Evaluación de rendimiento en atención al cliente: NLP mediante Bayes y Redes Neuronales Convolucionales en la detección de emociones de conversaciones cliente-empresa.

Carlos Flores (00329746), Jhonatan Quiroga (00330058)

Resumen

Este proyecto explora la aplicación de modelos de aprendizaje automático en la evaluación automatizada de chats de servicio al cliente como respuesta al creciente número de solicitudes de tickets en servicios web. Para esto, se hace uso de la clasificación de mensajes de texto en categorías emocionales (positivas y negativas) como medidor de rendimiento del servicio brindado por un asesor. Los modelos implementados con este objetivo fueron Naive Bayes Multinomial (Inferencia) y redes neuronales convolucionales (Predicción), entrenados en el dataset MELD, un conjunto de diálogos etiquetados por emoción y evaluados sobre interacciones reales de atención al cliente en Twitter.

Palabras clave: PLN, Análisis de sentimientos, Detección de emociones, Atención al cliente, Multinomial Naive Bayes, Redes Neuronales Convolucionales, Solapamiento de dominios, GloVe, TF-IDF.

1. Introducción

La era del consumo digital ha provocado una expansión sin precedentes del alcance de las compañías hacia los consumidores y los servicios que estas pueden ofrecer. Modelos de servicio como el streaming, publicidad, suscripciones, entre muchos otros, han generado economías de escala que han transformado la manera en que los consumidores se relacionan con el consumo. Se ha conseguido así nuevos flujos de ingresos y crecimiento económico dentro de internet.

No obstante, esta nueva dinámica ha implicado nuevos desafíos para las empresas: El crecimiento de los mercados ha sido acompañado por un aumento equivalente de solicitudes de servicio al cliente. Ante esta demanda, las empresas han reaccionado creando departamentos y puestos de asesores para atenderla; pero al ser tanto el volumen diario de interacciones, la misión de poder evaluar detalladamente el rendimiento de cada empleado y hacer las correcciones necesarias se ha vuelto imposible para muchas compañías.

Se ha vuelto imperativo en consecuencia el diseño de enfoques de evaluación en servicio al cliente que sean automatizables, a la par que altamente efectivos al evaluar la satisfacción del cliente. El siguiente trabajo busca cubrir esta necesidad mediante la **propuesta de una métrica** que cumpla con estas características. Así mismo se busca comprender los factores más importantes de las interacciones asesor-cliente exitosas y automatizar la calificación de nuevos inputs con modelos de machine learning.

Bajo esta premisa, se recurrió a diferentes fuentes de datos respecto a los factores que influyen al determinar si un cliente queda satisfecho después de haber recibido atención por parte de una empresa. El estudio concluyó con el descubrimiento de que, más allá de la resolución satisfactoria del requerimiento del cliente, las **emociones** positivas transmitidas por un asesor son uno de los factores más influyentes a la hora de determinar la fidelidad del cliente con la empresa (Pugh, 2001). Estas además tienen la ventaja de ser clasificables en categorías claras, como pueden ser felicidad, tristeza, enojo, etc. Consecuentemente, se pueden configurar modelos de Machine Learning categóricos que puedan automatizar la asignación e interpretación de estas mediante NLP.

Se propone de esta forma el uso de las emociones presentes en un chat, y su clasificación en positivas o negativas, como una medida de desempeño para determinar el rendimiento de una interacción en atención al cliente. La hipótesis central que guiará esta investigación en consecuencia es la siguiente:

"La evaluación efectiva de conversaciones de atención al cliente; mediante el uso de emociones, modelos de machine learning y datasets conversacionales; sí es factible".

2. Métodos

2.1. Modelos implementados

Con el objetivo de explorar efectivamente las capacidades de inferencia y predicción para la métrica propuesta, se implementarán 2 modelos tradicionalmente utilizados en NLP: Multinomial Naive Bayes y Red Neuronal Convolucional. La elección de estos fue realizada al considerar sus buenos rendimientos individuales y complementación mutua, ya que mientras que Naive bayes permitirá una comprensión inferencial de los datos, mientras que el modelo convolucional al pasar por el filtrado de features, posiblemente obtenga un mejor rendimiento predictivo, como ha sido notado anteriormente en la literatura científica (Kim, 2014).

Correspondiendo a estos modelos elegidos, se seleccionaron extractores de datos adecuados a cada modelo según estudios anteriores.

En el caso de Naive bayes se utilizo la técnica de vectorización TF-IDF, la cual cuenta el número de ocurrencias de tokens (palabras individuales transformadas a minusculas) en el texto para generar una matriz ponderada donde las palabras poco significativas (usualmente artículos) tienen menor peso (Salton & Buckley, 1988). Cumple de esta forma una función de removedor de ruido, pues le da menos importancia a las palabras más repetidas, como artículos, preposiciones y conjunciones. Se busca de esta forma beneficiar las capacidades del modelo al permitirle considerar exclusivamente los predictores más importantes, algo significativo al ser este un problema de alta dimensionalidad.

Para la red convolucional se procederá usando GloVe como técnica de vectorización. GloVe es un modelo que busca relaciones semánticas entre las palabras, convierte cada palabra en un vector de n dimensiones y busca las palabras parecidas en base a este vector (**penningtonglove**). Para esto se debe tokenizar previamente la oración con el objetivo de que pueda ser procesada. Se

decidió usar Glove debido a la habilidad de este para capturar la semántica y estructura de las oraciones que conforman una oración, algo indispensable para una red convolucional al depender de la extracción de las capas de características para poder ser efectivo.

