UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS EMPRESARIALES

TÉCNICAS DE AUMENTACIÓN DE TEXTO Y SU APLICACIÓN A CONJUNTOS DE TEXTOS EN ESPAÑOL PARA EL ANÁLISIS DE SENTIMIENTOS Y EMOCIONES

Tesis presentada por Rodrigo Andrés Gutiérrez Benítez
Para obtener el grado de Magíster en Ciencias de la Computación

Dirigido por Dra. Alejandra Segura Navarrete
Dr. Christian Vidal Castro

Agradecimientos

A mi familia por su apoyo, paciencia y sobre todo por enseñarme a ser una buena persona.

A Katherine, por estar a mi lado en todo momento y alentarme a perseguir este sueño después de mucho tiempo.

Abstract

Over the last decade, the use of social media as a massive communication medium has given people a tool to express their opinions. In it, they can write their thoughts about plenty of topics, such behavior generates massive amounts of data that can be analyzed by companies and researchers.

To analyze this data, Natural Language Processing tasks such as Emotion Analysis (EA) and Sentiment Analysis (SA) are used to classify the underlying emotions on texts and the polarity of it, respectively. To accomplish this tasks, one approach is to use Machine Learning (ML) and Deep Leaning (DL) techniques. However, to get good classification performance they require large datasets of labeled data for training. For researchers this is an open issue specially in Spanish where the labeled datasets are sparse, expensive in time and human effort to create. To solve the lack of large, labeled datasets problem, Data Augmentation (DA) techniques are used to artificially increase the size of small, labeled datasets.

This work contributes to Natural Language Processing (NLP) for Spanish text by experimenting with different Data Augmentation (DA) techniques over a set of small, labeled datasets created the SOMOS group of the Universidad del Bío-Bío and analyze the performance of the augmented datasets on the most used Machine Learning (ML) and Deep Learning (DL) classification algorithms for Sentiment and Emotion analysis with the goal of give a framework that guides the selection of a data augmentation technique and a classification model according to the features of a corpus.

Resumen

En la última década, el uso de las redes sociales ha entregado a las personas un lugar en el cual expresar sus opiniones. En ellas, las personas pueden escribir sus pensamientos sobre múltiples temas, este comportamiento genera gran cantidad de datos que pueden ser analizados tanto por empresas como por investigadores.

Para analizar estos datos, tareas del Procesamiento del Lenguaje Natural (NLP) tales como el Análisis de Emociones (EA) y Sentimientos (SA) son usados para clasificar las emociones subyacentes en el texto y la polaridad de este, respectivamente. Para completar dichas tareas, una de las aproximaciones es la utilización de técnicas de *Machine Learning* (ML) y *Deep Learning* (DL). Sin embargo, para obtener un buen rendimiento de clasificación, estas técnicas requieren de grandes cantidades de datos etiquetados para entrenar los algoritmos de clasificación. Para los investigadores este es un problema, especialmente en lenguajes como el español, donde la cantidad de datos etiquetados es escasa, costosa en tiempo y esfuerzo humano para su creación. Para resolver el problema de la falta de una gran cantidad de datos etiquetados, las técnicas de aumentación (DA) son usadas para incrementar artificialmente el tamaño de los conjuntos de datos con pocos datos etiquetados.

Este trabajo busca contribuye al Procesamiento de Lenguaje Natural (NLP) para textos en español mediante la experimentación con diversas técnicas de aumentación de datos sobre pequeños conjuntos de textos etiquetados creados el grupo SOMOS de la Universidad del Bío-Bío para luego analizar el rendimiento de clasificación pre y post aumentación en los modelos de clasificación de *Machine* y *Deep Learning* para el análisis de Sentimientos (SA) y Emociones (EA) con el objetivo de presentar un *framework* que guie la selección de una técnica de aumentación y modelo de clasificación de acuerdo con las características de un conjuntos de datos.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1 Introducción

El explosivo aumento del uso de las redes sociales como medio de comunicación masiva en la última década [1, 2], ha abierto nuevas aristas de investigación para el Procesamiento del Lenguaje Natural (NLP). Una de ellas, es la clasificación de textos con intencionalidad emotiva (EA) y la identificación de la polaridad del texto (SA). Entre los enfoques utilizados en estas tareas de NLP se encuentran aquellos basados en modelos de *Machine Learning* (ML) y *Deep Learning* (DL) [3]. Sin embargo, para lograr su cometido, estos modelos necesitan grandes cantidades de datos etiquetados [4, 5], lo cual supone un problema, ya que, la labor de etiquetado manual de los textos es costosa en tiempo y recursos [1, 6–8]. Es por ello, que se buscan alternativas que permitan obtener una cantidad de datos etiquetados de manera rápida, eficiente y que en la medida de lo posible no necesite de la intervención de humanos. En este sentido, las técnicas de aumentación de datos, que inicialmente fueron utilizadas en el análisis de imágenes, son ahora utilizadas en el análisis de texto para aumentar los conjuntos de datos etiquetados y así mejorar el rendimiento de los modelos de clasificación [3, 9].

Este trabajo se encuentra enmarcado dentro del análisis de la subjetividad de texto en español, como parte de las tareas de preprocesamiento de un conjunto de datos, con el propósito de analizar si el uso de las técnicas de aumentación de datos de texto mejora efectivamente la capacidad de clasificación de los modelos de *Machine Learning* (ML) y *Deep Learning* (DL). Para lograr dicho objetivo, luego de revisar el estado del arte de la aumentación de textos, se aplicarán las técnicas de aumentación de datos de textos más utilizadas a conjuntos de datos creados por la Universidad del Bío-Bío para evaluar el impacto de la aumentación en la clasificación de sentimientos y emociones en el idioma español con los algoritmos de ML o DL que más se utilizan de acuerdo con el Estado del Arte. Con dichos resultados, se busca contribuir al Procesamiento de Lenguaje Natural en español al proveer un framework que ayude en la selección de técnicas de aumentación de texto para SA y EA.

El resto de este trabajo está organizado de la siguiente forma: en el Capítulo 2 se presenta la propuesta de tesis con la hipótesis, objetivos y la metodología que sigue este trabajo, en el Capítulo 3 se presenta el marco conceptual y la revisión de literatura, luego, en el Capítulo 4 se describen los conjuntos de datos utilizados para la experimentación con las técnicas de aumentación, el Capítulo 5 presenta los resultados obtenidos en la experimentación, el

Capítulo 6 presenta una discusión de los resultados por técnica de aumentación, para finalizar con los Capítulos 7 y 8 que entregan una guía de selección, conclusiones y trabajo futuro.

Capítulo 2 Propuesta de tesis

2.1 Hipótesis

Es posible mejorar el rendimiento en el análisis de emociones y/o sentimientos de textos en español, basado en ML/DL a través de las técnicas de aumentación de datos.

2.2 Objetivo general

Proveer un framework que facilite la selección y aplicación de técnicas de aumentación de datos para el análisis de emociones y/o sentimientos de textos en español.

2.3 Objetivos específicos

- Revisar los últimos avances en el estado del arte de la aumentación de datos de texto, describiendo en qué consiste, qué técnicas existen y como se clasifican, cuáles son las técnicas más utilizadas y cuáles son las métricas con las que se evalúa el efecto de la aumentación de los conjuntos de datos en la clasificación de sentimientos y emociones.
- Probar mediante experimentación las técnicas más usadas de la aumentación de datos de texto, evaluando el impacto de la aumentación en el rendimiento de clasificación de los modelos de ML y DL.
- Definir un *framework* que ayude en la selección de una técnica de aumentación de acuerdo con el problema a abordar y el modelo de clasificación que se desea utilizar.
- Discutir sobre las ventajas, limitaciones y problemas enfrentados en la experimentación realizada con conjuntos de datos en español.

2.4 Alcance de la investigación

Esta tesis está enfocada en una de las etapas de preprocesamiento de datos para el análisis de emociones y sentimientos, por lo tanto, los algoritmos de clasificación utilizados para etiquetar datos serán los más utilizados en el estado del arte al momento de escribir este trabajo. Igualmente, no se crearán nuevos conjuntos de datos sobre los cuales generar aumentación, sino que, se emplearán conjuntos de datos existentes ya utilizados en investigaciones previas de análisis y subjetividad de textos en español realizada por el grupo SOMOS de la Universidad del Bío-Bío.

2.5 Metodología de trabajo

Se consideran las siguientes actividades para la metodología de trabajo.

1. Obtener el estado del arte de las técnicas de aumentación de datos de texto mediante la realización de una revisión de literatura de los últimos 5 años. Siguiendo la metodología sugerida por Biolchini et al [10].
2. Experimentar con las distintas técnicas de DA de la siguiente forma:
    a. Llevar a cabo la selección de los conjuntos de datos para análisis de texto en español que serán aumentados.
    b. Seleccionar las técnicas de aumentación con base en el estado del arte obtenido.
    c. Seleccionar los modelos de clasificación basados en ML y DL más utilizados en el estado del arte obtenido.
    d. Aplicar los modelos de ML y DL seleccionados a conjuntos de datos sin aumentar, para obtener el rendimiento de clasificación que será utilizado como línea base.
    e. Aplicar las técnicas de aumentación de datos de texto a los conjuntos de datos seleccionados, para aumentar la cantidad de datos etiquetados en cada uno de ellos.
    f. Aplicar los modelos de ML y DL seleccionados a conjuntos de datos aumentados para obtener el rendimiento de clasificación luego de la aumentación.

g. Evaluar el impacto en el desempeño de las técnicas de aumentación seleccionadas sobre los conjuntos de datos elegidos utilizando las métricas más utilizadas en el estado del arte.

3. Realizar una discusión sobre el impacto en el desempeño de la aumentación de datos de texto en los conjuntos de datos en español elegidos para experimentación.
4. Construir un *framework* que recoja los resultados obtenidos en la experimentación para ayudar en la selección de una técnica de aumentación de acuerdo con el problema a abordar y el modelo de clasificación que se desea utilizar.

Capítulo 3 Estado del Arte

Para una mejor comprensión de este documento, a continuación, se presentan los conceptos fundamentales y definiciones asociadas, junto con los resultados de la revisión sistemática de literatura realizada.

3.1 Marco Conceptual

Los conceptos fundamentales relacionados con esta investigación serán tratados de forma general para garantizar la comprensión de los contenidos y hallazgos posteriores.

3.1.1 Subjetividad de texto

Las categorías de la subjetividad de texto que se abordan en este trabajo son las siguientes:

Emotion Analysis (EA): Tarea del Procesamiento del Lenguaje Natural (NLP) que permite clasificar textos de acuerdo con las emociones que se encuentran presentes en estos, las categorías de clasificación de emociones abarcan normalmente la felicidad, tristeza, miedo, ira, sorpresa, entre otras [11].

Sentiment Analysis (SA): Es una de las tareas del Procesamiento del Lenguaje Natural (NLP) que permite extraer, analizar y procesar las características de un texto y luego las clasifica de acuerdo con su polaridad [1, 5, 12]. Las categorías usadas en SA pueden abarcar desde positivo y negativo; positivo, negativo y neutro; o 5 etiquetas que van desde muy negativo, negativo, neutro, positivo y muy positivo [12–15].

Para realizar la clasificación automática de EA y SA se pueden utilizar tres enfoques, el primero de ellos se basa en el uso de lexicones como recursos léxicos, el segundo utiliza modelos de ML y DL, el último de los enfoques corresponde a un enfoque híbrido, que combina lexicones y los modelos de ML/DL.

Dado que, el rendimiento de los enfoques de clasificación automática basados en ML/DL depende de la cantidad de datos etiquetados con los cuales son entrenados, este trabajo sigue la línea de investigación de la aumentación de textos para incrementar artificialmente la cantidad de datos de entrenamiento, específicamente textos, y de esta manera aportar en la

mejora del rendimiento de clasificación de los modelos de clasificación automática basados en ML/DL.

Al igual que todas las investigaciones en NLP, el rendimiento de las técnicas o modelos son relativos, y no solamente dependen del tipo de clasificación, sino que también dependen del idioma del conjunto de datos utilizado. Es en este último punto donde se centra nuestro foco de interés, concentrado específicamente en "Técnicas y/o Modelos de Aumentación de Datos Textuales en idioma español" basados en ML/DL para la clasificación de emociones y/o sentimientos.

3.1.2 Algoritmos de Machine Learning (ML)

El análisis de las emociones contenidas en texto considera la clasificación mediante el uso de técnicas de ML, algunas de estas técnicas son:

***Support Vector Machine* (SVM)**: Es uno de los algoritmos de ML más populares para realizar clasificación de texto, su funcionamiento se basa en la búsqueda del mejor hiperplano para la separación de las muestras [14], tiene buen rendimiento en conjuntos de datos con muchas dimensiones. Esta es una de las razones por las que necesita una gran cantidad de tiempo para determinar la función de kernel óptima [15]. En la Fig. 1 pueden observarse tanto los vectores de soporte como el hiperplano.

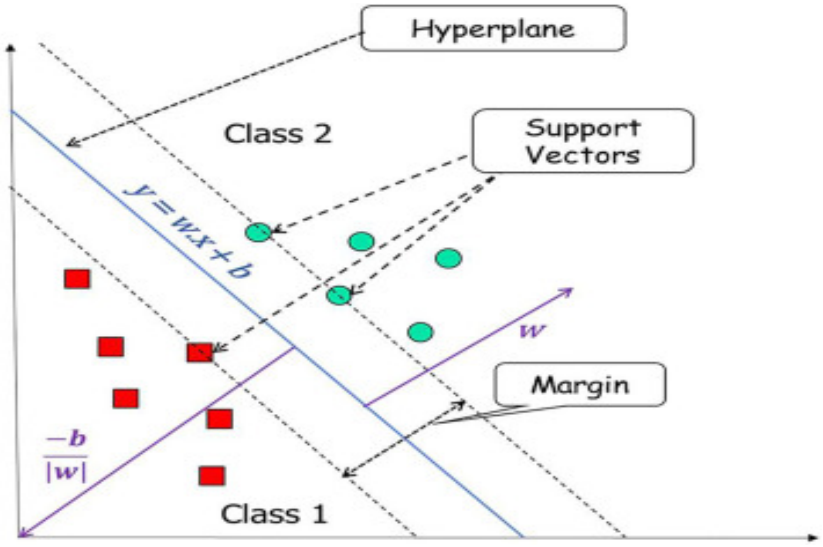

Fig. 1 Ejemplo de vectores de soporte e hiperplano[1]

[1] https://www.sciencedirect.com/topics/computer-science/support-vector-machine

***Random Forest* (RF)**: Para comprender de mejor forma RF, primero se debe definir el algoritmo Árbol de Decisión (*Decision Tree*). Este algoritmo realiza la clasificación mediante reglas del tipo *if – else*, tiene la habilidad de lidiar con grandes conjuntos de datos, pero sufre de inestabilidad en la clasificación. Para solucionar esto, RF produce un número individual de árboles de decisión y realiza la predicción final mediante la agregación de los resultados de cada árbol [15]. En la Fig. 2 se puede observar gráficamente el funcionamiento de RF.

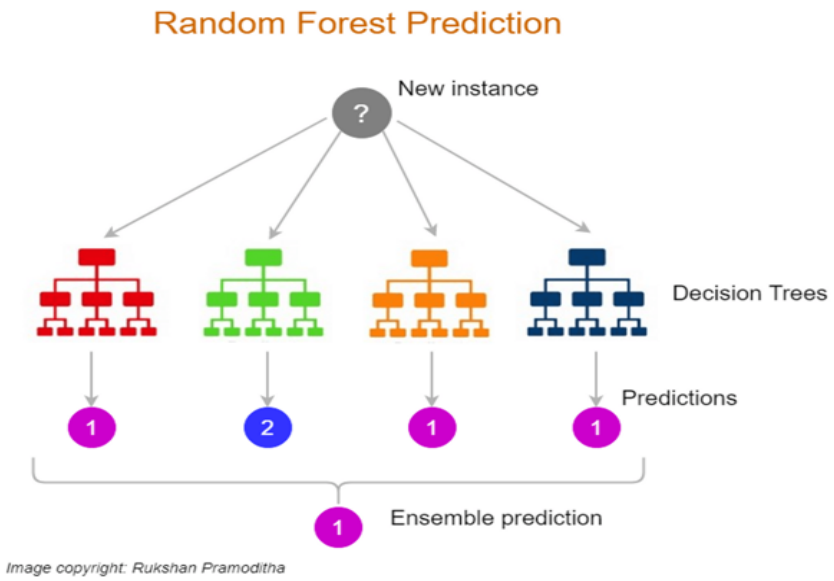

Fig. 2 Ejemplo de clasificador Random Forest[2]

Naive Bayes: Es un algoritmo de ML que se basa en el cálculo de la probabilidad de distribución de los datos para lograr la clasificación, en lugar de entrenar directamente la relación entre las clases de salida y los vectores de entrada [15, 16].

3.1.3 Algoritmos de Deep Learning (DL)

Inspirados en el funcionamiento del cerebro humano, los algoritmos de Deep Learning representa una red neuronal con varias capas. Este tipo de algoritmos ha demostrado buenos rendimientos en la clasificación de oraciones, generación de texto y representación de características [15]. A continuación se presentan 5 de las técnicas de DL utilizadas en la clasificación de subjetividad de texto:

[2] https://www.almabetter.com/bytes/tutorials/data-science/random-forest

***Convolutional Neural Network* (CNN)**: Es un tipo de algoritmo de DL que está compuesto de tres capas. Las primeras capas (*convolution y pooling*) se encargan de aprender las jerarquías espaciales de las características (entrada de datos en forma de embeddings), mientras que la última capa (*fully connected*) realiza un mapeo de las características extraídas hacia la salida final [15]. La Fig. 3 muestra un ejemplo de la arquitectura CNN.

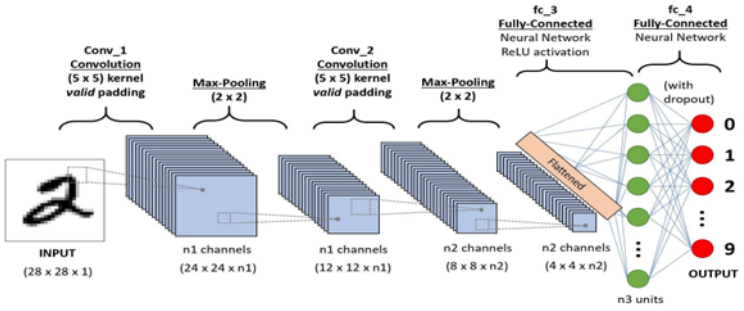

Fig. 3 Ejemplo de Convolutional Neural Network (CNN)[3]

***Recurrent Neural Network* (RNN)**: Son una clase de red neuronal que modela datos secuenciales y procesa predicciones basándose en un vector de palabras que trata con las relaciones de dependencia a largo plazo entre las palabras de un conjunto de datos [15, 17]. En la Fig. 4 puede observarse un ejemplo de la arquitectura RNN.

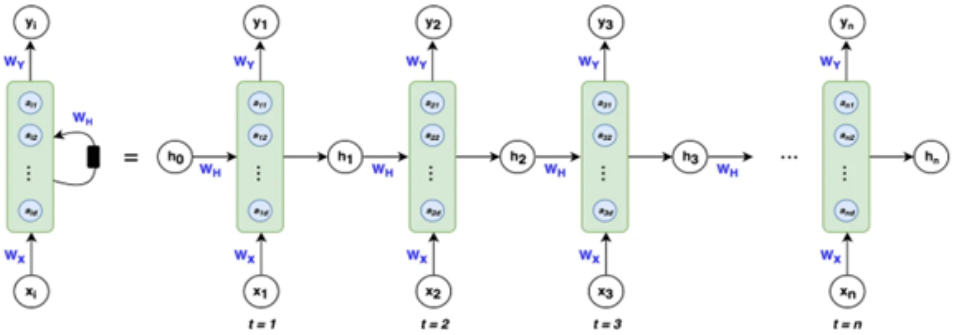

Fig. 4 Ejemplo de Recurrent Neural Network (RNN)[4]

***Long Short-Term Memory* (LSTM)**: Es una variante de las redes neuronales recurrentes (RNN) que introduce un mecanismo de puerta que resuelve los problemas de "*gradient disappearance*" y "*gradient explosion*". Cada LSTM está compuesta por una célula de memoria, una puerta de olvido que decide que elementos deben ser descartados de la célula

[3] https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/

[4] https://pub.towardsai.net/whirlwind-tour-of-rnns-a11effb7808f

y una puerta de salida que determina que estado de la célula debe ser entregado con valores entre -1 y 1 [5, 18–20].

***Gated Recurring Units* (GRU)**: Es un tipo de red neuronal recurrente (RNN) que está compuesta de dos elementos, la puerta de actualización y la puerta de reinicio. La primera de ellas determina que información debe ser conservada, mientras que la segunda controla cuanta información almacenada debe ser olvidada [18, 21].

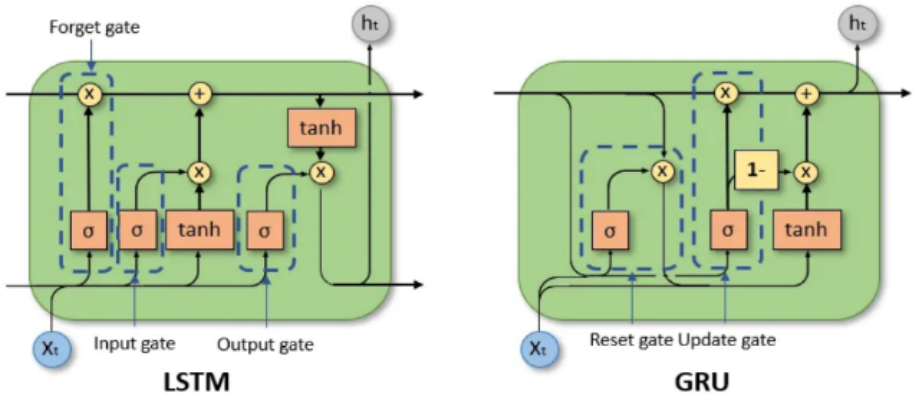

Fig. 5 Ejemplos de LSTM y GRU[5]

***Bidirectional Encoder Representations of Transformer* (BERT)**: Es un modelo pre-entrenado de NLP basado en redes neuronales, es de código abierto creado por Google y es considerado uno de los mayores avances en NLP. Está basado en el modelo *Transformer* [22] que procesa todas las oraciones paralelamente usando mecanismos de atención que le permiten capturar la semántica en estas, al realizar este tipo de procesamiento, se adapta mejor a entornos de *hardware* paralelo [23]. Tiene más de 100 millones de parámetros y requiere de un conjunto de datos de gran tamaño para su entrenamiento y ajuste fino [1, 11]. La Fig. 6 muestra la arquitectura del modelo BERT.

[5] https://towardsdatascience.com/a-brief-introduction-to-recurrent-neural-networks-638f64a61ff4

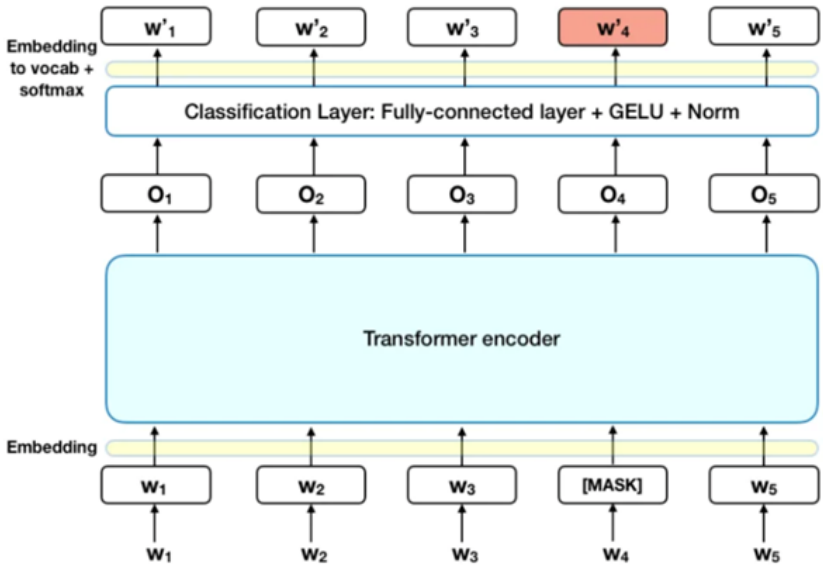

Fig. 6 Arquitectura BERT[6]

3.1.4 Métricas para la evaluación del rendimiento de clasificación

La evaluación de la influencia de la aumentación de un conjunto de datos está en directa relación con el aumento del rendimiento de clasificación de los algoritmos de clasificación que se utilicen para procesarlo. En ese sentido, para medir el rendimiento de los modelos de clasificación de sentimientos y/o emociones en conjuntos de datos sin aumentar y con aumentación aplicada, se utilizan las métricas *Accuracy, Recall, Precision y F-Measure*, entre otras [24].

