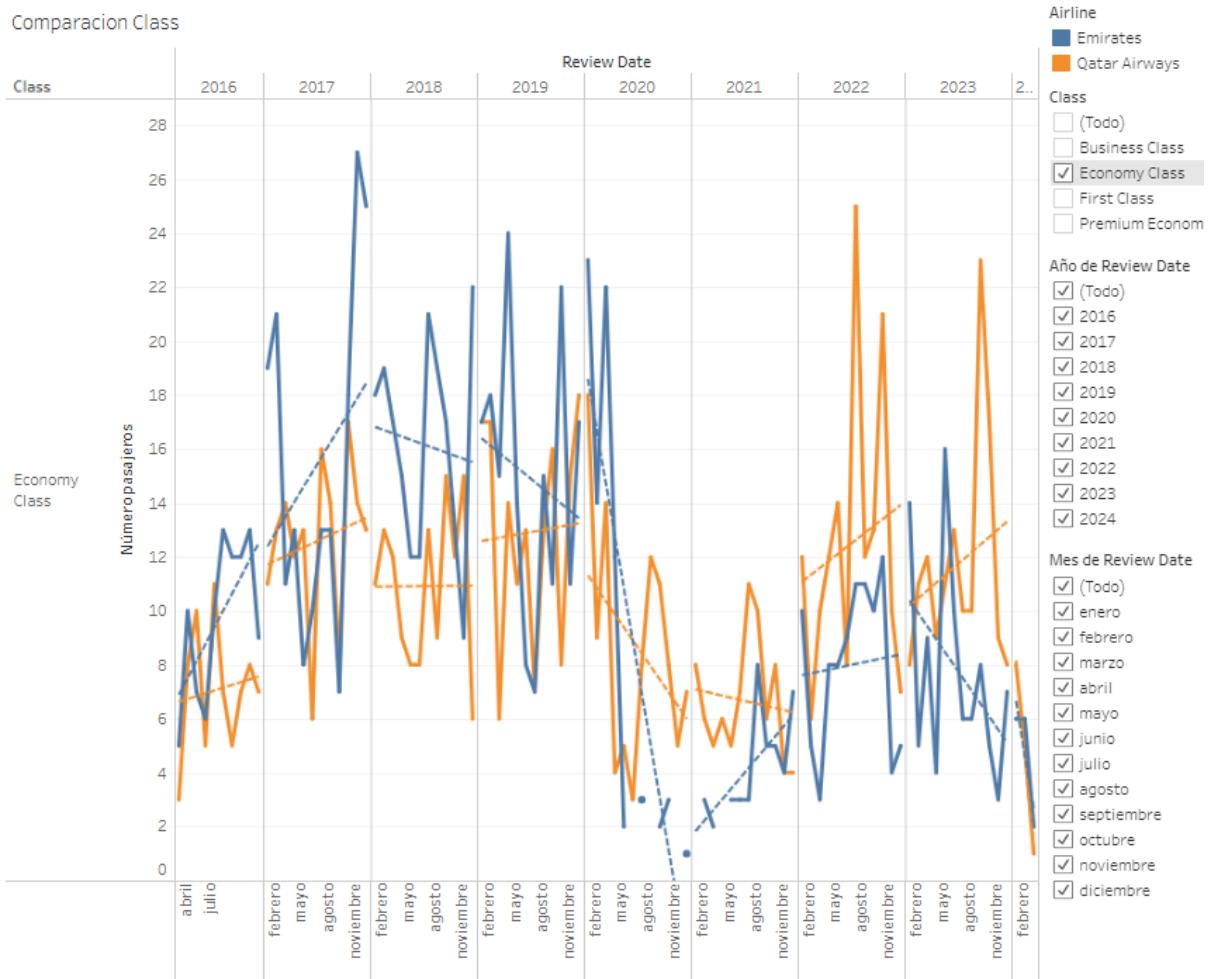
Ambas aerolíneas enfrentan descensos y ascensos en las valoraciones, pero la tendencia sugiere que Qatar Airways ha logrado mantener una percepción más estable y positiva del servicio al cliente. Este gráfico refleja la importancia de mantener un servicio al cliente de alta calidad y cómo esto impacta la percepción y satisfacción del pasajero en el tiempo.

A continuación hemos hecho una representación de los distintos tipos de billetes que viajan con nuestra compañía, el cual nos ayudara a determinar cuales son los viajeros mas comunes y para determinar en que sector debemos enfocar nuestro principal esfuerzo de mejora. Tambien hemos hecho un gráfico dinamico (clickable) que nos permite ver los distintos tipos de billetes y elegir el año que nos convenga. Este gráfico nos permite sacar conclusiones sobre el tipo de viajeros de nuestra aerolínea.



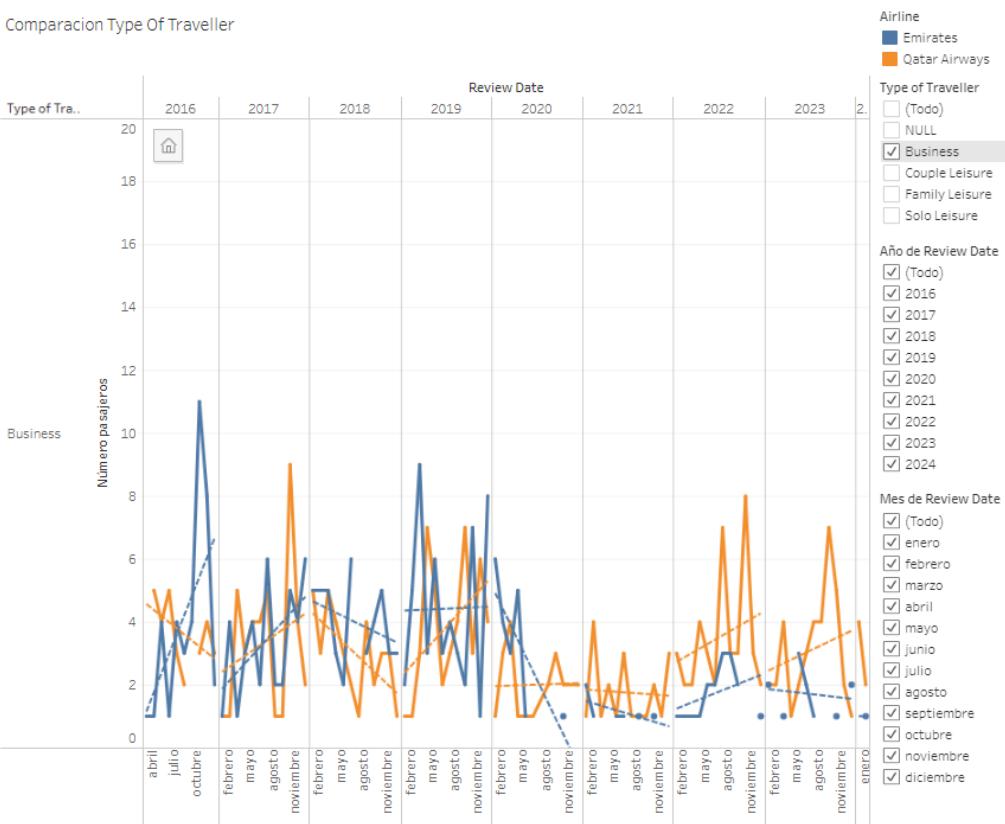
1.11. Figura: Business Class

Solo vamos a mostrar los pasajeros que viajan en Business y en Economy Class ya que pese a que las clases First Class y Premium Economy son muy importantes no contamos con las suficientes valoraciones realizadas por dichos pasajeros como para realizar un estudio en dichas clases.

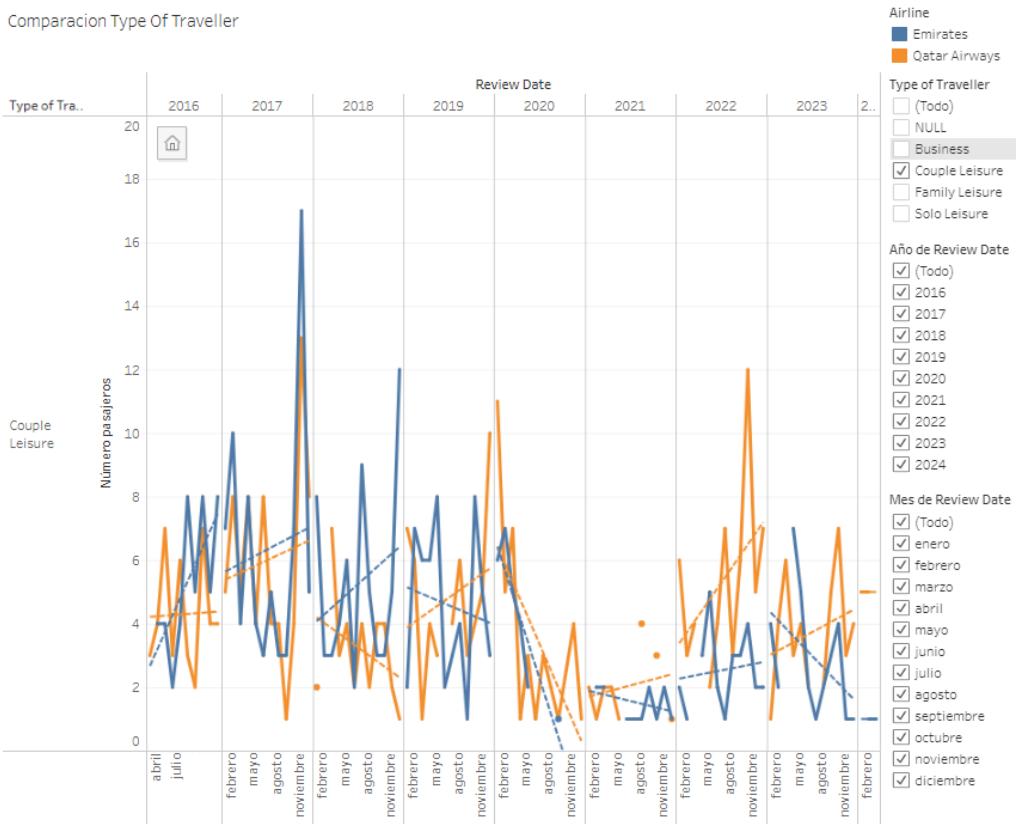


1.12. Figura: Economy Class

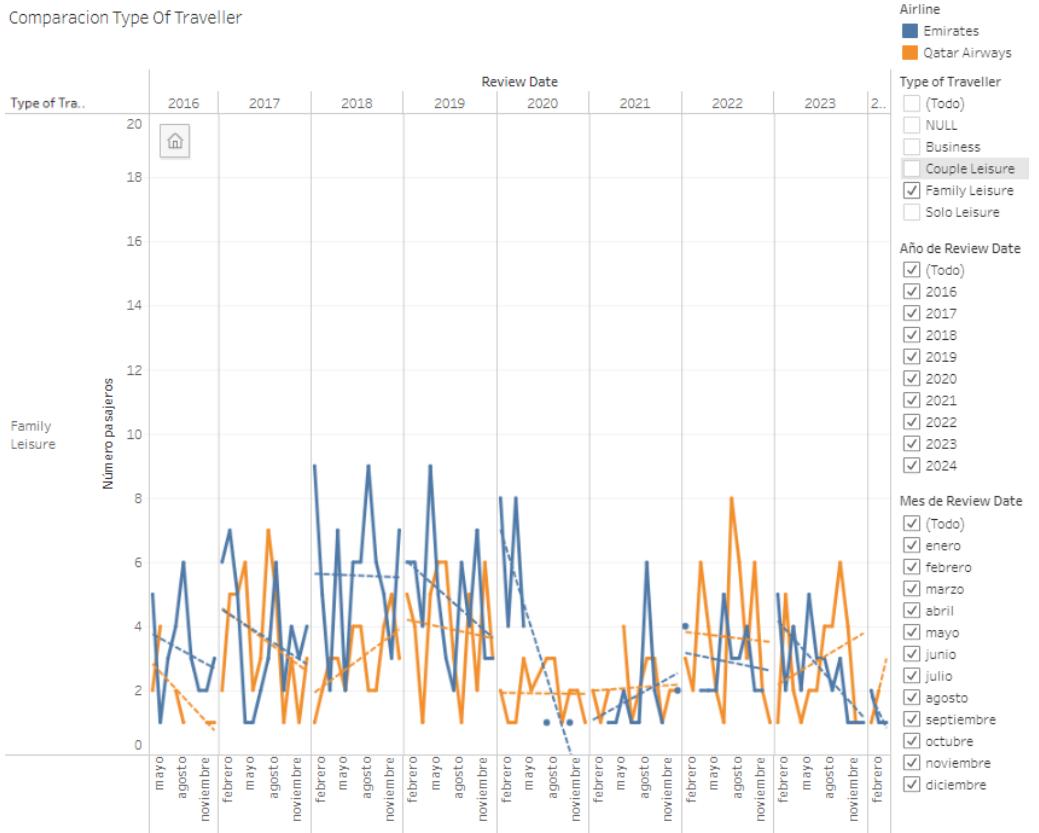
Algo parecido hemos hecho con los diferentes tipos de viajeros, lo hemos hecho con el mismo propósito que el de los billetes más vendidos y de ahí que este representado prácticamente igual.



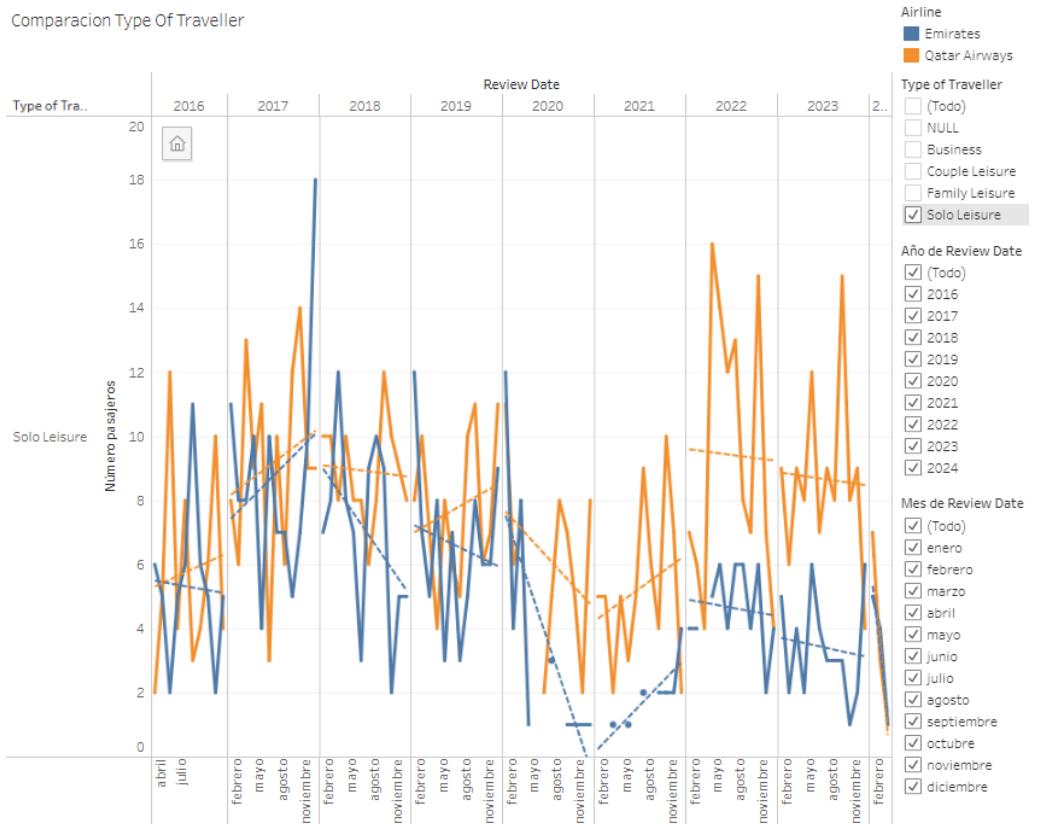
1.13. Figura: Business



1.14. Figura: Couple Leisure

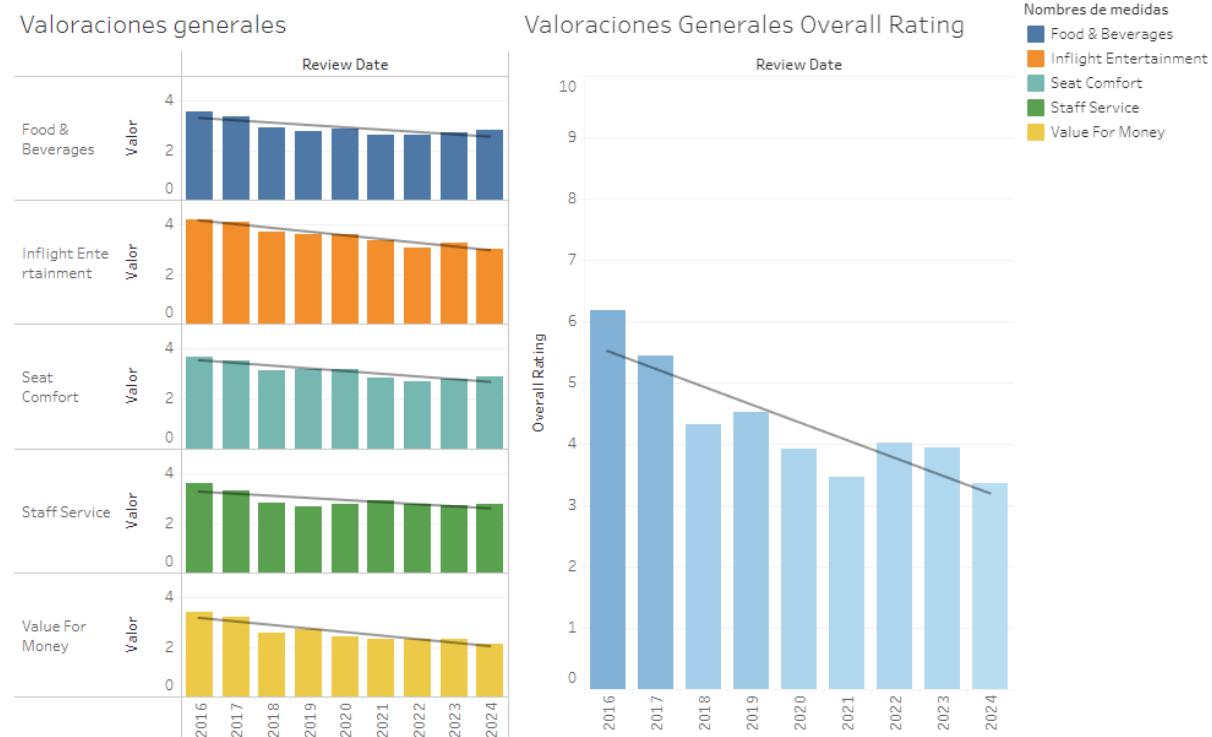


1.15. Figura: Family Leisure



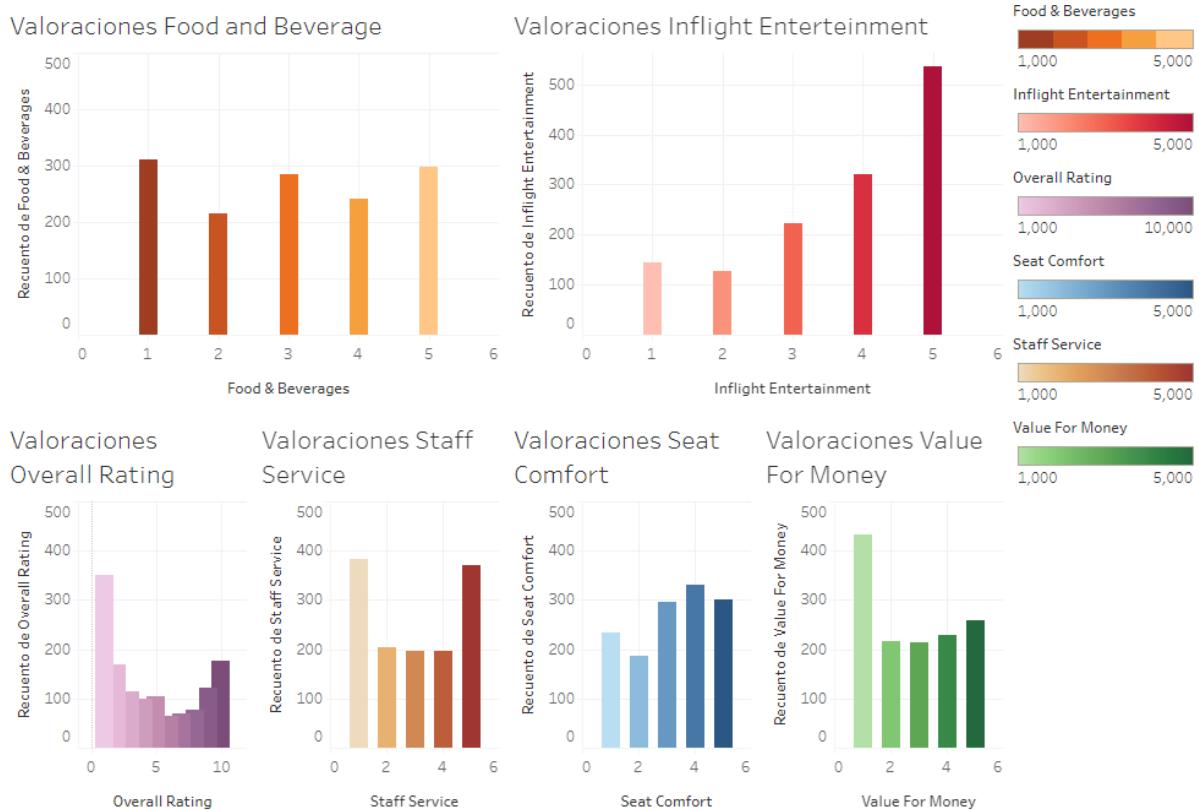
1.16. Figura: Solo Leisure

El gráfico proporciona una representación visual de las valoraciones y calificaciones generales para Emirates desde 2016 hasta 2024 en distintas categorías de servicio. La calificación global muestra una tendencia descendente significativa, lo que indica una disminución general en la satisfacción del cliente a lo largo del tiempo. Las categorías específicas, como comida y bebidas, entretenimiento en vuelo y confort de los asientos, exhiben una estabilidad relativa, sin grandes cambios en la percepción de los usuarios. Sin embargo, la categoría de valor por el dinero muestra una caída notable, particularmente desde 2021, sugiriendo que los pasajeros perciben que el servicio ofrecido no justifica el costo. Esto podría ser una de las causas principales de la disminución en la calificación global, afectando la imagen y posiblemente las decisiones de elección de los consumidores hacia Emirates.