2.2. Datasets seleccionados

El entrenamiento y evaluación de los modelos examinados fue posible gracias al uso de 2 fuentes de datos. El primero es el dataset MELD, compuesto por extractos de diálogos de la serie *FRIENDS* clasificados en siete emociones (alegría, tristeza, enojo, miedo, sorpresa, disgusto y neutralidad). Si bien estos provienen de un contexto ficticio y dominio diferente al de la atención al cliente, su estructura de dialogo y la calidad del etiquetado en una amplia cantidad de circunstancias hacen posible hipotetizar su suficiencia para entrenar modelos de detección emocional en texto. La validez de esta suposición se evaluará posteriormente en discusión. De igual forma vienen etiquetados los sentimientos de la oración como positivo, negativo y neutro.

El segundo dataset corresponde a interacciones reales entre clientes y empresas en Twitter obtenidas en el repositorio de Kaggle y usadas con el objetivo de desempeñar de test group (Axelbrooke, 2017). Para ello, los mensajes presentes en este fueron etiquetados con ayuda de un LLM y corregidos manualmente usando las mismas categorías que los presentes en el dataset MELD. Se posibilitará de esta forma verificar el nivel de precisión de los modelos al hacer predicciones.

2.3. Procedimiento realizado

Con los fundamentos de la investigación cubiertos, se procederá a describir los pasos seguidos para el desarrollo del trabajo:

2.3.1. Preprocesamiento y análisis de los datos

Antes de poder realizar cualquier acción de preprocesamiento, fue necesario etiquetar los datos de los chats del dataset de servicio al cliente para poder gestionarlos. Para esto, se hizo uso de un LLM, específicamente ChatGPT para poder facilitar el etiquetado de la gran cantidad de datos. El resultado de este procedimiento fue revisado y corregido de manera manual por los investigadores para asegurar el correcto clasificado del grupo de testeo, en caso de ser necesario se tradujeron las oraciones al inglés. Se tuvo también en consideración los escenarios donde los textos transmitían más de una emoción, o emociones por fuera de las categorías dadas; en estos casos se escogió la emoción predominante o la más cercana.

Una vez etiquetados todos los datos se procedió a convertir los diálogos y mensajes en vectores numéricos, para esto se tuvo que preparar previamente el texto con una limpieza. En este se incluyó la remoción de signos de puntuación, extensión de abreviaturas, eliminación de jergas y poner en minúscula a las palabras. Adicionalmente en el caso de GloVe se tuvo que dividir en tokens el texto, puesto que TF-IDF ya lo hacía automáticamente. Este proceso es vital para evitar problemas de rendimiento al entrenar los modelos (Kowsari et al., 2019). Cabe rescatar eso si que se evitó lematizar las palabras y eliminar las *stop words*, esto para no perder la capacidad de identificar patrones en los verbos en el caso de la red convolucional.

Acabado el saneamiento de los datos, se procedió a aplicar los vectorizadores correspondientes a cada modelo (TF-IDF y GloVe).

Para el caso de TF-IDF (Bag of Words), su aplicación sobre el dataset preprocesado resultó en un vector único para cada chat, en el cuál se especifica el peso de cada palabra dentro de la oración basado en la frecuencia de aparición (TF) e importancia real (IDF).

En el caso de GloVe (Word embedding) fue necesario utilizar un modelo preentrenado sobre un corpus de palabras estándar, debido a la extensivo trabajo requerido para obtener un buen modelo. Se seleccionó así el modelo del "Stanford University Natural Language Processing Group"(Pennington et al., 2014). Este se guardaría como un diccionario con el que se mapearían las palabras de cada una de las oraciones dentro del dataset MELD a un vector de embedding. Finalmente, todos los vectores pertenecientes a una palabra fueron promediados, obteniendo de esta forma un vector 300×1 con la información condensada de la oración.

Con los datos ya procesados, se procedió a realizar el análisis estadístico de los datos. Al haber extraído los grupos de entrenamiento y testeo de diferentes datasets, se tomó la decisión de realizar las siguientes pruebas estadísticas:

- **Análisis de la distribución de clases:** Se comparó la frecuencia de cada categoría en ambos datasets mediante gráficos de barras. Esto permitirá determinar si existirá sesgo por parte del modelo clasificador, si es necesario usar medidas como F1-ponderado y la naturaleza de los dominios comparados.

- **Análisis de las propiedades textuales:** Se analizaron las longitudes de los diálogos (número de tokens) en cada dataset mediante histogramas y métricas descriptivas (media, desviación estándar). Esto se hará así para verificar si los datasets tienen cercanía de longitud de palabras, algo influyente en la capacidad de encontrar patrones en CNN y la dispersión de la matriz TF-IDF. Además, estos datos permitirán determinar la complejidad de cada dominio.

- **Análisis léxico y de vocabulario:** Se evaluó la coincidencia o divergencia entre los vocabularios presentes en los dominios. Para ello, se calculó el porcentaje de palabras únicas del dataset de Twitter que estaban presentes en el vocabulario de MELD, se determino el tamaño de cada corpus y se extrajeron las 30 palabras más frecuentes de cada dataset. De esta manera, se determinó la proporción de solapamiento entre los dos dominios, un factor determinante para el exito tanto de naive bayes como de la red convolucional (La primera depende de la presencia de palabras de los datos de entrenamiento en el test, mientras que la segunda no podría identificar patrones dentro del dataset twitter)

2.3.2. Entrenamiento y Evaluación

Concluida la fase de preprocesamiento y análisis de los datasets seleccionados, se procederá a realizar el entrenamiento y posterior evaluación de los modelos. Para ello, se empezó realizando una búsqueda de hiperparámetros en el caso de Naive Bayes y estructurado de las capas en el caso del CNN:

- Para el caso de Naive Bayes multimodal se procedió a realizar, dentro del dataset Meld, k-folds cross validation con 5 folds, en búsqueda del valor $alfa$ entre 0.5 hasta 3. La métrica objetivo empleada para esto fue accuracy.
- Para el caso de Redes convolucionales se definión una estructura de la red convolucional basada en (Kowsari et al., 2019), bajo la cual se entrenaron los modelos de emociones y sentimientos. Las características de esta fueron:
    - Presencia de 3 capas convolucionales 1D, con kernels decrecientes (7, 5, 3) para una mejor exploración de los patrones dentro del texto.
    - Presencia de 3 capas Maxpooling con un tamaño relativamente pequeño (2), para poder obtener información de los datos sin recaer en una perdida elevada de información.
    - MLP de 2 capas fully connected para el procesamiento de datos con número de salida equivalente al número de clasificaciones (7 para emociones y 3 para sentimientos).