Accuracy: Corresponde a la proporción total del número de predicciones correctas sobre el total de casos examinados, calculada mediante la ecuación 1:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Siendo, TP = *true positive*, TN = *true negative*, FP = *false positive* y FN = *false negative* [15, 25].

[6] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Precision: Es el radio de resultados TP sobre el total de predicciones positivas del modelo, incluyendo verdaderos y falsos positivos [15].

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall: Corresponde a la proporción de casos positivos que son correctamente indentificados [15, 25].

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F-Measure: Es la media armónica entre las métricas de *precision* y *recall*, su rango varía entre 0 y 1. Entre mayor sea el valor de *F-Measure* mejor es el rendimiento del modelo [15, 25].

$$F - Measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{4}$$

3.2 Revisión de literatura

La aumentación de datos (DA) es un conjunto de métodos utilizados para generar nueva información a partir de un conjunto de datos etiquetados [15, 23, 26, 27]. En sus inicios fue principalmente utilizado para aumentar datos en el procesamiento de imágenes [28], pero con el tiempo se han generado técnicas de aumentación de datos para texto con el fin de mejorar el rendimiento de clasificación de los modelos de ML y DL. La aumentación de texto se utiliza para abordar los problemas de escasez de datos etiquetados y desbalance en las clases de un conjunto de datos [2], mediante la generación de nuevas oraciones a partir de las oraciones existentes en el conjunto de datos, tal como se muestra en la Tabla 1, donde se genera una nueva oración reemplazando palabras por su sinónimo. En esta Sección se ahondará en las distintas técnicas de aumentación de datos para texto.

Tabla 1 Ejemplo de aumentación mediante el reemplazo de sinónimos, Liu et al. en [19]

Original	Nueva a partir de reemplazo de sinónimos
I don't get any unusual code.	I don't get any strange code.
It's a good piece of work.	It's an effective part of work.
	It's a food nibble of work

El objetivo de la revisión de literatura realizada es conocer los avances en el estado del arte respecto de la aumentación de datos en texto. Para ello, se siguió la metodología propuesta por Biolchini et al. en [10]. Esta revisión se divide en 3 etapas: planificación, ejecución y reporte de resultados. Las preguntas que se busca resolver con esta revisión de literatura son: ¿En qué consiste la aumentación de datos de texto?, ¿Qué técnicas existen para realizar la aumentación de datos de texto y cuáles son las más utilizadas?, ¿Cuáles son las métricas empleadas para establecer la calidad de la aumentación de un conjunto de datos de texto? Y, por último, ¿Cuáles son los desafíos que enfrenta la aumentación de datos de texto?

Como fuentes de búsqueda para la extracción de artículos se seleccionaron las librerías digitales Web Of Science[7], ScienceDirect[8] y IEEEXplore[9]. Adicionalmente, se considera los artículos recomendados por los guías de tesis. Los resultados luego de la aplicación de los filtros definidos en la etapa de planificación fueron los siguientes:

Tabla 2 Resultados de búsqueda en los motores elegidos.

	Cantidad de artículos
Seleccionados	56
No superan los filtros de inclusión y exclusión	47
Duplicados	49
Retractados por revista	1
	153

3.2.1 Taxonomías de aumentación de datos de texto

Para clasificar las técnicas de aumentación de datos de texto fueron revisadas 2 taxonomías que permiten categorizarlas. La primera taxonomía, presentada por Bayer et al. en [29], define 2 grandes categorías, el espacio de datos (*data space*) que agrupa las técnicas que realizan aumentación directamente sobre los datos a nivel de caracteres, palabras, frases y documentos; mientras que en el espacio características (*feature space*) normalmente representado mediante *embeddings*, agrupa las técnicas que realizan aumentación a través de la manipulación de la representación vectorial de los datos. En la Fig. 7, puede observarse esta taxonomía con las subcategorías en el espacio de datos.

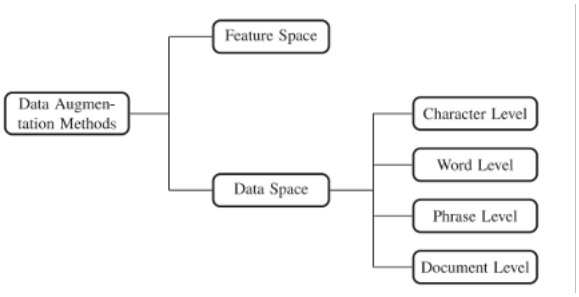

Fig. 7 Taxonomía de aumentación de datos propuesta por Bayer et al. en [29]

[7] https://www.webofscience.com

[8] https://www.sciencedirect.com

[9] https://ieeexplore.ieee.org/

La segunda taxonomía revisada, es la presentada en el trabajo de Abonizio et al. en [30]. En ella, se muestra una clasificación más completa de las técnicas de aumentación al clasificarlas mediante el tipo de manipulación que se realiza sobre el conjunto de datos, en la Fig. 8 puede observarse en detalle esta taxonomía.

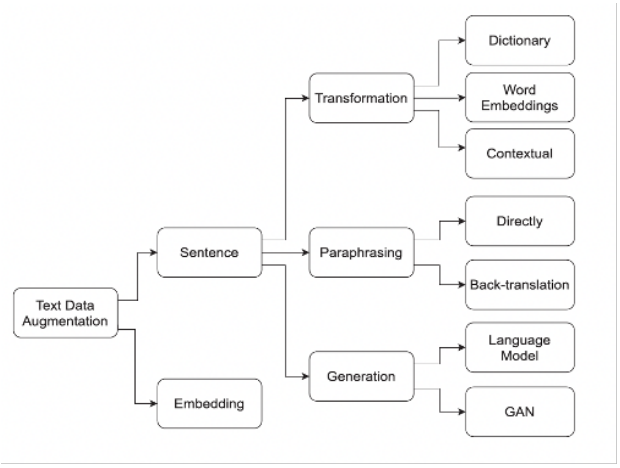

Fig. 8 Taxonomía de aumentación de datos propuesta por Abonizio et al. en [30]

De acuerdo con la definición de aumentación de datos de texto en la Sección 3.1.4 de este trabajo y las taxonomías presentadas en la Fig. 7 y Fig. 8, se utilizará la taxonomía propuesta por Abonizio et al. [30] ya que permite clasificar las técnicas de aumentación de acuerdo con el tipo de manipulación que se realiza sobre el conjunto de datos, A continuación, se profundiza en los detalles de cada clasificación propuesta en la taxonomía seleccionada.

Aumentación de datos mediante transformación de oraciones: se basa principalmente en las operaciones de reemplazo, inserción, intercambio y eliminación de palabras al azar [31]. En el caso del reemplazo y la inserción de palabras, la selección de estas puede realizarse mediante diccionarios tales como WordNet[10], *word embeddings* con Word2Vec [15] y la identificación de la importancia de las palabras en un conjunto de datos mediante técnicas como TF-IDF[11] para la conservación de la oración original [30]. En la Tabla 3, se muestran ejemplos de aumentación mediante reemplazo, inserción, intercambio y eliminación de palabras dentro de una oración.

[10] https://wordnet.princeton.edu

[11] https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

Tabla 3 Ejemplos de transformación de oraciones, Balakrishnan et al. [15]

Operación	Descripción	Ejemplo
Texto original	-	The quick Brown fox jumps over the lazy dog
Intercambio al azar	Se seleccionan 2 palabras al azar y son intercambiadas.	The ***lazy*** quick fox jumps over the dog ***brown***.
Eliminación al azar	Se elimina una palabra al azar de la sentencia	The quick Brown jumps over the lazy dog.
Inserción al azar	Introduce al azar una nueva palabra en la sentencia	The quick ***sluggish*** brown fox jumps over the lazy dog.
Reemplazo de sinónimo	Selecciona ***n*** palabras en la sentencia y las reemplaza con su sinónimo.	The quick sluggish ***umber*** fox jumps over the lazy dog.

Aumentación de datos mediante parafraseo: utiliza técnicas como *back-translation* consistente en la traducción de las oraciones a uno o más lenguajes intermedios para luego volver a traducirlos al lenguaje original, tal como puede observarse en la Fig. 9. Al realizar esta operación, se obtienen sentencias con ligeras modificaciones producidas por el efecto de la traducción, pero que mantienen la semántica de la oración original. Sin embargo, la calidad de las oraciones generadas mediante este método depende directamente de las herramientas de traducción utilizadas [8, 11, 30, 32].

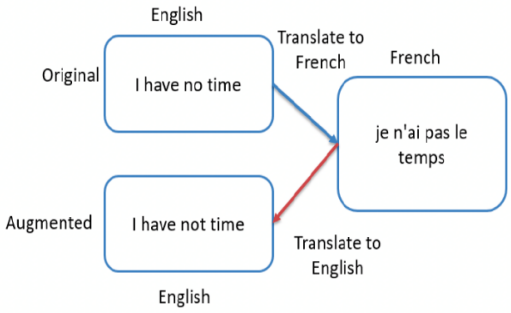

Fig. 9 Ejemplo de back-translation, GB y Jacob [8]

Aumentación de datos mediante generación: Utilizan modelos de generación que crean datos sintéticos a partir de un conjunto de datos existente. En este sentido, Generative Adversarial Networks (GAN) basan su funcionamiento en la utilización de un generador que crea oraciones a partir de un conjunto de datos existente y un discriminador, que juzga si las oraciones creadas por el generador son reales o falsas. Cuando el discriminador no puede diferenciar si una oración es real o falsa, estamos en presencia de una muestra para ser agregada al conjunto de datos aumentado. Este tipo de modelo generativo comúnmente utiliza modelos de DL para la implementación tanto del generador como el discriminador, por lo que dependen del conjunto de datos de entrenamiento para generar oraciones de buena calidad [18, 33–35]. La Fig. 10 muestra un ejemplo de arquitectura GAN usado por Luo et al. en [36].

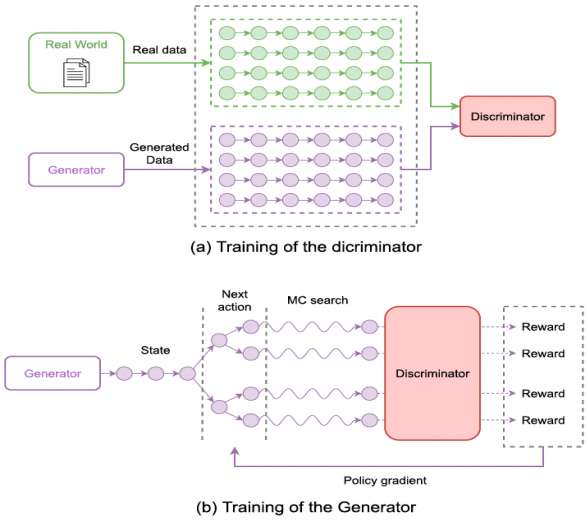

Fig. 10 Ejemplo de arquitectura GAN, Luo et al. [36]

Aumentación mediante manipulación del espacio vectorial: A diferencia de los métodos de manipulación de oraciones que trabajan directamente sobre el texto, la manipulación del espacio vectorial (*embedding*) opera en el nivel de los vectores representativos de las oraciones en el espacio vectorial del modelo utilizado. Dado este modo de funcionamiento, son muy dependientes del modelo a utilizar porque la arquitectura de este asume como las oraciones estarán representadas. Su implementación está basada en redes neuronales y existe menos investigación sobre ellos [30].

3.2.2 Técnicas de aumentación más utilizadas por los artículos

La distribución de los artículos de acuerdo con la técnica de aumentación utilizada se muestra en la Tabla 4, detallando la técnica de DA, cantidad de artículos y la identificación de estos.

Tabla 4 Distribución de artículos por método de aumentación

Técnica de aumentación	Cantidad de artículos	Identificación de artículos
Transformación	31	[1–7, 9, 15–17, 19–21, 23, 25–27, 37–49]
Parafraseo	5	[8, 11, 32, 50, 51]
Generación	9	[18, 33–36, 52–55]
Espacio vectorial	3	[24, 28, 56]
Híbrido	2	[14, 57]
No especifica	1	[12]
Comparativa	1	[30]

De la tabla anterior se puede desprender la Fig. 11, que muestra gráficamente que la categoría de técnicas de aumentación más utilizada es la transformación de oraciones con un 59%, seguida por la categoría de generación de oraciones con un 17%. En tercer lugar, se puede encontrar la categoría de parafraseo de oraciones con un 10%. De acuerdo con estos resultados, las técnicas de aumentación que se serán utilizadas para este trabajo son las antes mencionadas.

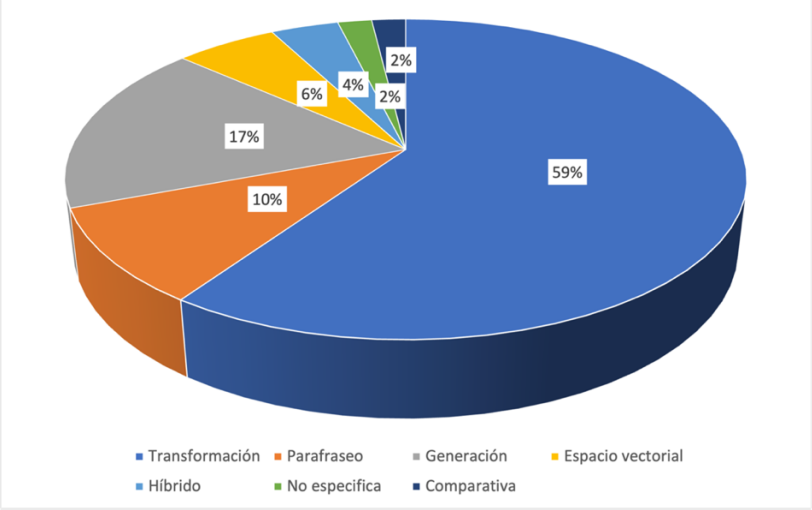

Fig. 11 Distribución de las categorías de aumentación según la revisión de literatura

3.2.3 Idioma de los conjuntos de datos utilizado por los artículos

La distribución de los idiomas de los conjuntos de datos utilizados en los trabajos revisados puede apreciarse en la Fig. 12, siendo el idioma inglés el más utilizado con un 57%. Cabe mencionar que el idioma español no es utilizado en los conjuntos de datos de los trabajos realizados. Solamente el trabajo de Bogoradnikova et al. en [14] menciona un conjunto de datos para la detección de mensajes de odio con 8.000 muestras etiquetadas en español, correspondientes a competencias de Google y Jigsaw del año 2020.

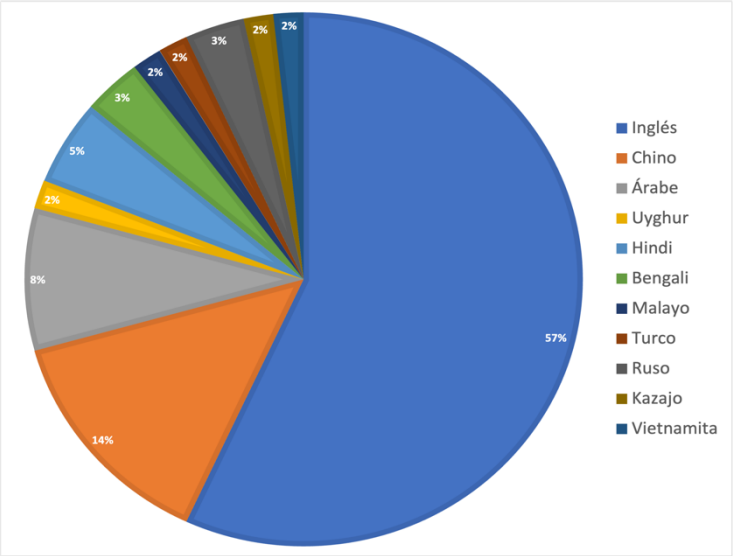

Fig. 12 Distribución de los idiomas en los conjuntos de datos en revisión de literatura

3.2.4 Modelos de clasificación más utilizados por los artículos

En cuanto a los clasificadores que fueron utilizados en los trabajos revisados, la Fig. 13 muestra la distribución considerando únicamente el clasificador principal de cada trabajo y no los clasificadores utilizados para comparar el rendimiento de un modelo en particular. El modelo de clasificación más utilizado es BERT y sus variantes con un 23%, seguido por LSTM con un 15%, mostrando una tendencia hacia los modelos de clasificación DL.

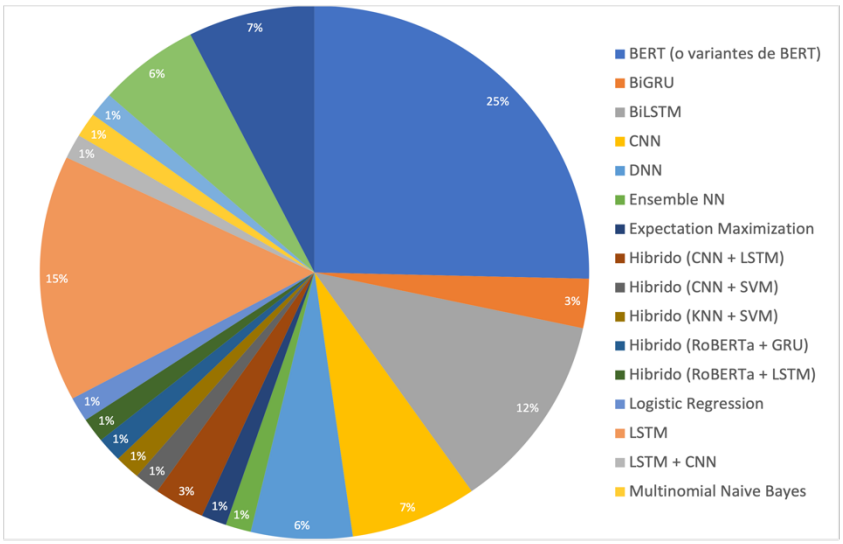

Fig. 13 Distribución de los modelos de clasificación en la revisión de literatura

3.2.5 Trabajos relacionados mediante transformación de oraciones

En [15] Balakrishnan et al. realizan la aumentación de un conjunto de datos mediante EDA (Easy Data Augmentation) [31], luego, hacen una comparativa entre los clasificadores de ML y DL. Los resultados reportados indican que los algoritmos de DL utilizados (CNN, RNN, BiLSTM y variantes de BERT) se comportan de mejor forma que los algoritmos de ML (*Logistic Regression*, *Naive Bayes, Decision Tree, Random Forest y Support Vector Machine*) para los conjuntos de datos original y aumentado. El rendimiento de los algoritmos de DL con los datos aumentados fue de 96% en *accuracy* y 91,1 %en *F1-Score.*

Por otra parte, en [3] los autores proponen una variación de EDA que considera la información que tienen los emojis en los *tweets* para preservar su semántica, para luego realizar SA mediante BiLSTM. Los resultados indican que el comportamiento del conjunto de datos original con el clasificador es deficiente por la baja cantidad de datos etiquetados que tienen, indican que, con la aumentación mediante EDA y emojis, sus resultados mejoraron considerablemente. Sus métricas de *accuracy* y *F1-Score* rondan el 72%. Yuan et al. en [16] proponen una nueva adaptación de EDA, que es utilizada para realizar la aumentación de un conjunto de datos con 4 categorías afectivas (*joy, angry, bored, sad*), para solucionar el problema de la pérdida de semántica, los autores proponen el uso de TF-IDF para identificar las palabras más importantes en el contexto. Luego, someten el conjunto de datos aumentado a clasificadores CNN, los resultados de su experimentación indican que en comparación con sus líneas base, el modelo propuesto mejora en 3% la métrica *accuracy*. Otro de los trabajos que se centra en la utilización de EDA es el de Lee et al. [41], en dicho trabajo, utilizan *knowledge graphs* para la representación de las oraciones y la modificación de ellas se basa en las técnicas que utiliza EDA, sus resultados en las métricas *ROUGE* no cambian significativamente al utilizar este método.

Dhiman et al [23], utiliza MOD-EDA para aumentar un conjunto de datos de *tweets* concernientes a los sentimientos de los usuarios con respecto a las políticas públicas de la India y la influencia que tienen en los periodos eleccionarios, luego mediante BERT realizan la clasificación de la polaridad de los *tweets*. Para medir el rendimiento de su modelo, realizan combinaciones entre el clasificador con y sin aumentación, obteniendo los mejores resultados con MOD-EDA + BERT con porcentajes de 71% en *F1-Score* y 70% en *accuracy*.

Para mejorar los resultados de clasificación mediante BERT en conjunto con in lexicón de modismos en inglés, Tahayna et al. [1] realiza el reemplazo de modismos en inglés de acuerdo con su significado, obteniendo así nuevas oraciones. Mediante la métrica *F1-Score* indican que su modelo obtuvo una mejora en la clasificación de más de 10% con respecto a la línea base (76,98%) en la clasificación de 150 *tweets*. Mientras, en el trabajo de Li et al. [9] propone sus propios algoritmos de aumentación basados en el reemplazo de sinónimos (PWSS) e intercambio del orden de las palabras dentro de una oración (DRAWS), logrando con esto aumentar 4 conjuntos de datos públicos referentes a la tarea de SA, como clasificadores para su trabajo utilizan LSTM. Los resultados reportados indican que la aumentación mediante DRAWS (11,49% *Macro-F1*) tiene mejor rendimiento que PWSS (2,9% *Macro-F1*) con el conjunto de datos utilizado. Para realizar SA en *tweets* en el idioma Turco, Shehu et al. [17] propone tres métodos de aumentación, *shift, shuffle* y una combinación de ambos, para luego someter el conjunto de datos aumentado a clasificadores de ML y DL. Lamentablemente, los autores no ahondan en las técnicas de aumentación, porque su foco es su modelo de clasificación propuesto (HAN). Sus resultados indican que la utilización de algoritmos de DL obtiene mejores resultados de clasificación respecto de los algoritmos de ML. Sin embargo, indican que, en cuanto a tiempos de ejecución y entrenamiento, ML se comporta de mejor manera.

En el trabajo de Liu et al. [19] se utiliza BiLSTM como modelo de clasificación de sentimientos en correos electrónicos y el reemplazo al azar de palabras mediante la combinación de WordNet y KNN (*K-nearest Neighbor*) para realizar la aumentación de datos del conjunto de datos con el objetivo de balancear las clases de polaridad. Algo interesante que destacan es que reemplazando el 20% de las palabras en una oración y aplicando LDA y TF-IDF [14] para mantener la semántica, se obtienen mejores resultados de aumentación. Los resultados reportados indican que con su modelo de aumentación y el uso de BiLSTM como clasificador logra una mejora en la clasificación entre 1,5% y 10% en la métrica *accuracy*. Uno de los trabajos que no obtuvo mejoras significativas con la aumentación del conjunto de datos para la clasificación de categorías de aspecto es el de Almasre [39], con resultados de 66,3% en *F1-Score* para la línea base y 66,1% en la misma métrica para el conjunto de datos aumentado. El método de aumentación utilizado se basa en el reemplazo de no más del 25% de palabras al azar en una oración y el empleo de la similaridad de coseno para determinar las palabras que serán usadas para reemplazo. Como clasificador, utilizan una variante de BERT para el idioma árabe. Uno de los problemas que enfrentaron en su modelo es la gran cantidad de categorías de aspecto que debían aumentar (34) y el desbalance en 11 de ellas.