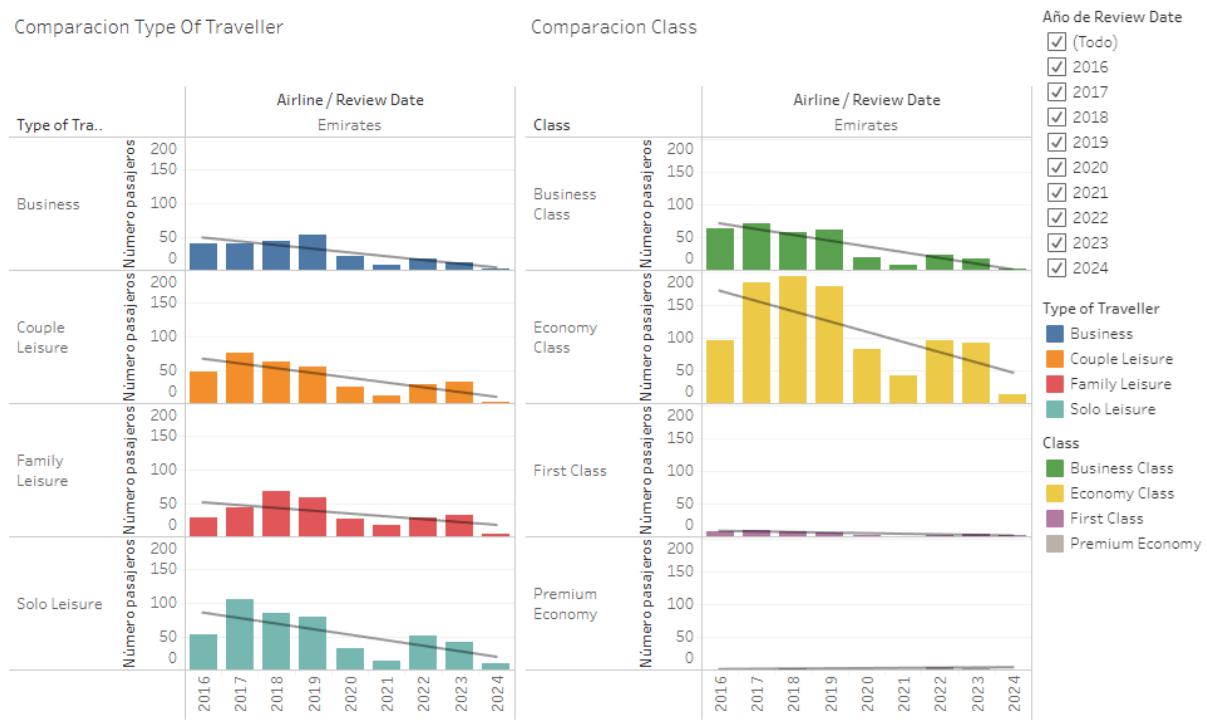
1.17. Figura: Valoraciones Generales

El gráfico detallado presenta un análisis específico de las valoraciones de los servicios de Emirates por categoría, mostrando el número de pasajeros que otorgan calificaciones de 1 a 5 para cada servicio ofrecido. En "Food and Beverages" e "Inflight Entertainment", las valoraciones tienden a ser positivas, con una concentración en los rangos medio-altos, destacando la satisfacción general en estas áreas. En "Overall Rating" y "Staff Service" se revela una distribución variada, lo que indica opiniones mixtas sobre la experiencia general de vuelo. En contraste, "Seat Comfort" muestra una mayor variabilidad en las calificaciones, mientras que "Value for Money" tiene una gran cantidad de negativas, sugiriendo áreas donde Emirates podría enfocar mejoras para aumentar la satisfacción del cliente y su percepción de valor.



1.18. Figura: Valoraciones Específicas

Continuamos nuestro análisis evaluando características clave de nuestros clientes, incluyendo el tipo de viaje (solo, en familia, en pareja, etc.) y la clase de servicio (Economy, Business, First Class o Premium Economy). Este enfoque nos permite entender las preferencias y necesidades de diferentes segmentos de viajeros, facilitando la optimización de nuestros servicios y estrategias de marketing. Este análisis se sustenta en datos obtenidos de encuestas y registros de compra, proporcionando una visión integral de las tendencias de consumo y comportamiento de nuestros clientes.



1.19. Figura: Comparacion Traveller And Class

Tipo de Viajero: Los gráficos muestran una tendencia general de disminución en el número de pasajeros de Emirates para todos los tipos de viajeros desde 2016 hasta 2024. Los viajeros de negocios han visto una caída notable, lo cual puede reflejar cambios en las prácticas de viaje corporativo. Las parejas que viajan por ocio y las familias también han disminuido, posiblemente influenciadas por factores económicos o globales como la pandemia. Los viajeros solos muestran una disminución menos marcada, pero aún evidente a lo largo de los años.

Clase de Viaje: En términos de clases de viaje, todas han experimentado una disminución en el número de pasajeros. Business Class muestra una reducción continua, alineándose con la caída en los viajeros de negocios. Economy Class, aunque también en declive, sigue siendo la opción más popular entre las clases. First Class y Premium Economy han visto disminuciones significativas, con la primera clase siendo particularmente afectada, posiblemente por su alto costo en comparación con los beneficios percibidos en tiempos económicos difíciles.

Mapas

Después de llevar a cabo el procesamiento previo de los datos y la desagregación de la columna Ruta en Origen, Destino y Escala procederemos a representar gráficamente todos nuestros puntos de origen, escala y destino. Posteriormente, nos enfocaremos en los aeropuertos que registran una mayor afluencia entre nuestros pasajeros, y realizaremos un análisis de las reseñas tanto positivas como negativas. Nos adentraremos especialmente en las negativas para investigar posibles relaciones con aquellas que no han sido verificadas.

Comenzamos con una representación general de nuestros aeropuertos de origen, de los cuales solo nos quedaremos con los que tienen un tráfico superior a 30 unidades debido a que representan la gran mayoría de afluencia de nuestra empresa. Entre ellos destacan: DXB (Aeropuerto Internacional de Dubái): 242 salidas LHR (Aeropuerto de Heathrow): 78 salidas BKK (Aeropuerto Internacional de Suvarnabhumi): 57 salidas MAN (Aeropuerto de Manchester): 37 salidas LGW (Aeropuerto de Gatwick, Londres): 34 salidas



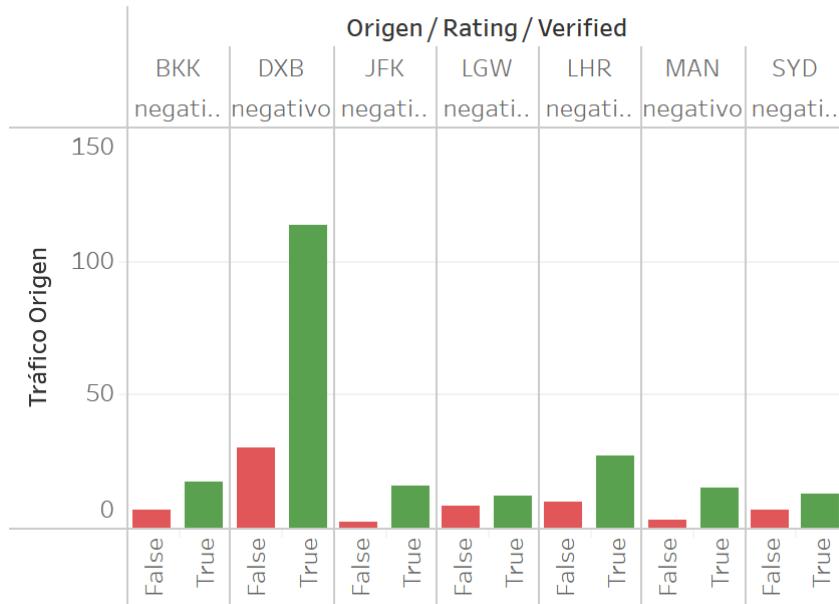
1.20. Figura: Mapa Origen Emirates

Teniendo en cuenta los aeropuertos más traficados comentados anteriormente, representamos el número de valoraciones negativas, neutras y positivas que nos han otorgado nuestros clientes para poder analizar su satisfacción en cada uno de ellos. Como podemos observar Dubai es el aeropuerto central de nuestra empresa por lo que es la que mayor número de valoraciones contiene. En el primer gráfico destaca frente al resto al tener un número superior de valoraciones, para ello representamos en el segundo gráfico la media del overall rating en las valoraciones negativas, neutras y positivas para obtener una visión de todas en una escala parecida. Aunque como podemos observar no hay tanta diferencia con respecto a las demás, el aeropuerto de Dubai supone un gran problema para nuestra empresa debido a su alto tráfico y que es uno de los aeropuertos con una media de valoración negativas más baja (1,909) y la valoración positiva tampoco es de las más altas (8,671).



1.21. Figura: Estimacion Origen Emirates

En el siguiente gráfico, representamos las valoraciones verificadas y no verificadas dentro de las valoraciones negativas para ver si el problema esta relacionado con valoraciones falsas, pero como podemos ver en el gráfico inferior no es el caso.



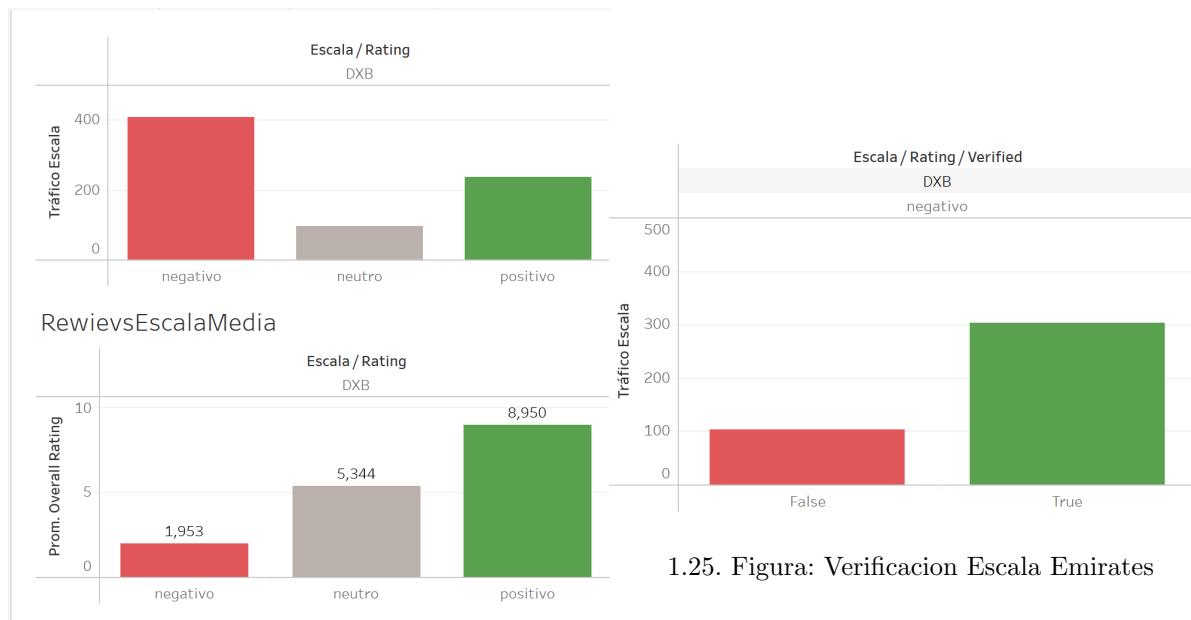
1.22. Figura: Verificadas Origen Emirates

Continuamos realizando lo mismo para la escala, de la cual solo tenemos en cuenta Dubai debido a su gran afluencia(740), el resto de aeropuertos tienen un numero de vuelos escaso.



1.23. Figura: Mapa Escala Emirates

Representamos el numero de valoraciones negativas neutras y positivas, así como la media de ellas y de las negativas las que están verificadas y las que no pero tampoco nos ayudan a llegar a ninguna conclusión.



1.24. Figura: Estimacion Escala Emirates

1.25. Figura: Verificacion Escala Emirates

En esta ocasión, nos adentraremos en los aeropuertos de destino, priorizando aquellos con mayor frecuencia de llegadas. Hemos establecido un criterio mínimo de 30 llegadas para la selección. A continuación, presentamos los aeropuertos destacados: DXB (Aeropuerto Internacional de Dubái): 235 llegadas LHR (Aeropuerto de Heathrow): 78 llegadas BKK (Aeropuerto Internacional de Suvarnabhumi): 67 llegadas SYD (Aeropuerto Internacional Kingsford Smith): 40 llegadas MAN (Aeropuerto de Manchester): 35 llegadas

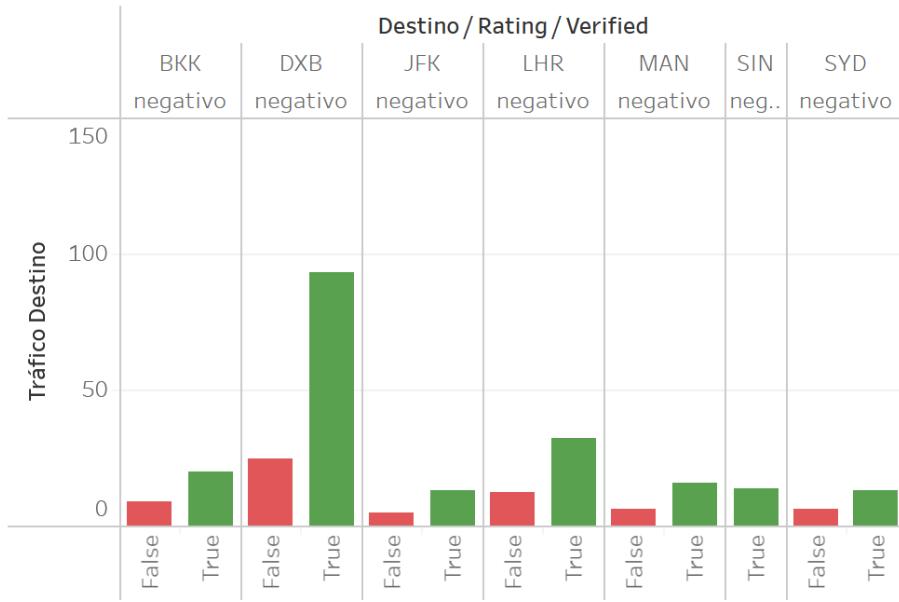


1.26. Figura: Mapa Destino Emirates

Representamos el numero de valoraciones negativas neutras y positivas, así como la media de ellas y de las negativas las que están verificadas como anteriormente, pero al no llegar a ninguna conclusión realizaremos una comparación de las reviews en los mismos aeropuertos de Emirates y Qatar para ver si el aeropuerto supone un factor determinante.



1.27. Figura: Estimacion Destino Emirates



1.28. Figura: Verificacion Destino Emirates

Para llegar a una conclusión de si los problemas de Emirates viene dado por los aeropuertos, hemos realizado una comparación contra nuestro competidor directo Qatar. Para ello vamos a realizar la comparación en los aeropuertos mas transitados que coinciden entre las dos aerolíneas. No vamos a tener en cuenta las escalas ya que en Emirates la mayoría son en Dubai mientras que en Qatar son en Doha.



1.29. Figura: Origen Emirates/Qatar

Los aeropuertos en los que analizaremos las valoraciones que coinciden en ambas aerolíneas son: BKK(Aeropuerto Internacional de Suvarnabhumi), JFK(Aeropuerto Internacional John F Kennedy), LHR(Aeropuerto de Heathrow), MAN(Aeropuerto de Manchester), SYD(Aeropuerto Internacional Kingsford Smith).



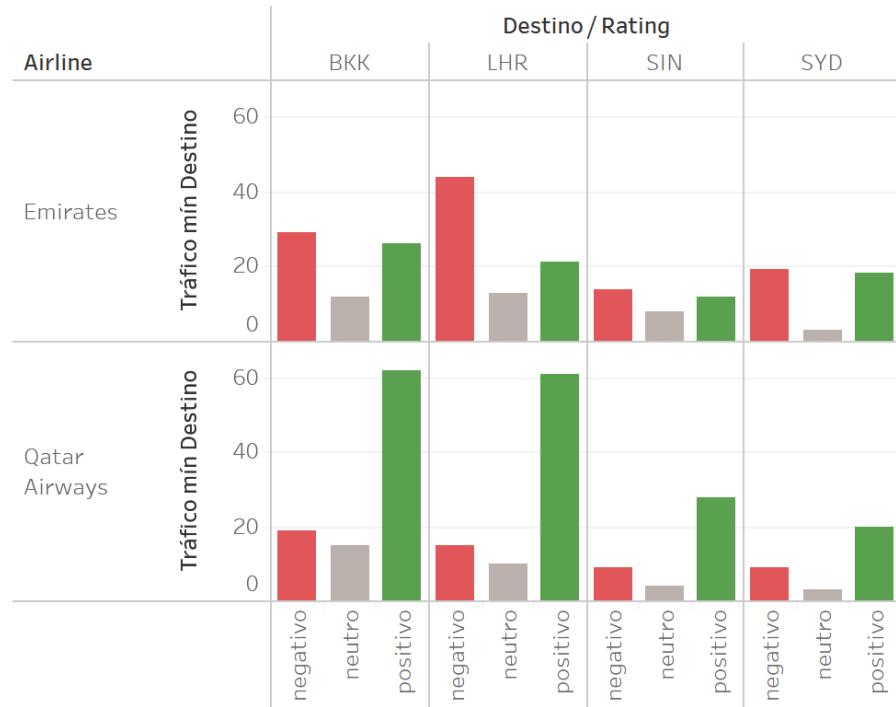
1.30. Figura: Destino Emirates/Qatar

En cuanto a los destinos analizaremos los siguientes aeropuertos: BKK(Aeropuerto Internacional de Suvarnabhumi), LHR(Aeropuerto de Heathrow), SIN(Aeropuerto Internacional de Singapur), SYD(Aeropuerto Internacional Kingsford Smith).

Ahora analizaremos el numero de valoraciones negativas, neutras y positivas en los diferentes aeropuertos comunes tanto de origen como de destino para llegar a alguna conclusión



1.31. Figura: Valoraciones Origen Emirates/Qatar

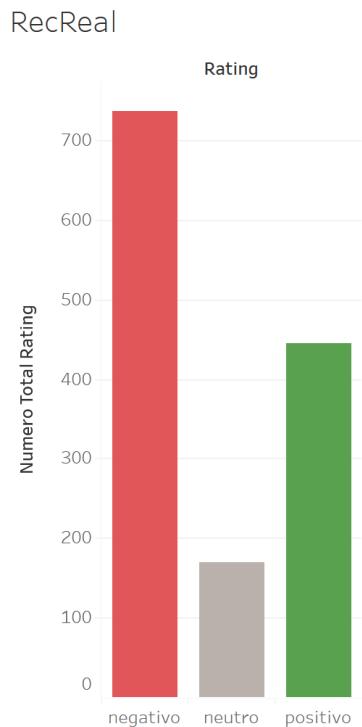


1.32. Figura: Valoraciones Destino Emirates/Qatar

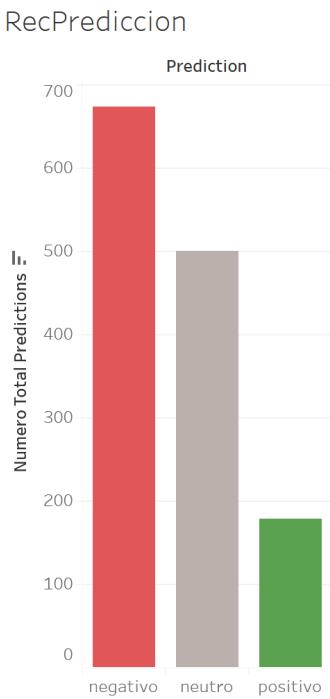
Tras analizar los aeropuertos en común mas transcurridos entre las dos aerolíneas podemos observar que los aeropuertos no son la causa del problema ya que si nos fijamos Emirates tiene una mayoría de valoraciones negativas en todos ellos mientras que nuestro competidor Qatar tiene una mayor tendencia a las valoraciones positivas. Esto ocurre tanto en los orígenes como en los destinos lo que a su vez nos indica que la causa de los problemas no tiene relación con los vuelos de llegada o de salida.

Concluyendo que el principal problema de nuestra aerolínea esta relacionado con los servicios ofrecidos; atención al cliente, entretenimiento, comida, puntualidad... Por lo que mediante el clustering trataremos de analizar profundamente las valoraciones negativas de nuestros servicios para llegar a unas conclusiones finales.

Debido a la clasificación realizada, llevaremos a cabo un mini análisis de la variación de valoraciones negativas, neutras y positivas que habría en el caso de necesitar predecirlas. Con los datos reales obtenemos la siguiente distribución:



1.33. Figura: Recuento valoraciones Reales



1.34. Figura: Recuento valoraciones Predicción

En ambos gráficos mostramos la distribución de las valores negativas, neutras y positivas. El análisis realizado hasta el momento es con los datos reales por lo que es totalmente preciso, sin embargo si usariámos las predicciones obtenidas a través de la clasificación estaríamos realizando un análisis con datos no reales. Para obtener el csv con las predicciones hemos utilizado para entrenar todas las aerolíneas excepto Emirates, que es la que se usa en el test para mantener el mismo numero de valoraciones y poder realizar una comparación precisa para ver como varia la distribución de las positivas, negativas y neutras con los diferentes f-score obtenidos.