Con la estructura de ambos modelos definida, se procedió a ajustarlos utilizando los datos del dataset MELD.

Por motivos que se verán más adelante en la sección de discusión, se tomó la decisión de entrenar dos grupos de modelos (tanto Naive Bayes como CNN) utilizando dos conjuntos de datos distintos: Un grupo basado en el datset original MELD, y un grupo basado en el dataset MELD sin la categoría "Neutral" tanto en emociones como sentimientos. Estos grupos serían evaluados de manera separada con el objetivo de comprobar si la remoción de estas clases mejoraba el rendimiento de los modelos. Cabe rescatar eso si que este proceso redujo considerable la cantidad de datos de ambos datasets, algo que se tomará en consideración al analizar los resultados.

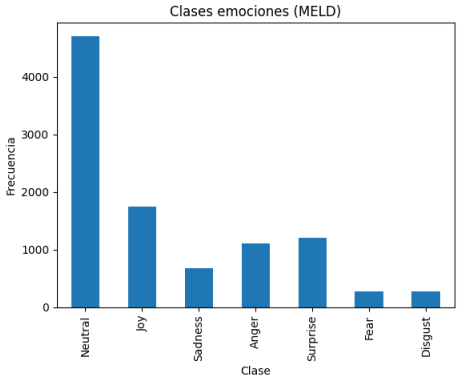
Finalmente, una vez entrenados los modelos se procedió a evaluarlos contra el dataset de atención al cliente de twitter. Para ello, se diseñaron matrices de confusión para verificar la correcta clasificación de los mensajes con la emoción-sentimiento esperado, y se midió el *accuracy* obtenido por parte del modelo.

3. Resultados

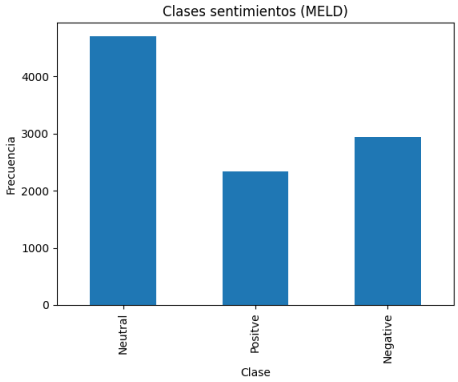
3.1. Análisis comparativo

3.1.1. Distribución de clases

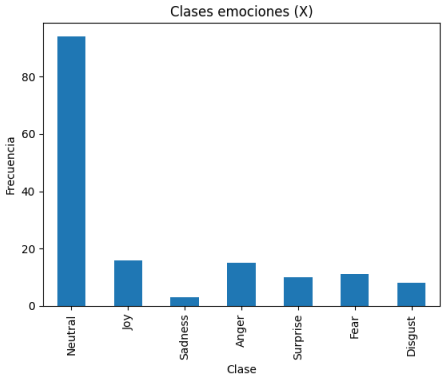
Se empezará observando los resultados obtenidos para la sección del análisis exploratorio de los datos. Se empezará visualizando la distribución de clases para ambos datasets (MELD y twitter):



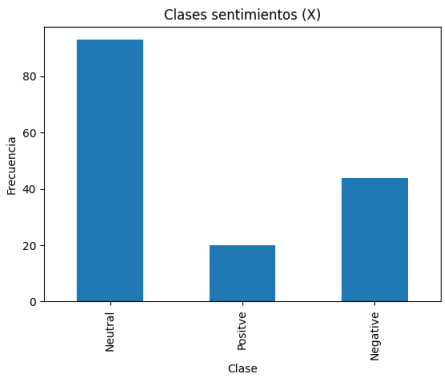
(a) MELD: Clases emoción



(b) MELD: Clases sentimiento



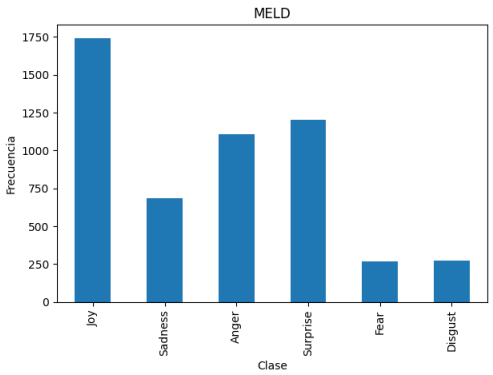
(c) Twitter: Clases emoción



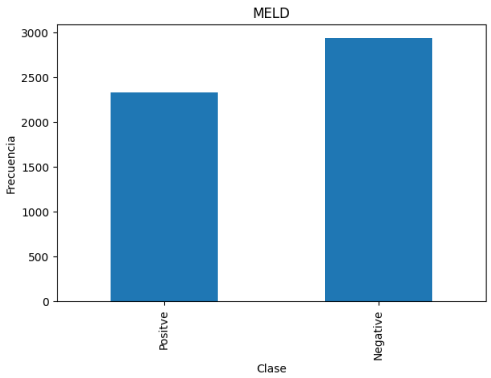
(d) Twitter: Clases sentimiento

Figura 1: Balance de clases para MELD y Twitter en emociones y sentimientos

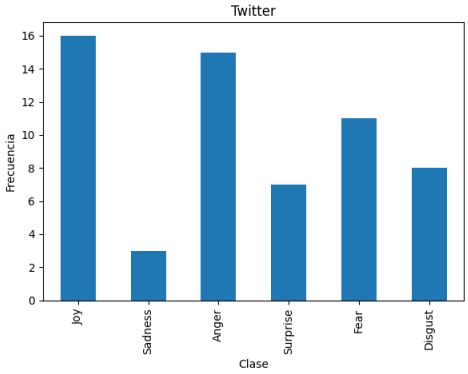
También se incluye la distribución de clases para los datasets después de removerse de estos la clase “Neutral”:



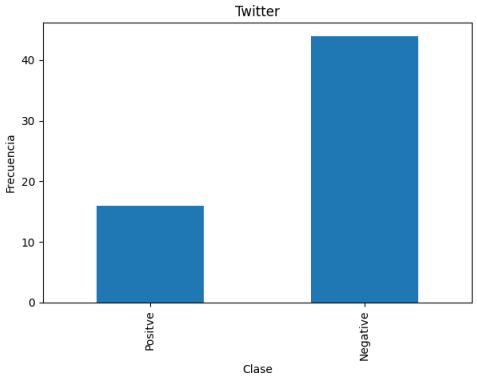
(a) MELD: Emoción sin neutral



(b) MELD: Sentimiento sin neutral



(c) Twitter: Emoción sin neutral



(d) Twitter: Sentimiento sin neutral

Figura 2: Distribución de clases MELD y Twitter sin clasificación neutral en emociones y sentimientos

3.1.2. Longitud de palabras

Siguiente a esto, se obtuvieron las gráficas descriptivas (caja de bigotes, histogramas y diagrama de violín) para cada dataset. Los resultados se incluyen más abajo

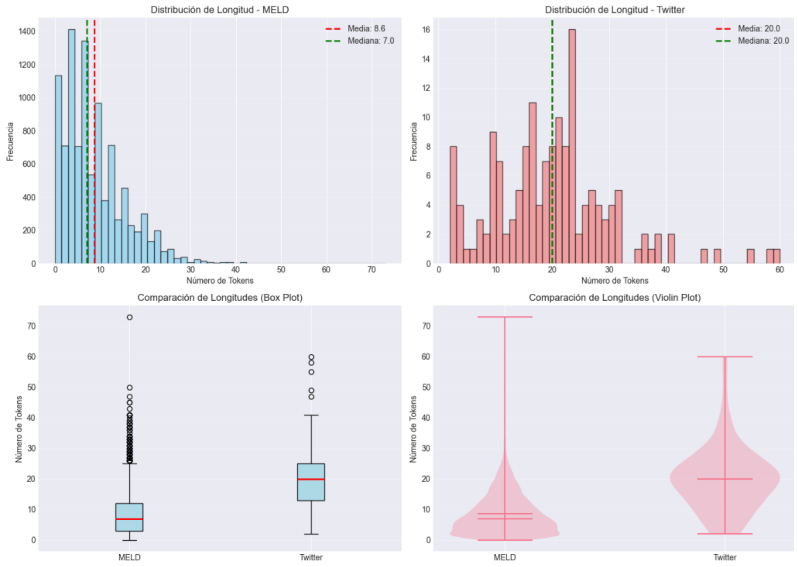


Figura 3: Tipo de palabras

3.1.3. Solapamiento entre ambos datasets

Para culminar la parte de análisis exploratorio se obtuvieron las gráficas referentes al solapamiento entre ambos datasets. Se graficó un histograma representando el solapamiento de palabras entre datasets, top 15 palabras por dataset, y una comparación del tamaño de ambos corpus vs la proporción de palabras compartidas. Los resultados se incluyen a continuación.