Otro ejemplo de la aplicación de la aumentación de datos para solucionar el problema de desbalance de un conjunto de datos es el propuesto por Ha et al. en [40], su modelo de aumentación utiliza el reemplazo de sinónimos basándose en *Paraphrase Database* para luego utilizar CNN para detectar 3 categorías (*support, other, y oppose*) en un conjunto de datos relativos a las opiniones hechas por ciudadanos de Estados Unidos concernientes a un plan de energía limpia. Luego de la aumentación, su conjunto de datos quedó balanceado en 2800 comentarios por categoría. Los resultados que obtuvieron en las métricas *accuracy* y *F1-Score* fueron de 84% y 71% respectivamente. Siguiendo con la utilización de la aumentación para el balanceo de conjuntos de datos, Lee et al [41] compara dos técnicas de aumentación, EDA y UDA, para balancear las muestras en *knowledge graphs* con carga de polaridad. La experimentación realizada nuestra que el uso de EDA en la categoría "*negative*" tiene un impacto muy bajo en el rendimiento de clasificación frente a los resultados obtenidos por UDA en la categoría "*positive*" para el mismo conjunto de datos.

Para realizar la clasificación de SA en textos chinos, Wang et al. [43] aumenta un conjunto de datos mediante el reemplazo de sinónimos extraídos desde un tesauro que ellos mismos construyen para su modelo, con el cual indican que se corrige el problema de reemplazo de sinónimos con bajo nivel de similaridad y que influye en la tarea de clasificación. Una vez aumentado el conjunto de datos, realizan SA mediante un modelo híbrido entre CharCNN (que extrae las características) y SVM (que realiza la clasificación). Los resultados reportados en su trabajo indican que con su modelo de aumentación y clasificación obtienen 95% de accuracy.

Para realizar el análisis de sentimientos en tweets que contienen modismos, Tahayna et al. [42] utiliza SliDE (lexicón de modismos de IBM) y un clasificador BERT. En su trabajo proponen un método aumentación de tweets mediante el reemplazo de modismos por su significado, con lo que reportan un rendimiento de clasificación 92% *F1-score* y una reducción de 16% en el error de clasificación con respecto al conjunto de datos sin aumentar.

Mediante la utilización de las técnicas de aumentación de texto mediante transformación, Qudar et al. [6] reemplaza el 5% de las palabras por su sinónimo, elimina el 10% de las palabras, inserta un 5% e intercambia un 5% de las palabras en una oración. Para luego aplicar un modelo semi supervisado de estudiante-profesor con el cual realizan el análisis de sentimientos. Los resultados obtenidos por su modelo son de 87,3% (F1-score) en el conjunto de datos *SemEval Aspect Sentiment Analysis* y 88,35% (F1-score) en un conjunto de datos de *Twitter*.

En un intento por mejorar la detección del sarcasmo en textos árabes cortos, Al -Jamal et al. [46] utiliza dos técnicas de aumentación de datos en texto, *random swap y random deletion* con el fin de balancear las clases de su conjunto de datos iSarcasmEval (SemEval, 2022). Para luego detectar sarcasmo mediante BERT. Sin embargo, sus resultados fueron menores a 60% en esta tarea para la métrica *F1-score*. Por lo que indican que es necesario construir conjuntos de datos etiquetados más robustos para mejorar la detección de comentarios sarcásticos.

Una aproximación distinta a la aumentación mediante transformación es la propuesta de Kraus et al. [47], donde realizan *random swap y random insertion* en conjunto de datos representado como un *rhetorical structure tree*. Para el análisis de sentimientos, utilizan un clasificador DL llamado *Discourse*-LSTM con el cual obtienen una mejora en el rendimiento con el conjunto de datos Rotten Tomatoes de 4,27% en la métrica *F1-score*.

Kelsingazin et al. [25] proponen dos implementaciones de algoritmos para la aumentación de datos en texto basándose en las operaciones *random insertion y random deletion*. Si embargo, no entregan detalles de la implementación de estos. En cuanto a los clasificadores utilizados, indican que usan SVM y LR, pero no entregan detalles de la implementación ni de los parámetros utilizados. Con relación a los resultados, muestran una mejora en el rendimiento de clasificación de 1% en la métrica *F1-score*.

En el caso de Iosifidis et al. [2], los autores proponen dos técnicas de aumentación. La primera de ellas realiza reemplazo de palabras en una sentencia seleccionando las más parecidas de acuerdo con la similaridad de coseno calculada desde *word embeddings*. La segunda, elimina palabras de la sentencia excepto aquellas que tengan peso sentimental y las que correspondan con negación, esto, con el fin de preservar la clase de la sentencia original. Los autores proponen la utilización de estas técnicas de aumentación en tiempo de entrenamiento de su modelo con el objetivo de corregir el desbalance de las clases al momento de entrenar. A lo largo de su experimentación indican que el método de reemplazo basado en *word embeddings* no se comporta igual de bien que el método de eliminación de palabras, ante esta situación, sospechan que al momento de realizar los reemplazos se cambian las clases de algunas de las muestras (es decir, se modifica el sentido de la frase).

Uno de los trabajos que reporta mayores mejoras en el rendimiento de clasificación es el de Duwairi [26], con una mejora de 42% de incremento en accuracy después de realizar la aumentación de un conjunto de datos con comentarios sobre productos en árabe. El método de aumentación propuesto se basa en el reemplazo de sinónimos desde Arabic-Wordnet y la aplicación de reglas sintácticas propias del idioma árabe para generar nuevas sentencias. En

su experimentación, evalúan los conjuntos de datos aumentados en tres clasificadores de ML (NB, KNN, SVM) y llegan a aumentar el conjunto de datos hasta diez veces, consiguiendo el aumento en el rendimiento de clasificación de más de 40% comparado con el conjunto de datos sin aumentar.

Con el objetivo de mantener la semántica de las sentencias aumentadas, Feng et al. [48] proponen dos algoritmos, el primero de ellos realiza reemplazo probabilístico de sinónimos seleccionando las palabras desde un lexicón. Mientras que el segundo, extrae el peso de las palabras en el contexto mediante el uso de TF-IDF y reemplaza la que tenga menor peso con respecto del contexto. Para la clasificación, utilizan un modelo basado en CNN, con lo cual logran una mejora de 5% en la clasificación con respecto a las líneas base definidas en la métrica *accuracy*.

Santoso et al. [27], proponen una extensión de EDA con la cual logran una mejora de entre 0,6% y 3,4% en *accuracy* con respecto a un conjunto de datos sin aumentar. La mejora a EDA consiste en la proposición de dos algoritmos para la aumentación, el primero mediante reemplazo manteniendo la información semántica de la sentencia y el segundo realizando desambiguación mediante el algoritmo Adapted Lesk. En su experimentación, tuvieron problemas con la clase neutral debido a que su modelo no puede identificar las palabras con mayor importancia en las sentencias con esa clase.

El trabajo presentado por Haralabopoulus et al. [7], abarca la clasificación de emociones y polaridad mediante LSTM y la aumentación por reemplazo de antónimos, inserción de negación y permutación. Su modelo de aumentación, al contrario de los otros, trata de aumentar el conjunto de datos mediante el cambio de clase en la sentencia aumentada con respecto a la sentencia original. Con este enfoque logran una mejora de rendimiento en accuracy de 4,1% con respecto a las líneas base. Uno de los hallazgos en este trabajo, es que no se preocupan de mantener la semántica de las sentencias, lo cual va en una dirección completamente distinta de los otros artículos.

Wei et al. [49] aumentan el conjunto de datos mediante las técnicas de reemplazo de sinónimos, inserción al azar, eliminación al azar y cambio al azar para luego entrenar un modelo de transferencia de aprendizaje con dos instancias de BERT, una como estudiante y otra como profesor. Los resultados de su modelo muestran que mantienen el rendimiento de BERT en *accuracy* para los conjuntos de datos SST2, YELP y Amazon.

Para finalizar, los desafíos que se vislumbran para llevar a cabo la aumentación de conjuntos de datos de textos en español, es precisamente la reducida cantidad de conjuntos de datos disponibles para la experimentación, como ejemplo de esto, en el estado del arte,

ninguno de los autores realiza aumentación en conjuntos de datos en español. Sin embargo, Pei et al. [5] mencionan que LaBSE de Google se encuentra entrenado en múltiples idiomas (149) incluido el español. Mientras que Bogoradnikova et al. [14] indica que, dentro de los conjuntos de datos revisados para la detección multilenguaje de toxicidad de mensajes de texto en ruso, existe un conjunto de datos en español con 8.000 mensajes anotados. Otro de los desafíos tiene relación con las técnicas utilizadas para la aumentación y como manejan la conservación de la semántica de la muestra original. Para solucionarlo, [36, 52] proponen el uso de TF-IDF, pLSA, *word embeddings*, entre otros.

3.2.6 Trabajos relacionados mediante generación de oraciones

Carrasco et al. en [35] realizan la aumentación de un conjunto de datos que abarca 5 dialectos del árabe mediante Sentimental GAN (SentiGAN), la composición de SentiGAN [58] se basa en múltiples generadores y un discriminador. Además, para mejorar la calidad de las oraciones generadas por su modelo de aumentación, utilizan una función de recompensa en lugar de la función de penalidad usada en los modelos GAN. Adicionalmente, con el objetivo de asegurar la novedad y diversidad del conjunto de datos aumentado, proponen dos funciones, la primera para asegurar que las nuevas oraciones creadas por el generador son diferentes de las oraciones originales. Mientras que la segunda, se encarga de medir que tan distintas son las nuevas oraciones de la original. Los resultados de clasificación con su modelo de generación aplicado a distintos clasificadores tanto de ML como DL superan el 80% en la métrica *accuracy* para el conjunto de datos MADAR.

Con el objetivo subsanar el desbalance de clases en un conjunto de datos en el idioma hindi, Rafi-Ur-Rashid et al. en [18] utilizan un modelo de generación GAN para aumentar la clase con menor cantidad de muestras etiquetadas. Su modelo obtiene resultados mínimos de *F1-Score* de 90%, mientras que para *accuracy* consiguen un rendimiento que no supera el 65%. Finalizan su trabajo indicando que intentaran aplicar nuevos modelos de generación de texto para aumentar datos de texto en sus próximas investigaciones.

Continuando con los trabajos que utilizan modelos generativos para aumentar datos de texto, Shang et al. [33] proponen un modelo GAN mejorado con *Transformers* que logra mejorar los resultados en la métrica *accuracy* llegando a 98,44% para un conjunto de datos en el idioma chino utilizando un clasificador de sentimientos basado en BERT. Uno de los puntos a rescatar es que destacan los beneficios de la utilización de las técnicas de aumentación de

datos de texto, ya que ayudan a balancear conjuntos de datos y evitar el problema de *overfitting*[12].

Rahul Gupta [34] propone un modelo GAN compuesto por 3 componentes. Un generador basado en redes neuronales que genera oraciones, un clasificador basado también en redes neuronales que se encarga de evaluar la calidad de las oraciones creadas por el generador y por último un discriminador que evalúa si la oración es falsa o no. Sin embargo, su modelo necesita ser entrenado previamente con una gran cantidad de datos para obtener buenos resultados en la generación de oraciones de calidad. El rendimiento de clasificación obtenido logra una mejora de alrededor de 10% en *accuracy*, llevando de 64,5% para el conjunto de datos sin aumentar a 74% luego de aumentar el conjunto de datos con su modelo *conditional* GAN (cGAN).

En [36], Luo et al. utilizan una variante de GAN (SeqGAN [59]) para aumentar conjuntos de datos con textos largos, además, corrigen el problema que tiene SegGAN [59] al generar oraciones que pueden perder la información semántica del conjunto de datos de origen. Para agregar nuevas oraciones al conjunto de datos, primero entrenan un clasificador con los datos sin aumentar y luego comparan los resultados de la clasificación con los datos aumentados, solamente las oraciones que tienen correctamente asignada la etiqueta de clase para el análisis de sentimientos son agregadas. Una de las razones por las cuales eligieron SeqGAN [59] como modelo de generación de sentencias es que no necesita ser pre-entrenado con una gran cantidad de datos.

Por otra parte, Liu et al. [52] proponen un modelo de aumentación basado en un modelo de lenguaje que conserva la información de la etiqueta sentimental al momento de hacer la aumentación de datos de texto, esto, lo hacen enmascarando las palabras que no tienen la carga afectiva y la reemplazan por sus sinónimos. Al igual que Carrasco et al. [35], indican que los modelos generativos pueden generar sentencias que son iguales a las ya existentes en el conjunto de datos. Los resultados obtenidos por su modelo para un conjunto de datos en idioma chino llegan a 94,25%.

Pandey et al. [54] utiliza GAN para realizar pruebas de aumentación para 12 conjuntos de datos logrando un rendimiento de clasificación en análisis de sentimientos de 94,5% en *accuracy*. Esto lo logran al utilizar PosGAN para mantener la estructura sintáctica de las oraciones generadas con respecto a las que se encuentran en un conjunto de datos de referencia. Es de los pocos trabajos que realizan análisis de emociones además de análisis

[12] https://aws.amazon.com/what-is/overfitting/

de sentimientos. Un punto interesante es que proveen acceso tanto al conjunto de datos y el código de su modelo de aumentación.

3.2.7 Trabajos relacionados mediante parafraseo de oraciones

El trabajo realizado por Krishnan et al. [32] realiza la aumentación mediante la traducción de oraciones desde el idioma inglés al hindi, con el fin de aumentar la cantidad de oraciones que son utilizadas para su modelo de clasificación, que consiste en un modelo profesor-alumno mediante la utilización de mBERT y XLM-R. Los autores no profundizan en la técnica de aumentación, solamente indican que puede utilizarse cualquier herramienta de traducción o transliteración. Los resultados de su modelo en el conjunto de datos de pruebas logran preservar o mejorar su rendimiento con respecto de la línea base obteniendo para mBERT 61,35% y 66,24% para los lenguajes hindi y malayo en la métrica *F1-Score* y 62,23% y 76,46% para los mismos lenguajes utilizando XLM-R.

Por otra parte, Tang et al. en [11], utilizan la aumentación mediante *back-translation* traduciendo textos desde chino al inglés mediante el api Baidu para aumentar el número de muestras del conjunto de datos de entrenamiento que consisten en textos de *micro-blogs* chinos, con textos en inglés, chino y japonés pertenecientes a la tarea compartida 1 de NLPCC 2018. Lo anterior como parte de su modelo BERT-MSAUC que clasifica las emociones felicidad, tristeza, miedo, ira y sorpresa. Los resultados obtenidos por su modelo en la métrica *F1-Score* superan en dos (miedo e ira) de las cinco clases emotivas mencionadas con anterioridad a un modelo BERT(M) utilizado como línea base. El rendimiento de clasificación por cada clase es:

Tabla 5 Resultados de clasificación con F1-Score Tang et al.[11]

Clase emotiva	BERT(M)	BERT-MSAUC
Felicidad	77,3%	76,9%
Tristeza	63,4%	61,4%
Miedo	65,8%	70,6%
Ira	45,2%	51,3%
Sorpresa	50,1%	49,7%

Para terminar con los trabajos que realizan manipulación de oraciones mediante parafraseo, tenemos el de Bogoradnikova et al. [14], que realiza análisis de sentimientos,

detección de comentarios tóxicos y detección de porciones de texto tóxicos mediante en el idioma ruso. En su investigación, comparan un modelo SVM con el api Perspective[13] (api utilizado para moderación de contenido), en primera instancia el rendimiento de clasificación del modelo SVM es de 61,83% en *accuracy* luego de aumentar el conjunto de datos. Luego, utilizan EDA y *back-translation* para aumentar el conjunto de datos, pero no entregan detalles del proceso de aumentación. Luego de aumentar, los resultados de clasificación con su modelo SVM en la tarea de análisis de sentimientos mejora su rendimiento en 10% llegando a 95% de rendimiento de clasificación en la métrica *accuracy*.

[13] https://www.perspectiveapi.com

Capítulo 4 Conjuntos de datos utilizados

A lo largo de este Capítulo, se revisarán los conjuntos de datos producidos por actividades docentes y utilizados en experimentos realizados por el grupo SOMOS de la Universidad del Bío-Bío que fueron aumentados con las técnicas de aumentación de texto por transformación, generación y parafraseo. Adicionalmente, por cada conjunto de datos revisado se agregan las estadísticas extraídas desde la bolsa de palabras.

4.1 Dieciocho de Octubre

El conjunto de datos “18 de Octubre” está compuesto de comentarios recopilados de *Twitter* en el contexto del estallido social que se produjo en Chile en el año 2019. El conjunto se encuentra etiquetado con 8 categorías para *Emotion Analysis*. La Fig. 14 muestra las palabras que tienen más frecuencia dentro de este conjunto de datos.

Fig. 14 Nube de palabras 18Octubre

Mientras que, la Tabla 6 muestra las estadísticas extraídas para el conjunto de datos 18 de Octubre.

Tabla 6 Estadística 18 de Octubre

Descripción	Valor
Cantidad total de palabras	3.336
Cantidad de palabras únicas	1.680
Promedio de palabras por oración	17

4.2 Agresividad

El conjunto de datos de Agresividad utilizado en el trabajo de Ledez et al. [60] está compuesto de comentarios de *Twitter* usados para detectar *cyberbulling* en español. Las palabras que aparecen con mayor frecuencia en este conjunto de datos pueden apreciarse en la Fig. 15

Fig. 15 Nube de palabras Agresividad

Las estadísticas extraídas desde el conjunto de datos de Agresividad son mostradas a través de la Tabla 7.

Tabla 7 Estadística Agresividad

Descripción	Valor
Cantidad total de palabras	33.143
Cantidad de palabras únicas	8.685
Promedio de palabras por oración	22

4.3 Emoji

El conjunto de datos de Emoji está compuesto por comentarios de *Twitter* para *Emotion Analysis*, el objetivo de este conjunto de datos es demostrar que el emoji en un texto está asociado a la emoción en el texto. Para este trabajo se utilizó el conjunto de datos de Emoji limpio, esto quiere decir, solamente con los comentarios textuales de cada tweet. En la Fig. 16 se muestran las palabras que aparecen con mayor frecuencia en este conjunto.

Fig. 16 Nube de palabras Emoji

Las estadísticas extraídas desde el conjunto de datos de Emoji son mostradas a través de la Tabla 8.

Tabla 8 Estadística Emoji

Descripción	Valor
Cantidad total de palabras	11.120
Cantidad de palabras únicas	4.234
Promedio de palabras por oración	11

4.4 Encuesta Docente

Este conjunto de datos es una recolección de los comentarios realizados por alumnos en la encuesta docente realizada en la Universidad del Bío-Bío en el año 2018. Este conjunto a diferencia de los otros se encuentra etiquetado con la finalidad de realizar clasificación a 4 distintas categorías por cada muestra (Afecto, Agresividad, Polaridad, Seriedad). Es por ello, que fue dividido en 4 conjuntos de datos, 1 por cada categoría. La Fig. 17 muestra las palabras más frecuentes en el conjunto de datos.

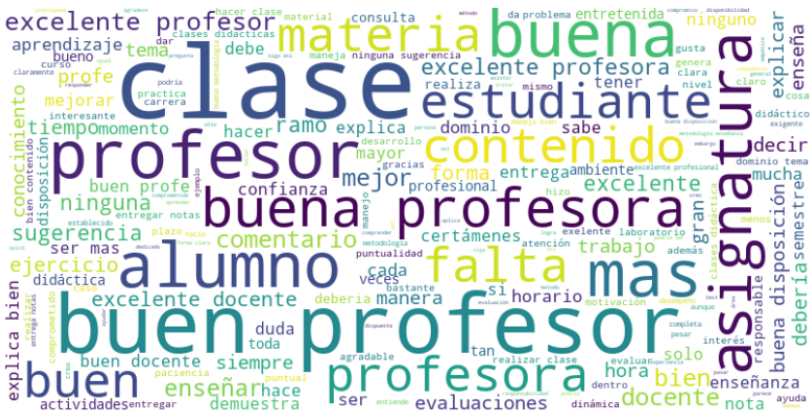

Fig. 17 Nube de palabras Encuesta Docente

La Tabla 9 muestra las estadísticas extraídas del conjunto de datos.

Tabla 9 Estadística Encuesta Docente

Descripción	Valor
Cantidad total de palabras	8.645
Cantidad de palabras únicas	2.119
Promedio de palabras por oración	5

4.5 Titulares de diario

El conjunto de datos de Titulares de diario fue creado para el trabajo de Martínez-Araneda et al. [61] con el objetivo de analizar el sesgo de los titulares de diarios chilenos entre los años 2014 y 2015. La Fig. 18 refleja las palabras con mayor frecuencia en el conjunto de datos.

Fig. 18 Nube de palabras Titulares Diarios

La Tabla 10 muestra las estadísticas extraídas para este conjunto de datos.

Tabla 10 Estadística Titulares Diarios

Descripción	Valor
Cantidad total de palabras	24.245
Cantidad de palabras únicas	8.447
Promedio de palabras por oración	12

4.6 Violencia de Género

Creado para el trabajo de Martínez-Araneda et al. [62], este conjunto de datos agrupa canciones en español de distintos géneros musicales con el objetivo de identificar aquellas que contengan violencia hacia la mujer. La Fig. 19 muestra gráficamente las palabras más frecuentes en este conjunto.

Fig. 19 Nube de palabras Violencia Género

La Tabla 11 muestra las estadísticas extraídas para este conjunto de datos.

Tabla 11 Estadística Violencia Género

Descripción	Valor
Cantidad total de palabras	44.024
Cantidad de palabras únicas	8.843
Promedio de palabras por oración	44

4.6.1 Clasificación de conjuntos de datos

Para realizar la clasificación de los conjuntos de datos elegidos para ser aumentados, se utilizan los criterios de la

Tabla 12. Esta clasificación busca poder comparar los conjuntos de datos de acuerdo con sus características con el fin de analizar los resultados de los experimentos de aumentación que serán realizados sobre ellos.

Tabla 12 Criterios para clasificación de conjuntos de datos

Ítem	**Valor**	**Descripción**
Balanceado (B)	SI / NO	Indica si la cantidad de muestras es la misma por cada clase.
Texto formal (TF)	SI / NO	Indica que el ámbito en el que se escribe el texto tiene restricciones en cuanto al tipo de palabras y expresiones utilizadas.
Largo promedio por oración (LPO)	Número	Cantidad de palabras que conforman una oración
Tamaño (T)	Tamaño <= 500 = S Tamaño > 500 y <= 1000 = M Tamaño > 1000 = L	Representación en categorías de la cantidad total de muestras por conjunto de datos
Cantidad de muestras (CM)	Número	Cantidad total de muestras por conjunto de datos
Tarea clasificación (TC)	Sentiment analysis = SA Emotion analysis = EA	Tarea de clasificación que busca resolver el conjunto de datos
Cantidad de clases (CC)	Número	Cantidad de clases en un conjunto de datos
Tipo de contenido (TC)	Comentarios Canciones Titulares	Tipo de contenido que conforma el conjunto de datos

Ahora que se encuentran definidos los criterios de clasificación para los conjuntos de datos, la Tabla 13 refleja cada conjunto de datos con su clasificación correspondiente.