Matriz de confusión:			
	Predicted negative	Predicted neutral	Predicted positive
Actual negative	562	60	60
Actual neutral	68	552	62
Actual positive	67	31	584

Métricas agregadas:

Average Type	Precision	Recall	F1-Score
Micro	0.829912	0.829912	0.829912
Macro	0.830661	0.829912	0.82993
Weighted	0.830661	0.829912	0.82993

1.35. Figura: Medidas Train

Matriz de confusión:

	Predicted Negative	Predicted Neutral	Predicted Positive
Actual Negative	531	176	29
Actual Neutral	74	76	19
Actual Positive	67	248	130

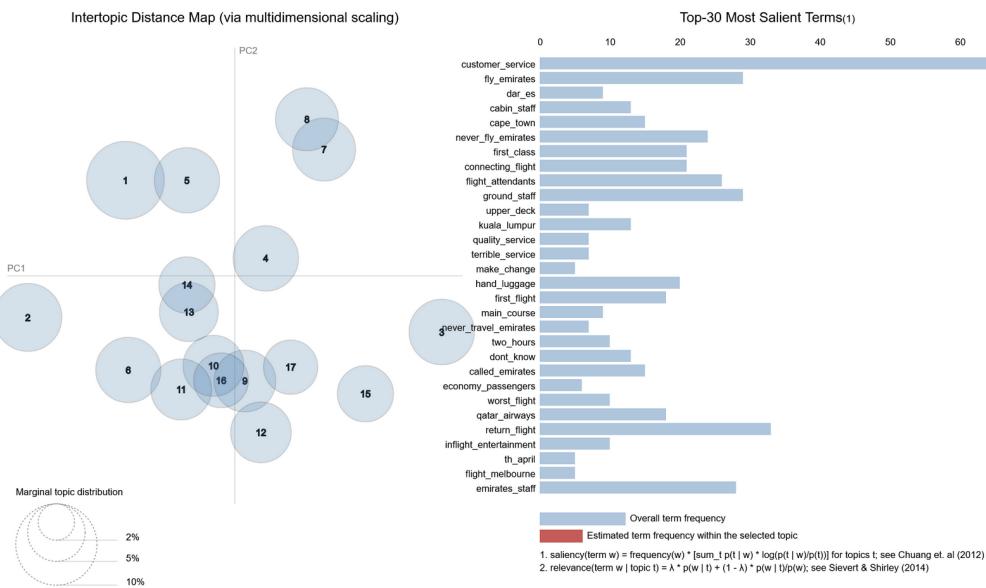
Métricas Detalladas:

Average Type	Precision	Recall	F1-Score
Micro	0.545926	0.545926	0.545926
Macro	0.557505	0.487769	0.466267
Weighted	0.690563	0.545926	0.577221

1.36. Figura: Medidas Test

Topic Modelling

Ahora mediante Topic Modelling vamos a profundizar en los motivos reales de las causas de nuestras valoraciones. Para ello vamos a comenzar analizando las valoraciones negativas de Emirates:



1.37. Figura: Malo Nuestro

Análisis de Temas en Comentarios sobre Emirates Negativos

Tema 1: Problemas con el servicio al cliente

Temas Relacionados: Topic 0, Topic 1, Topic 6, Topic 8, Topic 13, Topic 16

Descripción: Este tema aborda los problemas relacionados con el servicio al cliente de Emirates. Los clientes expresan insatisfacción con aspectos como la atención al cliente, la capacidad de resolver problemas y la calidad del servicio.

Ánalisis: Los clientes señalan diversos problemas en el servicio al cliente de Emirates, incluyendo falta de atención, respuestas inadecuadas a problemas y percepción general de servicio deficiente.

Palabras Representativas: customer_service, flight_attendants, cabin_crew, ground_staff, online_complaint_system, emirates_customer_service.

Tema 2: Experiencia desfavorable en business class

Temas Relacionados: Topic 1, Topic 2, Topic 4, Topic 10, Topic 11, Topic 16

Descripción: Este tema aborda las experiencias negativas relacionadas con la clase business de Emirates. Los clientes expresan descontento con aspectos como el servicio, la comodidad y la calidad general de la experiencia en business class.

Ánalisis: Los clientes comentan sobre diversos problemas experimentados en la clase business de Emirates, incluyendo servicio insatisfactorio, falta de comodidad y percepción de baja calidad.

Palabras Representativas: business_class, disappointed_service, fly_emirates, service_quality.

Tema 3: Problemas de vuelo y gestión de vuelos

Temas Relacionados: Topic 3, Topic 4, Topic 5, Topic 9, Topic 10, Topic 15, Topic 16

Descripción: Este tema se centra en los problemas relacionados con los vuelos y su gestión por parte de Emirates. Los clientes expresan frustración con aspectos como retrasos, cancelaciones y problemas de gestión de vuelos.

Análisis: Los clientes comparten experiencias negativas relacionadas con vuelos, incluyendo retrasos, cancelaciones y pérdida de vuelos, lo que afecta negativamente su percepción de la aerolínea.

Palabras Representativas: missed_flight, never_fly_emirates, flights_cancelled, baggage_claim, baggage_services.

Tema 4: Experiencias desfavorables en la cabina y con la tripulación

Temas Relacionados: Topic 2, Topic 4, Topic 5, Topic 6, Topic 11, Topic 12, Topic 14

Descripción: Este tema aborda las experiencias negativas en la cabina y con la tripulación de Emirates. Los clientes expresan descontento con aspectos como el servicio de la tripulación y la calidad de la experiencia en la cabina.

Análisis: Los clientes comparten diversas quejas relacionadas con la tripulación y la experiencia en la cabina, incluyendo actitudes groseras, servicio insatisfactorio y percepción general de baja calidad.

Palabras Representativas: disappointed_service, ground_staff, crew_rude, cabin_crew, flight_attendants.

Tema 5: Problemas con la logística y la gestión de equipajes

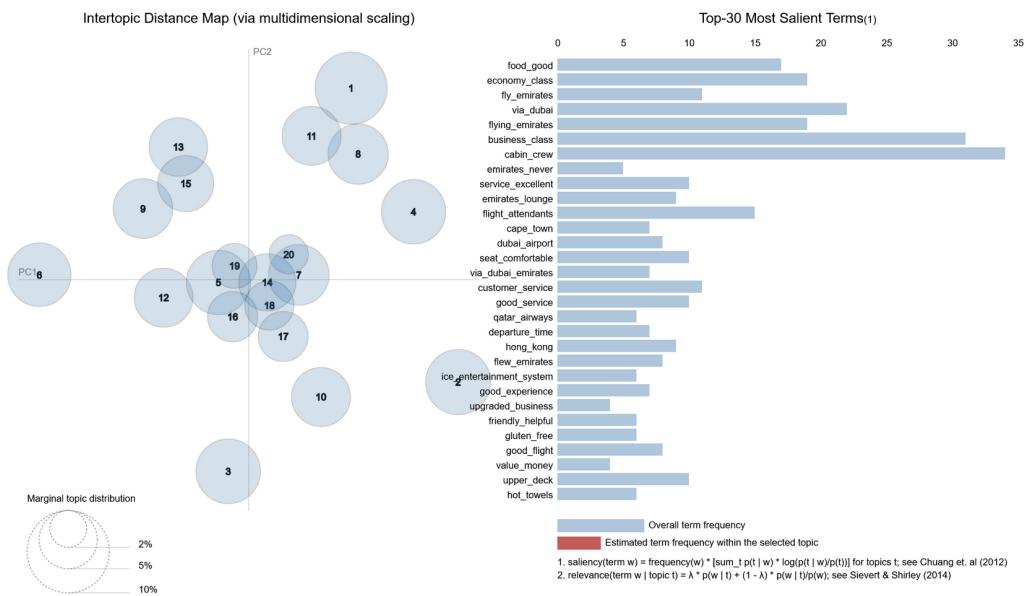
Temas Relacionados: Topic 4, Topic 6, Topic 8, Topic 9, Topic 12, Topic 16

Descripción: Este tema aborda los problemas relacionados con la logística y la gestión de equipajes de Emirates. Los clientes expresan frustración con aspectos como la gestión de equipajes, cancelaciones de vuelos y problemas de logística.

Análisis: Los clientes señalan diversas dificultades relacionadas con la gestión de equipajes y la logística de vuelos, lo que genera molestias y afecta su experiencia de viaje.

Palabras Representativas: baggage_claim, flights_cancelled, left_stranded, baggage_services.

Seguimos realizando un análisis de las valoraciones positivas de Emirates para llegar a unas conclusiones eficaces:



1.38. Figura: Bueno Nuestro

Análisis de Temas en Comentarios sobre Emirates Positivos

Tema 1: Experiencia de vuelo en business class

Temas Relacionados: Topic 0, Topic 1, Topic 2, Topic 8, Topic 10, Topic 16, Topic 17

Descripción: Este tema se centra en la experiencia de vuelo en la clase ejecutiva de Emirates, abordando aspectos como la calidad de la comida, el servicio de la tripulación de cabina y las comodidades a bordo.

Ánalisis: Los clientes expresan opiniones positivas sobre la experiencia en la clase ejecutiva de Emirates, elogiando la calidad de la comida, la amabilidad de la tripulación de cabina y la comodidad general del vuelo.

Palabras Representativas: business_class, food_good, flight_attendants_friendly, special_meal, upgraded_business, outstanding_pair.

Tema 2: Comparaciones con otras aerolíneas

Temas Relacionados: Topic 4, Topic 14, Topic 5, Topic 11, Topic 18

Descripción: En este tema, los clientes comparan la experiencia de volar con Emirates con la de otras aerolíneas, discutiendo aspectos como el servicio al cliente, la comida a bordo y la comodidad de los vuelos.

Ánalisis: Los clientes destacan las diferencias entre Emirates y otras aerolíneas, señalando aspectos como la calidad del servicio, la comida y la comodidad. Algunos elogian a Emirates en comparación con otras aerolíneas.

Palabras Representativas: qatar_airways, one_worlds, via_dubai, fly_emirates, departing_kuwait, ive_flown.

Tema 3: Experiencia de vuelo en economy class

Temas Relacionados: Topic 3, Topic 12, Topic 15

Descripción: Este tema aborda la experiencia de volar en la clase económica de Emirates, incluyendo aspectos como la comodidad de los asientos, el servicio de la tripulación de cabina y la calidad del entretenimiento a bordo.

Análisis: Los clientes comentan sobre su experiencia en la clase económica de Emirates, destacando aspectos como la comodidad de los asientos, la amabilidad de la tripulación de cabina y la calidad del entretenimiento a bordo.

Palabras Representativas: economy_class, crew_members, entertainment_system, legs_journey, seat_comfortable, service_good.

Tema 4: Servicio al cliente y experiencia general

Temas Relacionados: Topic 9, Topic 13, Topic 6, Topic 7, Topic 18

Descripción: Este tema se centra en el servicio al cliente de Emirates y la experiencia general de vuelo, abordando aspectos como la amabilidad del personal, la puntualidad y la resolución de problemas.

Análisis: Los clientes expresan opiniones variadas sobre el servicio al cliente de Emirates y su experiencia general de vuelo. Se elogia la amabilidad del personal y la eficacia del servicio.

Palabras Representativas: customer_service, service_excellent, arrived_time, good_service, pleased_emirates.

Tema 5: Comodidades a bordo y entretenimiento

Temas Relacionados: Topic 1, Topic 4, Topic 5, Topic 12, Topic 19

Descripción: En este tema, se discuten las comodidades a bordo y la calidad del entretenimiento proporcionado por Emirates durante el vuelo.

Análisis: Los clientes comentan sobre las comodidades a bordo, como el entretenimiento y la calidad de los servicios. Se elogian la variedad de opciones de entretenimiento y la comodidad de los asientos.

Palabras Representativas: ice_entertainment_system, food_beverages, seat_pitch, inflight_entertainment, entertainment_system, onboard_food.

Conclusiones Negativas

Basándonos en los temas identificados en el topic modeling y en los gráficos proporcionados, podemos extraer algunas conclusiones sobre áreas específicas en las que Emirates puede necesitar mejorar:

Servicio al Cliente

Existen quejas consistentes sobre el servicio al cliente, incluyendo falta de atención, respuestas inadecuadas a problemas y percepción general de servicio deficiente. Esto indica la necesidad de mejorar la capacitación del personal y la eficacia en la resolución de problemas para aumentar la satisfacción del cliente. Esto lo podemos observar ya que la mayoría de nuestras valoraciones de servicio al cliente varían entre muy malas y muy buenas. El motivo de las malas condiciones podría deberse a retrasos en los aviones y falta de comprensión ante los clientes.

Experiencia en Clase Business

Los clientes expresan descontento con la experiencia en la clase business de Emirates, señalando problemas de servicio, comodidad y calidad general. La aerolínea debería revisar y mejorar la calidad de su servicio en esta clase para garantizar una experiencia más satisfactoria para los pasajeros de negocios. Lo que puede verse reflejado ya que a lo largo de los años se va disminuyendo el uso de la clase business en Emirates.

Problemas de Vuelo y Gestión

Las quejas relacionadas con retrasos, cancelaciones y problemas de gestión de vuelos indican deficiencias en la logística y gestión operativa de Emirates. Es crucial mejorar la puntualidad y la eficiencia en la gestión de vuelos para evitar inconvenientes y frustraciones para los pasajeros.

Experiencia en la Cabina y con la Tripulación

Se reportan experiencias negativas en la cabina y con la tripulación, incluyendo actitudes groseras, servicio insatisfactorio y percepción general de baja calidad. Emirates debe enfocarse en mejorar la actitud y el servicio de su tripulación, así como en garantizar una experiencia más cómoda y placentera en la cabina.

Logística y Gestión de Equipajes

Los problemas relacionados con la gestión de equipajes, cancelaciones de vuelos y logística indican deficiencias en la gestión operativa y logística de Emirates. La aerolínea debe mejorar la gestión de equipajes y optimizar sus operaciones para evitar inconvenientes y garantizar un servicio más confiable.

Conclusiones Positivas

Basándonos en la información proporcionada por los gráficos implementados y el análisis de topic modeling, podemos extraer algunas conclusiones sobre la experiencia de vuelo de Emirates y su comparación con otras aerolíneas, específicamente Qatar Airways, así como sobre la percepción de los clientes en diferentes aspectos de su servicio:

Experiencia en Clase Ejecutiva (Business Class)

Este tema se alinea con las valoraciones mostradas en los gráficos, donde se observa una tendencia general de mejora en la calidad de servicio a lo largo de los años, aunque con algunas fluctuaciones notables.

Comparaciones con Otras Aerolíneas (Especialmente Qatar Airways)

Los gráficos muestran una tendencia a favor de Qatar Airways en algunos aspectos, como la valoración del entretenimiento a bordo y la relación calidad-precio, lo que indica que Emirates podría enfrentar una competencia cercana en ciertos aspectos de su servicio.

Experiencia en Clase Económica (Economy Class)

Los gráficos muestran una tendencia positiva en la valoración de la comodidad de los asientos y el servicio al cliente, aunque con algunas fluctuaciones a lo largo del tiempo.

Servicio al Cliente y Experiencia General

Los gráficos muestran una mejora general en la valoración del servicio al cliente a lo largo de los años, aunque con algunas caídas notables durante la pandemia de COVID-19.

Comodidades a Bordo y Entretenimiento

Los gráficos muestran una competencia cercana con Qatar Airways en términos de valoración del entretenimiento a bordo, esto es verídico ya que Emirates ha ganado los premios a mejor entretenimiento en 2023, aunque Emirates enfrenta desafíos en la percepción de la relación calidad-precio.

Consideraciones a tener en cuenta

Tras analizar los datos disponibles y planificar una serie de mejoras, es importante destacar que contábamos únicamente con 1,350 valoraciones de Emirates para realizar nuestro análisis, frente a los más de 100 millones de pasajeros que la aerolínea asegura transportar anualmente. Esto significa que las 1,350 valoraciones representan solo el 0.00135 % de las opiniones sobre la empresa, lo que podría limitar la representatividad de nuestros hallazgos.

En el contexto de los premios de 2023 otorgados por Skytrax, los World Airline Awards, se observó una destacada presencia asiática en los primeros puestos, con Singapore Airlines liderando como la mejor aerolínea del mundo, seguida por Qatar Airways y ANA All Nippon Airways. Emirates se situó en el cuarto lugar, siendo también reconocida en categorías específicas como la mejor primera clase del mundo, el mejor asiento de primera clase del mundo y las mejores comodidades de primera clase del mundo.

Además, Emirates fue galardonada como la "Mejor Aerolínea del Mundo." en los Premios ULTRAs 2023, resaltando su competitividad en el sector. En términos financieros, la aerolínea reportó ingresos totales de 107,400 millones de AED (aproximadamente 29,300 millones de USD) durante el último ejercicio fiscal, un incremento del 81 % respecto al período anterior, y experimentó un beneficio récord de 10,600 millones de AED (aproximadamente 2,900 millones de USD). Estos resultados subrayan el sólido crecimiento y la robusta posición de mercado de Emirates, reflejando su compromiso continuo con la innovación y la excelencia en el servicio.

Finalmente, en portales de reseñas como TripAdvisor, los usuarios destacan aspectos específicos del servicio de Emirates. Los comentarios frecuentemente elogian la calidad del servicio a bordo, las opciones de entretenimiento, y la comodidad de los asientos, especialmente en clases premium. Sin embargo, también hay menciones ocasionales de áreas con margen de mejora, como la gestión de equipajes y la atención al cliente en ciertos aeropuertos. En TripAdvisor, Emirates ha recibido una calificación general de 4.0 sobre 5 basada en 65,133 opiniones. Además, ha ganado múltiples premios Travellers' Choice en los años 2017, 2018, 2019 y 2020, lo que indica un reconocimiento constante de su calidad de servicio.

Estas consideraciones proporcionan una visión clara de la posición de Emirates en la industria, destacando tanto su éxito en captar la aprobación del cliente como su impresionante desempeño financiero y reconocimientos globales.

2. Datos para clasificación: Análisis, Preproceso y Experimentación

2.1. Datos

2.1.1. División entre Train Dev y Test

Conjunto De Datos	% de instancias	Num. de instancias
Train	64	5730
Dev	16	1432
Test Final	20	1797

2.1. Cuadro: División Train, Dev y Test

2.1.2. Distribución de las clases en cada conjunto

Conjunto De Datos	Clase Neg	Clase Neutra	Clase Pos.
Train	2530	2484	505
Dev	632	621	126
Test Final	805	749	171

2.2. Cuadro: Distribución Train, Dev y Test

2.1.3. Descripción del preproceso

Para poder realizar el preproceso lo primero que hacemos es dividir las columnas del csv en numéricas, categóricas y textuales, para poder aplicar un preprocesado diferente a cada una de estas, y después realizamos los ajustes dependiendo de lo elegido en el config.json. Podemos borrar las columnas faltantes, llenar las columnas faltantes de valores numéricos o categóricos y después reescalamos los valores numéricos. Para los textuales usamos la librería nltk para poder tokenizar, eliminar stopwords y lematizar.

También es importante destacar que hemos utilizado pickle (transformadores) para guardar cada uno de estos preprocesos y poder reutilizarlos a la hora de realizar la predicción.

2.1.4. Primeros resultados

Los resultados que obtenemos se guardan en un txt, un ejemplo de esto se puede observar en la figura ValoresDev.

Matriz de confusión:			
	Predicted negative	Predicted neutral	Predicted positive
Actual negative	458	117	58
Actual neutral	79	439	115
Actual positive	35	29	568

Precision:
[0.8006993 0.75042735 0.76653171]
Recall:
[0.7235387 0.69352291 0.89873418]
F1-score:
[0.76016598 0.72085386 0.82738529]
Accuracy:
0.7718651211801897

Metricas agregadas:			
Average Type	Precision	Recall	F1-Score
Micro	0.771865	0.771865	0.771865
Macro	0.772553	0.771932	0.769468
Weighted	0.772556	0.771865	0.769438

2.1. Figura: ValoresDev

2.1.5. Descripción del Proceso de Submuestreo o Sobremuestreo

Para hacer undersampling y oversampling hemos seguido dos métodos diferentes.

Para el undersampling el código selecciona filas al azar y elimina tantas como sean necesarias para igual la cantidad de valores diferentes en el TARGET.

Para el oversampling tenemos dos formas, la "normal", en la cual aleatoriamente seleccionamos filas y las repetimos para igualar el numero de valores de los TARGET, y la forma "smote", la cual genera nuevas filas mezclando valores de otras filas que sean parecidas, para esta forma utilizamos el método RandomOversampler de la librería imbalanced-learn.

2.2. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados

2.2.1. Experimentación: Algoritmos empleados y Breve Descripción

Hemos empleado los siguientes algoritmos con los siguientes hiper-parámetros.

■ Multinomial Naive Bayes:

- Descripción breve: Clasificador probabilístico basado en el teorema de bayes asumiendo independencias entre los predictores.
- Pros:
 - Simple rapido y especialmente util para grandes datasheets.
 - Funciona bien con un gran número de características y es eficaz en problemas de clasificación multinomial.
 - Menos propenso al overfitting.
- Contras:
 - La suposición de independencia en las características rara vez ocurre en la practica, lo que puede disminuir su rendimiento.
 - Puede ser superado por modelos más complejos si la relación entre atributos es importante.
- Hiperparámetros:
 - **alpha (Laplace smoothing parameter):** valor por defecto es 1.0. Ayuda a manejar las características no vistas en el modelo, ajustando la estimación de probabilidad. Valores más altos aplican una mayor suavización.
- Link: Sklearn MultinomialNB

■ Logistic Regression:

- Descripción breve: Modelo de regresión utilizado para predecir la probabilidad de una categoría o clase..
- Pros:
 - Proporciona probabilidades y es más interpretable.
 - Maneja bien modelos con múltiples clases.
- Contras:
 - No maneja bien grandes interacciones no lineales entre características sin transformación.
 - Requiere selección de características para evitar la multicolinealidad.
- Hiperparámetros:
 - **C:** valor por defecto es 1.0. Controla la regularización: valores más pequeños indican una regularización más fuerte.
 - **max iterator:** valor por defecto 100. Especifica el número máximo de iteraciones que el algoritmo de optimización puede tener para encontrar los parámetros óptimos del modelo.
 - **solver:** valor por defecto lbfsgs. Algoritmo de optimización; otros valores incluyen newton-cg, sag, saga. newton-cg, lbfsgs son buenos para datasets más pequeños o con regularización l2 y sag y saga son más rápidos para grandes datasets y saga también soporta penalizaciones l1.
 - **penalty:** valor por defecto l2: Especifica el tipo de penalización (norma) a aplicar. l1 conduce a modelos más dispersos (con coeficientes cero), lo cual puede ser útil para la selección de características. l2 penaliza el cuadrado de los coeficientes y es el estándar para evitar el sobreajuste. elasticnet es una combinación de L1 y L2.
 - **multi class:** valor por defecto auto. Puede usar ovr si son solo binarios o multinomial si se lo permite. Define si el algoritmo debe utilizar un enfoque binario (ovr) o multinomial para problemas de clasificación con más de dos clases. multinomial es más adecuado para clasificaciones directas entre múltiples clases
- Link: Sklearn Logistic Regression

■ Linear SVM:

- Descripción breve: Linear SVM busca el hiperplano de margen máximo que mejor divide las clases en el espacio de características.
- Pros:
 - Muy efectivo en espacios de alta dimensión y con un margen de separación claro.
 - Robusto ante los outliers en la clasificación.
- Contras:
 - No es adecuado para datasets con más ruido, donde las clases se superponen.
 - Requiere cuidadosa selección del parámetro de regularización y no es escalable a grandes volúmenes de datos.
- Hiperparámetros:
 - **C:** valor por defecto es 1.0. Controla la regularización: valores más pequeños indican una regularización más fuerte.
 - **penalty:** valor por defecto l2: Especifica el tipo de penalización (norma) a aplicar. l1 conduce a modelos más dispersos (con coeficientes cero), lo cual puede ser útil para la selección de características. l2 penaliza el cuadrado de los coeficientes y es el estándar para evitar el sobreajuste. elasticnet es una combinación de L1 y L2.
 - **max iterator:** valor por defecto 100. Especifica el número máximo de iteraciones que el algoritmo de optimización puede tener para encontrar los parámetros óptimos del modelo.
 - **dual:** valor por defecto true. Determina si se debe resolver el problema de optimización en su forma dual o primal. La forma dual es generalmente más eficiente cuando el número de muestras es menor que el número de características; de lo contrario, es mejor usar la forma primal. Primal: Se usa especialmente cuando hay un mayor número de características que de muestras. Dual: Más adecuado para situaciones donde hay más muestras que características.