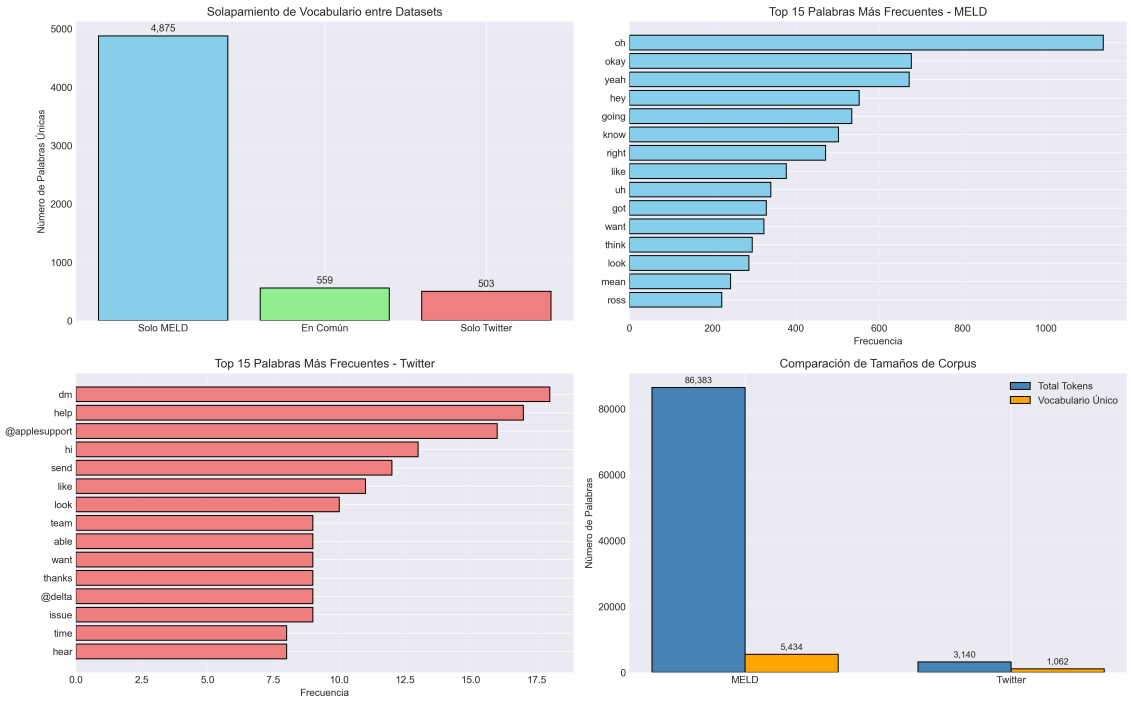


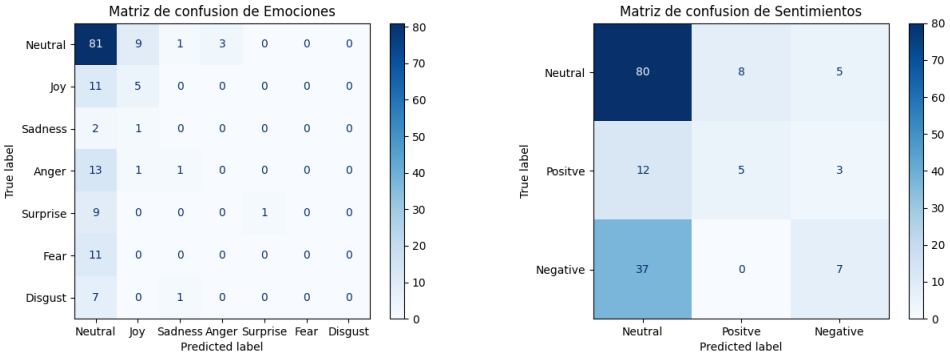
Figura 4: Solapamiento de palabras

3.2. Evaluación

3.2.1. Dataset MELD completo (todo el conjunto de clases)

Naive Bayes en dataset completo

Siguiente al análisis de datos se revisarán los resultados obtenidos después de la evaluación de los modelos. Se empezará examinando las matrices de confusión obtenidas para Naive Bayes.



(a) Matriz emociones

(b) Matriz sentimientos

Figura 5: Matriz de confusión Bayes

También se anexan en la siguiente tabla los resultados de *accuracy* obtenidos tanto en emociones como sentimientos:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.60	0.86	0.71	94
Joy	0.31	0.31	0.31	16
Sadness	0.00	0.00	0.00	3
Anger	0.00	0.00	0.00	15
Surprise	1.00	0.10	0.18	10
Fear	1.00	0.00	0.00	11
Disgust	1.00	0.00	0.00	8
Accuracy			0.55	157
Macro Avg	0.56	0.18	0.17	157
Weighted Avg	0.58	0.55	0.47	157

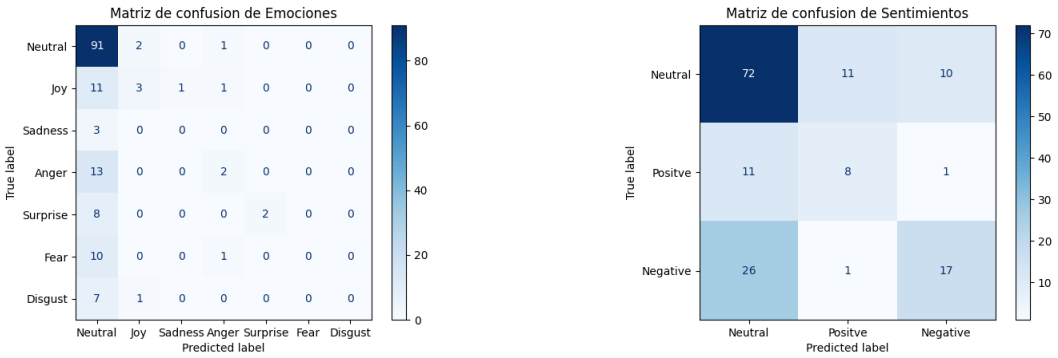
Cuadro 1: Reporte de Naive Bayes - Emociones (Completo)

Class	Precision	Recall	F1-Score	Support
Neutral	0.62	0.86	0.72	93
Positive	0.38	0.25	0.30	20
Negative	0.47	0.16	0.24	44
Accuracy			0.59	157
Macro Avg	0.49	0.42	0.42	157
Weighted Avg	0.55	0.59	0.53	157

Cuadro 2: Reporte de Naive Bayes - Sentimientos (Completo)

Red neuronal convolucional en dataset completo

Siguiente a esto se procederá a enlistar los resultados de la red convolucional. Las matrices de confusión obtenidas para emociones y sentimientos se anexan a continuación:



(a) Matriz emociones

(b) Matriz sentimientos

Figura 6: Matriz de confusión CNN

Class	Precision	Recall	F1-Score	Support
Neutral	0.64	0.97	0.77	94
Joy	0.50	0.19	0.27	16
Sadness	0.00	0.00	0.00	3
Anger	0.40	0.13	0.20	15
Surprise	1.00	0.20	0.33	10
Fear	1.00	0.00	0.00	11
Disgust	1.00	0.00	0.00	8
Accuracy			0.62	157
Macro Avg	0.65	0.21	0.22	157
Weighted Avg	0.65	0.62	0.53	157

Cuadro 3: Reporte de Redes convolucionales - Emociones (Completo)

Finalmente, el *accuracy* obtenido al clasificar las emociones y sentimientos se incluye a continuación:

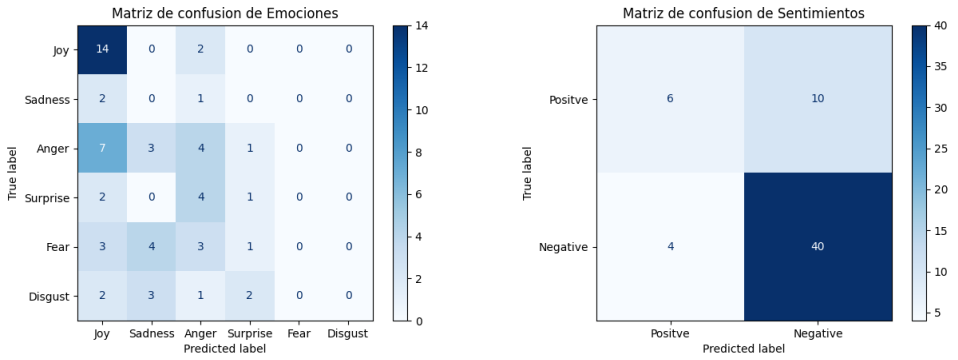
Class	Precision	Recall	F1-Score	Support
Neutral	0.66	0.77	0.71	93
Positive	0.40	0.40	0.40	20
Negative	0.61	0.39	0.47	44
Accuracy			0.62	157
Macro Avg	0.56	0.52	0.53	157
Weighted Avg	0.61	0.62	0.61	157

Cuadro 4: Reporte de Redes convolucionales - Sentimientos (Completo)

3.2.2. Dataset MELD modificado (Eliminación *Neutral*)

Naive Bayes en dataset sin "Neutral"

Una vez finalizada la examinación de los resultados obtenidos al entrenar los modelos con el dataset MELD completo, se procederá a enlistar los datos obtenidos al remover de las clases disponibles la clasificación "Neutral". Se empezará examinando las matrices de confusión obtenidas para el modelo Naive bayes.



(a) Matriz emociones sin neutral

(b) Matriz sentimientos sin neutral

Figura 7: Matriz de confusión Bayes sin neutral

También se anexan en la siguiente tabla los resultados de accuracy obtenidos tanto en emociones como sentimientos:

Class	Precision	Recall	F1-Score	Support
Joy	0.47	0.88	0.61	16
Sadness	0.00	0.00	0.00	3
Anger	0.27	0.27	0.27	15
Surprise	0.20	0.14	0.17	7
Fear	1.00	0.00	0.00	11
Disgust	1.00	0.00	0.00	8
Accuracy			0.32	60
Macro Avg	0.49	0.21	0.17	60
Weighted Avg	0.53	0.32	0.25	60

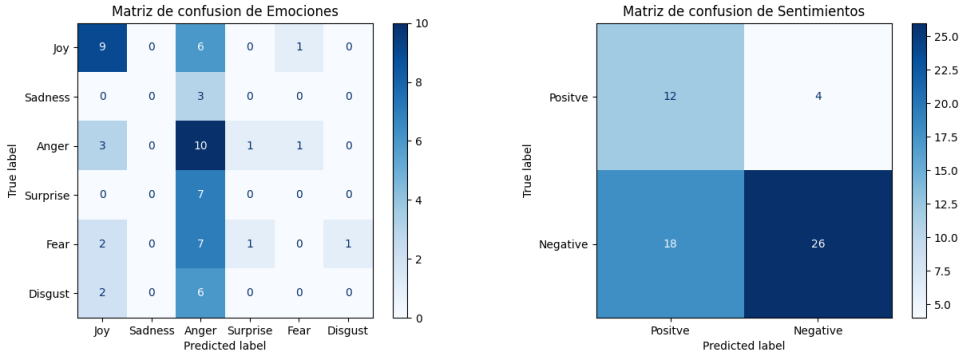
Cuadro 5: Classification Report

Class	Precision	Recall	F1-Score	Support
Positive	0.60	0.38	0.46	16
Negative	0.80	0.91	0.85	44
Accuracy			0.77	60
Macro Avg	0.70	0.64	0.66	60
Weighted Avg	0.75	0.77	0.75	60

Cuadro 6: Classification Report

Red neuronal convolucional en dataset sin "Neutral"

Siguiente a esto, se procederá a enlistar los resultados obtenidos para la red neuronal convolucional sin "Neutral". Las matrices de confusión obtenidas para emociones y sentimientos se anexan a continuación:



(a) Clases emoción sin neutral

(b) Clases sentimiento sin neutral

Figura 8: Matriz de confusión CNN sin neutral

Finalmente, el accuracy obtenido al clasificar las emociones y sentimientos se incluye a continuación:

Class	Precision	Recall	F1-Score	Support
Joy	0.56	0.56	0.56	16
Sadness	1.00	0.00	0.00	3
Anger	0.26	0.67	0.37	15
Surprise	0.00	0.00	0.00	7
Fear	0.00	0.00	0.00	11
Disgust	0.00	0.00	0.00	8
Accuracy			0.32	60
Macro Avg	0.30	0.20	0.16	60
Weighted Avg	0.26	0.32	0.24	60

Cuadro 7: Reporte de Redes convolucionales - Emociones (Sin Neutral)

Class	Precision	Recall	F1-Score	Support
Positive	0.40	0.75	0.52	16
Negative	0.87	0.59	0.70	44
Accuracy			0.63	60
Macro Avg	0.63	0.67	0.61	60
Weighted Avg	0.74	0.63	0.65	60

Cuadro 8: Reporte de Redes convolucionales - Sentimientos (Sin Neutral)

4. Discusión

4.1. Análisis exploratorio

Al revisarse los hallazgos obtenidos del análisis exploratorio se descubrieron varios indicadores de inconsistencias entre ambos datasets que en gran medida revelaron las múltiples dificultades que se iba a encontrar el experimento.

Empezando con la distribución de clases, se encontró que, si bien ambos datasets compartían ciertas características (como una gran cantidad de datos neutrales tanto en emociones como sentimientos), la realidad era que ambos datasets poseían balances bastante diferentes. Esto se pudo observar por ejemplo en la elevada cantidad de datos clasificados como ira en las emociones de Twitter, así como la mayor presencia de sentimientos negativos dentro del mismo dataset. Se reveló de esta forma que los modelos posiblemente presentarían problemas al clasificar en Twitter, ya que las elevadas proporciones de emociones y sentimientos positivos de MELD sesgaría las predicciones con tendencias ausentes en Twitter.