Tabla 13 Clasificación conjuntos de datos

Conjunto de datos	B	TF	LPO	T	CM	TC	CC	TC
18 Octubre	NO	NO	17	S	195	EA	8	Comentarios
Agresividad	SI	NO	22	L	1.470	SA	2	Comentarios
Emoji	NO	NO	11	M	1.000	EA	11	Comentarios
Encuesta Docente Afecto	NO	SI	5	L	1.632	EA	7	Comentarios
Encuesta Docente Agresividad	NO	SI	5	L	1.632	SA	2	Comentarios
Encuesta Docente Polaridad	NO	SI	5	L	1.632	SA	3	Comentarios
Encuesta Docente Seriedad	NO	SI	5	L	1.632	SA	2	Comentarios
Titulares Diario	NO	SI	12	L	2.031	EA	9	Titulares
Violencia Género	NO	NO	44	M	1.000	SA	2	Canciones

Capítulo 5 Experimentación

Este Capítulo presenta el detalle de los experimentos llevados a cabo para realizar la evaluación de la influencia de las técnicas de aumentación sobre los conjuntos de datos descritos en el Capítulo 4. En primer lugar, se definen los experimentos, para luego presentar los resultados de las diferentes técnicas de aumentación por cada uno de los clasificadores seleccionados.

5.1 Descripción del hardware utilizado

Los experimentos fueron llevados a cabo en un servidor con las características descritas a continuación:

- Sistema operativo Debian GNU/Linux 12 (*bookworm*)
- 2 x Intel Xeon® CPU E5-2683 v4 2.10GHz, 2 *sockets*, 16 *cores*, 2 *threads/core* con *hyperthreading* activado para un total de 64 *threads*.
- 256 GB de memoria principal.

La Fig. 20 muestra en detalle las características del hardware utilizado para la realización de los experimentos.

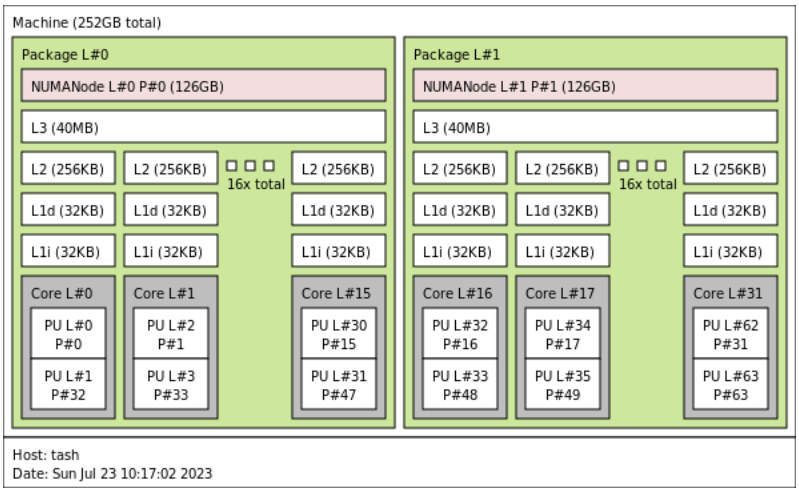

Fig. 20 Descripción del hardware utilizado en experimentos extraída con comando lstopo

5.2 Formato de registro de resultados

El registro de los resultados de los experimentos con las distintas técnicas de aumentación que fueron utilizadas en este trabajo es llevado a cabo de acuerdo con la Tabla 14.

Tabla 14 Plantilla presentación de resultados

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Conjunto datos	Base					
	Técnica DA 1					
	Técnica DA 2					
	Técnica DA n					

5.3 Clasificación sin aumentación

Las siguientes secciones describen el procedimiento utilizado para obtener el rendimiento de clasificación que fue utilizado como base de comparación para los experimentos de clasificación de los conjuntos de datos posterior a la aumentación.

5.3.1 Procedimiento de clasificación sin aumentación

Como líneas base de clasificación de sentimiento y emociones, fueron utilizados los siguientes algoritmos de clasificación que representan los más utilizados en la revisión de literatura.

- CNN
- LSTM
- BiLSTM
- BERT (BETO)
- SVM

Las muestras disponibles en los conjuntos de datos fueron divididas en 70% para entrenamiento y 30% para validación. Los subconjuntos de validación fueron los mismos para

todos los clasificadores tanto para obtener la línea base como para la obtención del rendimiento de clasificación posterior a la aumentación.

Los hiper parámetros utilizados por cada uno de los clasificadores puede observarse en la Tabla 15. Estos hiper parámetros son los mismos utilizados para realizar la clasificación luego de llevar a cabo la aumentación con el fin de obtener la comparación del rendimiento de los modelos de ML y DL con las mismas condiciones en el conjunto de datos original y aumentado.

Tabla 15 Hiper parámetros modelos de clasificación

Modelo	Hiper parámetros
CNN	Embedding_dim = 20, Conv1D(32, 3, activation = relu, padding = same), 4 MaxPooling1D(3, padding = same), Conv1D(64, 3, activation = relu, padding = same), Conv1D(64, 3, activation = relu, padding = same), 2 Conv1D(128, 3, activation = relu, padding = same), dropout = 0.5, Dense(11, activation = softmax)
LSTM	Embedding_dim = 16, salidas LSTM = 32, 1 unidad LSTM, 1 Dense(19, activation = relu), 1 Dense(11, activation = softmax), loss = sparse_categorical_crossentropy, optimizer = ADAM, metrics = accuracy)
BiLSTM	Embedding_dim = 100, 2 unidades LSTM(100, dropout = 0.3), 1 Conv1D(100, 5, activation = relu), 1 Dense(16, activation = relu), 1 Dense(11, activation = softmax)
BERT	model_name = 'dccuchile/bert-base-spanish-wwm-cased' batch_size = 16, optimizer = ADAM, learning_rate = 2e-5, epochs = 10
SVM	param_range = [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0] param_grid = [{'svc__C': param_range, 'svc__kernel': ['linear']}, {'svc__C': param_range, 'svc__gamma': param_range, 'svc__kernel': ['rbf']}] GridSearchCV(estimator=pipe_svm, param_grid=param_grid, scoring='accuracy', cv=10, n_jobs=-1)

5.3.2 Resultados de línea base para la comparación

Los resultados de clasificación de los modelos de ML y DL sobre los conjuntos de datos sin aumentar se reflejan en la Tabla 16. En ella pueden observarse los resultados con las métricas *accuracy* y *f1-score* para cada conjunto de datos por cada uno de los modelos de clasificación.

Tabla 16 Resultados de clasificación sin aumentar

Conjunto de datos	Clasificador									
	SVM		CNN		LSTM		BiLSTM		BERT	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
18 Octubre	56%	10%	56%	40%	56%	40%	56%	40%	**63%**	**70%**
Agresividad	90%	90%	82%	82%	84%	83%	58%	43%	**90%**	**91%**
Emoji	**29%**	32%	22%	10%	21%	25%	23%	9%	28%	**40%**
Encuesta Docente Afecto	57%	53%	52%	53%	**59%**	54%	33%	17%	56%	**55%**
Encuesta Docente Agresividad	95%	92%	95%	92%	95%	92%	95%	92%	**96%**	**94%**
Encuesta Docente Polaridad	71%	69%	75%	75%	72%	73%	56%	40%	**80%**	**79%**
Encuesta Docente Seriedad	**79%**	70%	76%	76%	**79%**	70%	**79%**	70%	77%	**73%**
Titulares de Diario	39%	**43%**	26%	23%	31%	22%	27%	11%	**44%**	**43%**
Violencia de Género	**86%**	**86%**	76%	66%	83%	81%	76%	66%	83%	83%

De los resultados expuestos, puede apreciarse que el modelo de clasificación que obtiene mejores resultados ya sea en *accuracy* o *F1-Score* es BERT (BETO). Sin embargo, en el conjunto de datos de Violencia de Género el modelo SVM es el que obtiene los mejores resultados en ambas métricas con una diferencia de 3% con respecto a su seguidor más cercano (BERT).

5.4 Procedimientos de aumentación

En las siguientes secciones, se detallan los procedimientos asociados a las distintas técnicas de aumentación que serán utilizadas para aumentar artificialmente los conjuntos de datos descritos en el Capítulo 4.

5.4.1 Procedimiento de aumentación con técnicas de transformación

La técnica más representativa de esta categoría es EDA [31] que utiliza los métodos de reemplazo de sinónimos, cambio de orden de palabras, eliminación de palabras y agregación de palabras en una oración. Para los experimentos llevados a cabo en este trabajo no se consideró el método de eliminación, ya que, según los autores de EDA, la aplicación de dicho método empeora los resultados de clasificación en los conjuntos de datos aumentados.

Para aumentar los conjuntos de datos definidos en el Capítulo 4, el subconjunto de entrenamiento es aumentado en dos dimensiones. La primera de ellas es el porcentaje de modificación sobre una oración, mientras que el segundo es la cantidad de aumentación realizadas sobre el conjunto de datos. La Tabla 17 muestra la distribución porcentaje/cantidad aplicada.

Tabla 17 Parámetros de aumentación EDA

Porcentaje de modificación	Cantidad de aumentaciones
5	1,2,3,4,5,6,7,8,9
10	1,2,3,4,5,6,7,8,9
20	1,2,3,4,5,6,7,8,9
30	1,2,3,4,5,6,7,8,9

El segundo experimento corresponde a la selección del mejor porcentaje de modificación luego de aumentar con EDA y clasificar, para luego aumentar la cantidad de muestras de las clases minoritarias en los conjuntos de datos utilizando dicho porcentaje para realizar balanceo del conjunto de datos.

5.4.2 Procedimiento de aumentación con técnicas generativas

Los experimentos de aumentación con técnicas generativas fueron realizados mediante redes adversariales generativas (GAN en inglés) a través del modelo SentiGAN propuesto por Wang et al. en [58]. Para realizar la aumentación con este modelo, se separaron las clases de cada conjunto de datos en archivos independientes para luego aumentarlos uno a uno. La aumentación con SentiGAN [58] se realizó de esta forma ya que a pesar de que el modelo dice estar preparado para generar aumentación para múltiples clases presentes en un conjunto de datos, en la práctica solamente puede generar para 1 sola clase.

Los parámetros de configuración utilizados para la aumentación con SentiGAN se muestran en la Tabla 18.

Tabla 18 Parámetros de configuración SentiGAN

Generador	
EMB_DIM	200
HIDDEN_DIM	200
MAX_SEQ_LENGTH	Promedio palabras por oración en conjunto de datos
BATCH_SIZE	Cantidad de oraciones por clase
Discriminador	
DISM_EMBEDDING_DIM	64
DIS_FILTER_SIZES	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15]
DIS_NUM_FILTERS	[100, 200, 200, 200, 200, 100, 100, 100, 100, 100]
DIS_DROPOUT_KEEP_PROB	0.75
DIS_L2_REG_LAMBDA	0.2
DIS_BATCH_SIZE	Cantidad de oraciones por clase
TOTAL_BATCH (proceso adversarial)	100

5.4.3 Procedimiento de aumentación con técnicas de parafraseo

Los experimentos de aumentación de los conjuntos de datos mediante las técnicas de parafraseo serán llevados a cabo con la utilización del servicio de traducción de Google[14], esto, debido a que las API para traducción de texto normalmente son de pago (por ejemplo, el api de Google).

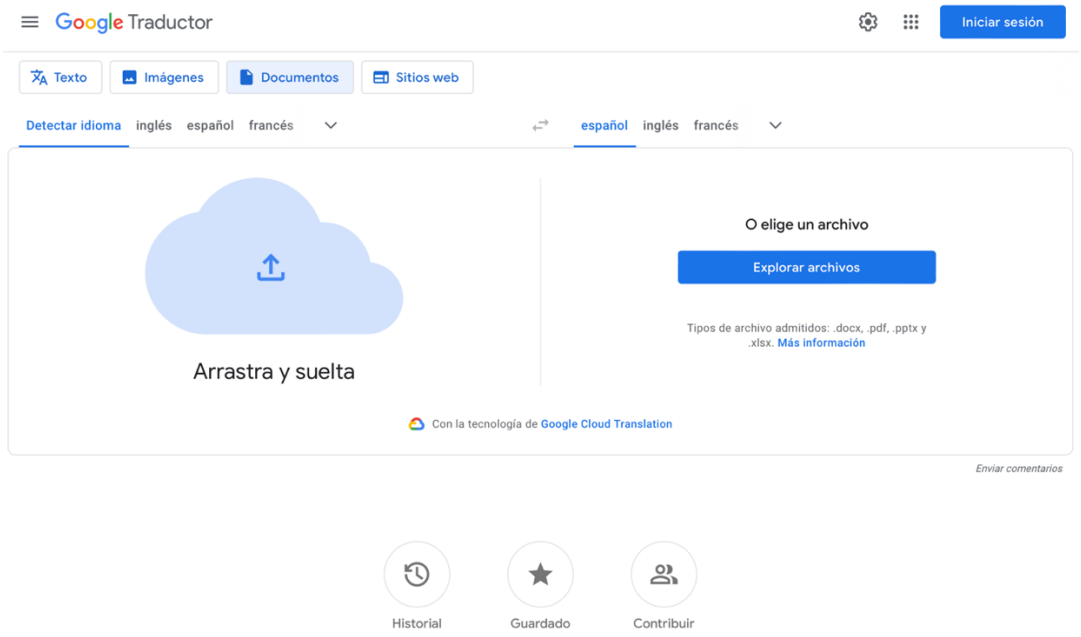

Fig. 21 Servicio Google Traductor

El procedimiento para obtener las traducciones se basa en la opción de Google Traductor que permite traducir documentos completos (Fig. 21). Para comenzar se suben los conjuntos de datos en formato xlsx al sitio, para luego descargar la traducción en 3 idiomas (inglés, alemán y francés). El siguiente paso consiste en subir los documentos traducidos y realizar la traducción de vuelta al español. Obteniendo así, oraciones ligeramente modificadas que son agregadas al conjunto de datos de entrenamiento para los modelos de clasificación utilizados.

Los conjuntos de datos son aumentados incrementalmente con los resultados obtenidos de la traducción, de acuerdo con la siguiente tabla.

[14] https://translate.google.com

Tabla 19 Niveles de aumentación con back translation

Aumentaciones	Idiomas
1	Base (español) + inglés
2	Base (español) + inglés + alemán
3	Base (español) + inglés + alemán + francés

5.5 Resultados de clasificación con técnicas de transformación

Los experimentos de aumentación y posterior clasificación sobre los conjuntos de datos aumentados con EDA [31] fueron llevados a cabo tal como se indicó en el punto 5.2.2. Los resultados por cantidad de aumentaciones y porcentaje de modificación de oraciones son mostrados a través de mapas de calor por cada uno de los conjuntos de datos, para posteriormente representar los mejores resultados a través de la plantilla propuesta en la Tabla 14.

5.5.1 Ejemplos de aumentación con EDA

Para realizar la aumentación con EDA [31], fue necesario realizar modificaciones al código fuente del modelo para que fuese compatible con el idioma español. En primer lugar, se definió explícitamente español como lenguaje de WordNET [63]. La siguiente modificación realizada fue cambiar el listado de *stopwords* en inglés por el listado de *stopwords* en español que proporciona la librería NLTK[15].

Como resultado de lo anterior la

Tabla 20 muestra una oración sin aumentar y una con la aumentación mediante EDA [31] correspondiente al conjunto de datos de Titulares de Diarios.

Tabla 20 Ejemplo de aumentación de texto con EDA

Oración original	Al menos dos policías muertos y cinco heridos en emboscada guerrillera en Perú (policía)

[15] https://www.nltk.org

Oración aumentada con EDA 30% de modificación	heridos menos dos polic cinco muertos y en al en emboscada polic as per guerrillera a aluminum basketball team guerrillera en per polic a

5.5.2 Resultados para Dieciocho de Octubre

Los resultados de clasificación medidos con la métrica *accuracy* luego de aumentar el conjunto de datos 18 Octubre son presentados a continuación.

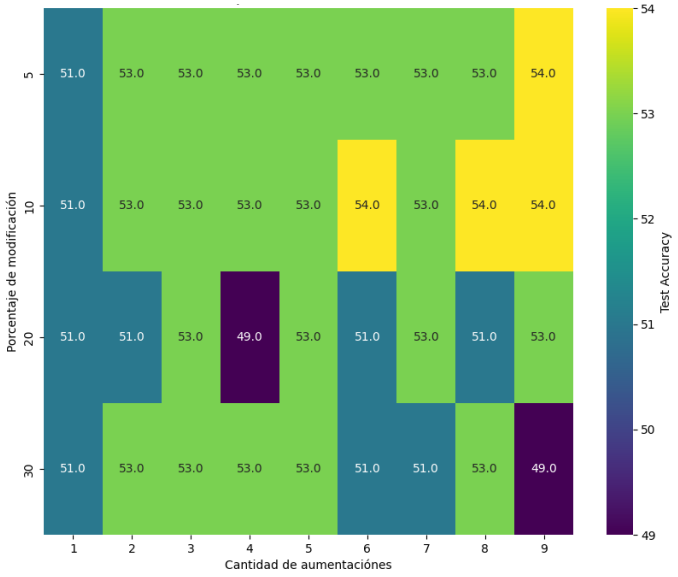

Fig. 22 Resultados 18Octubre, EDA, SVM

La Fig. 22 muestra los resultados de clasificación con el modelo SVM, en ella se puede apreciar que el mejor resultado de clasificación (54%) se produce cuando el conjunto de datos es aumentado en 6, 8 y 9 veces con porcentajes entre 5% y 10% de modificación de palabras por oración. Mientras que el peor resultado de clasificación (49%) se produce al aumentar el conjunto de datos 4 y 9 veces con porcentajes de modificación de palabras por oración de 20% y 30% respectivamente.

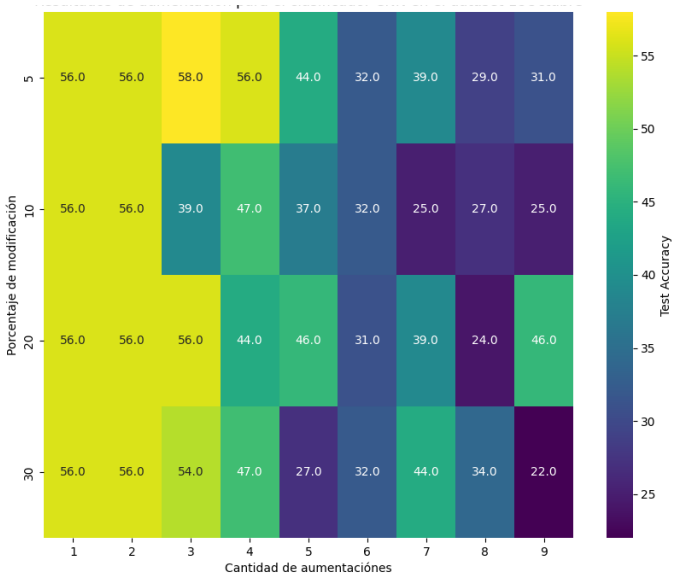

Fig. 23 Resultados 18Octubre, EDA, CNN

En el caso del clasificador CNN, los resultados mostrados en la Fig. 23 muestran que el mejor rendimiento de clasificación (58%) es obtenido cuando el conjunto de datos se aumenta en 3 veces con un porcentaje de modificación de palabras por oración de 5%. Por otro lado, el peor rendimiento se produce cuando el conjunto de datos es aumentado 9 veces y el porcentaje de modificación de palabras por oración aplicado es 30%. Otro punto que destacar es que a medida que se aumenta el conjunto de datos el porcentaje de clasificación baja hasta llegar al punto más bajo en la novena aumentación.

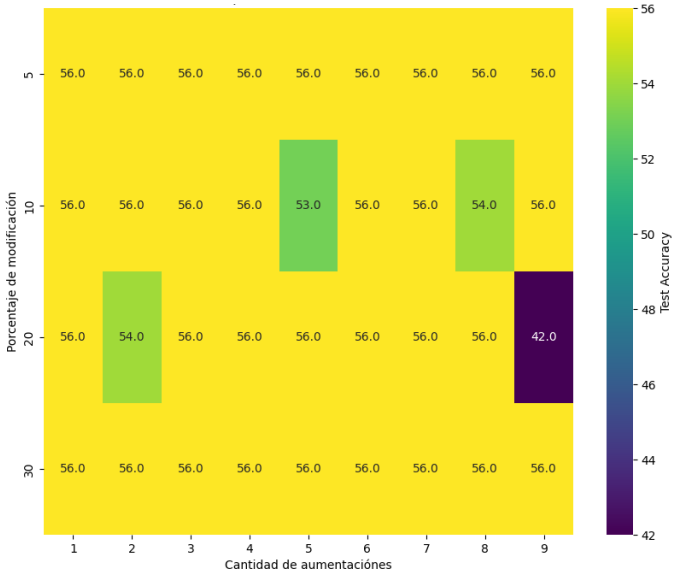

Fig. 24 Resultados 18Octubre, EDA, LSTM

En el caso del clasificador LSTM, los resultados expresados en la Fig. 24 indican que el mejor rendimiento de clasificación (56%) se consigue en la mayor parte de los experimentos con aumentaciones de distintos niveles. Mientras que el peor rendimiento es obtenido cuando el conjunto de datos es aumentado 9 veces y se aplica un porcentaje de modificación de palabras por oración de 20%.

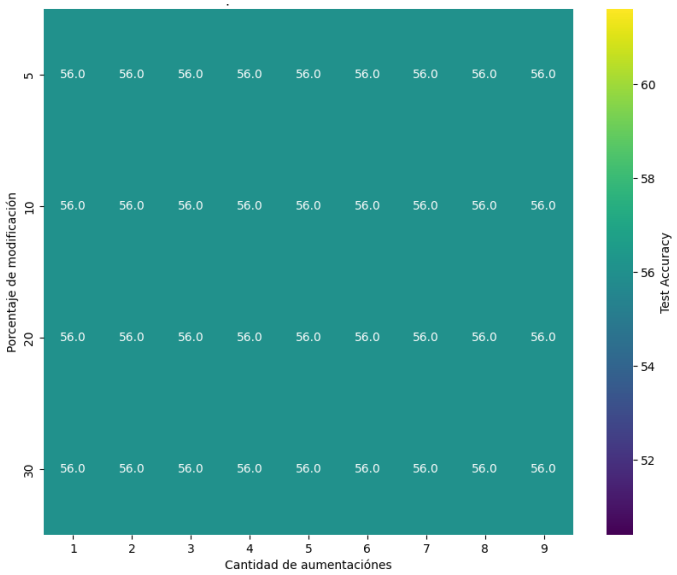

Fig. 25 Resultados 18Octubre, EDA, BiLSTM

Los resultados obtenidos al clasificar el conjunto de datos 18Octubre con el modelo BiLSTM de la Fig. 25, indican que sea cual sea el nivel de aumentación y modificación de palabras en una oración, se obtienen los mismos resultados (56%)

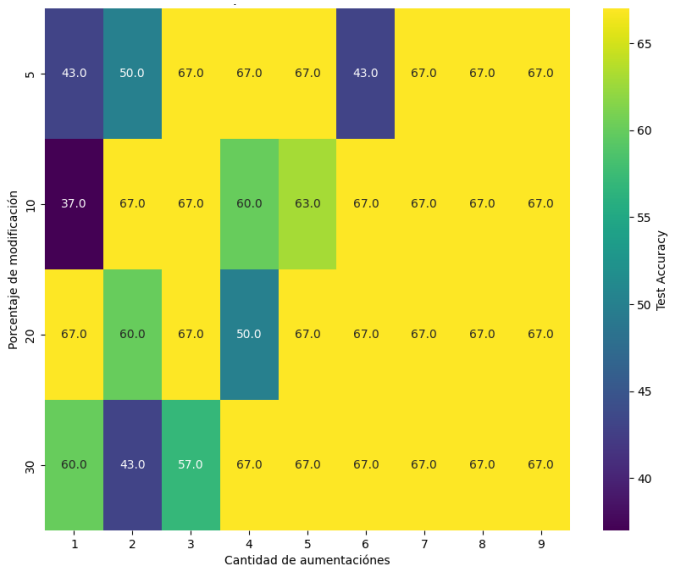

Fig. 26 Resultados 18Octubre, EDA, BERT

Los resultados de clasificación con el modelo BERT representados en la Fig. 26, muestran que el mejor resultado de clasificación (67%) se produce en la mayor parte de los escenarios de aumentación propuestos, siendo el que requiere menor manipulación del conjunto de datos el que aumenta 1 vez el conjunto de datos y aplica un 20% de modificación de palabras por oración. Mientras que, el menor rendimiento de clasificación (37%) el que modifica el 10% de modificación de palabras por oración y aumenta el conjunto 1 vez.