- **loss:** valor por defecto squared hinge. Especifica la función de pérdida que se utiliza en el modelo. Para LinearSVC, las opciones son 'hinge' o 'squared hinge'. 'hinge': Es la función de pérdida estándar utilizada para el SVM. Esta es la misma función de pérdida que se usa en el SVM clásico y está destinada a clasificación "max-margin". 'squared hinge': Es el cuadrado de la función de pérdida hinge. Penaliza más los errores marginales, lo cual puede llevar a un modelo con mayor margen, pero potencialmente más sensible a outliers y a sobreajuste en comparación con el hinge simple.
- Link: Sklearn Linear SVM
- **XGBoost:**
 - Descripción breve: Es una implementación optimizada de gradient boosting que utiliza árboles de decisión como aprendices base.
 - Pros:
 - Más complejo de sintonizar debido a la gran cantidad de hiperparámetros.
 - Puede ser propenso al sobreajuste si no se configura adecuadamente.
 - Contras:
 - La suposición de independencia en las características rara vez ocurre en la práctica, lo que puede disminuir su rendimiento.
 - Puede ser superado por modelos más complejos si la relación entre atributos es importante.
 - Hiperparámetros:
 - **max depth:** valor por defecto 6. Profundidad máxima de cada árbol, controla la complejidad del modelo.
 - **learning rate:** valor por defecto 0.3. Reduce la contribución de cada árbol, previniendo el sobreajuste.
 - **n estimators:** valor por defecto 100. Número de árboles de decisión.
 - **gamma:** valor por defecto 0.0. Especifica el valor mínimo de reducción de pérdida necesaria para hacer una partición adicional en un nodo del árbol. Un valor más alto de gamma hace que el algoritmo sea más conservador, lo que significa que se necesitan reducciones más significativas de la pérdida para justificar divisiones adicionales. Esto resulta en árboles más simples y un modelo general menos complejo, ayudando a evitar el sobreajuste.
 - **max child weight:** valor por defecto 1.0. Define el peso mínimo (suma de los pesos de las instancias) necesario para un niño. En el contexto de la regresión, se refiere a la cantidad mínima de instancias necesarias para formar un nuevo nodo en el árbol. Si el peso de las instancias en un nodo es menor que min child weight, entonces el proceso de construcción de árbol se detendrá más adelante en ese nodo. Un valor más alto previene el modelo de aprender relaciones demasiado específicas, reduciendo así el sobreajuste pero podría llevar a un subajuste si se configura demasiado alto.
 - Link: XGBoost

2.2.2. Resultados sobre el Development

En esta sección se presentan los resultados obtenidos para el development. Importante, todos estas operaciones se han calculado utilizando oversampling.

2.2.2.1. Optimizando los resultados de la clase negativa

Algoritmo	Combinación hyperparámetros	Prec	Rec	F-score
MultinomialNaiveBayes	alpha=0,1	0,833	0,833	0,833

2.3. Cuadro: Resultados Clase Negativa

2.2.2.2. Optimizando los resultados de la clase positiva

Algoritmo	Combinación hyperparámetros	Prec	Rec	F-score
MultinomialNaiveBayes	alpha=0,75	0,782	0,782	0,782

2.4. Cuadro: Resultados Clase Positiva

2.2.2.3. Sin optimizar ninguna clase en particular

Algoritmo	Combinación hyperparámetros	Prec	Rec	F-score
MultinomialNaiveBayes	alpha=1	0,771	0,771	0,711

2.5. Cuadro: Resultados generales

2.2.3. Discusión sobre el Sentiment Analysis

La combinación **alpha=0,1** obtiene mejores resultados para negativo. La razón puede ser que las palabras comunes pierden peso, lo que le da mayor peso a las palabras únicamente negativas.

La combinación **alpha=0,75** obtiene mejores resultados para positivo. Pero esta mejora no es muy significativa, a diferencia de lo que ocurre con alpha=0,1 en los negativos.

La combinación **alpha=1** no favorece en especial a ningun resultado. La razón puede ser que las palabras equilibran sus valores.

Se han presentado los siguientes problemas con la implemetnación de n-gramas, que se intentó probar pero sucedió la RAM del ordenador no soportaba hacerlo con oversampling, lo que nos obligaba a usar undersampling y nos daba peores resultados que con oversampling y no pudo acabarse la prueba.

2.2.4. Conclusión sobre el Sentiment Analysis

Para la optimización de los resultados sobre la clase negativa se ha seleccionado Multinomial Naive Bayes por eficiencia en tiempos de entrenamiento y que las características son independientes de otras.

Para la optimización de los resultados generales se ha seleccionado finalmente se ha seleccionado Multinomial Naive Bayes por eficiencia en tiempos de entrenamiento y que las características son independientes de otras.

Las ejecuciones de los otros tres algoritmos (Logistic Regression, Linear SVM y XGBoost) dan mejorar resultados en dev que los resultados de Multinomial Naive Bayes, pero ninguno mejora los resultados de test. Para realizar estas pruebas hemos hecho una division del .csv original, siendo un 1/5 para el test y 4/5 para el train, esto es importante ya que es la manera que hemos utilizado para saber si los resultados de dev eran buenos o malos.

Creemos que uno de los principales motivos de que ningún metodo supere a Multinomial Naive Bayes es el tamaño del .csv, ya que de por si no es demasiado amplio y además lo hemos reducido aun mas. A continuación se mostraran los resultados una ejecución de cada método para que se pueda observar la superioridad en este caso de Multinomial Naive Bayes. El resto de métodos deberían de funcionar mejor con una mayor muestra de datos, ya que son más complejos y utilizan más procesos que Multinomial Naive Bayes.

Naive Bayes

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted negative | Predicted neutral | Predicted positive |
+-----+-----+-----+
| Actual negative |        491 |          83 |         59 |
+-----+-----+-----+
| Actual neutral |         50 |         533 |          50 |
+-----+-----+-----+
| Actual positive |         51 |          22 |        559 |
+-----+-----+-----+

Precision:
[0.82939189 0.8354232 0.83682635]

Recall:
[0.77567141 0.84202212 0.88449367]

F1-score:
[0.80163265 0.83870968 0.86      ]

Accuracy:
0.8340358271865121

Metricas agregadas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro       | 0.834036 | 0.834036 | 0.834036 |
+-----+-----+-----+
| Macro       | 0.83388 | 0.834062 | 0.833447 |
+-----+-----+-----+
| Weighted    | 0.833879 | 0.834036 | 0.833433 |
+-----+-----+-----+

```

2.2. Figura: ValoresTrain Naive Bayes

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted Negative | Predicted Neutral | Predicted Positive |
+-----+-----+-----+
| Actual Negative |        565 |          144 |         96 |
+-----+-----+-----+
| Actual Neutral |         68 |          63 |          40 |
+-----+-----+-----+
| Actual Positive |        135 |         282 |        412 |
+-----+-----+-----+

Informe de clasificación:
      precision  recall  f1-score  support
negativo     0.74   0.70   0.72    805
neutro       0.15   0.37   0.22    171
positivo     0.75   0.55   0.64    749

accuracy          0.60    1725
macro avg       0.55   0.54   0.52    1725
weighted avg    0.69   0.60   0.63    1725

Metricas Detalladas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro       | 0.602899 | 0.602899 | 0.602899 |
+-----+-----+-----+
| Macro       | 0.547179 | 0.540117 | 0.523642 |
+-----+-----+-----+
| Weighted    | 0.68503 | 0.602899 | 0.63263 |
+-----+-----+-----+

Precision Detallada:
[0.73567708 0.15403423 0.75182482]

Recall Detallado:
[0.70186335 0.36842105 0.55006676]

F1-Score Detallado:
[0.71837254 0.21724138 0.63531226]

Accuracy:
0.6028985507246377

```

2.3. Figura: ValoresTest Naive Bayes

Valores de los hiperparámetros utilizados:

- alpha: 0.1

Logistic Regression

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted negative | Predicted neutral | Predicted positive |
+-----+-----+-----+
| Actual negative |      517 |          70 |        46 |
+-----+-----+-----+
| Actual neutral |       24 |        580 |        29 |
+-----+-----+-----+
| Actual positive |      48 |         36 |      548 |
+-----+-----+-----+

Precision:
[0.87775891 0.84548105 0.87961477]

Recall:
[0.81674566 0.91627172 0.86708861]

F1-score:
[0.84615385 0.87945413 0.87330677]

Accuracy:
0.8667017913593256

Metricas agregadas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro       | 0.866702 | 0.866702 | 0.866702 |
+-----+-----+-----+
| Macro       | 0.867618 | 0.866702 | 0.866305 |
+-----+-----+-----+
| Weighted    | 0.867612 | 0.866702 | 0.866301 |
+-----+-----+-----+

```

2.4. Figura: ValoresTrain LogisticRegression

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted Negative | Predicted Neutral | Predicted Positive |
+-----+-----+-----+
| Actual Negative |      217 |          410 |        178 |
+-----+-----+-----+
| Actual Neutral |       55 |         61 |        55 |
+-----+-----+-----+
| Actual Positive |      283 |         79 |      387 |
+-----+-----+-----+

Informe de clasificación:
      precision  recall  f1-score  support
negativo     0.39   0.27   0.32    805
neutro       0.11   0.36   0.17    171
positivo     0.62   0.52   0.57    749

accuracy           0.39    1725
macro avg       0.38   0.38   0.35    1725
weighted avg    0.46   0.39   0.41    1725

Metricas Detalladas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro       | 0.385507 | 0.385507 | 0.385507 |
+-----+-----+-----+
| Macro       | 0.375365 | 0.380993 | 0.351234 |
+-----+-----+-----+
| Weighted    | 0.464484 | 0.385507 | 0.411183 |
+-----+-----+-----+

Precision Detallada:
[0.39099099 0.11090909 0.62419355]

Recall Detallado:
[0.26956522 0.35672515 0.51668892]

F1-Score Detallado:
[0.31911765 0.16920943 0.56537619]

Accuracy:
0.3855072463768116

```

2.5. Figura: ValoresTest LogisticRegression

Valores de los hiperparámetros utilizados:

- C: 1.0
- max iterator: 1000
- solver: lbfgs
- penalty: l2
- multi class: auto

LinearSVM

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted negative | Predicted neutral | Predicted positive |
+-----+-----+-----+
| Actual negative |      533 |        46 |       54 |
+-----+-----+-----+
| Actual neutral |       6 |      624 |        3 |
+-----+-----+-----+
| Actual positive |      58 |       31 |     543 |
+-----+-----+-----+

Precision:
[0.89279732 0.89015692 0.905      ]

Recall:
[0.84202212 0.98578199 0.85917722]

F1-score:
[0.86666667 0.93553223 0.88149351]

Accuracy:
0.8956796628029505

Metricas agregadas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro | 0.89568 | 0.89568 | 0.89568 |
+-----+-----+-----+
| Macro | 0.895985 | 0.89566 | 0.894564 |
+-----+-----+-----+
| Weighted | 0.89598 | 0.89568 | 0.894571 |
+-----+-----+-----+

```

2.6. Figura: ValoresTrain LinearSVM

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted Negative | Predicted Neutral | Predicted Positive |
+-----+-----+-----+
| Actual Negative |      319 |        275 |       211 |
+-----+-----+-----+
| Actual Neutral |       65 |        46 |       68 |
+-----+-----+-----+
| Actual Positive |     232 |        49 |      468 |
+-----+-----+-----+

Informe de clasificación:
precision    recall   f1-score  support
negativo      0.52      0.40      0.45      885
neutro        0.12      0.27      0.17      171
positivo      0.63      0.62      0.63      749

accuracy          0.48      1725
macro avg       0.43      0.43      0.42      1725
weighted avg    0.53      0.48      0.50      1725

Metricas Detalladas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro | 0.482899 | 0.482899 | 0.482899 |
+-----+-----+-----+
| Macro | 0.425157 | 0.430037 | 0.416022 |
+-----+-----+-----+
| Weighted | 0.528967 | 0.482899 | 0.499509 |
+-----+-----+-----+

Precision Detallada:
[0.51785714 0.12432432 0.63328823]

Recall Detallado:
[0.39627329 0.26900585 0.62483311]

F1-Score Detallado:
[0.44897959 0.17005545 0.62903226]

Accuracy:
0.48289855072463767

```

2.7. Figura: ValoresTest LinearSVM

Valores de los hiperparámetros utilizados:

- C: 0.75
- penalty: 'l2'
- max iterator: 100
- dual: False
- loss: squared hinge

XGBoost

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted negative | Predicted neutral | Predicted positive |
+-----+-----+-----+
| Actual negative |      546 |          39 |        48 |
+-----+-----+-----+
| Actual neutral |       6 |         624 |         3 |
+-----+-----+-----+
| Actual positive |      58 |          30 |      544 |
+-----+-----+-----+

Precision:
[0.89508197 0.9004329 0.91428571]

Recall:
[0.86255924 0.98578199 0.86075949]

F1-score:
[0.87851971 0.94117647 0.88671557]

Accuracy:
0.903058482613277

Metricas agregadas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro | 0.903056 | 0.903056 | 0.903056 |
+-----+-----+-----+
| Macro | 0.903267 | 0.903034 | 0.902137 |
+-----+-----+-----+
| Weighted | 0.903261 | 0.903056 | 0.902145 |
+-----+-----+-----+

```

2.8. Figura: ValoresTrain XGBoost

```

Matriz de confusión:
+-----+-----+-----+
|       | Predicted Negative | Predicted Neutral | Predicted Positive |
+-----+-----+-----+
| Actual Negative |      274 |          308 |        223 |
+-----+-----+-----+
| Actual Neutral |       65 |          59 |        47 |
+-----+-----+-----+
| Actual Positive |      293 |          149 |      307 |
+-----+-----+-----+

Informe de clasificación:
      precision    recall   f1-score  support
negativo       0.43     0.34     0.38     805
neutro         0.11     0.35     0.17     171
positivo       0.53     0.41     0.46     749

accuracy           0.37     1725
macro avg       0.36     0.37     0.34     1725
weighted avg    0.44     0.37     0.40     1725

Metricas Detalladas:
+-----+-----+-----+
| Average Type | Precision | Recall | F1-Score |
+-----+-----+-----+
| Micro | 0.371014 | 0.371014 | 0.371014 |
+-----+-----+-----+
| Macro | 0.359983 | 0.365094 | 0.338719 |
+-----+-----+-----+
| Weighted | 0.444678 | 0.371014 | 0.396046 |
+-----+-----+-----+

Precision Detallada:
[0.4335443 0.11434109 0.53206239]

Recall Detallado:
[0.34037267 0.34502924 0.40987984]

F1-Score Detallado:
[0.38135003 0.17176128 0.46304676]