Sin embargo, incluso más importante que este hallazgo fue encontrar que los datos neutrales opacan al resto de sentimientos y emociones de interés. Para mitigar este riesgo, se tomó la decisión vista en metodología, se realizaron 2 conjuntos de entrenamiento y prueba: Uno con el dataset completo y otro retirando la clasificación "Neutral" tanto de emociones como de sentimientos. Como se verá a continuación, esta propuesta no mejoró significativamente los resultados, pero permitió comprender mejor el comportamiento de los modelos, logrando de esta forma una mejor descripción de los datos. Cabe destacar que este procedimiento redujo considerablemente la cantidad de datos que poseídos, siendo que al eliminar la neutralidad se borraron 4710 datos de MELD y 94 de Twitter. Este hecho posiblemente impacte negativamente a los modelos y por ende a los resultados obtenidos.

Siguiente a la distribución de los datos estuvo la exploración de longitud de palabras de ambos datsets. En este se pudo encontrar que Twitter, al presentar situaciones con contextos mucho más técnicos que una conversación casual, presentó en promedio palabras con extensiones el doble de largas que las vistas en MELD. De hecho si se explora las gráficas de caja de bigotes y violín obtenidas también se encuentra que MELD presenta una elevada concentración de palabras por debajo de la media, mientras que la dispersión de Twitter es mucho más focalizada alrededor de la media. Se halló de esta forma que los datasets posiblemente divergen en cuanto al tipo de palabras usadas y la complejidad de las acotaciones, algo sumamente negativo para ambos modelos.

No obstante, el estudio estadístico que verdaderamente reveló la extensión de esta problemática fue "Solapamiento de ambos datasets". Este mostró que los corpus de palabras presentes en cada

dataset compartían muy pocas palabras entre si (solo 559), que estos no poseían palabras comunes entre sus top 15 más frecuentes, y que las pocas palabras compartidas que poseían representaban una fracción pequeña de cada dataset. Todos estos descubrimientos terminaron por confirmar lo supuesto: Los dominios de los datos se solapan en muy baja medida, teniendo contextos, nomenclaturas y patrones muy diferentes entre si. Quedó claro consecuentemente que los modelos que se iban a poder entrenar y aplicar iban a resultar deficientes para la tarea prevista, puesto que Naive Bayes requiere la presencia de palabras de su entrenamiento en el grupo de testeo para poder clasificar, mientras que los patrones encontrados por las redes convolucionales tambien necesitan de la presencia de palabras comunes o similares.

4.2. Evaluación de los modelos

4.2.1. Dataset completo - Naive bayes

Los resultados para Naive bayes con el dataset sin modificaciones presentaron un anticipo de lo que cabría esperar después del análisis exploratorio. Por ejemplo, las matrices de confusión procedentes de sentimientos y emociones mostraron un rápido vista inicial de los modelos: La mayoría de clasificaciones realizadas tanto para sentimientos como emociones fueron categorizadas dentro de "Neutral", indicando caso omiso de las características de cada oración y las características del resto de clasificaciones. Las diagonales de las matrices también se encuentran mayoritariamente vacías.

Se verificó de esta forma que:

- Los problemas de falta de solapamiento entre datasets tuvieron resultados negativos sobre la capacidad de reconocimiento de los modelos
- La falta de balanceo de clases entre los datasets resultó en los modelos clasificando para la clase con mayor concentración (neutralidad).

Se encuentra que el *accuracy* general para emociones fue de 55 %, mientras que para sentimientos fue de 59 %. En ambos casos este se puede considerar como un puntaje bajo, y este número en si mismo se encuentra inflado, ya que al categorizar la gran mayoría de predicciones como neutrales el modelo predice correctamente los datos neutrales, aunque sea por los motivos incorrectos.

Cabe resaltar eso sí que en el caso de los sentimientos que el modelo logra clasificar ligeramente mejor, teniendo al menos un valor numérico diferente de 0 dentro de la diagonal de la matriz de confusión. Se hipotetiza que esto pueda ser al encontrarse ante un conjunto más balanceado de clases. En cuanto a la precisión el mejor siguen siendo los sentimientos neutrales, incluso con un *recall* y *F1-Score* bastante altos.

4.2.2. Dataset completo - Red neuronal convolucional

El caso de la red convolucional es muy similar al de Naive Bayes. Su sesgo por predecir la clase neutral parece incluso más exacerbado que en el del caso anterior. Esto resulta de igual forma en un rendimiento deficiente, por los motivos ya mencionados anteriormente.

Comenzando con el *accuracy* de las emociones, CNN obtuvo un puntaje general de 62 %, una mejoría respecto a bayes; pero igualmente bajo y en buena medida debido al aumento de aciertos

al tener un mayor sesgo al clasificar la clase neutral (acierta todos los neutros). El modelo no cataloga *sadness* y tiene dificultades con *fear*, *surprise* y *disgust*. En la sección de los sentimientos, se obtuvo un *accuracy* de 62 %, siendo mejor que Naive Bayes.

4.2.3. Dataset sin neutralidad - Naive bayes

Ante el gran desafío que representó la evaluación de modelos entrenados con clasificaciones de neutralidad, se decidió probar modelos entrenados en datasets sin esta clasificación tanto para el caso de sentimientos como emociones.

En el caso de Naive bayes sin neutralidad, se observó dentro de la matriz de confusión para emociones que, si bien los datos procedieron a presentar una clasificación algo más dispersa por la ausencia del sesgo neutral, este sesgo se traslado hacia la siguiente clase mayoritaria (*Joy*), provocando de esta forma nuevamente un mal rendimiento al clasificar la mayoría de emociones en este tipo.

No obstante, más significativo que lo anterior es que se pudo observar la existencia de un ligero sesgo en el modelo a clasificar los datos dentro de *sadness*, *anger*, y *surprise*, mientras que no clasifica *fear* ni *disgust*. Este comportamiento se alinea a las frecuencias observadas en la gráfica de distribución de clases del dataset MELD sin neutralidad. Se revela de esta forma que el modelo, ante su incapacidad de encontrar las palabras conocidas de su entrenamiento con MELD para poder clasificar, gravita a predecir la clase de una oración en base a la distribución de clases que vio en el entrenamiento.