Para finalizar con la revisión de resultados para el conjunto 18Octubre aumentado, se muestra la Tabla 21 que entrega un resumen de los resultados de clasificación y los compara con la línea base.

Tabla 21 Resultados consolidados 18Octubre / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
18 Octubre	Base	56%	56%	56%	56%	63%
	EDA	54%	58%	56%	56%	67%

5.5.3 Resultados para Agresividad

Los resultados para el conjunto de datos de agresividad son representados en las siguientes figuras.

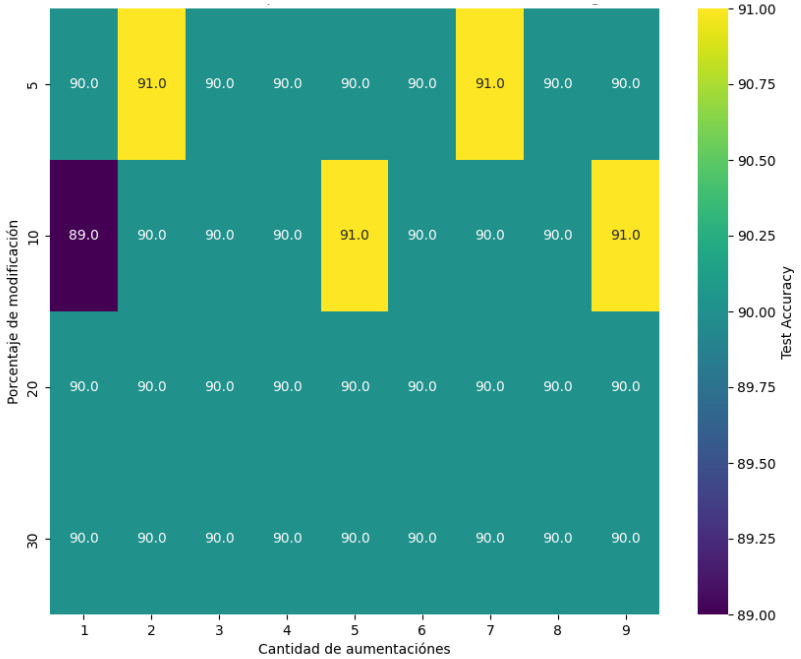

Fig. 27 Resultados Agresividad, EDA, SVM

Los resultados para el clasificador SVM en el conjunto de datos de Agresividad aumentado (Fig. 27) muestran que no existe gran variación en la clasificación. A pesar de que se obtiene el mismo nivel de clasificación con otras combinaciones de aumentación/porcentaje de modificación, el mejor resultado (91%) se da cuando el conjunto de datos es aumentado 2 veces y el porcentaje de modificación de palabras en cada oración es de 5% ya que implica una menor manipulación del conjunto de datos. Por otra parte, el rendimiento de clasificación más bajo (89%) se registra cuando el conjunto de datos es aumentado 1 vez y se modifica el 10% de las palabras por oración.

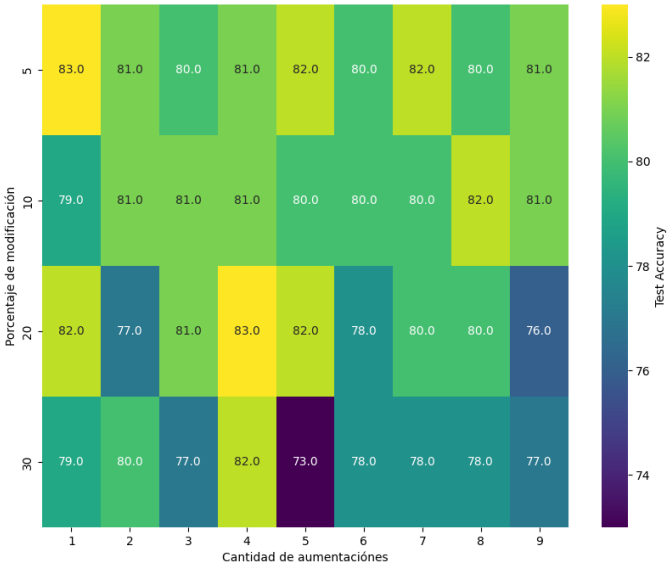

Fig. 28 Resultados Agresividad, EDA, CNN

Los resultados de clasificación para el modelo CNN muestran que el mejor rendimiento de clasificación que además realiza la menor cantidad de modificaciones al conjunto de datos es mostrado en la Fig. 28. En ella, se registra que se obtiene 83% cuando es aumentado 1 vez y se modifica el 5% de las palabras en cada oración. Mientras que el rendimiento de clasificación más bajo (73%) se registra cuando se aumenta el conjunto de datos en 5 veces con un porcentaje de modificación de palabras por oración de 30%.

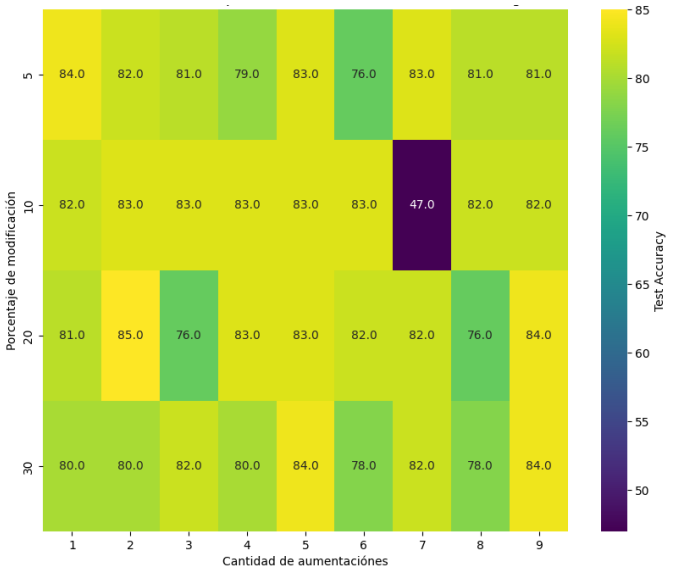

Fig. 29 Resultados Agresividad, EDA, LSTM

La Fig. 29 refleja los resultados de clasificación al utilizar el modelo LSTM, en ella puede observarse que el mejor resultado de clasificación es de 85% y se registra cuando el conjunto de datos es aumentado 2 veces y se modifica el 20% de las palabras pertenecientes a cada oración. Por otra parte, cuando el conjunto de datos es aumentado 7 veces y las palabras por oración son modificadas en un 10% se registra el rendimiento de clasificación más bajo, llegando solamente a un 47%.

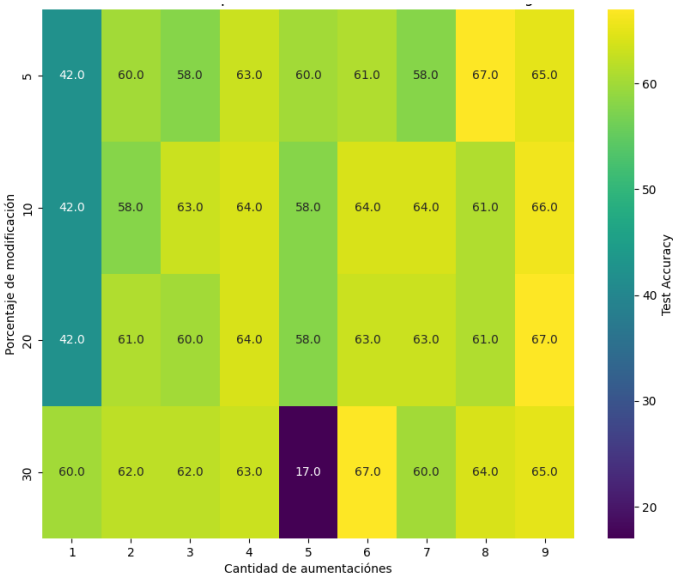

Fig. 30 Resultados Agresividad, EDA, BiLSTM

Los resultados registrados para el clasificador BiLSTM, reflejados en la Fig. 30, muestran que el mejor rendimiento de clasificación se registra al aumentar el conjunto de datos 6 veces

con un porcentaje de modificación de palabras de 30%. Mientras que el rendimiento de clasificación más bajo registrado es de 17% cuando el conjunto de datos se aumenta 5 veces y se modifica el 30% de las palabras en cada oración.

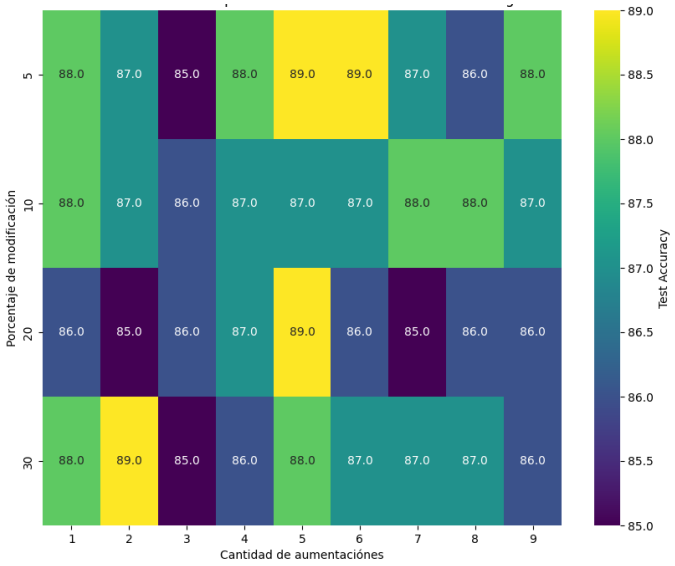

Fig. 31 Resultados Agresividad, EDA, BERT

El mejor resultado de clasificación con el modelo BERT (89%) de acuerdo con la Fig. 31, se da cuando el conjunto de datos es aumentado 2 veces y se modifica el 30% de las palabras en cada oración. Mientras que el rendimiento más bajo es de 85% y se da con distintas combinaciones de aumentación/porcentaje de modificación.

Para finalizar con la revisión de los resultados de clasificación al aumentar con EDA para el conjunto de datos de Agresividad, se presenta la Tabla 22, que muestra tanto el rendimiento de línea base y el rendimiento con el conjunto de datos aumentado.

Tabla 22 Consolidado resultados Agresividad / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Agresividad	Base	90%	82%	84%	58%	90%
	EDA	91%	83%	85%	67%	89%

5.5.4 Resultados para Emoji

Las siguientes figuras grafican los resultados obtenidos luego de aumentar el conjunto de datos Emoji.

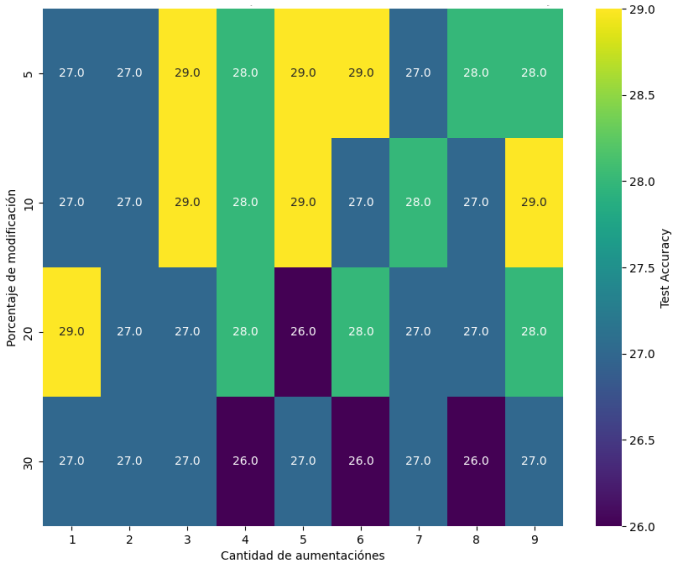

Fig. 32 Resultados Emoji, EDA, SVM

En la Fig. 32, puede apreciarse que el mejor resultado de clasificación (29%) se obtiene cuando el conjunto de datos es aumentado 1 vez y se modifica el 20% de las palabras por oración. Adicionalmente, es el que menor manipulación sobre el conjunto de datos realiza. Mientras que, el menor rendimiento de clasificación (26%) se presenta al aumentar el conjunto de datos 4, 5, 6 y 7 veces con porcentajes de modificación de palabras de 20% y 30%.

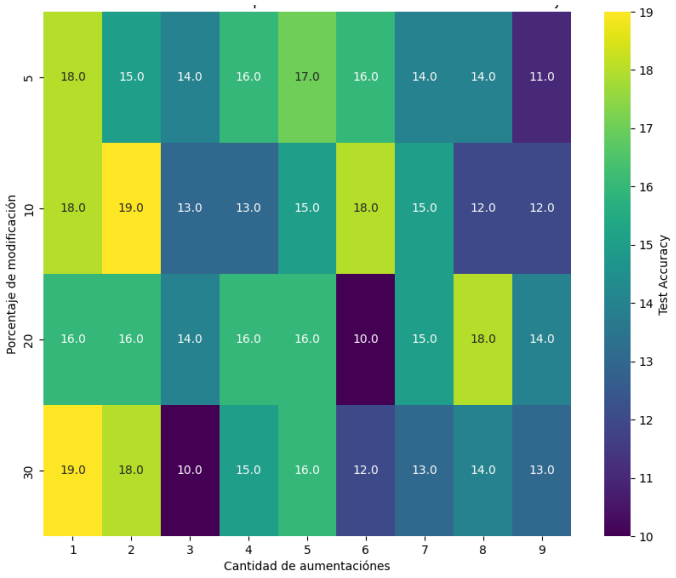

Fig. 33 Resultados Emoji, EDA, CNN

Los resultados de clasificación para el modelo CNN representados en la Fig. 33, muestran que el mejor rendimiento de clasificación (19%) se registra cuando el conjunto de datos es aumentado 1 y 2 veces con porcentajes de modificación de palabras por oración de 10% y 30% respectivamente. Por otra parte, el peor rendimiento de clasificación (10%) se obtiene al aumentar el conjunto de datos 3 y 6 veces con porcentajes de modificación de palabras por oración de 30% y 20% respectivamente.

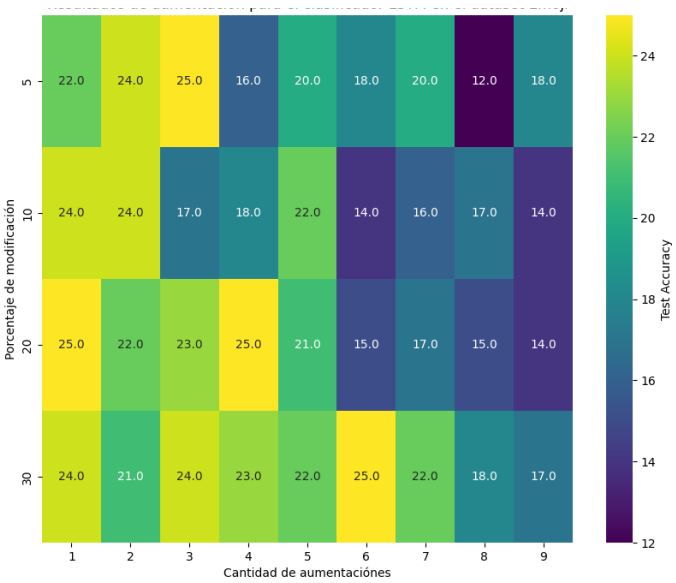

Fig. 34 Resultados Emoji, EDA, LSTM

La Fig. 34 Muestra los resultados obtenidos para el clasificador LSTM, donde el mejor resultado de clasificación (25%) que manipula menos el conjunto de datos se logra cuando el

conjunto de datos es aumentado 1 vez y se aplica un porcentaje de modificación de palabras por oración de 20%. Por el contrario, el menor resultado de clasificación (12%) es obtenido cuando el conjunto de datos es aumentado 8 veces y se modifica el 5% de las palabras por oración.

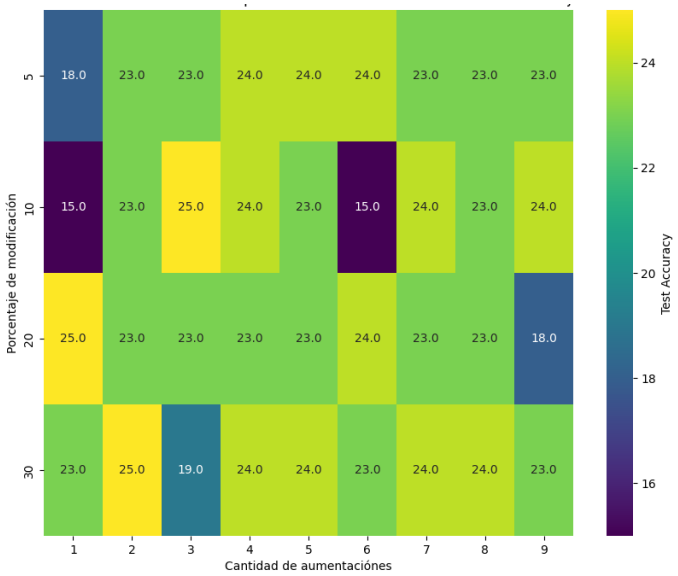

Fig. 35 Resultados Emoji, EDA, BiLSTM

Pasando a los resultados del siguiente clasificador, la Fig. 35 muestra los resultados para el modelo BiLSTM registrando que el mejor resultado de clasificación (25%) que manipula en menor medida el conjunto de datos se da cuando se aumenta el conjunto 1 vez y se modifica el 20% de las palabras por oración. Mientras que, el peor resultado de clasificación (15%) se presenta cuando el conjunto de datos es aumentado 1 vez y se modifica el 10% de las palabras por oración.

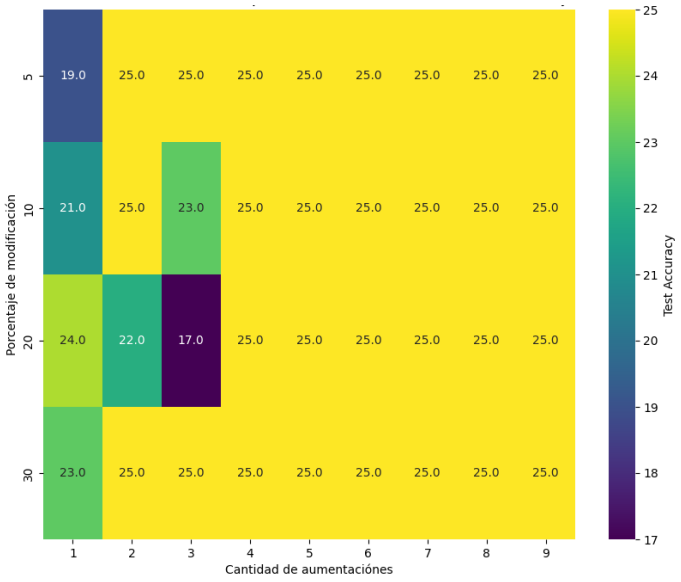

Fig. 36 Resultados Emoji, EDA, BERT

Por último, la Fig. 36 presenta los resultados de clasificación para el modelo BERT. En ella se muestra que el peor rendimiento de clasificación (17%) se obtiene al aumentar 3 veces el conjunto de datos modificando el 20% de las palabras que componen cada oración. Ahora bien, el mejor resultado de clasificación (25%) se da desde aumentar 4 veces el conjunto de datos sin importar el porcentaje de modificación de palabras por oración. Siendo el escenario que menos manipula el conjunto de datos y que obtiene el mejor rendimiento de clasificación aquel que aumenta 2 veces el conjunto de datos y modifica el 5% de las palabras por oración.

Para finalizar, se presenta la Tabla 23 que resume los resultados de clasificación usando la métrica *accuracy* por modelo de clasificación y los compara con la línea base.

Tabla 23 Resultados consolidados Emoji / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Emoji	Base	29%	22%	21%	23%	28%
	EDA	29%	19%	25%	25%	25%

5.5.5 Resultados para Encuesta Docente Afecto

Los resultados para el subconjunto de datos de Encuesta Docente correspondientes a las categorías de Afecto se presentan a continuación.

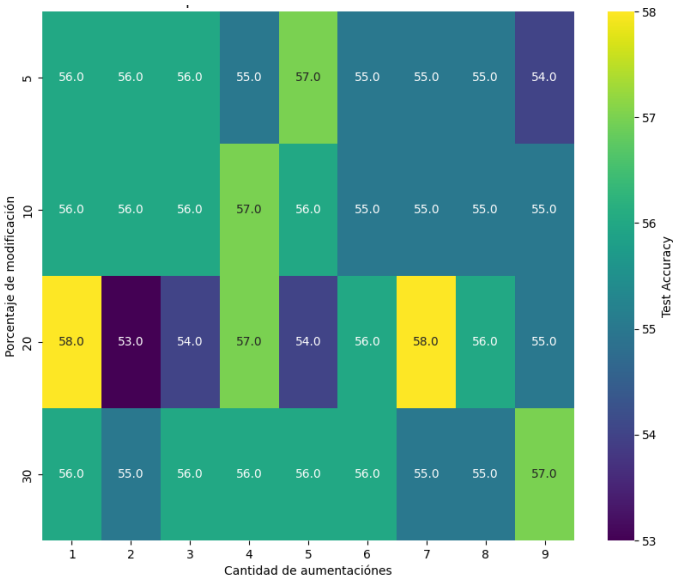

Fig. 37 Resultados Encuesta Docente Afecto, EDA, SVM

En el caso del modelo SVM (Fig. 37), el rendimiento de clasificación del conjunto de datos aumentado muestra que al aumentar 1 vez y modificar el 20% de las palabras dentro de una oración logra el mejor resultado con un 58%. Por otra parte, el rendimiento de clasificación más bajo se registra cuando el conjunto de datos es aumentado en 2 veces y se modifica el 20% de las palabras por oración.

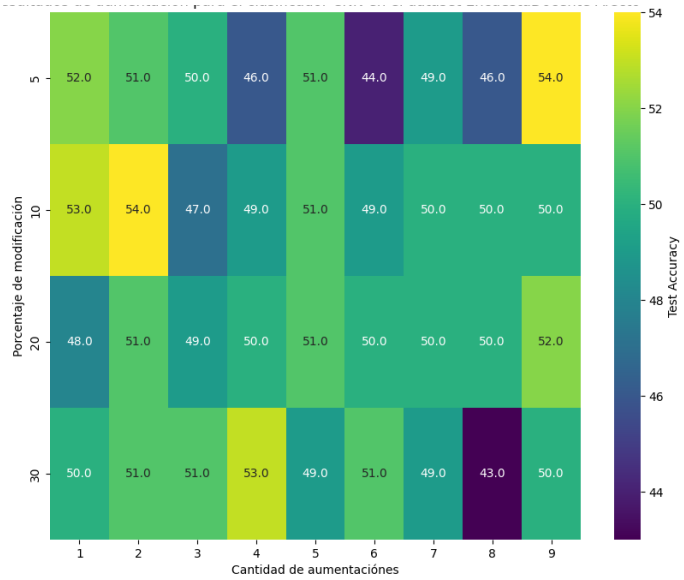

Fig. 38 Resultados Encuesta Docente Afecto, EDA, CNN

En el caso del modelo de clasificación CNN (Fig. 38), el mejor rendimiento (54%) se obtiene cuando el conjunto de datos es aumentado 2 veces y se modifica un 10% de las palabras que componen cada oración. En cambio, cuando se aumenta el conjunto de datos en 8 veces y se modifica el 30% de las palabras en una oración, se registra el rendimiento más bajo con solo un 43%.