Accuracy:
0.3710144927536232

```

2.9. Figura: ValoresTest XGBoost

Valores de los hiperparámetros utilizados:

- max depth: 5
- learning rate: 0.
- gamma: 0.5
- n estimators: 100
- max child weight: 5

2.2.5. Como ejecutar

Todas las llamadas de nuestro código son: python clasificacion.py config.json

config.json tiene el control de las variables globales, su funcionamiento se puede comprobar en cualquier momento con el comando python clasificacion.py -h, pero para realizar los entrenamientos y las predicciones lo más importantes en la variable MODO poner o entrenar o predecir y en INPUT FILE el csv destinado al Train y Test, estos csv están incluidos en la carpeta pero se pueden obtener con el método separarTrainTest.py.

En el json es importante distinguir si vas a predecir o entrenar, esto se cambia en la variable MODO y por cómo está pensado el código cuando entranas en INPUT FILE pones Train.csv y cuando predices pones Test.csv.

El json también es importante que hace falta descartar la columna Overall Rating, ya que es casi la misma que la columna TARGET, por lo que para ello si seleccionas en la variable CUANTOS ATRIBUTOS SELECCIONAS menos serán todas las columnas menos las que aparezcan en la variable CUALES DESCARTAS O ELIGES, y si en CUANTOS ATRIBUTOS SELECCIONAS eliges pocos, serán solo los que elijas en CUALES DESCARTAS O ELIGES, importante si se usa esta opción en CUALES DESCARTAS O ELIGES se deberá incluir la variable que se haya seleccionado en TARGET NAME.

En caso de por ejemplo solo querer usar una empresa puedes usar las variables COLUMNA FILA ELEGIDA y FILA ELEGIDA, y si por ejemplo querías que el programa use solo las filas que hablas de Emirates, pondrías en FILA ELEGIDA Emirates y en columna fila elegida la columna en la que sale que en este caso es Airline.

La variable RUTA MODELO que indica dónde está el .sav, está por defecto en el formato de rutas de windows, si se usa linux será necesario modificarlo.

La variable ALGORITMO sirve para elegir el algoritmo de entrenamiento, siendo las opciones NaïveBayes, LogisticRegression, LinearSVM y XGBoost.

Por último, el resto de variables son para poder modificar los hiperparámetros de los métodos.

3. Datos para el Topic Modeling: Experimentación

3.1. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados

Para este apartado, hemos utilizado el algoritmo de Asignación Latente de Dirichlet (Latent Dirichlet Allocation, LDA), el cual funciona mediante el agrupamiento de palabras que frecuentemente aparecen juntas en los textos, lo que nos permite identificar patrones y agrupar las palabras en lo que se conoce como tópicos.

En la implementación de este método, hemos empleado la biblioteca Gensim: (link a la documentación). Dentro de esta biblioteca, los argumentos con los que hemos experimentado son:

- **num_topics:** Especifica el número de temas que el modelo intenta identificar dentro del conjunto de datos. Un mayor número de temas puede permitir una mayor granularidad en la diferenciación de los temas, pero también puede llevar a una sobre-especificación si el número es demasiado alto para el volumen de datos disponible. Para encontrar la cantidad óptima debemos de buscar mediante un barrido el valor en el que se encuentre un codo con los mejores valores posibles.
- **passes:** Este parámetro controla cuántas veces el algoritmo pasa por el corpus durante el entrenamiento. Un número mayor de pasadas puede mejorar la calidad del modelo, permitiendo que el algoritmo aprenda mejor de los datos, aunque a costa de un tiempo de entrenamiento más largo.
- **alpha:** En la distribución de Dirichlet, este parámetro controla cómo se distribuyen los temas entre los documentos. Un alpha alto conduce a una mayor concentración de probabilidades, lo que significa que cada documento abarca una variedad más amplia de temas. En contraste, un alpha bajo resulta en una menor concentración, con cada documento concentrándose en unos pocos temas principales.
- **eta:** Este parámetro también utiliza la distribución de Dirichlet y afecta la variedad de palabras dentro de cada tema. Un eta alto indica que los temas son inclusivos, abarcando una amplia gama de palabras. Por otro lado, un eta bajo sugiere que los temas son más específicos, con pocas palabras dominando cada tema.

3.1.1. Experimentación: Algoritmos empleados y Breve Descripción

Descripción del Proceso:

Nestro programa realiza un análisis de topic modeling utilizando el algoritmo LDA o NMF. El programa es muy variable gracias al bien pensado diseño de configuración mediante JSON que nos da la posibilidad de ejecutar de muchas maneras el codigo. En este JSON podremos ajustar si queremos que se apliquen unigramas, bigramas y/o trigramas, tambien decidiremos la coherencia que vayamos a usar pudiendo escoger entre: φ -nmpi”, φ -v”, φ -mass.” φ -uci”. A continuación, se puede decidir los tópicos mínimos y máximos a estudiar así como el intervalo que se quiere aplicar en los mismos,por ejemplo, una ejecución de dos en dos entre 20 y 40 tópicos. De estos tópicos se puede decidir el numero de palabras que quieres que aparezcan de cada uno que al final de la ejecución se guardaran en un topics.txt con sus respectivas frecuencias. Por ultimo para facilitar el estudio hemos implementado que tipo de datos se quieren estudiar, las reseñas y titulos buenos y/o malas, nuestras y/o de los competidores(BuenoNosotros,MaloNosotros,BuenoCompetidores,MaloCompetidores). Para finalizar se puede escoger si se quieren visualizar en forma de mapa las visualizacion en un formato html con gran interactividad. Todo esto nos da muchisima flexibilidad a la hora de hacer las pruebas de estudio.

El programa principal comienza cargando el conjunto de datos (airlinesreviewsTrainDev.csv) y separándolo por aerolínea entre nosotros(Emirates) y los competidores(Qatar airways). A su vez estos se separan cada uno por la reviews donde tenemos las malas, neutras y buenas. Se ejecutara el algoritmo solo para las reviews que hayamos decidido en el JSON (BuenoNosotros,MaloNosotros,BuenoCompetidores, MaloCompetidores). Luego aplica un preprocesamiento donde convertimos el texto a minusculas,lo tokeniza, elimina las stop words, luego se aplican unigramas,bigramas y/o trigramas en funcion de lo que hayamos definido en el json. A continuación, aplica el modelo LDA o NMF a las revisiones de cada aerolínea y clasificación por bueno y malo utilizando la coherencia escogida en el JSON. Los resultados incluyen métricas como la perplexity y la coherence del modelo LDA o NMF para diferentes números de topicos y se guardan en archivos CSV, donde tendremos los CSV de bueno y malo para nuestra aerolinea y de bueno y malo para la aerolinea competitora.

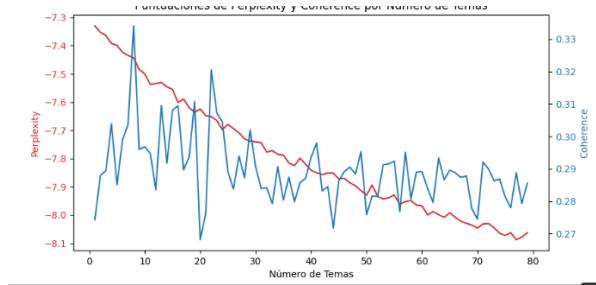
Finalmente, el programa genera gráficos para visualizar los resultados y facilitar su interpretación. Estos gráficos muestran cómo varían la perplexity y la coherence del modelo LDA con respecto al número de tópicos. Si hemos escogido la opcion de visualizar del JSON podremos tener la visualización mencionada anteriormente. La coherence nos dice cuánta coherencia tiene en el número de temas el resultado que nos haya dado y la perplexity nos dice cómo de bueno es el modelo. Cuanto más bajo sea el valor de la perplexity, mejor.

3.1.2. Resultados

Malo Nosotros

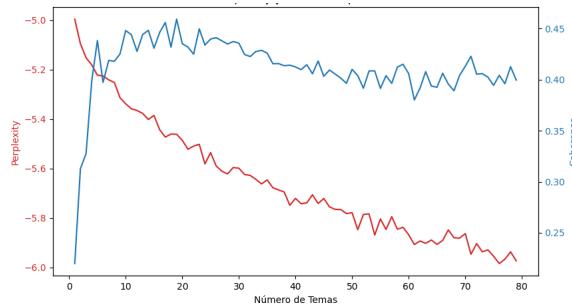
Para malo nosotros,intuimos que la forma correcta sería con bigramas y trigramas debido a que es mas descriptivo que con unigramas pero graficamos con todas las posibilidades igualmente

Unigramas:



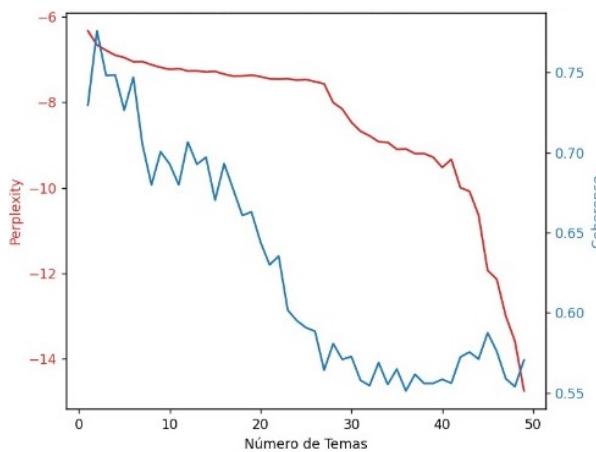
3.1. Figura: Perplexity y coherence c v en malo nosotros

Bigramas:



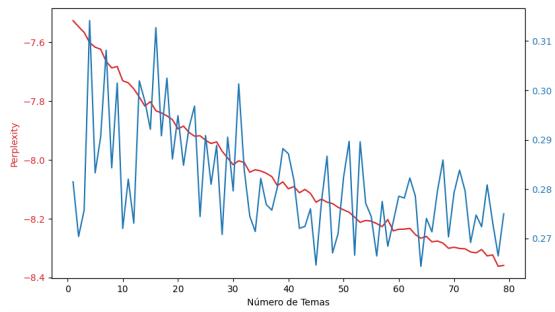
3.2. Figura: Perplexity y coherence c v en malo nosotros - Bigramas

Trigramas:



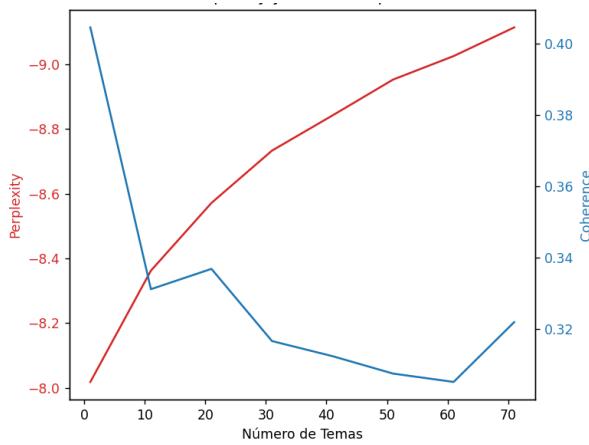
3.3. Figura: Perplexity y coherence c v en malo nosotros - Trigramas

Unigramas y Bigramas



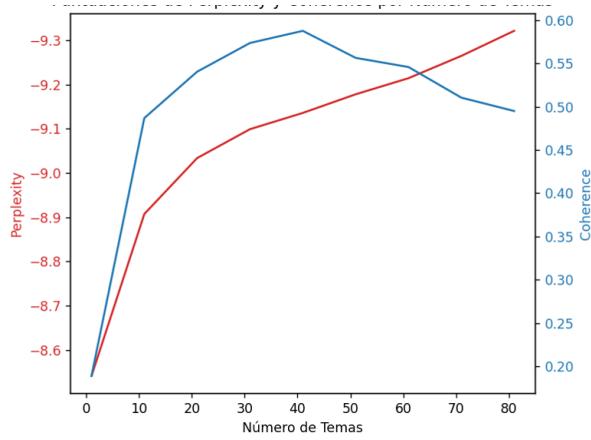
3.4. Figura: Perplexity y coherence c v en malo nosotros - Unigramas y Bigramas

Unigramas y Trigramas



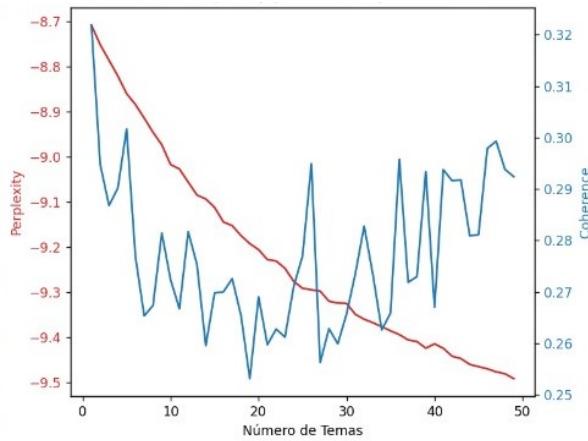
3.5. Figura: Perplexity y coherence c v en malo nosotros - Unigramas y Trigramas

Bigramas y Trigramas



3.6. Figura: Perplexity y coherence c v en malo nosotros - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas

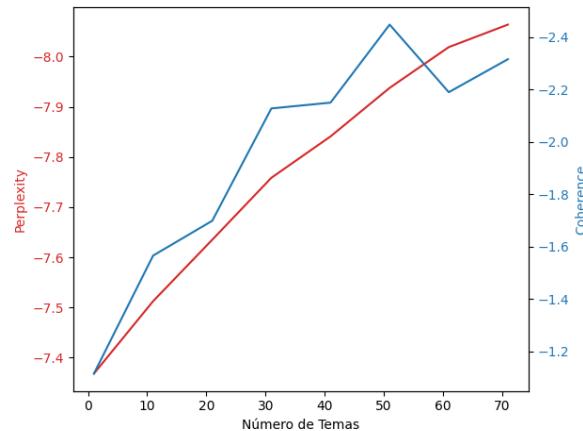


3.7. Figura: Perplexity y coherence cv en malo nosotros - Unigramas, Bigramas y Trigramas

Conclusiones:

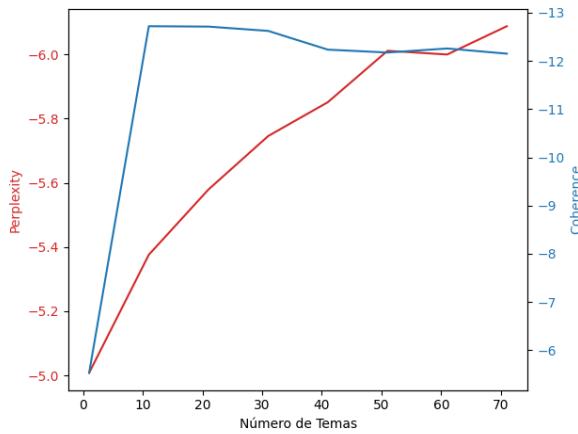
Decidimos no depender únicamente de la coherencia cv debido a sus limitaciones. Optamos por complementarla con la coherencia U-Mass, que considera la co-ocurrencia de palabras en el corpus, proporcionando así una evaluación más completa de la calidad de los temas generados

Unigramas:

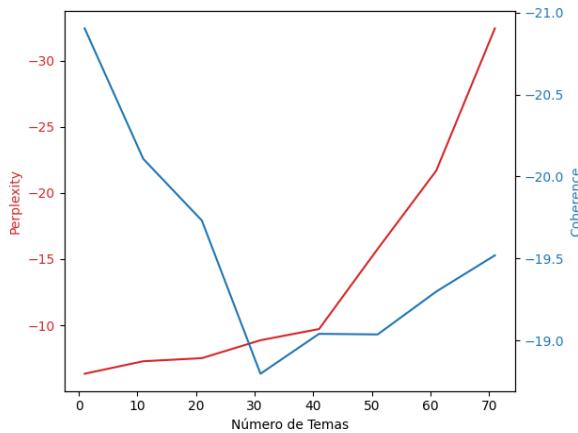


3.8. Figura: Perplexity y coherence u-mass en malo nosotros

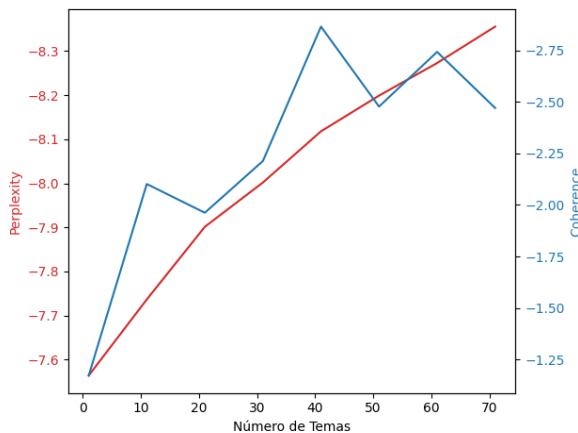
Bigramas:



3.9. Figura: Perplexity y coherence u-mass en malo nosotros - Bigramas

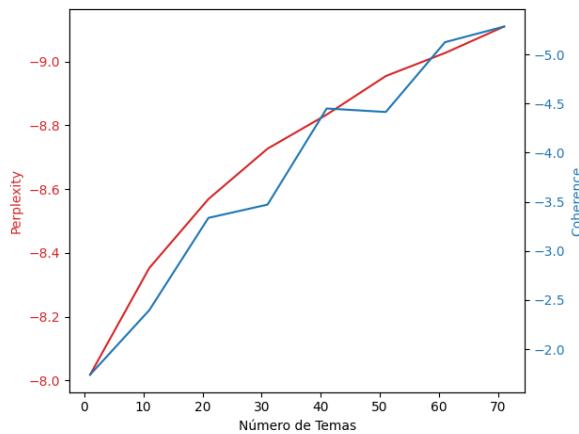
Trigramas:

3.10. Figura: Perplexity y coherence u-mass en malo nosotros - Trigramas

Unigramas y Bigramas

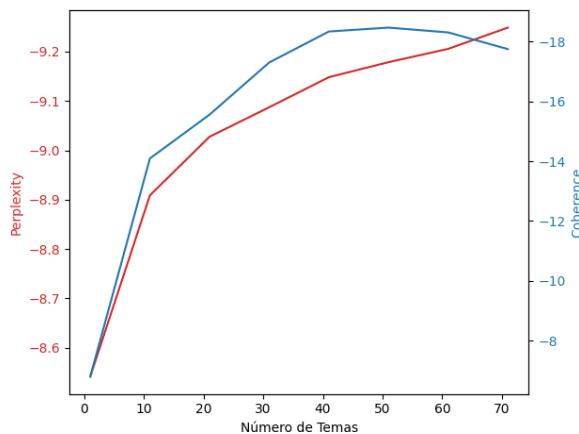
3.11. Figura: Perplexity y coherence u-mass en malo nosotros - Unigramas y Bigramas

Unigramas y Trigramas