Para el caso de la matriz de confusión para sentimientos, se observó una mejora considerable en la clasificación de sentimientos. Al igual que en la matriz anterior, el modelo tiende a predecir datos como negativos al ser esta la segunda clase más numerosa. No obstante, en este caso la distribución de clases de ambos datasets concuerda mejor, provocando de esta manera una mejor coincidencia.

Esto se puede comprobar de esta forma viendo el *accuracy* general, en el caso de emociones se obtuvo un valor de 32 %, observando un empeoramiento del modelo. Sin embargo, por la parte de sentimientos se obtuvo un *accuracy* del 77 %, una mejora considerable, aunque paupérrima al no basarse en una verdadera comprensión de los datos por parte de los modelos.

4.2.4. Dataset sin neutralidad - Red neuronal convolucional

Al inspeccionar la matriz de confusión para las emociones, se identificó un cambio significativo en el sesgo del modelo en comparación con Naive Bayes. Mientras que el modelo probabilístico tendía a clasificar erróneamente hacia la alegría (*joy*), la CNN mostró predominancia a predecir la clase *anger*. Esto sugiere que los patrones léxicos presentes en el dataset de Twitter pueden presentar mayor intensidad a los vistos en MELD, provocando de esta forma que el modelo prediga constantemente *anger*. Por otro lado, la matriz de confusión en sentimientos exhibió un comportamiento peor al de naive bayes, resultando considerablemente disperso . Respecto al *accuracy*, el modelo obtuvo un valor general de 32 % en emociones, un valor idéntico al de Naive Bayes. Por su parte, en la clasificación de sentimientos, la CNN alcanzó un *accuracy* general del 63 %. Estos valores inferiores a los de sin neutralidad indican que, ante la reducción drástica del volumen de datos de entrenamiento, falta de solapamiento de vocabulario y eliminación de sesgo neutral que beneficiaba al modelo, la red neuronal profunda sufrió mayores dificultades.

4.2.5. Discusión de los resultados

El estudio permitió observar las limitaciones de aplicar modelos entrenados en un dominio a otro dispar: Ambos modelos probados rindieron ineficientemente, presentando una tendencia hacia la clase más frecuente. Si bien esto elevó artificialmente el *accuracy* global, la precisión para varias clases se volvió nula. Se observó así una incapacidad de los modelos para discriminar características distintivas.

El quitar la categoría "Neutral" permitió una evaluación más clara de la capacidad de generalización de los modelos sin el sesgo proveniente de neutro, algo que si bien mejoró el rendimiento de los modelos, en la práctica no cambió significativamente el resultado.

La causa subyacente de este comportamiento, como se pudo revisar en cada una de las secciones anteriores, es la ausencia de solapamiento lingüístico y contextual de los datasets empleados. MELD se compone de diálogos de cortos, casuales, propios de conversaciones. En contraste, el dataset de Twitter presenta textos técnicos, donde los usuarios detallan problemas y las empresas ofrecen soluciones igual de detalladas. Esta falta de concordancia implica necesariamente que buena parte de las palabras del dataset de Twitter se encuentren inexistentes dentro de MELD, esto implicó que los modelos no pudiesen identificar patrones con los datos.

Se extrae de todo esto que si se desea trabajar entre datasets con dominios diferentes, o bien se debe optar por un dataset con mejor coincidencia, o se deben implementar técnicas adicionales para la adaptación entre dominios, como pueden ser mapeo entre palabras; para que así los modelos entrenados puedan resultar efectivos en escenarios reales.

5. Conclusiones

En conclusión el dataset MELD basado en la serie *FRIENDS* no fue un buen entrenamiento para compararlo con el dataset de Twitter sobre atención al cliente. Podemos ver esto en sus deficientes resultados y solo obteniendo buenos puntajes en sentimientos tras la eliminación de los datos "Neutral", el poco solapamiento entre los datasets influye en el fallo. Cabe destacar que el modelo que mejor rindió antes del cambio fueron las CNN y después fue Multinomial Naive Bayes.

No fue posible validar directamente la hipótesis sobre el uso de emociones y sentimientos como métrica para poder evaluar la calidad de la atención al cliente recibida. Sin embargo, se pudo verificar la imposibilidad de evaluar datasets con modelos entrenados con contenido diferente (casual vs. formal). También importante fue que se pudo profundizar en el procesamiento de lenguaje natural (NLP), así como 2 técnicas de preprocesamiento de palabras (TD-IDF y Glove), determinando de esta forma las ventajas y limitaciones que tienen.

Para investigaciones futuras, se recomienda el uso de un dataset de entrenamiento que tenga contenido similar al dataset de prueba, o en su defecto una manera de mapear palabras entre datasets para adaptar el modelo al nuevo dominio. Adicionalmente, aumentar el número de datos de prueba y asegurar un balanceo de clases. Finalmente, se sugiere explorar diferentes combinaciones para las CNN y probar diferentes parámetros de la misma, de igual forma buscar otras arquitecturas más complejas como pueden ser los Transformers.

Referencias

Axelbrooke, S. (2017). Customer Support on Twitter. https://doi.org/10.34740/KAGGLE/DSV/8841

Chen, S., Hsu, C., Kuo, C., Huang, T. K., & Ku, L. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. *CoRR*, *abs/1802.08379*. http://arxiv.org/abs/1802.08379

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4). https://doi.org/10.3390/info10040150

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *CoRR*, *abs/1810.02508*. http://arxiv.org/abs/1810.02508

Pugh, S. D. (2001). Service with a smile: Emotional contagion in the service encounter. *Academy of management journal*, *44*(5), 1018-1027.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.