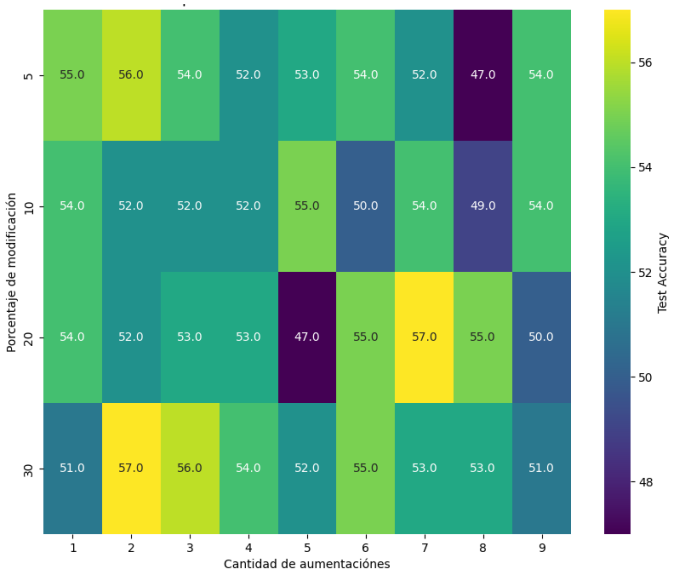

Fig. 39 Resultados Encuesta Docente Afecto, EDA, LSTM

Los resultados para el clasificador LSTM, representados en el mapa de calor de la Fig. 39, indica que el mejor rendimiento de clasificación (57%) se registra cuando el conjunto de datos es aumentado 2 veces y se modifica el 30% de las palabras por oración. En cambio, el

rendimiento de clasificación más bajo (47%) se presenta al aumentar en 5 veces el conjunto de datos y modificar el 20% de las palabras que componen cada oración.

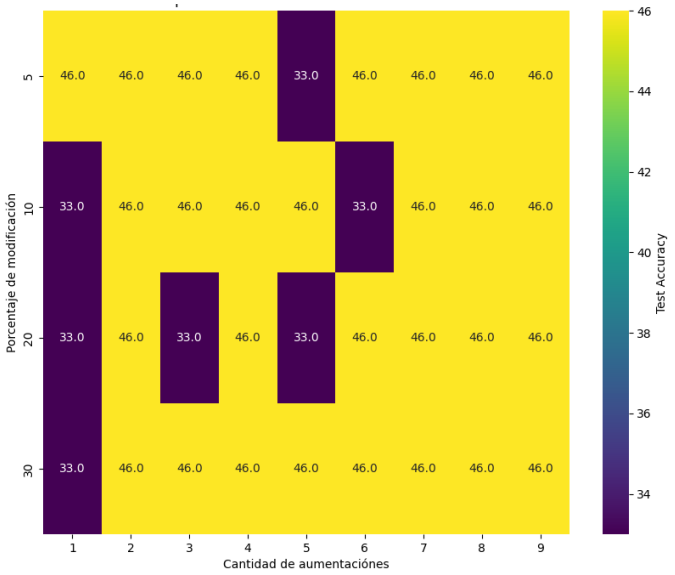

Fig. 40 Resultados Encuesta Docente Afecto, EDA, BiLSTM

Los resultados para el clasificador BiLSTM mostrados en la Fig. 40 indican que el peor rendimiento de clasificación es de 33% y se produce al aumentar 1, 3 5 y 6 veces el conjunto de datos con porcentajes de modificación entre 10% y 30% cuando se aumenta 1 vez, 20% cuando se aumenta 3 veces, 5% y 10% cuando se aumenta 5 veces y 10% cuando se aumenta 6 veces. El resto de los experimentos logra un rendimiento de 46% en la métrica accuracy.

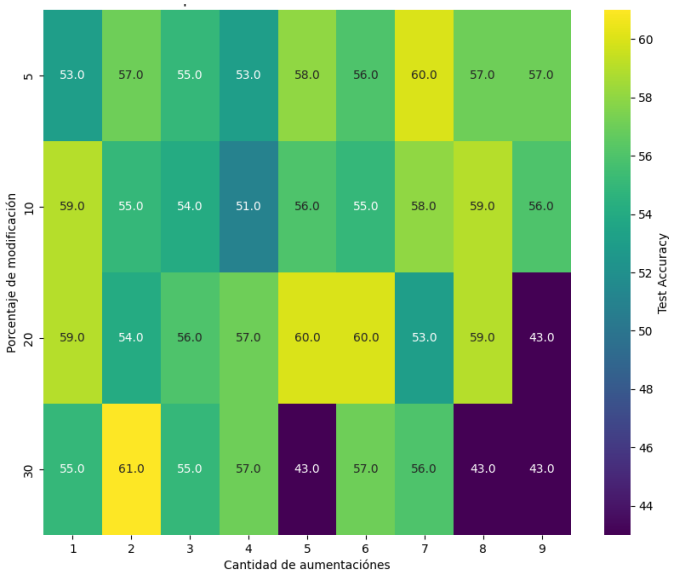

Fig. 41 Resultados Encuesta Docente Afecto, EDA, BERT

En el mapa de calor de la Fig. 41 se muestra que el mejor resultado de clasificación con BERT (61%) se produce cuando el conjunto de datos es aumentado 2 veces y se modifica el 30% de las palabras pertenecientes a cada oración. Mientras que el resultado de clasificación más bajo con el mismo modelo (43%) se produce al aumentar en 5 y 8 veces el conjunto de datos modificando el 30% de las palabras que componen una oración y cuando el conjunto de datos es aumentado 9 veces modificando en 20% y 30% de las palabras en una oración.

Para finalizar con este conjunto de datos y los resultados de aumentación con EDA, se presenta la Tabla 24 que permite visualizar los resultados con EDA y la línea base.

Tabla 24 Resultados consolidados Encuesta Docente Afecto / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	EDA	58%	54%	57%	46%	61%

5.5.6 Resultados para Encuesta Docente Agresividad

Los resultados para el conjunto de datos de Encuesta Docente enfocado en las categorías de Agresividad son presentados a continuación.

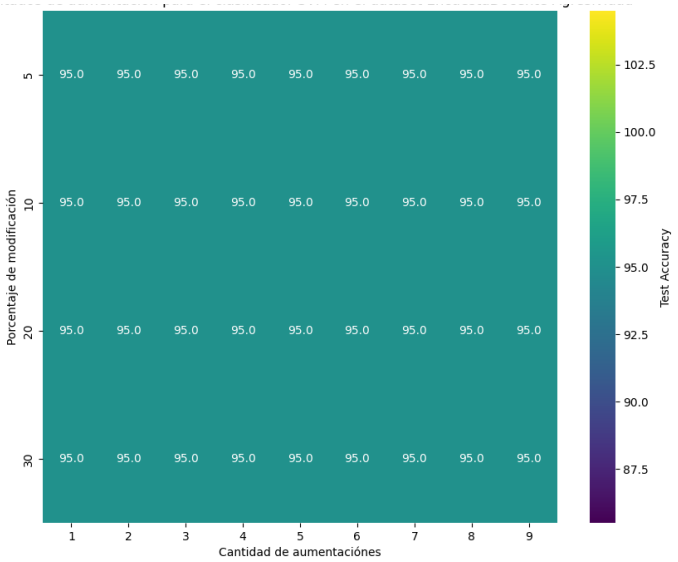

Fig. 42 Resultados Encuesta Docente Agresividad, EDA, SVM

La Fig. 42 representa los resultados del primer clasificador aplicado al conjunto de datos de Agresividad. Dichos resultados indican que la aumentación no tuvo impacto en la clasificación con SVM.

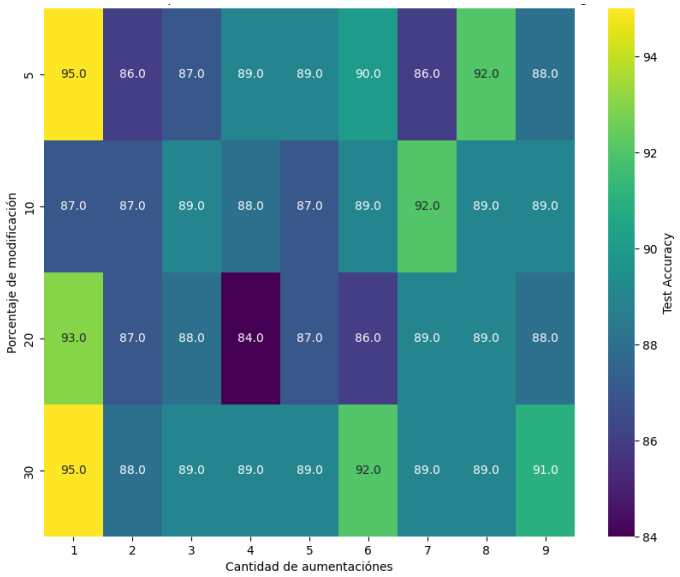

Fig. 43 Resultados Encuesta Docente Agresividad, EDA, CNN

Pasando a los resultados de clasificación con el modelo CNN, los resultados plasmados en la Fig. 43 indican que el mejor rendimiento (95%) al clasificar se logra al aumentar 1 vez el conjunto de datos y modificar el 5% de las palabras por oración. Por otra parte, el rendimiento más bajo (84%) de clasificación se produce cuando se aumenta 4 veces el conjunto de datos y se modifica el 20% de las palabras por oración.

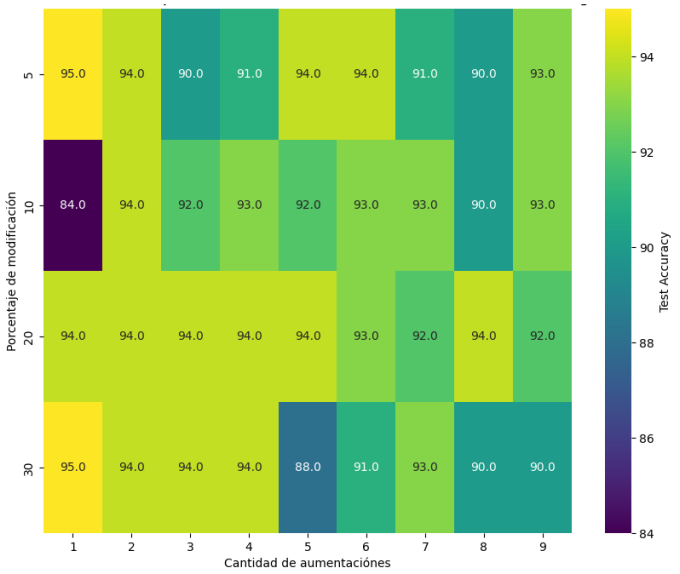

Fig. 44 Resultados Encuesta Docente Agresividad, EDA, LSTM

Los resultados de clasificación para el modelo LSTM mostrados en la Fig. 44 indican que el mejor rendimiento de clasificación (95%) se produce al aumentar 1 vez el conjunto de datos y modificar el 5% de las palabras en cada oración. Mientras que el rendimiento más bajo (84%) se registra cuando también se aumenta el conjunto de datos 1 vez, pero se modifica el 10% de las palabras por oración.

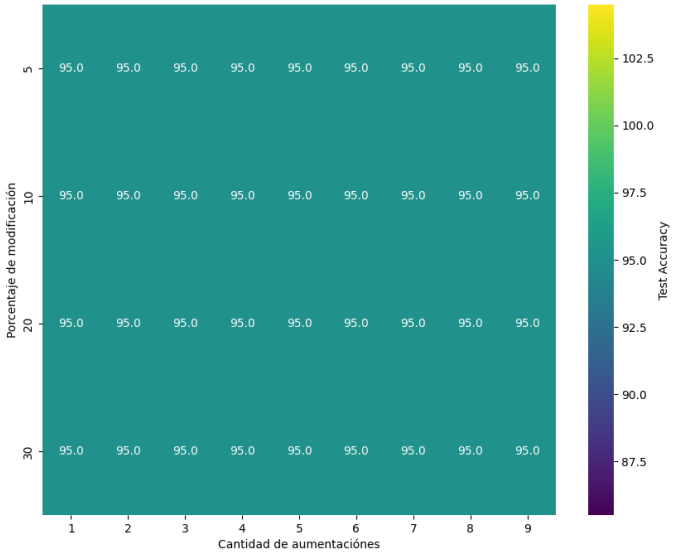

Fig. 45 Resultados Encuesta Docente Agresividad, EDA, BiLSTM

En la Fig. 45 se muestran los resultados de clasificación con el modelo BiLSTM sobre el conjunto de datos aumentado. Puede apreciarse que el rendimiento de clasificación es siempre 95%, no importando la cantidad de veces que sea aumentado el conjunto de datos ni el porcentaje de modificación de palabras por oración.

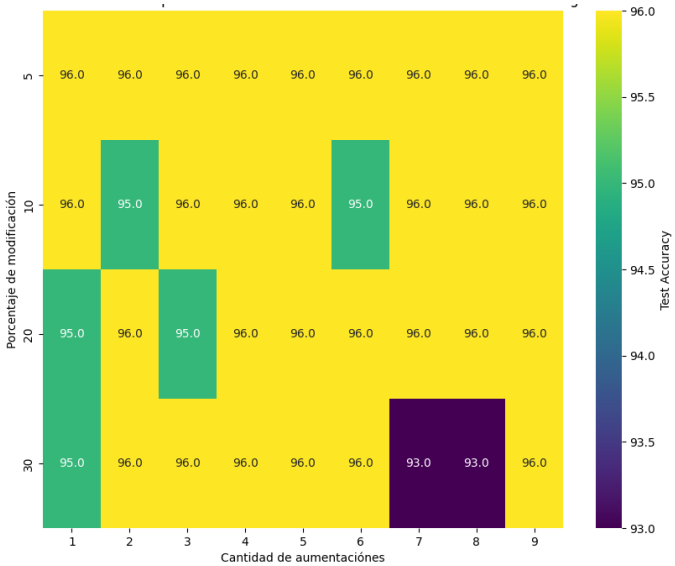

Fig. 46 Resultados Encuesta Docente Agresividad, EDA, BERT

En el caso del clasificador BERT (Fig. 46), gran parte de los experimentos de aumentación produjeron el mejor resultado (96%). Sin embargo, en términos de modificación del conjunto de datos el que obtiene el mejor rendimiento es 1 aumentación modificando el 5% de las palabras por oración. Mientras que, el rendimiento de clasificación más bajo (93%) se registró cuando el conjunto de datos es aumentado 7 y 8 veces, modificando el 30% de las palabras por oración en ambos casos.

Para finalizar con la revisión de resultados del conjunto de datos Encuesta Docente Agresividad se presenta la tabla Tabla 25 con el resumen de los resultados obtenidos.

Tabla 25 Resultados consolidados Encuesta Docente Agresividad / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	EDA	95%	95%	95%	95%	96%

5.5.7 Resultados para Encuesta Docente Polaridad

En las siguientes figuras se presentan los resultados de clasificación obtenidos al aumentar el conjunto de datos Encuesta Docente Polaridad con EDA.

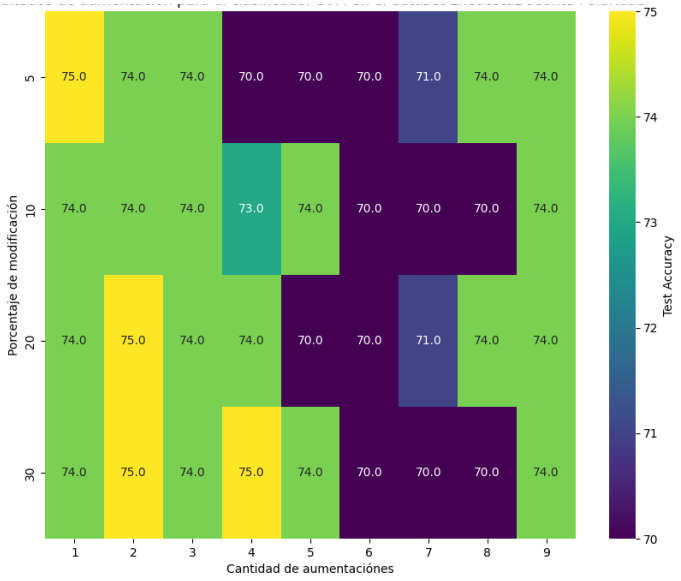

Fig. 47 Resultados Encuesta Docente Polaridad, EDA, SVM

La Fig. 47 muestra que el mejor rendimiento (75%) al clasificar con SVM se logra aumentando el conjunto de datos 5 veces y modificando el 5% de las palabras por oración. Por otro lado, el rendimiento más bajo (70%) se obtiene al aumentar el conjunto de datos entre 4 y 9 veces con distintos porcentajes de modificación de palabras. Cabe resaltar que dentro de los rendimientos más bajos se encuentra aumentar el conjunto de datos 6 veces, no importando el porcentaje de modificación de palabras por oración aplicado.

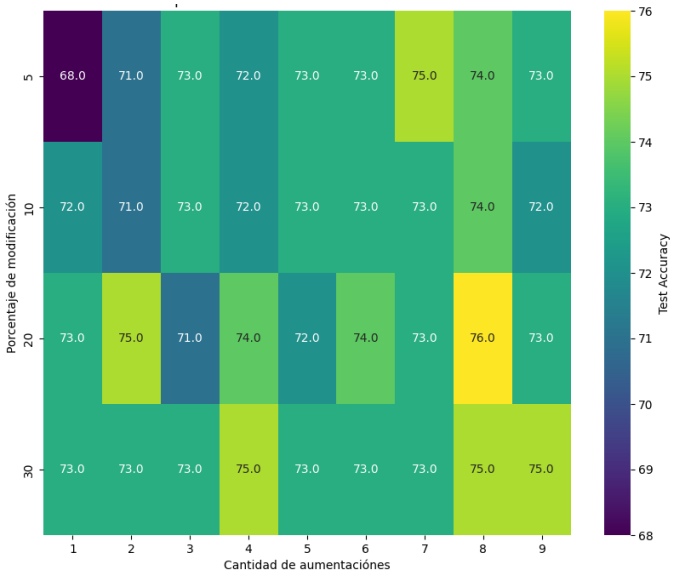

Fig. 48 Resultados Encuesta Docente Polaridad, EDA, CNN

El mejor rendimiento de clasificación (76%) con el modelo CNN se logra cuando el conjunto de datos 8 veces y se modifica el 20% de las palabras por oración, según lo representado en la Fig. 48. Mientras que el rendimiento más bajo (68%) es obtenido al aumentar 1 vez el conjunto de datos y modificar el 5% de las palabras que componen cada oración.

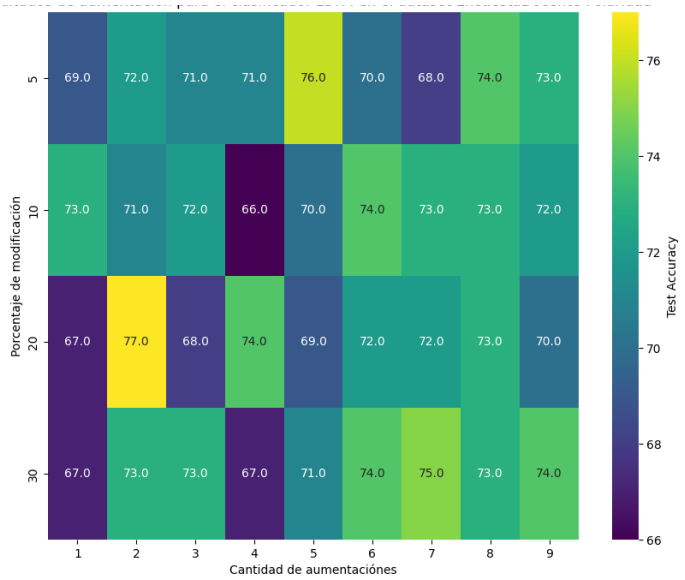

Fig. 49 Resultados Encuesta Docente Polaridad, EDA, LSTM

Continuando con la presentación de resultados de clasificación luego de aumentar con EDA, la Fig. 49 muestra el rendimiento del clasificador LSTM. El mejor resultado con este clasificador (77%) se obtiene cuando el conjunto de datos es aumentado 2 veces y se modifica el 20% de las palabras por oración. Mientras que, el rendimiento más bajo (66%) se obtiene cuando el conjunto de datos es aumentado 4 veces y se modifica el 10% de las palabras que componen cada una de las oraciones presentes en el conjunto de datos.

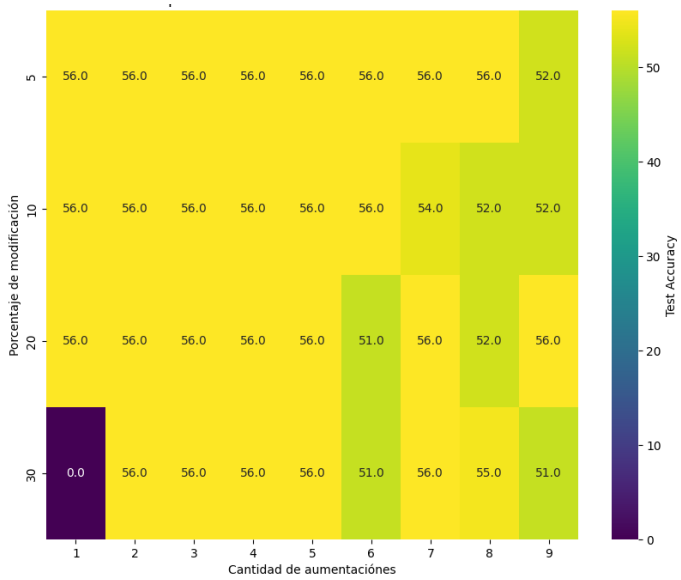

Fig. 50 Resultados Encuesta Docente Polaridad, EDA, BiLSTM

En cuanto al clasificador BiLSTM (Fig. 50), los resultados de clasificación son bastante homogéneos durante los experimentos, siendo el mejor de ellos (56%) registrado cuando el

conjunto de datos se aumenta 1 vez y se modifica el 5% de las palabras que componen cada oración. Por otra parte, el rendimiento más bajo (0%) se produce al aumentar el conjunto de datos 1 vez y modificar el 30% de las palabras por oración.

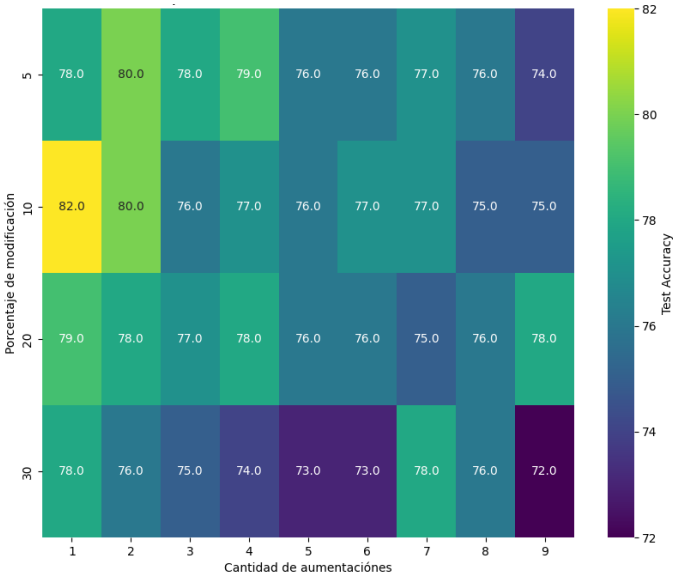

Fig. 51 Resultados Encuesta Docente Polaridad, EDA, BERT

El último de los clasificadores empleados sobre el conjunto de datos es BERT (Fig. 51), el mejor rendimiento con este clasificador (82%) se logra cuando el conjunto de datos es aumentado 1 vez y el porcentaje de modificación de palabras por oración es de 10%. Mientras que el menor rendimiento (72%) es obtenido cuando el conjunto de datos es aumentado 9 veces y se modifica el 30% de las palabras que componen cada oración.