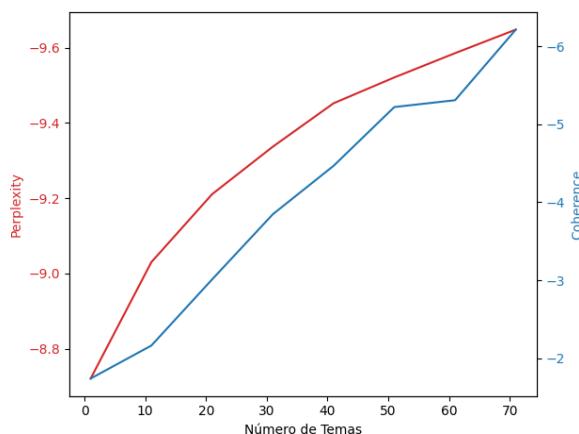
3.12. Figura: Perplexity y coherence u-mass en malo nosotros - Unigramas y Trigramas

Bigramas y Trigramas



3.13. Figura: Perplexity y coherence u-mass en malo nosotros - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas

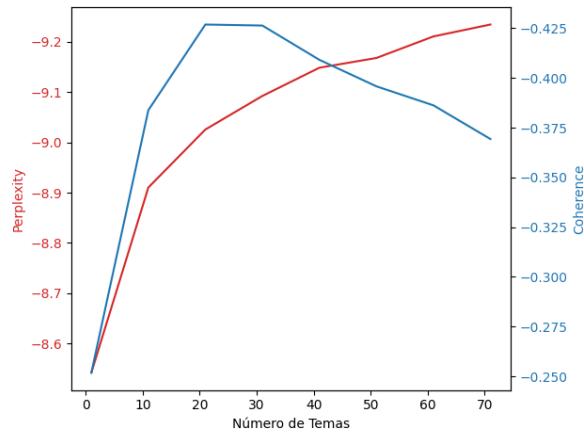


3.14. Figura: Perplexity y coherence u-mass en malo nosotros - Unigramas, Bigramas y Trigramas

Conclusiones:

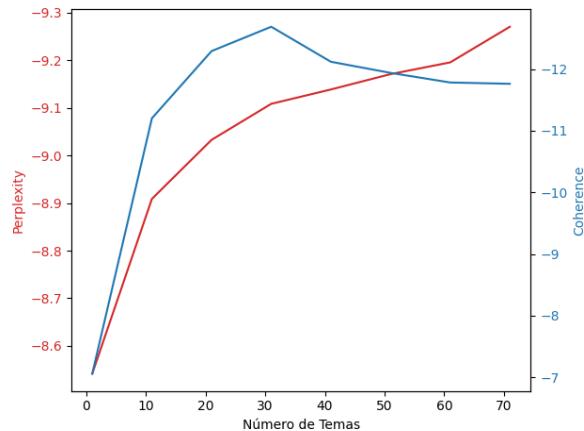
Dado que los codos más evidentes se encuentran en los bigramas y trigramas, exploramos las medidas de coherencia C-NPMI y C-UCI para estimar el número de temas. Según las evaluaciones de CV y U-Mass, parece que el número óptimo de temas está en el rango de 30 a 40.

C-NMPI



3.15. Figura: Perplexity y coherence C-NMPI en malo nosotros - Bigramas y Trigramas

C-UCI

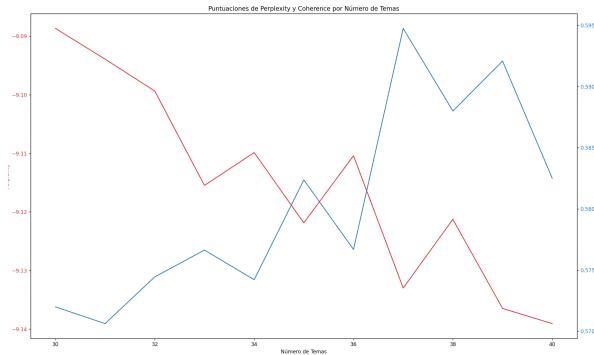


3.16. Figura: Perplexity y coherence C-UCI en malo nosotros - Bigramas y Trigramas

Conclusiones:

Gracias a las coherencias C-NPMI y C-UCI, pudimos afirmar con seguridad que el codo se encontraba entre el 30 y 40 así que ejecutamos entre 30 y 40 con coherencia C-V uno a uno

Topics 30 a 40

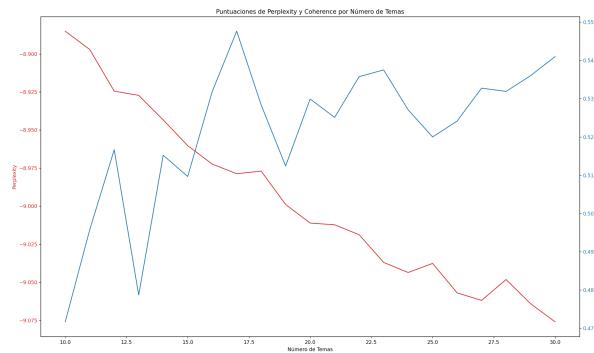


3.17. Figura: c-v de 30 a 40 tópicos en malo nosotros

Conclusiones:

Después de estimar que el número ideal de temas estaría entre 35 y 40, procedimos con múltiples ejecuciones para visualizar los resultados utilizando LDAvis. Tras realizar numerosas iteraciones y análisis visual exhaustivo, no llegamos a ninguna conclusión ya que habían demasiados temas y no tenían ningún tipo de correlación.

Visto esto procedimos a hacer otro análisis pero buscando el codo que estaba antes. Para esto hicimos un análisis de c-v de 10 a 30



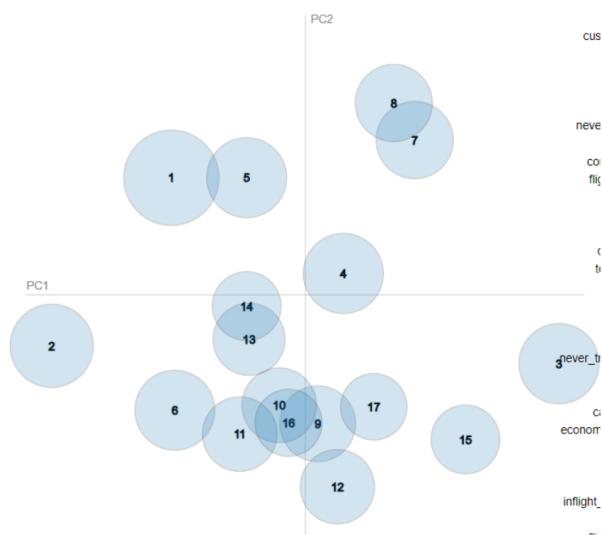
3.18. Figura: Barrido de 10 a 30 tópicos

Vimos que había un dramático codo en 17. Esto nos demostró que pesar de que en los barridos de 10 en 10 que habíamos hecho antes nos demostraron la tendencia, había más a examinar.

Hicimos varios tests y ejecuciones y con 17 temas este fue el mejor mapa:

Visualización final:

Intertopic Distance Map (via multidimensional scaling)



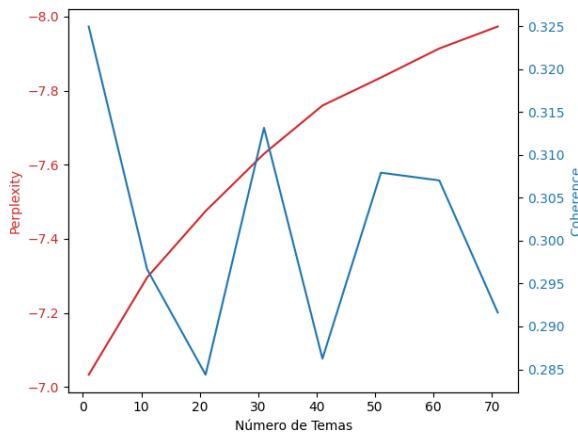
3.19. Figura: Mejor visualización conseguida

Bueno Nosotros

Para bueno nosotros, hemos graficado todas las posibilidades en busqueda de nuevas conclusiones. Para ello usamos distintos tipos de coherence las cuales son c-v, c-npmi, u-mass y c- uci.

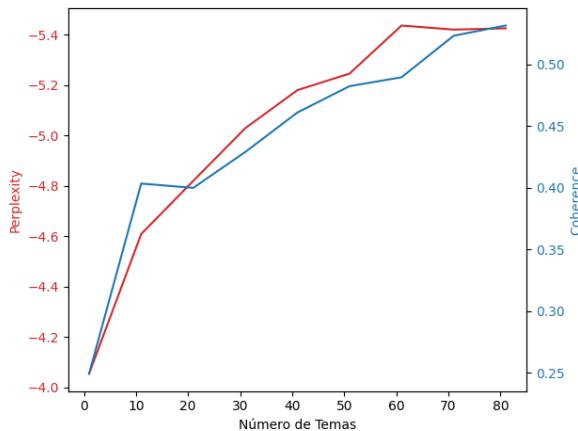
C-V:

Unigramas:



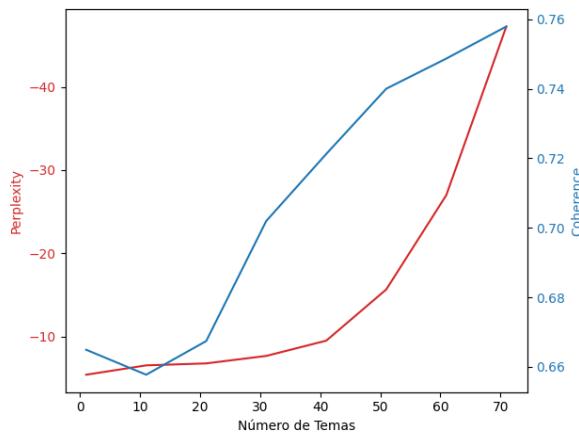
3.20. Figura: Perplexity y coherence c-v en bueno nosotros - Unigramas

Bigramas:



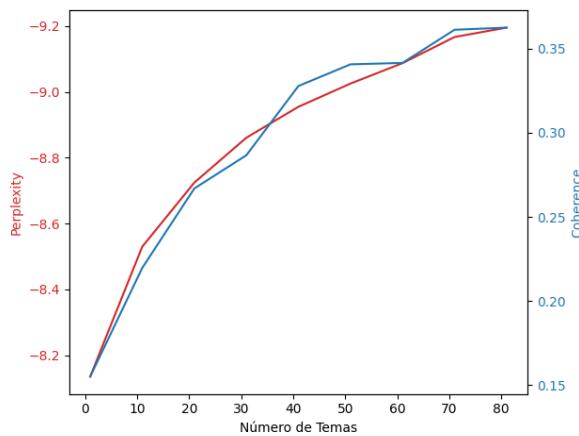
3.21. Figura: Perplexity y coherence c-v en bueno nosotros - Bigramas

Trigramas:



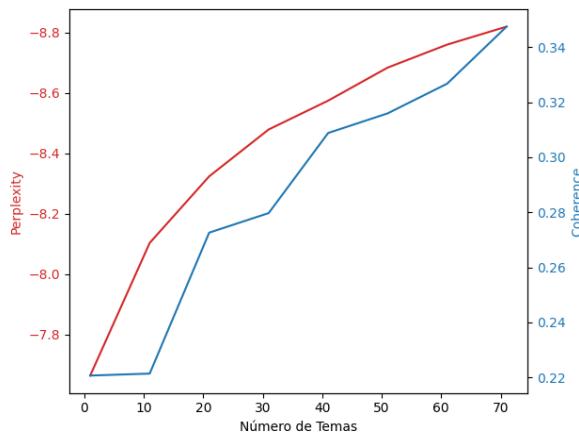
3.22. Figura: Perplexity y coherence c-v en bueno nosotros - Trigramas

Unigramas y Bigramas



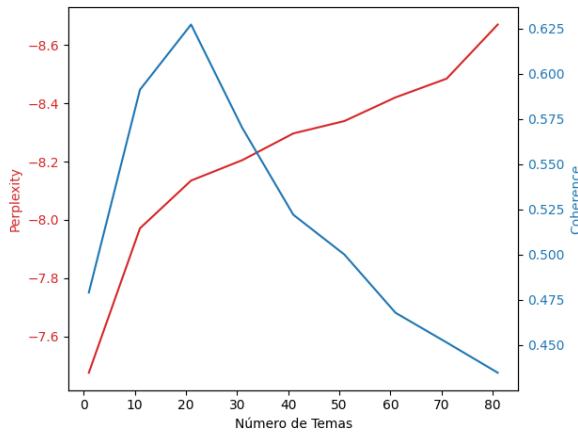
3.23. Figura: Perplexity y coherence c-v en bueno nosotros - Unigramas y Bigramas

Unigramas y Trigramas



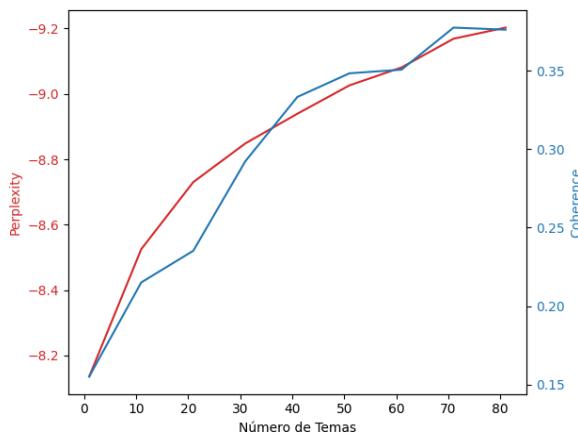
3.24. Figura: Perplexity y coherence c-v en bueno nosotros - Unigramas y Trigramas

Bigramas y Trigramas



3.25. Figura: Perplexity y coherence c-v en bueno nosotros - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas



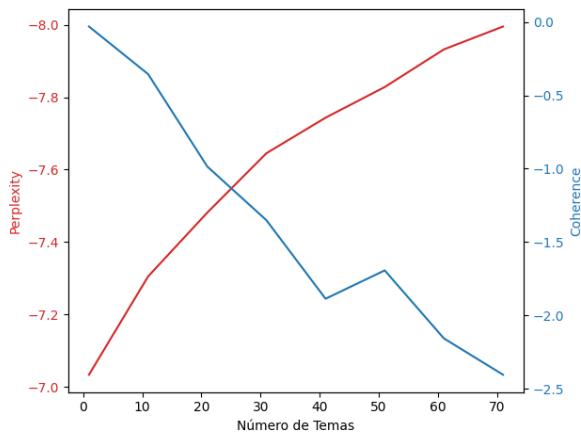
3.26. Figura: Perplexity y coherence c-v en bueno nosotros - Unigramas, Bigramas y Trigramas

Conclusiones:

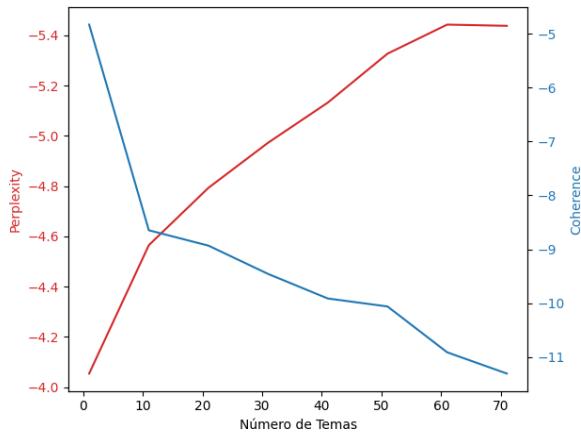
Como con malo nosotros decidimos no solo utilizar la coherencia c-v. Por ello en este caso hemos considerado optar por otra coherencia la cual es c-uci para poder hacer una mayor evaluación de los resultados y obtener mejores conclusiones.

C-UCI:

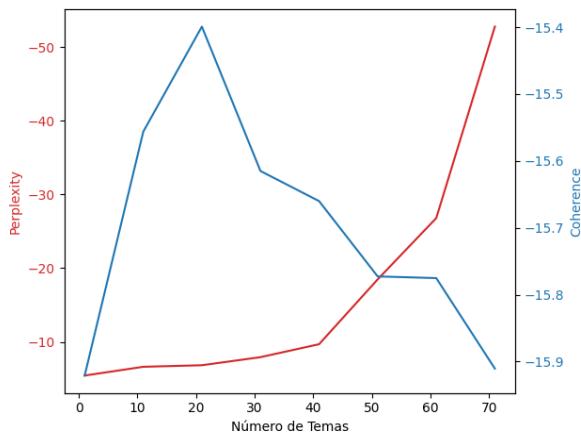
Unigramas:



3.27. Figura: Perplexity y coherence c-uci en bueno nosotros - Unigramas

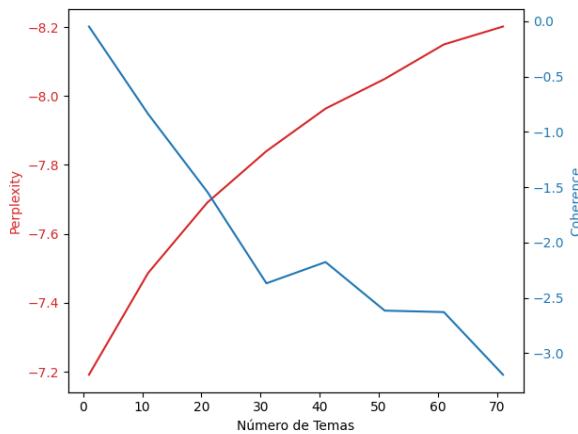
Bigramas:

3.28. Figura: Perplexity y coherence c-uci en bueno nosotros - Bigramas

Trigramas:

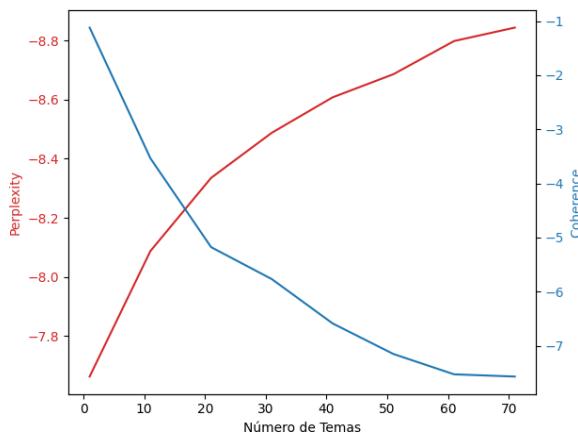
3.29. Figura: Perplexity y coherence c-uci en bueno nosotros - Trigramas

Unigramas y Bigramas



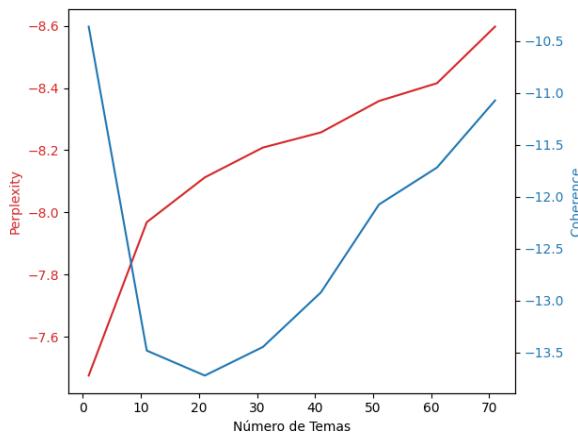
3.30. Figura: Perplexity y coherence c-uci en bueno nosotros - Unigramas y Bigramas

Unigramas y Trigramas



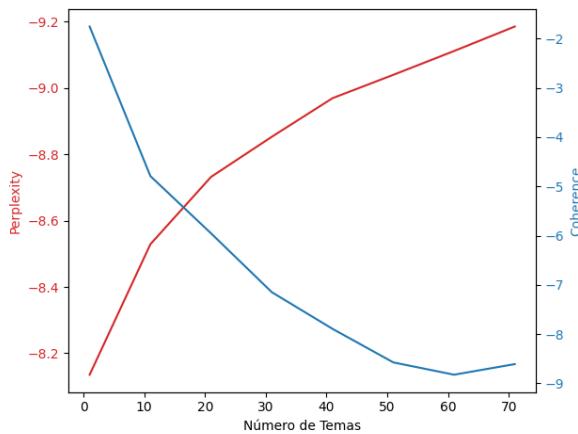
3.31. Figura: Perplexity y coherence c-uci en bueno nosotros - Unigramas y Trigramas

Bigramas y Trigramas



3.32. Figura: Perplexity y coherence c-uci en bueno nosotros - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas

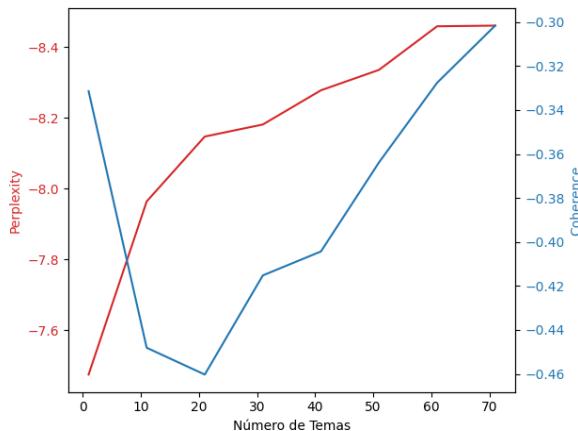


3.33. Figura: Perplexity y coherence c-uci en bueno nosotros - Unigramas, Bigramas y Trigramas

Conclusiones:

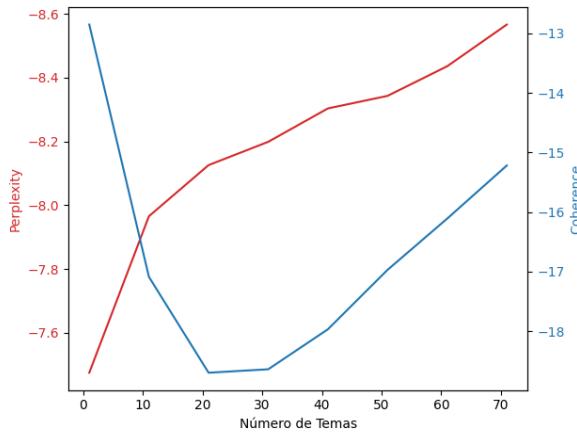
Como en malos nosotros los codos más evidentes de bueno nosotros se encuentran en los bigramas y trigramas, segun los resultados que nos muestran las graficas de c-v y c-uci. Para fortalecer esta conclusion decidimos implementar ,a continuacion, las otras dos coherencias u-mass y c-nmpi pero solamente en bigramas y trigramas. En esta primera estimacion los codos nos dan entre 15 y 30 y con la ayuda de las dos siguientes coherencias nos aseguraremos.

C-NMPI



3.34. Figura: Perplexity y coherence C-NMPI en bueno nosotros - Bigramas y Trigramas

C-UCI

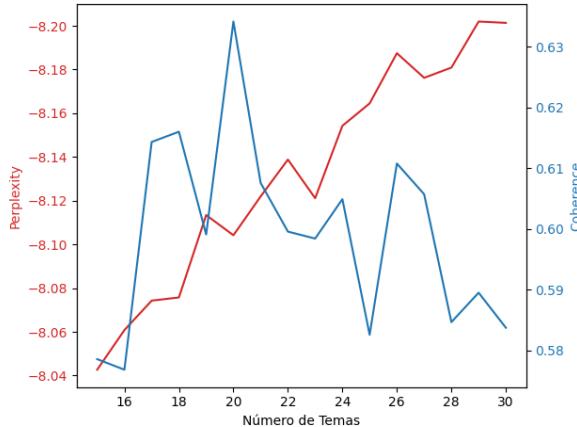


3.35. Figura: Perplexity y coherence U-MASS en bueno nosotros - Bigramas y Trigramas

Conclusiones:

Tras obtener los resultados de estas dos coherencias podemos afirmar que el codo se encuentra entre 15 y 30. Para obtener el numero de temas ideal volvemos a aplicar LDA pero solo con c-v de 15 a 30.

Topicos 15 a 30



3.36. Figura: c-v de 15 a 30 tópicos en malo nosotros

Conclusiones:

Observando la grafica estimamos que el numero de topicos ideal esta entre 18 y 22 y hacemos varias ejecuciones entre estos dos para estimar el mejor numero de topicos. Finalmente tras varias ejecuciones los resultados no varian en gran medida y estimamos que el mejor numero de topicos es de 20.

Visualización final:

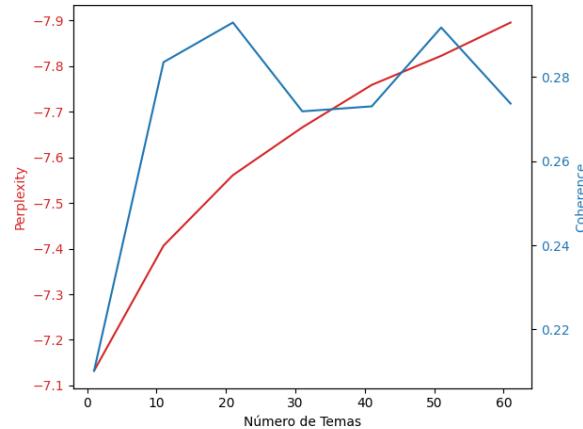


3.37. Figura: Mejor visualización conseguida Bueno Nosotros

Malo Competidor

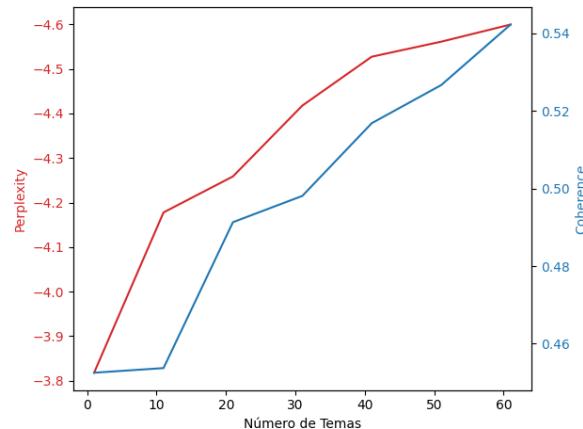
Al tener una estructura parecida, supusimos que el mejor caso de malo competidor sería parecido al de malo nosotros pero graficamos aun así todo para buscar el mejor punto posible.

Unigramas:



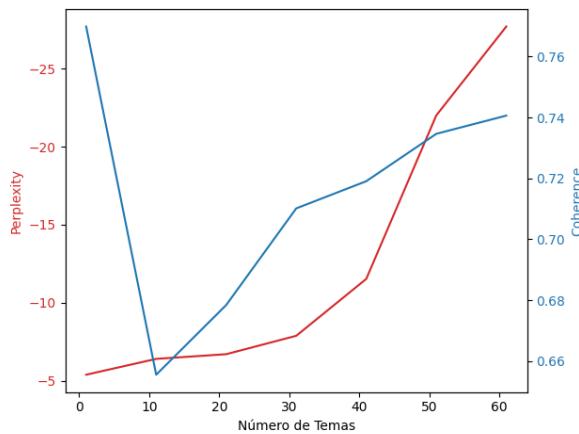
3.38. Figura: Perplexity y coherence c v en malo competidor

Bigramas:



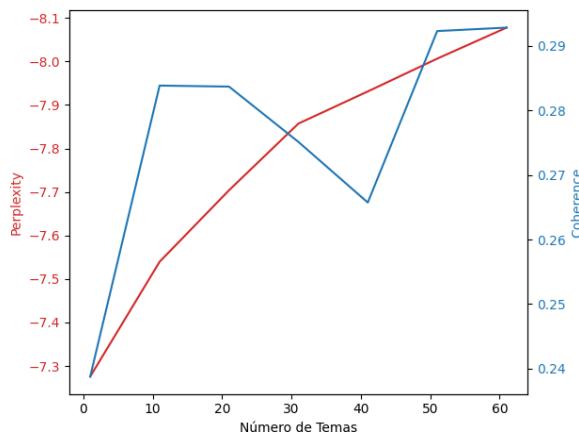
3.39. Figura: Perplexity y coherence c v en malo competidor - Bigramas

Trigramas:



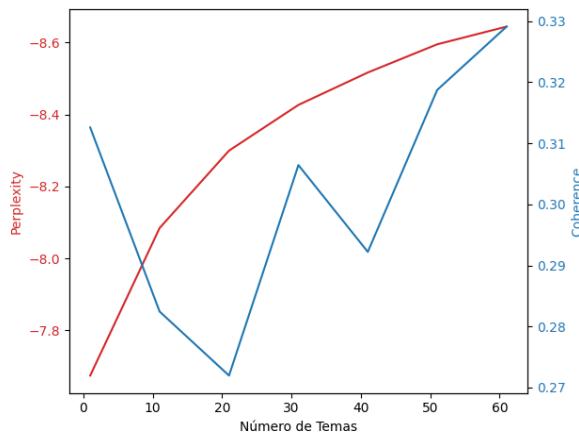
3.40. Figura: Perplexity y coherence c v en malo competidor - Trigramas

Unigramas y Bigramas



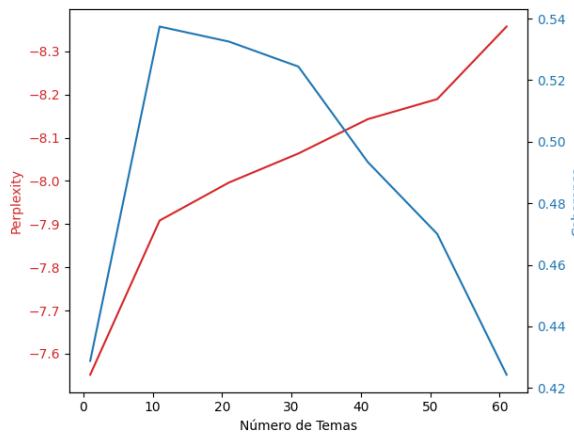
3.41. Figura: Perplexity y coherence c v en malo competidor - Unigramas y Bigramas

Unigramas y Trigramas



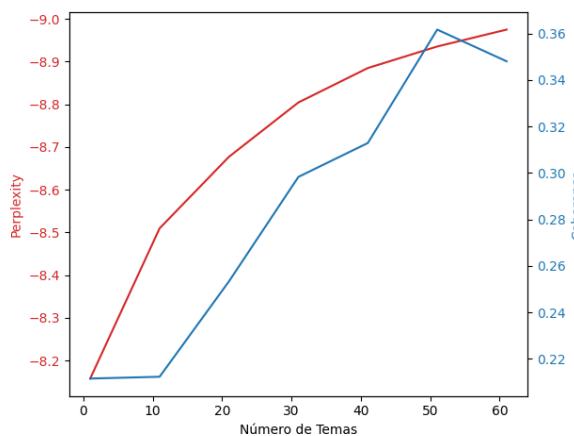
3.42. Figura: Perplexity y coherence c v en malo competidor - Unigramas y Trigramas

Bigramas y Trigramas



3.43. Figura: Perplexity y coherence c v en malo competidor - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas

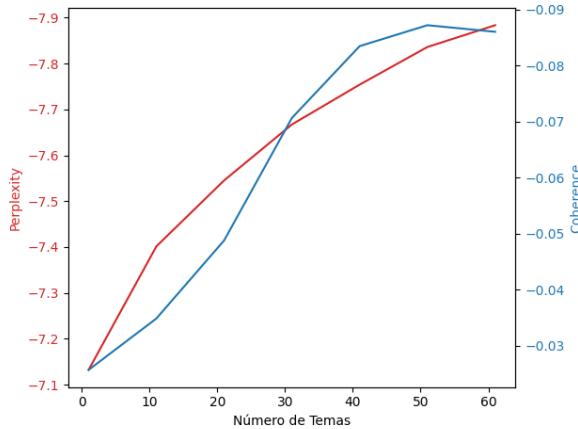


3.44. Figura: Perplexity y coherence c v en malo competidor - Unigramas, Bigramas y Trigramas

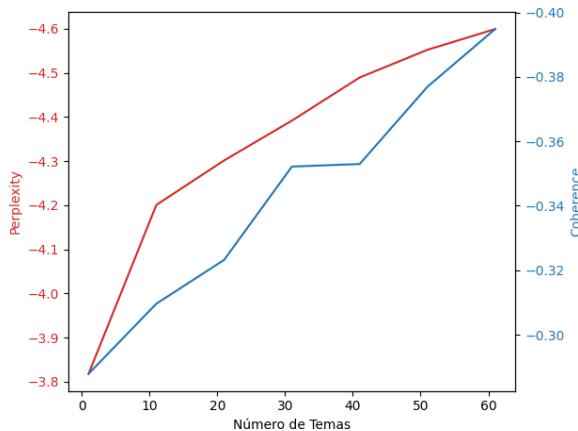
Conclusiones:

Nos sorprendió y fue en contra de nuestra intuición darnos cuenta que el pico estaba sobre 10 temas en bigramas y trigramas así que decidimos graficar con c-npmi para tener un segundo punto de vista.

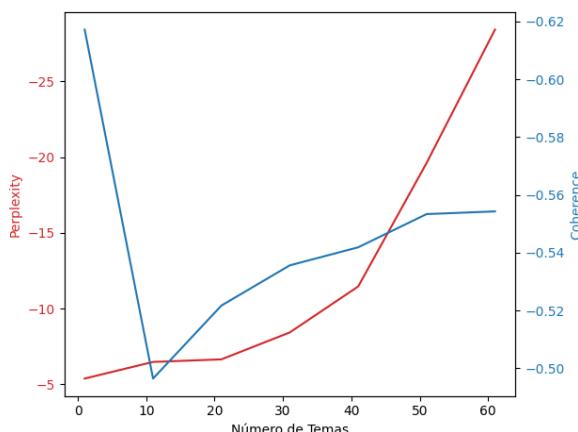
Unigramas:



3.45. Figura: Perplexity y coherence c-npmi en malo competidor

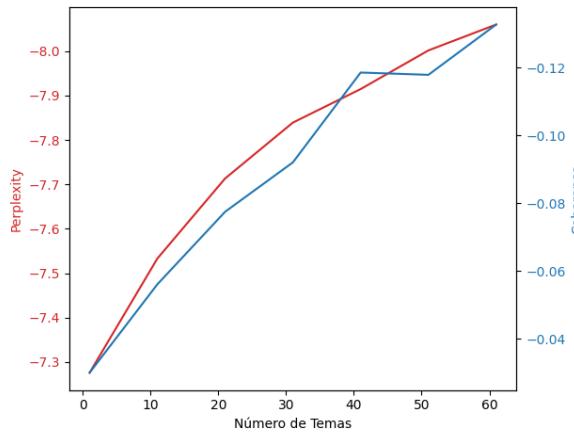
Bigramas:

3.46. Figura: Perplexity y coherence c-npmi en malo competidor - Bigramas

Trigramas:

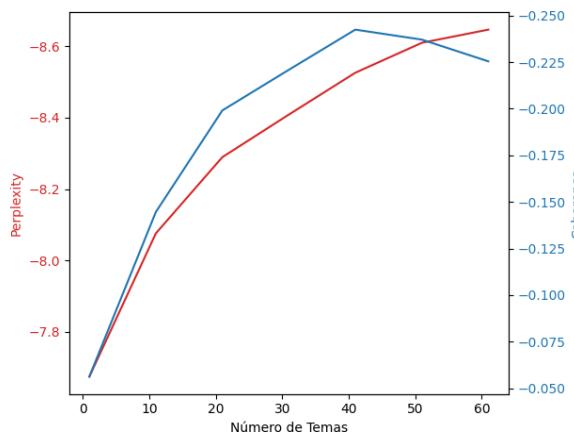
3.47. Figura: Perplexity y coherence c-npmi en malo competidor - Trigramas

Unigramas y Bigramas



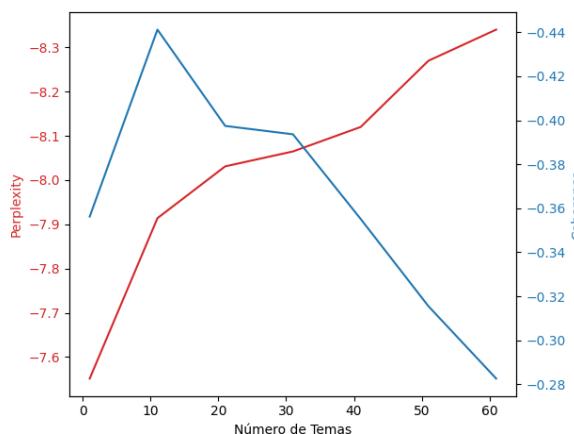
3.48. Figura: Perplexity y coherence c-npmi en malo competidor - Unigramas y Bigramas

Unigramas y Trigramas



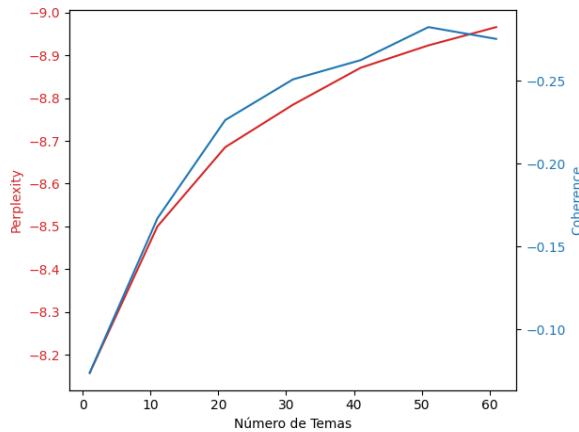
3.49. Figura: Perplexity y coherence c-npmi en malo competidor - Unigramas y Trigramas

Bigramas y Trigramas



3.50. Figura: Perplexity y coherence c-npmi en malo competidor - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas

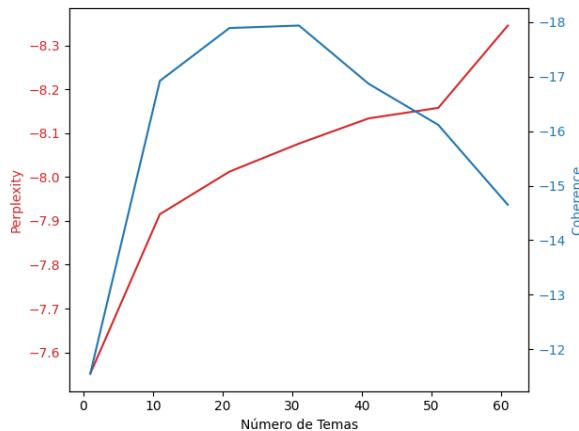


3.51. Figura: Perplexity y coherence c-npmi en malo competidor - Unigramas, Bigramas y Trigramas

Conclusiones:

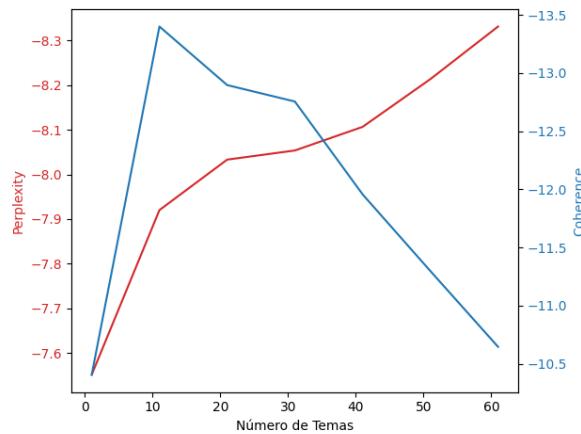
Nos decidimos por bigramas y trigramas en aproximadamente 10 topicos pero aun asi decicimos asegurar mediante u-mass y c-uci

U-MASS



3.52. Figura: Perplexity y coherence U-MASS en malo competidor - Bigramas y Trigramas

C-UCI

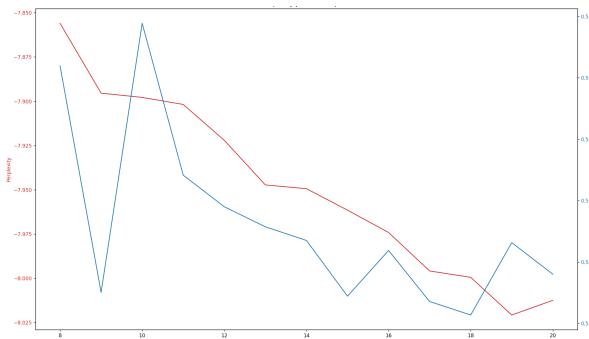


3.53. Figura: Perplexity y coherence C-UCI en malo competidor - Bigramas y Trigramas

Conclusiones:

Nos dimos cuenta que el mejor punto estaba entre 10 y 20 tópicos asique decidimos ver mediante c-v los resultados de 8 a 20 y asi resolver todas las dudas de cual es el mejor punto

Topicos de 8 a 20

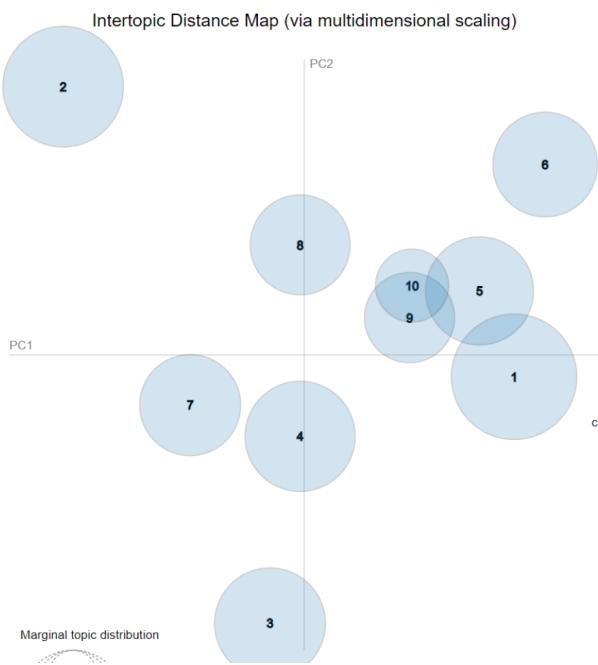


3.54. Figura: c-v de 8 a 20 tópicos en malo competidor

Conclusiones:

Se puede ver un dramático pico en 10 tópicos asique decidimos analizar ese numero exactamente. Finalmente después de varias ejecuciones, este fue el mejor mapa:

Visualización final:



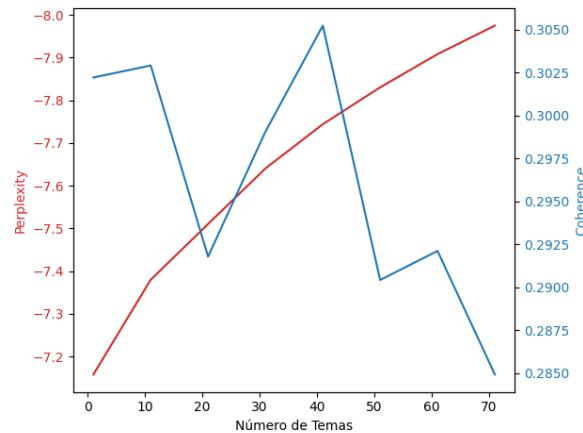
3.55. Figura: Mejor visualización conseguida para malo competidor

Bueno Competidor

En bueno competidor como en los demás graficamos todo para buscar los mejores resultados posibles.

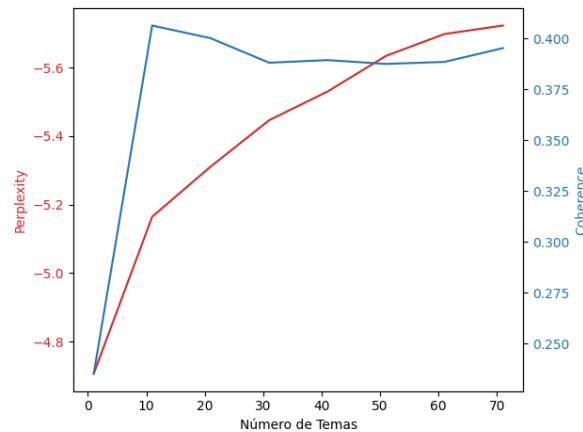
C-V:

Unigramas:



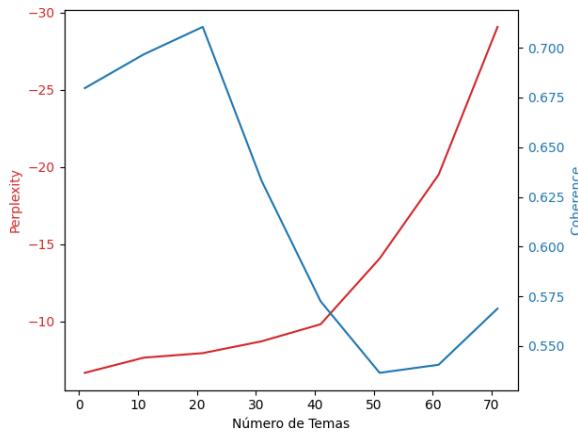
3.56. Figura: Perplexity y coherence c-v en bueno competidor - Unigramas

Bigramas:



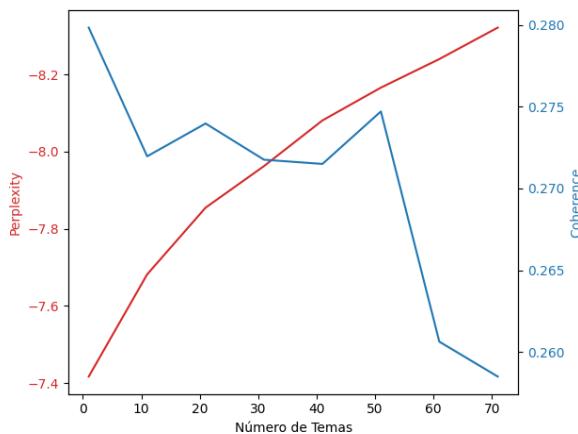
3.57. Figura: Perplexity y coherence c-v en bueno competidor - Bigramas

Trigramas:



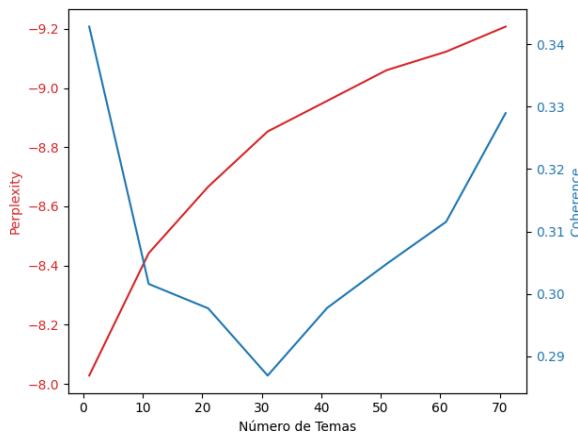
3.58. Figura: Perplexity y coherence c-v en bueno competidor - Trigramas

Unigramas y Bigramas



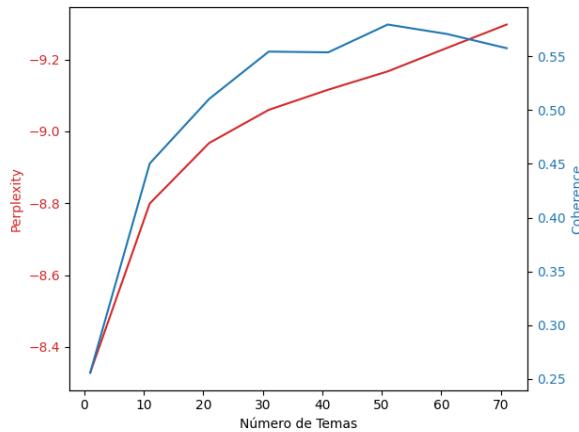
3.59. Figura: Perplexity y coherence c-v en bueno competidor - Unigramas y Bigramas

Unigramas y Trigramas



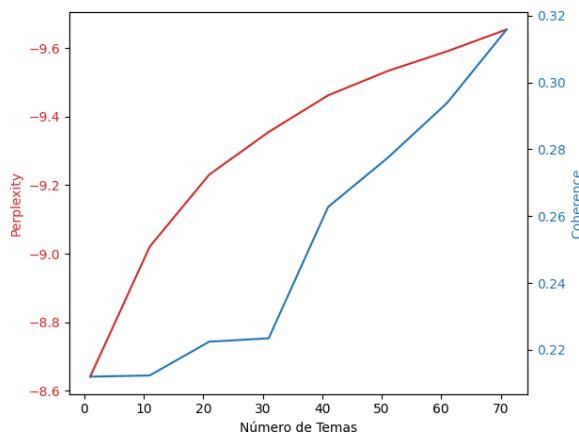
3.60. Figura: Perplexity y coherence c-v en bueno competidor - Unigramas y Trigramas

Bigramas y Trigramas



3.61. Figura: Perplexity y coherence c-v en bueno competidor - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas



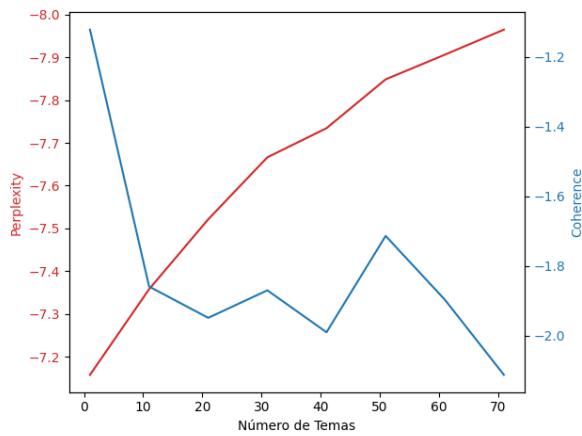
3.62. Figura: Perplexity y coherence c-v en bueno competidor - Unigramas, Bigramas y Trigramas

Conclusiones:

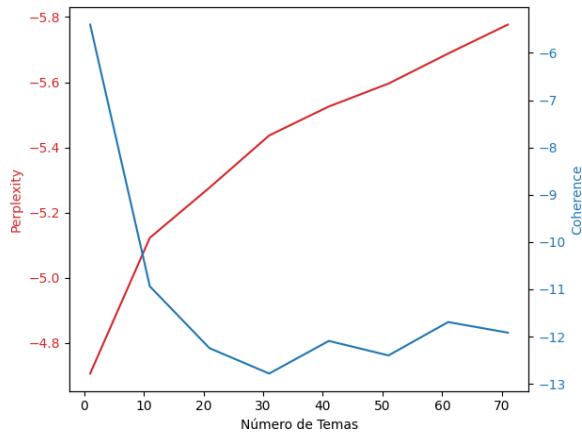
Observando las graficas nos quedamos con la union de bigramas y trigramas como en los demás. Despues de analizarlo los mejores topicos rondan entre los 30 o los 50,no obstante, para asegurarnos hacemos analisis con una coherencia u-mass.

U-MASS:

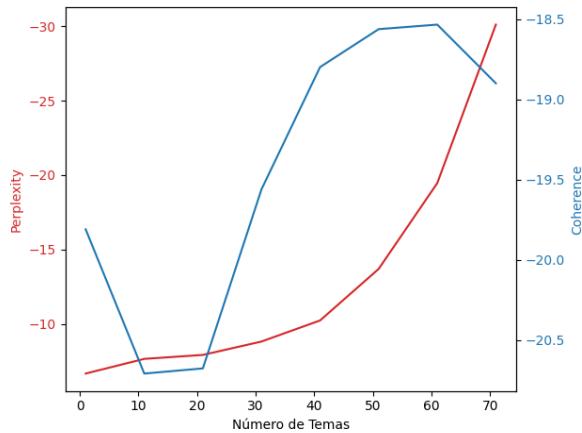
Unigramas:



3.63. Figura: Perplexity y coherence u-mass en bueno competidor

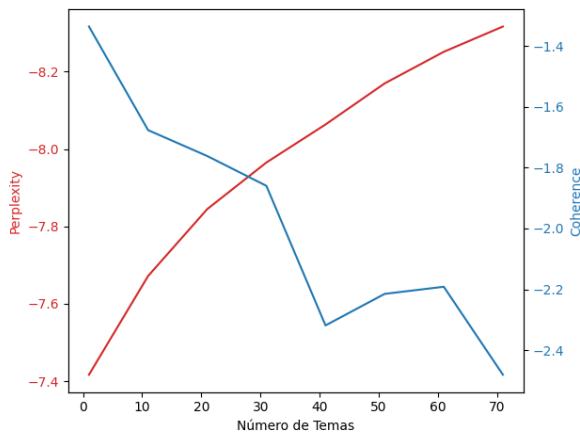
Bigramas:

3.64. Figura: Perplexity y coherence u-mass en bueno competidor - Bigramas

Trigramas:

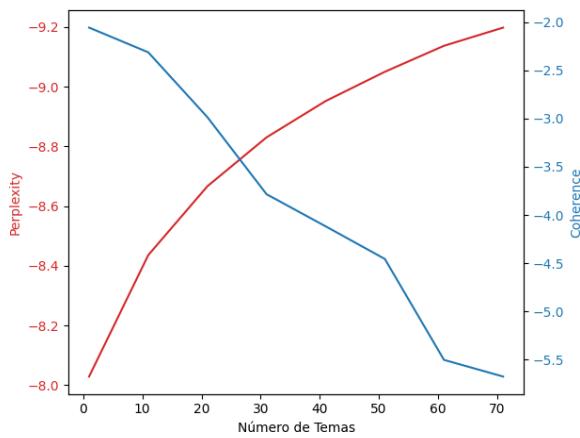
3.65. Figura: Perplexity y coherence u-mass en bueno competidor - Trigramas

Unigramas y Bigramas



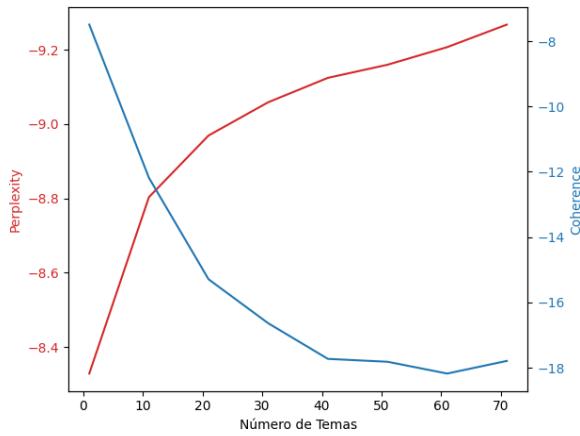
3.66. Figura: Perplexity y coherence u-mass en bueno competidor - Unigramas y Bigramas

Unigramas y Trigramas



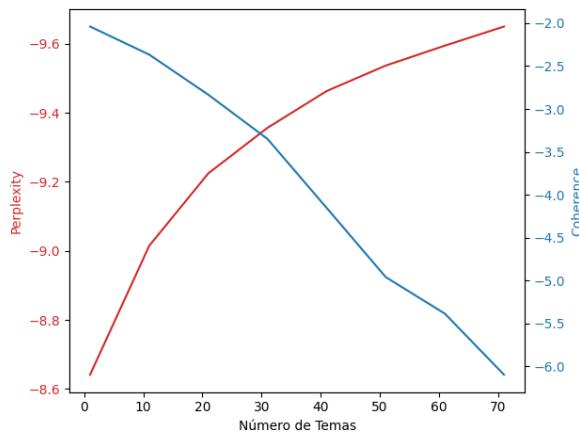
3.67. Figura: Perplexity y coherence u-mass en bueno competidor - Unigramas y Trigramas

Bigramas y Trigramas



3.68. Figura: Perplexity y coherence u-mass en bueno competidor - Bigramas y Trigramas

Unigramas, Bigramas y Trigramas

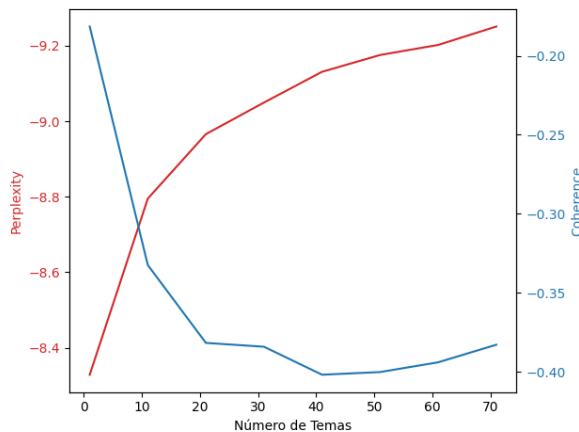


3.69. Figura: Perplexity y coherence u-mass en bueno competidor - Unigramas, Bigramas y Trigramas

Conclusiones:

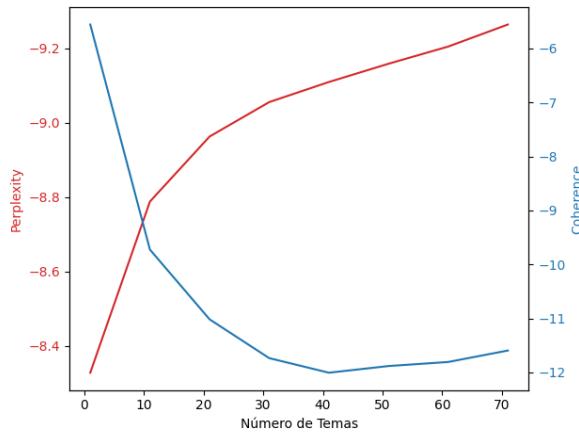
Una vez analizadas estas posibilidades seguimos con duda de cuantos topicos pueden ser los ideales para bigramas y trigramas puesto que 50 topicos pueden ser demasiados. Buscamos otros análisis con una coherencia c-npmi y c-uci.

C-NPMI



3.70. Figura: Perplexity y coherence C-NPMI en bueno competidor - Bigramas y Trigramas

C-UCI

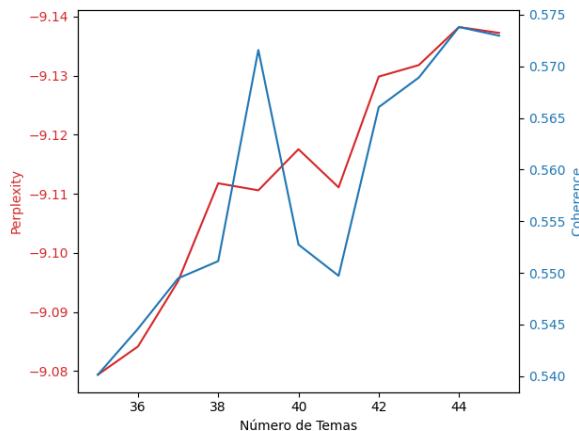


3.71. Figura: Perplexity y coherence C-UCI en bueno competidor - Bigramas y Trigramas

Conclusiones:

Gracias a estos dos ultimos analisis nos queda mas claro cual puede ser el numero de topicos optimo. Vemos un pico en 40 por lo que decidimos hacer un ultimo analisis entre 35 y 45 topicos con coherencia c-v para quedarnos con el mejor numero de topicos.

Topicos de 35 a 45

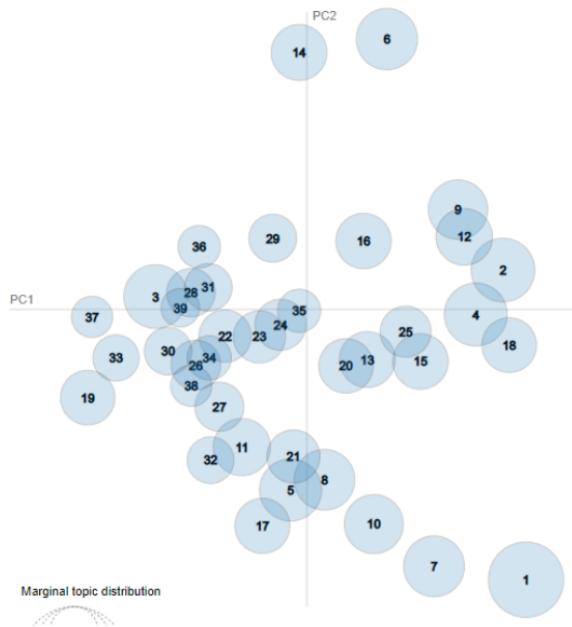


3.72. Figura: c-v de 35 a 45 tópicos en bueno competidor

Conclusiones:

Observando esta ultima gráfica y tras varias ejecuciones entre estos tópicos nos quedamos con 39 tópicos el cual es el que mas coherencia tiene. Este nos da una visualización donde los tópicos tienen una predominación parecida ya que el tamaño de las burbujas es muy parecida.

Visualización final:



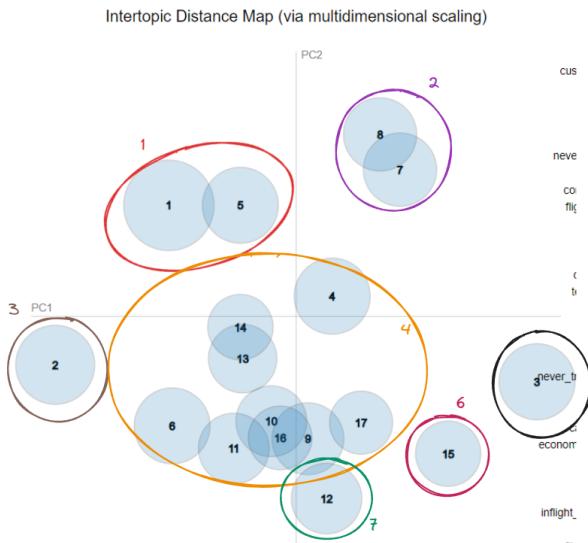
3.73. Figura: Mejor visualización conseguida para bueno competidor

3.1.3. Discusión sobre los descubrimientos realizados en la tarea de Topic Modeling

Malo Nosotros:

Una vez conseguida una visualización legible, pasamos a separar en subtemas:

Visualización final dividida en subtemas:



3.74. Figura: Mejor visualización conseguida separada por subtemas malo nosotros

Explicaciones:

Tema 1: Las palabras principales son "business-class", "cabin-crew" lo cual nos puede decir que habría que mejorar el trato que tiene la tripulación en la zona del avión donde la gente más paga, "business-class-seat" también sigue el tema y nos explica que están descontentos con los asientos. Hay más ngramas dignos de mencionar como "hidden-terms-conditions", "rude-crew", "worst-flight" que indican que el viajar con nuestra aerolínea no se siente bien y la reputación de la aerolínea podría estar sufriendo debido a problemas recurrentes en la calidad del servicio y la experiencia del pasajero. Una palabra que tiene una gran importancia en este tópico es "gluten-free" que sabiendo que es una orientación negativa podría significar que hay problemas con la comida de la aerolínea.

Tema 2: Se vuelve a mencionar "customer-service", "response-broken" dejando claro que a los pasajeros no les satisface la atención al cliente. En este caso se quejan de un "connection-flight" que se refiere lo más seguro a una escala que no se ha cumplido ya que hay otro bigrama "left-stranded", "emirates-failed" lo cual es una falta muy grave de nuestra aerolínea.

Tema 3: En este tema se mencionan las palabras "first-class", "flight-dubai", "tier-miles" lo que se podría significar que los vuelos a destinos de renombre como Dubai y los programas de acumulación de millas, la experiencia proporcionada por la aerolínea no cumple con las expectativas.

Tema 4: A pesar de ser un subtema muy grande, tienen muchos términos en común como "flight-attendant", "inflight-service-poor", "take-responsability", "cancel-flight", "service-desk", "duty-free", "sit-together", "crew-rude", "lack-proffesionalism", "worst-airline-deal" lo cual refleja la visión general de las malas opiniones de nuestra aerolínea, destacando problemas recurrentes en el servicio al cliente, la actitud del personal y la gestión de vuelos, lo que indica una necesidad urgente de mejorar estos aspectos para recuperar la confianza y satisfacción de los pasajeros.

Tema 5: Se vuelve a mencionar el “customer-service” pero tambien se mencionan palabras importantes como ”seat-recline.”, ”food-average” lo que podría significar que habría que comprobar más a menudo los asientos de los aviones y mejorar la comida

Tema 6: Las palabras mas representativas de este tema son: ”poor-form”, ”dissapointment-experience”, ”carry-laptop-bag”, ”hard-productz” ”hope-see”, lo que podría indicar que un pasajero ha perdido su ordenador y a parte de haberle decepcionado el trato, cree que no recuperará el objeto.

Tema 7: Las palabras específicas de este tema son ”connection-counter”, ”lost-found”, ”nothing-retrievedz”, ”aircraft-cleared”, lo que podría significar que los pasajeros enfrentan problemas al tratar de recuperar objetos perdidos o equipaje después de que un avión haya sido despejado, enfrentándose a la ineficacia de los mostradores de escala y de objetos perdidos donde no logran recuperar sus pertenencias, es parecido al anterior tema pero no igual.

¿Y si están separados en 7 subtemas, porque no hacer directamente 7 topicos?

Este razonamiento es intuitivamente válido, en este caso podríamos haber intentado hacer directamente 7 tópicos pero la coherencia era menor a la mitad de la que se conseguía con 17 tópicos.

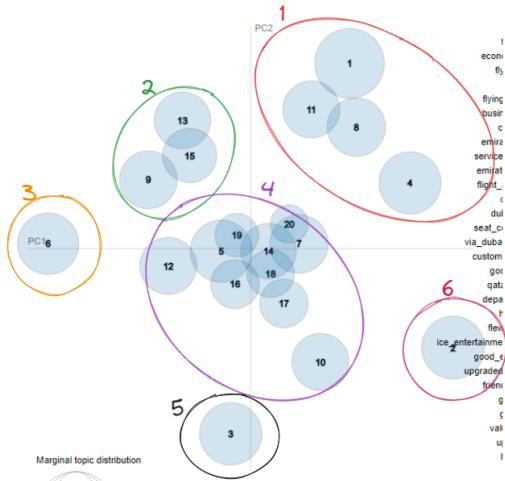
Además, cuando un tópico contiene un número elevado de palabras, existe el riesgo de que palabras importantes sean omitidas, dado que generalmente solo se consideran las más frecuentes. Por este motivo, subdividir los temas en subtemas más descriptivos y específicos puede proporcionar un análisis más detallado y útil.

No obstante, es cierto que en ocasiones, tener demasiados tópicos puede llevar a múltiples interpretaciones del mismo resultado, dependiendo del nivel de tolerancia empleado en el análisis. Esto puede complicar la comprensión y la aplicación práctica de los resultados obtenidos.

Bueno Nosotros:

Una vez conseguida una visualización legible, pasamos a separar en subtemas:

Visualización final dividida en subtemas:



3.75. Figura: Mejor visualización conseguida separada por subtemas bueno nosotros

Explicaciones:

Tema 1: En este tema se mencionan entre otras a destacar debido a su frecuencia de aparicion : "flew-emiratesfood-goodfood-okbar-areaupper-deck". Estas palabras aparecen en todos los topics que incluye el tema lo que nos lleva a ciertas conclusiones. Por lo que se puede analizar la comida que se ofrece en la aerolinea esta bien al igual que su zona de bar por lo que el servicio de comida en este tema es de gran importancia. Tambien con la palabra upper-deck concluimos que la cubierta alta del avion esta bien valorada. Ademas encontramos en algun topico palabras como *incredible-meals.*^{en} relacion a la comida.

Tema 2: En este tema se mencionan palabras como "flight-attendantsstrongly-recommendflight-attendants-friendlyflight-dubai", "excellent-service". Esto nos lleva a pensar que el tema principal es la buena experiencia y amabilidad de los asistentes de vuelo es decir del servicio.

Tema 3: A diferencia de los otros temas en este se menciona con gran frecuencia:"departing-kuwait", ".online-checking.^ademas de "good-expecting.^{en}tre otras. Esto nos dan a entender buenas experiencias con el check-in online y los viajes a Kuwait.

Tema 4: El tema 4 es un tema muy grande con muchos topics no obstante vemos muchos terminos en comun asi encontramos en casi todos sus topics palabras como:entertainment-systemservice-good hot-towelstop-notchfriendly-helpful". Lo que nos dice que en este extenso tema se habla mucho de los servicios, del sistema de entretenimiento de la aerolinea ,que es de gran calidad, y se recalcan pequeños detalles como que las toallas estan calientes para la conformidad de los clientes. Tambien se habla de que la aerolinea es de primera categoria("top notch").

Tema 5: Las palabras principales de este tema no representadas en gran medida en el resto de temas son "hong-kongeverything-gooeverytnig-smooth". Esto nos da a entender que todo esta muy bien en cuanto a experiencias de viaje con Hong-Kong.Ademas se menciona "familyfriendly-service"

Tema 6: En este ultimo observamos palabras que no salen en otros como "self-checking", "pleased-emirates.^ademas de las habituales en las buenas reseñas como "good-flight".Concluimos que se habla de un buen auto check in ademas de estar complacido con Emirates.

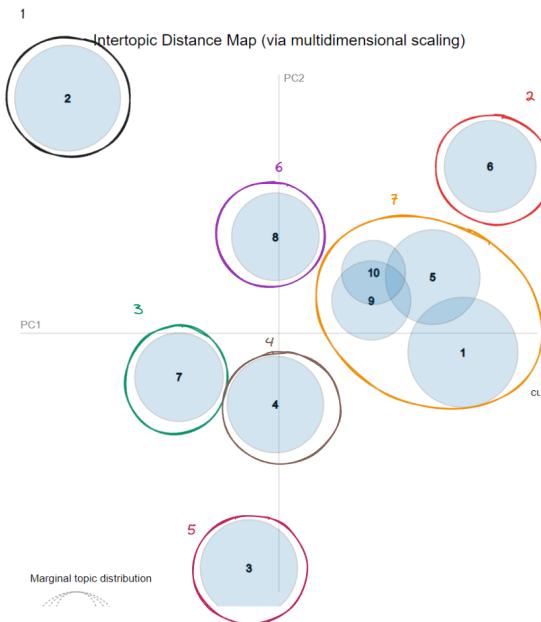
Conclusiones de la división en subtemas en Bueno Nosotros

En cuanto a esta visualización las burbujas tienen un tamaño parecido lo que nos indica que los tópicos tienen una predominación parecido, los únicos mas pequeños son unos pocos del tema 4. Además los tópicos no se solapan en gran medida lo que nos indica que es un buen modelo. El objetivo de la división de subtemas es explicar los temas que se hablan en los tópicos de manera mas general. La conclusión que nos lleva es que al margen de que los tópicos que están en el mismo subtema den información parecida sigue habiendo una diferencia entre ellos, por lo que cada topico habla de un tema diferente.

Malo Competidor:

Una vez conseguida una visualización legible, pasamos a separar en subtemas:

Visualización final dividida en subtemas:



3.76. Figura: Mejor visualización conseguida separada por subtemas malo competidor

Explicaciones:

Tema 1: Se menciona "hours-later", "customer-service", "call-centre", "didnt-offer" lo cual podría significar que ha habido un retraso en un vuelo y que los pasajeros se vieron obligados a lidiar con una atención al cliente deficiente, sin ofrecer compensación.

Tema 2: Se menciona principalmente "via-doha", "flight-doha", puesto que doha es la sede de ésta aerolinea es malo para esta aerolinea. Tambien se menciona "tasted-awful", "gluten-free", "meals-worst", lo que sugiere que las comidas ofrecidas durante el vuelo, especialmente las opciones sin gluten, fueron de baja calidad y posiblemente desagradables para los pasajeros.

Tema 3: Las palabras más importantes son "partial-refund", ".attempts-obtain", "paid-initial", "three-months", "ticket-doesnt", lo que sugiere que un cliente ha experimentado dificultades para obtener un reembolso parcial después de cancelar un billete, a pesar de haber pagado inicialmente y haber intentado obtener la devolución durante tres meses. Esto indica un problema con el servicio al cliente o el proceso de reembolso de la aerolínea.

Tema 4: Se pueden recalcar las palabras "couldnt-change", "capacity-label", ".experienced-bad", "service-board" lo que podría significar que ha habido un caso de overbooking en uno de los vuelos y el cliente ha tenido problemas con la reclamación del importe pagado. Este tema probablemente va de la mano con el tema anterior.

Tema 5: Los temas principales son "business-class", "upper-deck", ".exit-row-seats", ".absolutely-horrible", lo que sugiere que los compradores de la clase ejecutiva están descontentos con el espacio para guardar el equipaje de mano y los asientos cercanos a las salidas de emergencia.

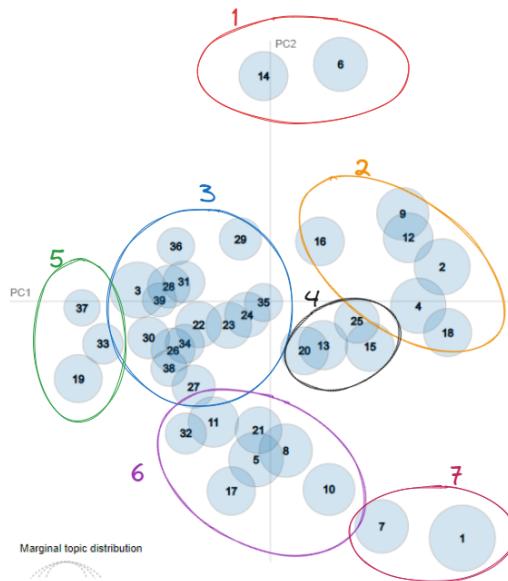
Tema 6: Las palabras más remarcables son "missing-bag", ".^apparently-traced", "praying-bagz" "lost-luggage", lo que significa que quien ha hecho la reseña ha experimentado un problema de pérdida de equipaje con la compañía y ha recibido un servicio deficiente en respuesta a esta situación.

Tema 7: Este tema es bastante variopinto y incluye muchos comentarios que se mencionan dentro del subtema, tales como "ignored-complaint", ".^online-website", "seats-awful", "shortests legroom", "terms-conditions", ".^affected-passengers", "technical-problemz" "denied-entry", lo que recopila desde que un pasajero experimenta dificultades técnicas o malas condiciones durante su vuelo hasta que se enfrenta a problemas de servicio al cliente como la denegación de entrada y quejas ignoradas pasando por que los asientos son pequeños e incomodos.

Bueno Competidor:

Una vez conseguida una visualización legible, pasamos a separar en subtemas:

Visualización final dividida en subtemas:



3.77. Figura: Mejor visualización conseguida separada por subtemas bueno nosotros

Explicaciones:

Tema 1: Estos dos tópicos del subtema son los mas separados de todos, en ello encontramos en común palabras como "staff-friendlyqatar-airways -staffgreat-staff" lo que nos indica que los dos hablan del buen personal de qatar.

Tema 2: Aquí nos encontramos varios topics poco solapados. En la mayoría de ellos nos encontramos con varias palabras que no aparecen en el resto de subtemas: "first-timeground-staffstaff-attentiveservice-excellentinflight-entertainment." entre otras. Esto nos da entender que el personal de tierra es de gran calidad. También se habla de la buena experiencia de primer vuelo así como de el entretenimiento dentro del vuelo. Estos son varios temas que se hablan dentro de este subtema en los diferentes tópicos ya que como mencionado anteriormente estan poco solapados.

Tema 3: Este subtema tiene muchos tópicos con cierta solapacion. A diferencia del resto se habla mucho de la comida. Pues vemos que aparecen en la mayoria o todos de los topics palabras como: "food-served." "food-excellent". Ademas en varios topics se hablan de comidas mas especificas , cranberry-juicered-wine", como el jugo de arandano o el vino tinto. Al tener muchos topics tambien se hablan de otros temas como el servicio o las pantallas táctiles del avión.

Tema 4: Nos encontramos con 4 topics ligeramente solapados. En estos nos volvemos a encontrar que en muchos de ellos se habla de inflight-entertainment. el entretenimiento a bordo. También aparece en gran medida : "seats-comfortable" por lo que los asientos son de gran comodidad y aparece "top-notch" que nos dice que la compañía es de primera categoría.

Tema 5: Observamos tres topics los cuales comparten palabras como "pleasantly-surprisedmade-sure abu-dhabigenuinely-pleasant". Esto nos indica la tranquilidad de viajar en esta compañía que esta asegurada y sobre todo en los viajes a Abu-Dhabi.

Tema 6: Se puede recalcar palabras que no aparecen en el resto de temas como: "covid-policyexceptional-service-facilitiescape-town". Nos lleva a que se recalca las políticas covid llevada así como del facilidades que da el servicio, tambien se menciona la Ciudad Del Cabo.

Tema 7: Este subtema contiene dos topics que estan bastante separados. De estos cabe destacar varias palabras que comparten y que no se observan en el resto de subtemas: "best-servicebest-business-classamazing-service". Concluimos que el tema principal de estos dos subtemas es el servicio de gran calidad en la clase business.

3.1.4. Conclusión de la tarea de Topic Modeling

El ejercicio de topic modelling ha sido útil para identificar algunos hallazgos significativos, aun así, hemos enfrentado numerosos desafíos en la obtención de resultados consistentes. En nuestro proceso, hemos tenido que aplicar muchas veces los algoritmos LDA y NMF hasta conseguir encontrar resultados satisfactorios, y hemos experimentado varios cambios en el camino, a pesar de no mostrar en esta documentación, experimentamos con NFM al igual que con todo tipo de coherencias, solo que no nos pareció necesario mostrarlo.

Durante este proceso, hemos descubierto tanto temas útiles como otros menos relevantes. Esto ha sido especialmente notable en la sección de las reviews de nuestra aerolínea, donde creemos que la cantidad de comentarios disponibles no ha sido tan amplia como para hacer esta tarea de una forma precisa. Como resultado, los tópicos generados no siempre han sido precisos o representativos y han sido muy generales, lo que nos ha llevado a realizar una subdivisión en subtemas para poder adecuar lo maximo posible la decisión del numero de tópicos de las graficas a algo legible.

Juntandolo con la tarea de Tableau, creemos que estos resultados les van a ser utiles pero hasta cierto punto ya que va a ser difícil de representar gráficamente nuestros descubrimientos y algunos apartados son difíciles de interpretar.

4. Bibliografía

4.1. Tableau

Link: Premios SkyTrax

Link: TripAdvisor Emirates

Link: Emirates

Link: IATA Codes

Link: Tableau Documentacion

4.2. Sentiment Analysis

Link: Algoritmos Sentiment Analysis

4.3. Topic Modelling

Link: Lda de Gensim

Link: Coherencia de Gensim

Link: NMF de Gensim