Para finalizar con la revisión de los resultados con el conjunto de datos Encuesta Docente Polaridad, se presenta la Tabla 26 que muestra un resumen de los resultados comparados con la línea base.

Tabla 26 Resultados consolidados Encuesta Docente Polaridad / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Polaridad	Base	71%	75%	72%	56%	80%
	EDA	75%	76%	77%	56%	82%

5.5.8 Resultados para Encuesta Docente Seriedad

El último de los subconjuntos de datos pertenecientes a Encuesta Docente es el que evalúa la seriedad de los comentarios expresados por los alumnos, las siguientes figuras muestran cuales fueron los resultados de clasificación sobre este conjunto de datos al ser aumentado con EDA.

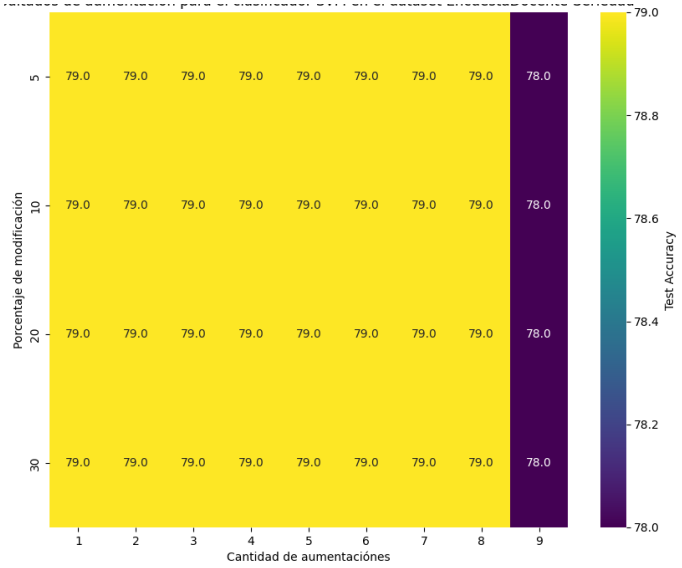

Fig. 52 Resultados Encuesta Docente Seriedad, EDA, SVM

Los resultados de clasificación con el modelo SVM (Fig. 52) muestran que no existe una gran variación en el rendimiento, por lo que el mejor resultado de clasificación (79%) se presenta en la mayor parte de los experimentos, tal como se puede apreciar en la Fig. 52. En cuanto al rendimiento más bajo (78%) de clasificación, este se presenta cuando el conjunto de datos es aumentado 9 veces no importando que porcentaje de modificación de palabras por oración se haya aplicado.

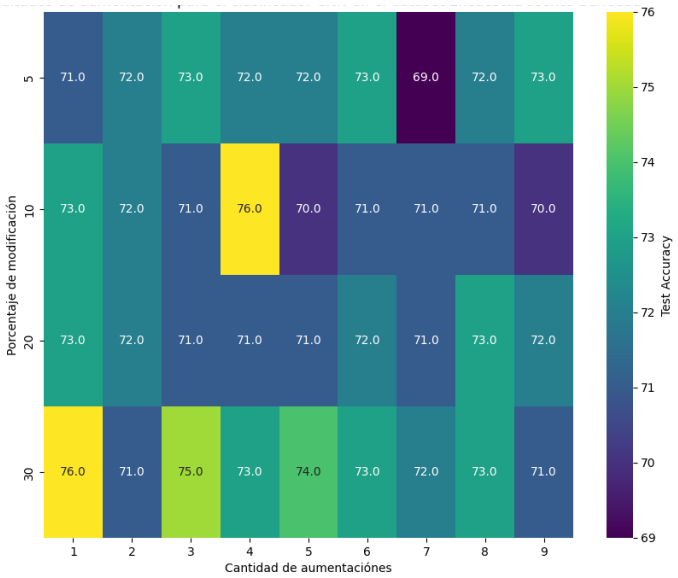

Fig. 53 Resultados Encuesta Docente Seriedad, EDA, CNN

La Fig. 53 muestra que le mejor resultado de clasificación (76%) con el modelo CNN se obtiene bajo dos escenarios. El primero de ellos es cuando el conjunto de datos es aumentado 1 vez modificando el 30% de las palabras por oración. El segundo escenario, se produce cuando el conjunto de datos es aumentado 4 veces y el porcentaje de modificación por palabras aplicado es de 10%. Por otra parte, el rendimiento de clasificación más bajo (69%) se da cuando el conjunto de datos es aumentado 7 veces y el porcentaje de modificación de palabras por oración aplicado es de 5%.

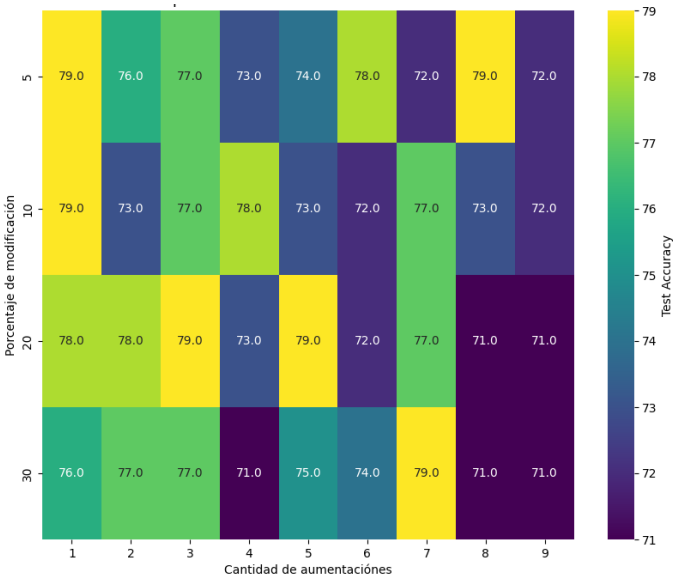

Fig. 54 Resultados Encuesta Docente Seriedad, EDA, LSTM

Los resultados con el modelo LSTM representados en la Fig. 54, muestran que el mejor rendimiento de clasificación (79%) se da en varios escenarios. Siendo el que realiza una menor manipulación del conjunto de datos aquel que aumenta el conjunto de datos en 1 vez y modifica el 5% de las palabras que componen una oración. Mientras que, el menor rendimiento (71%) se da cuando se aumenta el conjunto de datos 4 veces y se aplica un 30% de modificación de palabras por oración, los otros escenarios en los que se da el menor rendimiento de clasificación, es cuando el conjunto de datos es aumentado entre 8 y 9 veces con porcentajes de modificación de palabras por oración entre 20% y 30%.

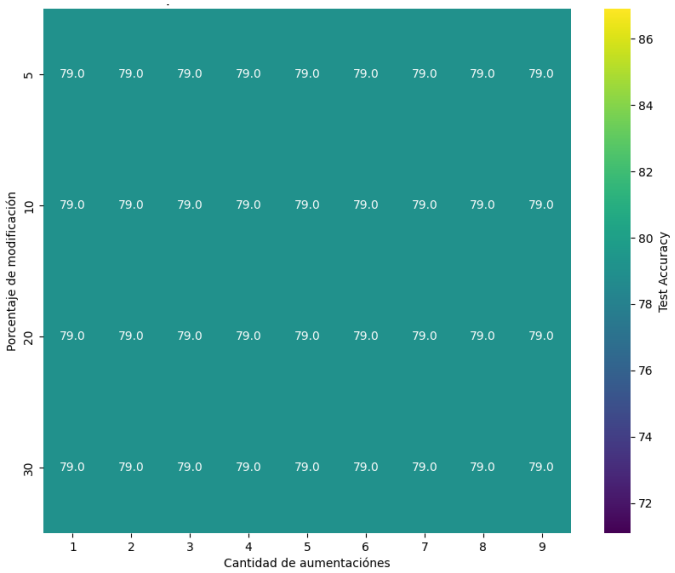

Fig. 55 Resultados Encuesta Docente Seriedad, EDA, BiLSTM

Los resultados de clasificación con el modelo BiLSTM, representados en la Fig. 55, indican que no hubo variaciones en la clasificación a lo largo de las aumentaciones realizadas, llegando todos los resultados al 79% de accuracy.

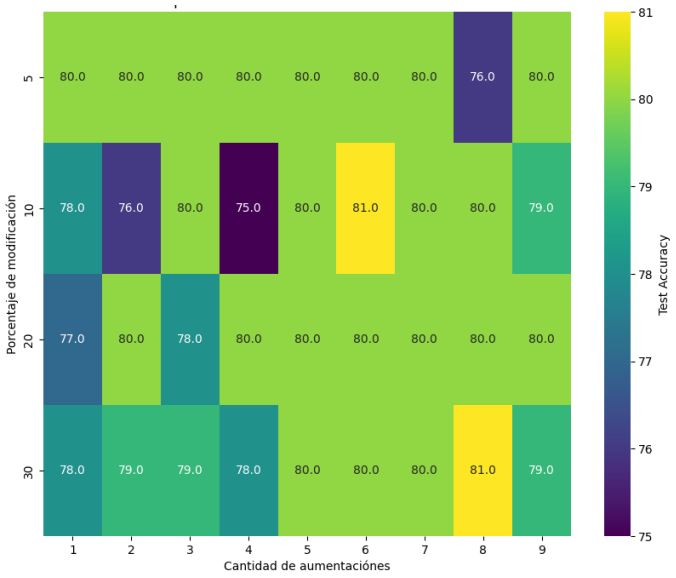

Fig. 56 Resultados Encuesta Docente Seriedad, EDA, BERT

El ultimo clasificador utilizado con este conjunto de datos es BERT (Fig. 56), donde puede observarse que el mejor rendimiento de clasificación (81%) se logra cuando el conjunto de datos es aumentado 6 veces y se aplica un porcentaje de modificación de palabras por oración de 10%. En cambio, cuando el conjunto de datos es aumentado 4 veces y se aplica un 10% de modificación de palabras por oración se obtiene el menor rendimiento (75%).

Para finalizar con la revisión de resultados en el conjunto de datos Encuesta Docente Seriedad aumentado, se presenta la Tabla 27, que muestra un resumen de los resultados por clasificador y los compara con la línea base.

Tabla 27 Resultados consolidados Encuesta Docente Seriedad / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	EDA	79%	76%	79%	79%	81%

5.5.9 Resultados para Titulares de Diarios

Los resultados para el conjunto de datos de Titulares de Diarios son presentados en las siguientes figuras.

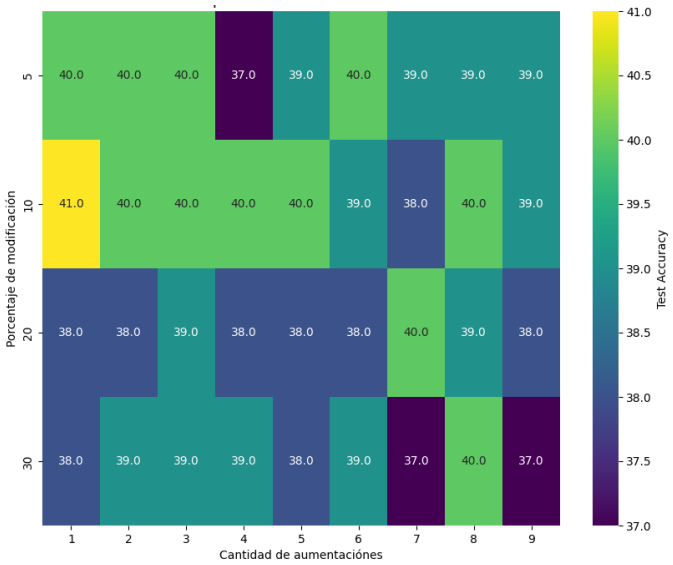

Fig. 57 Resultados Titulares Diarios, EDA, SVM

El mapa de calor de la Fig. 57 muestra los resultados de clasificación obtenidos con SVM. En este mapa se puede apreciar que el mejor rendimiento (41%) se obtiene cuando el conjunto de datos es aumentado 1 vez y se modifica el 10% de las palabras pertenecientes a cada oración. Por otra parte, el rendimiento más bajo de clasificación (37%) se da cuando el conjunto de datos es aumentado 4, 7 y 9 veces, con porcentajes de modificación de palabras por oración de 5% y 30%.

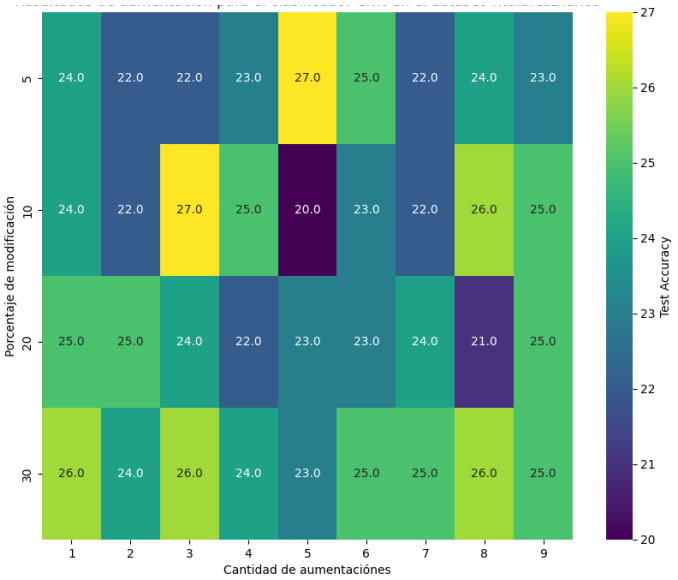

Fig. 58 Resultados Titulares Diarios, EDA, CNN

Los resultados de clasificación para el modelo CNN (Fig. 58) muestran que existen el mejor rendimiento de clasificación (27%) es obtenido cuando el conjunto de datos es aumentado 3 veces con un porcentaje de modificación de palabras por oración de 10%. Mientras que el rendimiento más bajo (20%) se presenta al aumentar 5 veces el tamaño del conjunto de datos y modificar el 10% de las palabras en una oración.

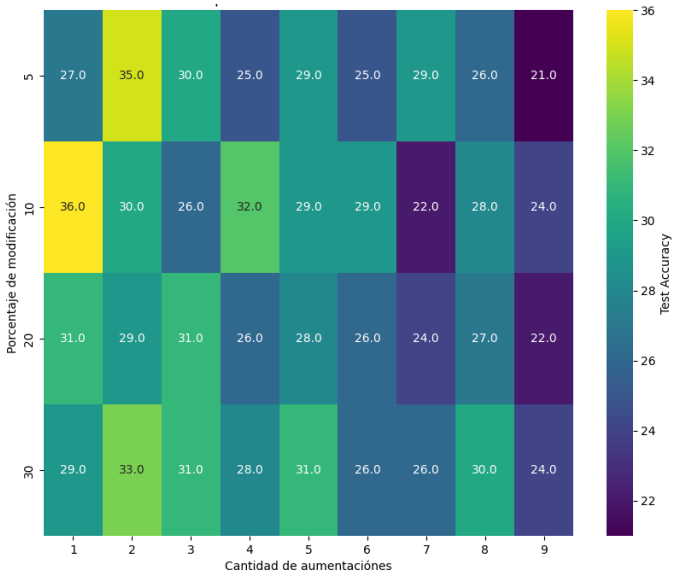

Fig. 59 Resultados Titulares Diarios, EDA, LSTM

Los resultados de clasificación con el modelo LSTM sobre el conjunto de datos aumentado, representados en la Fig. 59 muestran que el mejor rendimiento (36%) se obtiene al aumentar 1 vez el conjunto de datos y modificar el 10% de las palabras por oración. Mientras que, al

aumentar 9 veces el tamaño del conjunto de datos modificando el 5% de las palabras por oración obtiene el menor rendimiento de clasificación con un 21%.

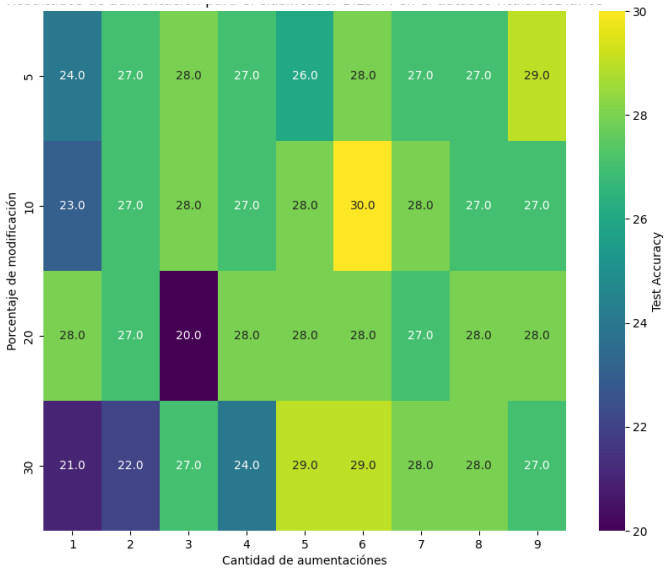

Fig. 60 Resultados Titulares Diarios, EDA, BiLSTM

El rendimiento de clasificación con el modelo BiLSTM para el conjunto de datos aumentado que se presenta en la Fig. 60, reflejan que al aumentar 6 veces el conjunto de datos y modificando el 10% de palabras por oración obtiene el mejor rendimiento (30%). El mapa de calor también refleja que el rendimiento más bajo (20%) se registra al modificar el 20% de las palabras por oración y aumentar 3 veces el tamaño del conjunto de datos.

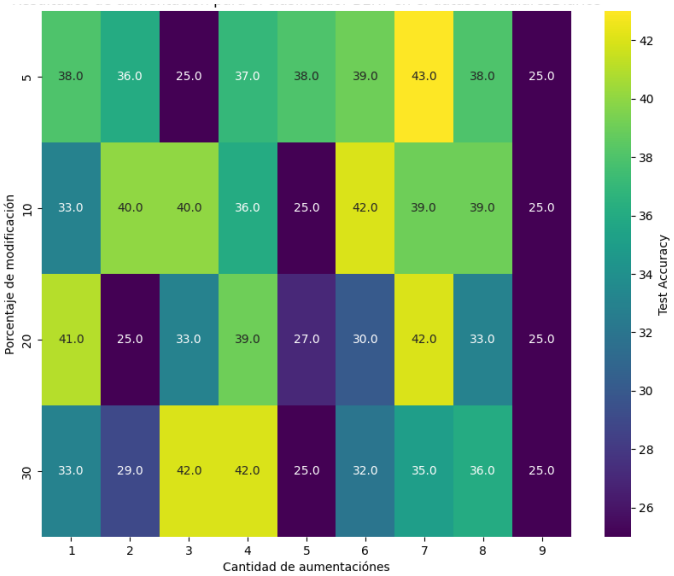

Fig. 61 Resultados Titulares Diarios, EDA, BERT

Los resultados obtenidos al clasificar con BERT el conjunto de datos aumentado entrega que el mejor rendimiento de clasificación (43%) se obtiene cuando el conjunto de datos es aumentado 7 veces y se modifica un 5% de las palabras presentes en cada oración. Mientras que, el menor rendimiento de clasificación (25%) se obtiene con diversas combinaciones de aumentación/porcentaje de modificación. Lo anteriormente descrito se muestra en la Fig. 61, un punto a destacar en la figura es que cuando se aumenta 9 veces el conjunto de datos se llega al menor rendimiento de clasificación, no importando el porcentaje de modificación de palabras por oración.

Para finalizar con el análisis de resultados para el conjunto de datos de Titulares de Diarios aumentado, se presenta la Tabla 28 que resume los resultados para todos los clasificadores.

Tabla 28 Consolidado resultados Titulares Diarios / EDA

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
Titulares Diarios	Base	39%	26%	31%	27%	44%
	EDA	41%	27%	36%	30%	43%

5.5.10 Resultados para Violencia de Género

Las siguientes figuras representan los resultados de clasificación con la métrica accuracy luego de realizar la aumentación del conjunto de datos de Violencia de Género.

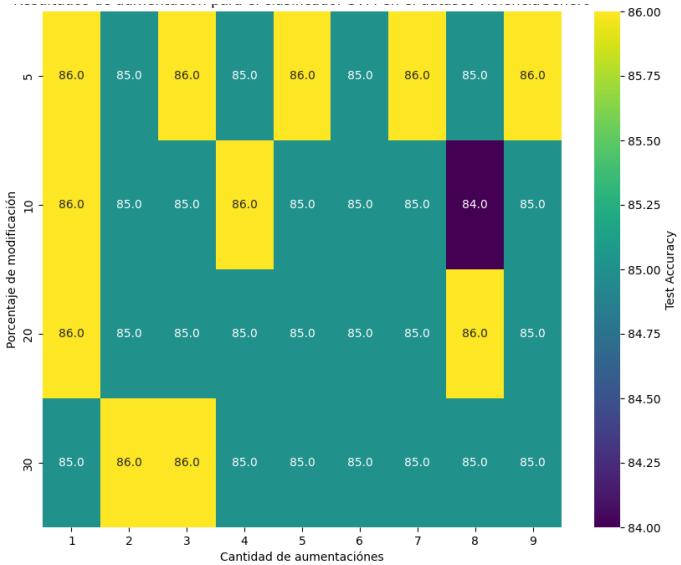

Fig. 62 Resultados Violencia Género, EDA, SVM

De acuerdo con el mapa de calor de la Fig. 62, el menor resultado de clasificación se da cuando se el conjunto de datos es aumentado 8 veces con un 10% de modificación de cada oración. Mientras que el mejor resultado se logra con distintas combinaciones de cantidad de aumentaciones y porcentaje de modificación. En términos de modificación al conjunto de datos, el mejor resultada se da cuando se aumenta en 1 vez el conjunto de datos y se modifica un 5% de las palabras pertenecientes a cada oración.

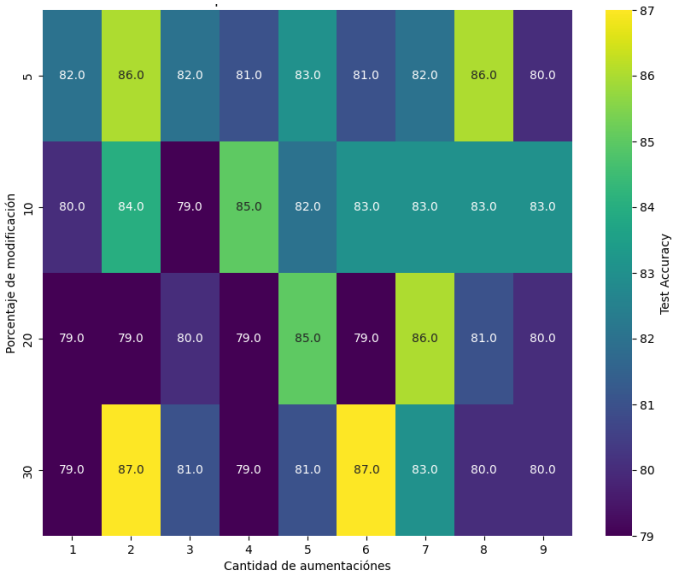

Fig. 63 Resultados Violencia Género, EDA, CNN

La Fig. 63 muestra los resultados de clasificación con el modelo CNN luego de aumentar el conjunto de datos. El mejor resultado de clasificación (87%) corresponde a un porcentaje de modificación de 30% y realizando 2 o 6 aumentaciones del conjunto de datos. Por otra parte, el resultado de clasificación más bajo (79%) se repite cuando se realizan entre 1 y 6 aumentaciones con distintos porcentajes de modificación de palabras pertenecientes a una oración en el conjunto de datos.

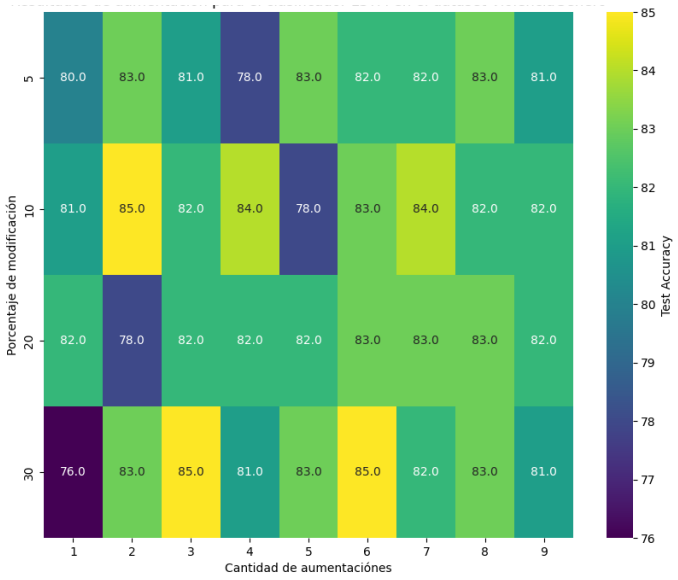

Fig. 64 Resultados Violencia Género, EDA, LSTM

La Fig. 64 muestra los resultados de clasificación con el modelo LSTM luego de aplicada la aumentación. El resultado más bajo de clasificación (76%) se obtiene cuando se aumenta en 1 vez el conjunto de datos y se modifica el 30% de las palabras de cada oración. Mientras que el mejor resultado (85%), con la menor cantidad de aumentaciones y modificaciones, es obtenido al aumentar 2 veces con una modificación del 10% de las palabras por oración.

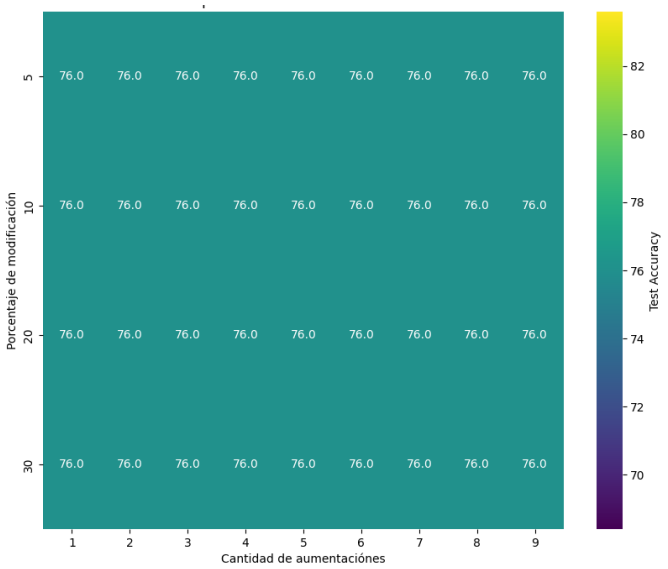

Fig. 65 Resultados Violencia Género, EDA, BiLSTM

En la Fig. 65 se presentan los resultados con el modelo de clasificación BiLSTM, dónde los distintos experimentos con EDA no afectan los resultados de clasificación con el modelo.

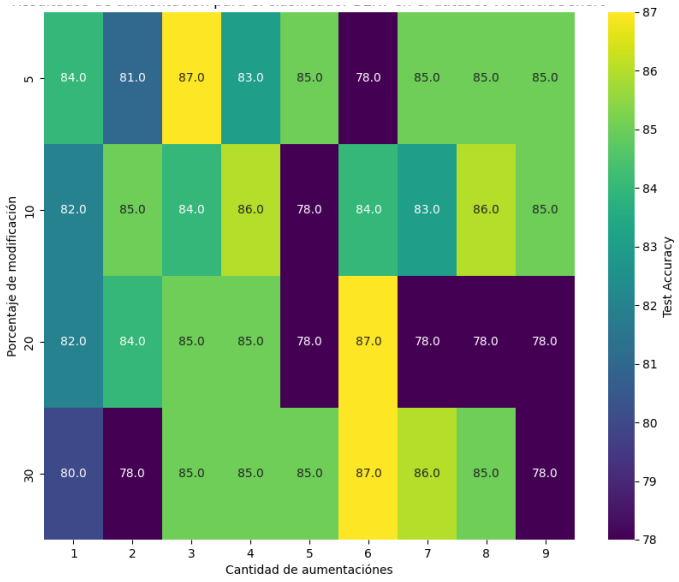

Fig. 66 Resultados Violencia Género, EDA, BERT

La Fig. 66 presenta los resultados con el clasificador BERT (BETO), el cual, luego de las aumentaciones realizadas obtiene como mejor resultado 87% al aumentar el conjunto de datos 3 veces modificando el 5% de las palabras dentro de una oración. Por otra parte, se observa que el rendimiento de clasificación más bajo es de 78% en distintos experimentos con diversos porcentajes de modificación y cantidad de aumentaciones.

Para finalizar, se presenta la Tabla 29 con el resumen de los mejores resultados para el conjunto de datos Violencia de Género.

Tabla 29 Consolidado resultados Violencia Género / EDA

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Violencia Género	Base	86%	76%	83%	76%	83%
	EDA	86%	87%	85%	76%	87%

5.5.11 Resultados consolidados EDA

Para finalizar con los resultados con la técnica de aumentación por transformación de oraciones, se presenta la Tabla 30 que consolida los resultados con EDA para cada conjunto de datos. En la tabla se encuentran resaltados en rojo los rendimientos de clasificación que están por debajo de la línea base y en azul el mejor resultado para el conjunto de datos en términos de puntos porcentuales por sobre la línea base.

Tabla 30 Resultados consolidados EDA

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
18 Octubre	Base	56%	56%	56%	56%	63%
	EDA	**54%**	58%	56%	56%	**67%**
Agresividad	Base	90%	82%	84%	58%	90%
	EDA	91%	83%	85%	**67%**	**89%**
Emoji	Base	29%	22%	21%	23%	28%
	EDA	29%	**19%**	**25%**	25%	25%
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	EDA	58%	54%	**57%**	**46%**	61%
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	EDA	95%	95%	95%	95%	96%
Encuesta Docente Polaridad	Base	71%	75%	72%	56%	80%
	EDA	75%	76%	**77%**	56%	82%
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	EDA	79%	76%	79%	79%	**81%**
Titulares Diarios	Base	39%	26%	31%	27%	44%
	EDA	41%	27%	**36%**	30%	**43%**
Violencia Género	Base	86%	76%	83%	76%	83%
	EDA	86%	**87%**	85%	76%	87%

5.5.12 Resultados consolidados EDA clases balanceadas

Los resultados de clasificación para el segundo experimento con EDA consistentes en el balanceo de las clases en los conjuntos de datos son presentados en la Tabla 31. Los conjuntos de datos 18Octubre y EncuestaDocente-Afecto no fueron balanceados debido a las pocas muestras de las clases minoritarias. Igualmente, el conjunto de datos de Agresividad tampoco fue balanceado ya que la cantidad de muestras por cada clase es similar.

Tabla 31 Resultados consolidados EDA Balanceado

Conjunto de datos	Técnica DA	Clasificador				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Emoji	Base	29%	22%	21%	23%	28%
	EDA-B	**24%**	**8%**	**5%**	**3%**	**5%**
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	EDA-B	93%	91%	90%	**5%**	**4%**
Encuesta Docente Polaridad	Base	71%	75%	72%	56%	80%
	EDA-B	**75%**	**69%**	**71%**	56%	**78%**
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	EDA-B	**76%**	**75%**	**75%**	**21%**	**20%**
Titulares Diarios	Base	39%	26%	31%	27%	44%
	EDA-B	38%	**33%**	25%	**8%**	25%
Violencia Género	Base	86%	76%	83%	76%	83%
	EDA-B	86%	**85%**	**73%**	76%	**84%**

En la tabla anterior, se registra que al balancear los conjuntos de datos con EDA produce resultados negativos en la mayor parte de ellos, siendo el conjunto de datos Emoji el más afectado con disminuciones de rendimiento que bordean el 20%. Por otra parte, solo un pequeño universo de experimentos entrega buenos resultados luego de balancear los conjuntos de datos, siendo la clasificación con CNN del conjunto de datos Violencia Género el que obtiene el mejor resultado con una mejora de 9%.

5.6 Resultados de clasificación con técnicas generativas

Los experimentos de aumentación y clasificación sobre los conjuntos de datos aumentados mediante redes generativas adversariales (GAN en inglés), más específicamente mediante SentiGAN[58] fueron llevados a cabo tal como se indicó en el punto 5.2.3. Los resultados de clasificación post aumentación son mostrados a través de tablas según el punto 5.2.5.

5.6.1 Ejemplos de aumentación con GAN

Para realizar la aumentación de datos con SentiGAN [58], fueron aplicados los parámetros descritos en el punto 5.2.3 de acuerdo con las características de cada conjunto de datos.

Como resultado, la Tabla 32 muestra una oración sin aumentar y una con la aumentación mediante SentiGAN[58] correspondiente al conjunto de datos Agresividad.

Tabla 32 Ejemplo de aumentación de texto con SentiGAN

Oración original	agarre el guante del novio tonta hueona A128
Oración creada con SentiGAN	La weona es tonta pero no le <UNK> <UNK> con la <UNK> <UNK> es wea de <UNK> al fin y al y

5.6.2 Resultados para dieciocho de Octubre

Los resultados de clasificación luego de la aumentación con SentiGAN[58] representados en la Tabla 33, muestran que solamente el clasificador SVM ve incrementado el rendimiento luego de la aumentación (+2%) con respecto a la línea base. Por otra parte, el clasificador BERT decae considerablemente en su rendimiento al bajar (-60%).

Tabla 33 Resultados consolidados 18 Octubre / SentiGAN

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
18Octubre	Base	56%	56%	56%	56%	63%
	SentiGAN	**58%**	44%	42%	17%	**3%**

5.6.3 Resultados para Agresividad

Los resultados de clasificación para el conjunto de datos de Agresividad presentados en la Tabla 34, indican que hubo mejoras significativas con respecto a la línea base para todos los clasificadores, alcanzando el máximo rendimiento de clasificación con el modelo CNN (100%).

Tabla 34 Resultados consolidados Agresividad / SentiGAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Agresividad	Base	90%	82%	84%	58%	90%
	SentiGAN	99%	100%	99%	64%	94%

5.6.4 Resultados para Encuesta Docente Afecto

Los resultados correspondientes al conjunto Encuesta Docente Afecto aumentado con GAN mostrados en la Tabla 35, indican que todos los clasificadores vieron negativamente afectado su rendimiento de clasificación. Siendo el más perjudicado BERT con una baja de más de 50% con respecto a la línea base.

Tabla 35 Resultados consolidados Encuesta Docente Afecto / SentiGAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	SentiGAN	43%	43%	25%	13%	4%

5.6.5 Resultados para Encuesta Docente Agresividad

Los resultados de clasificación luego de aplicar aumentación mediante SentiGAN presentados en la Tabla 36, muestran que en la mayor parte de los clasificadores se mantiene el rendimiento de clasificación, solamente en los clasificadores SVM y CNN se marca una disminución en el porcentaje de clasificación, siendo SVM el que más disminuye con un 5% con respecto de la línea base.

Tabla 36 Resultados consolidados Encuesta Docente Agresividad / GAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	SentiGAN	90%	92%	95%	95%	96%

5.6.6 Resultados para Encuesta Docente Polaridad

Los resultados de la Tabla 37 correspondientes a la clasificación del conjunto de datos Encuesta Docente Polaridad, muestran que el único clasificador que incrementó su rendimiento es SVM (+2). Mientras que, el clasificador que vio más afectado su rendimiento es BiLSTM que bajó en 25% su rendimiento.

Tabla 37 Resultados consolidados Encuesta Docente Polaridad / GAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
Encuesta Docente Polaridad	Base	71%	76%	72%	56%	80%
	SentiGAN	73%	71%	66%	31%	72%

5.6.7 Resultados para Encuesta Docente Seriedad

En cuanto a los resultados de clasificación para el conjunto de datos Encuesta Docente Seriedad luego de la aumentación con SentiGAN mostrados en la

Tabla 38. Puede apreciarse que el único clasificador que aumentó su rendimiento con respecto a la línea base es BERT (+3%). Mientras que, el resto de los clasificadores mantiene el rendimiento con respecto a la línea base.

Tabla 38 Resultados consolidados Encuesta Docente Seriedad / GAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
	Base	79%	76%	79%	79%	77%

Encuesta Docente Seriedad	SentiGAN	79%	79%	79%	79%	80%

5.6.8 Resultados para Titulares Diarios

Los resultados de clasificación luego de realizar la aumentación SentiGAN representados en la Tabla 39, indican que solamente el clasificador CNN mantuvo el rendimiento de clasificación con respecto de la línea base. Mientras que, los clasificadores restantes bajan su rendimiento, siendo SVM y BiLSTM los que disminuyen en un mayor porcentaje (-5%) su rendimiento.

Tabla 39 Resultados consolidados Titulares Diarios / SentiGAN

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Titulares Diarios	Base	39%	26%	31%	27%	44%
	SentiGAN	34%	26%	29%	22%	42%

5.6.9 Resultados para Violencia Género

La clasificación del conjunto de datos Violencia de Género luego de aumentar con SentiGAN representados en la Tabla 40, indican que solamente el clasificador CNN aumenta su rendimiento (+9%). Mientras que, el resto de los clasificadores disminuye sus resultados de clasificación, siendo BiLSTM el que más disminuye su rendimiento (-26%).

Tabla 40 Resultados consolidados Violencia Género / GAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
Violencia Género	Base	86%	76%	83%	76%	83%
	SentiGAN	85%	85%	81%	50%	78%

5.6.10 Resultados consolidados SentiGAN

Para finalizar con la revisión de resultados de clasificación de conjuntos de datos aumentados con modelos generativos, se presenta la tabla Tabla 41 que consolida los resultados de clasificación con SentiGAN para cada conjunto de datos y ofrece una comparación de los resultados con la línea base. Los resultados que sobrepasan la línea base son destacados en azul, mientras que los resultados por debajo la línea base son destacados en rojo.

Tabla 41 Resultados consolidados SentiGAN

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
18Octubre	Base	56%	56%	56%	56%	63%
	SentiGAN	**58%**	**44%**	**42%**	**17%**	**3%**
Agresividad	Base	90%	82%	84%	58%	90%
	SentiGAN	99%	**100%**	99%	64%	94%
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	SentiGAN	**43%**	**43%**	**25%**	**13%**	**4%**
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	SentiGAN	**90%**	**92%**	95%	95%	96%
Encuesta Docente Polaridad	Base	71%	76%	72%	56%	80%
	SentiGAN	**73%**	**71%**	**66%**	**31%**	**72%**

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	SentiGAN	79%	**79%**	79%	79%	**80%**
Titulares Diarios	Base	39%	26%	31%	27%	44%
	SentiGAN	**34%**	26%	**29%**	**22%**	**42%**
Violencia Género	Base	86%	76%	83%	76%	83%
	SentiGAN	**85%**	**85%**	**81%**	**50%**	**78%**

5.7 Resultados de clasificación con técnicas de parafraseo

Los experimentos de aumentación y posterior clasificación sobre los conjuntos de datos aumentados mediante "back translation" (uno de los métodos de transformación por parafraseo, fueron llevados a cabo tal como se indicó en el punto 5.2.4. Los resultados de clasificación post aumentación son mostrados a través de mapas de calor por cada uno de los conjuntos de datos, para posteriormente representar los mejores resultados a través de la plantilla propuesta en Tabla 14.

5.7.1 Ejemplos de aumentación con back translation

La Tabla 42 muestra un ejemplo de aumentación mediante *back-translation* correspondiente al conjunto de datos 18 Octubre. En ella, se muestra la variación en la oración producto de la traducción de español a inglés y la posterior traducción del inglés al español.

Tabla 42 Ejemplo de aumentación con back-translation

Base (español)	Traducción al inglés	Traducción al español
Vuelve la Burra al trigo... #nomasencapuchados	The Donkey returns to wheat... #nomasencapuchados	El Burro vuelve al trigo... #nomasencapuchados

5.7.2 Resultados para dieciocho Octubre

Los resultados de clasificación para el conjunto de datos 18Octubre por cada modelo de clasificación son mostrados en las siguientes figuras.

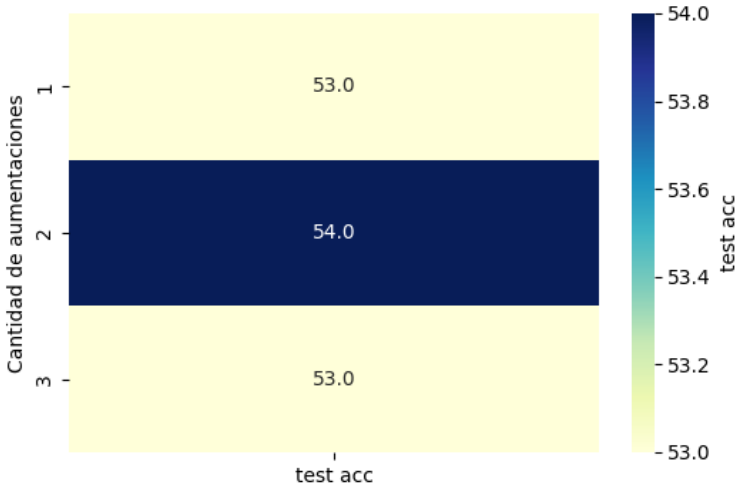

Fig. 67 Resultados 18Octubre, BT, SVM

El primero de los modelos de clasificación revisados es SVM, en la Fig. 67 puede apreciarse que el mejor rendimiento de clasificación se presenta al aumentar el conjunto de datos en 2 veces, llegando al 54% en la métrica *accuracy*.

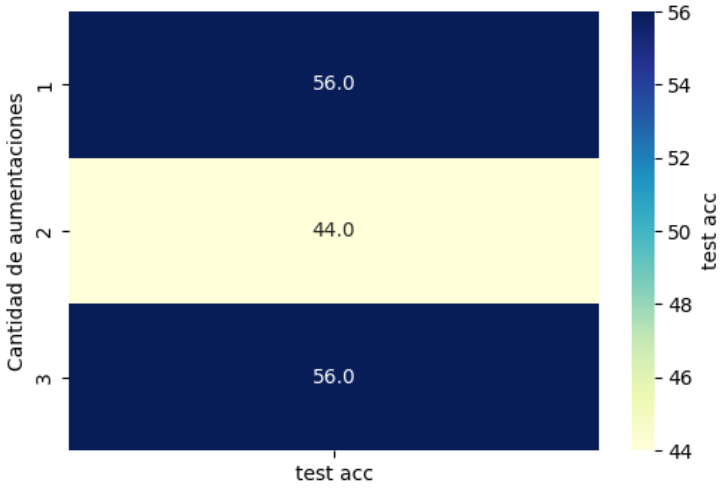

Fig. 68 Resultados 18Octubre, BT, CNN

En el caso del clasificador CNN, la Fig. 68 nos indica que el rendimiento de clasificación es igual para 1 y 3 aumentaciones (56%). Sin embargo, cuando el conjunto de datos es

aumentado 2 veces, el rendimiento de clasificación alcanza su punto más bajo con solamente un 44% en la métrica *accuracy*.

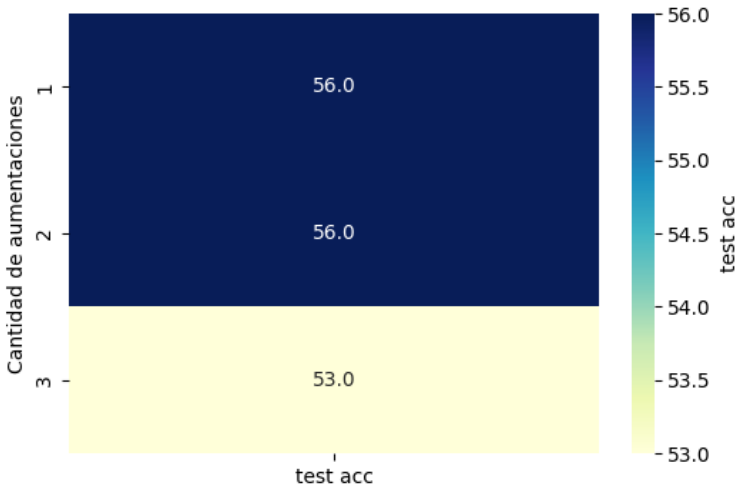

Fig. 69 Resultados 18Octubre, BT, LSTM

Mientras que para el clasificador LSTM, los resultados en la Fig. 69 indican que los mejores rendimientos de clasificación (56%) se obtienen al aumentar el conjunto de datos hasta 2 veces. Si el conjunto es aumentado una tercera vez, el rendimiento de clasificación disminuye a 53%.

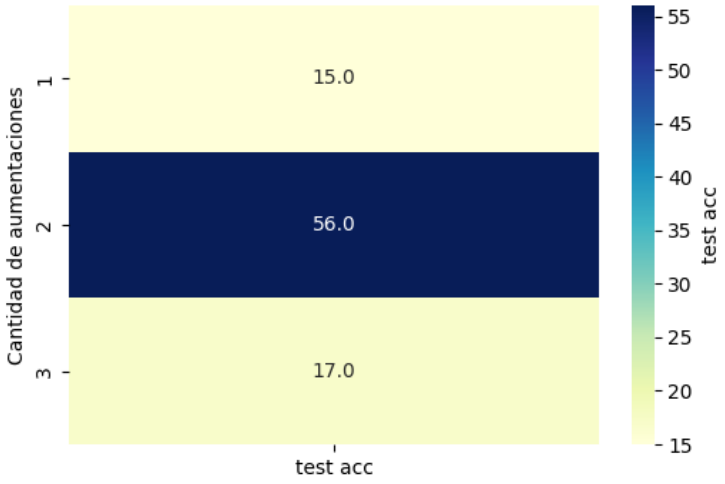

Fig. 70 Resultados 18Octubre, BT, BiLSTM

Por otra parte, los resultados para el modelo BiLSTM representados en la Fig. 70 indican que la clasificación alcanza su nivel más alto (56%) cuando el conjunto de datos es aumentado

2 veces. En el resto de los casos el rendimiento disminuye considerablemente, llegando a un mínimo de 15% cuando el conjunto de datos es aumentado 1 vez.

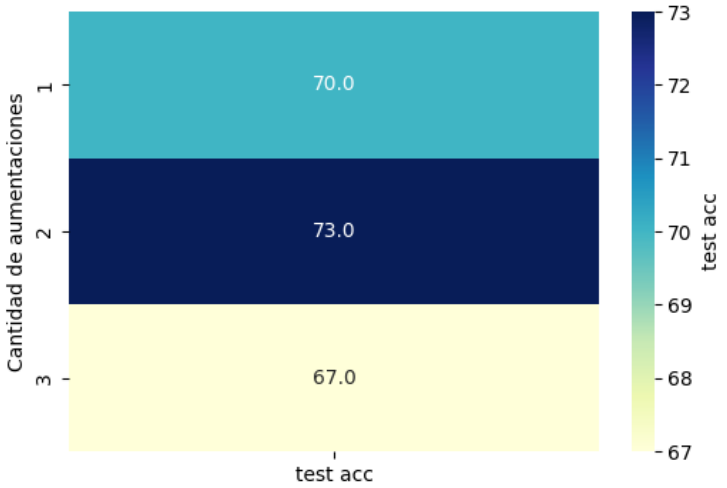

Fig. 71 Resultados 18 Octubre, BT, BERT

El último clasificador aplicado al conjunto de datos aumentado es BERT (Fig. 71), sus resultados indican que el mejor rendimiento (73%) es obtenido cuando el conjunto de datos se aumenta 2 veces. Mientras que el rendimiento más bajo se consigue cuando el conjunto es aumentado 3 veces.

Para finalizar con la revisión del conjunto de datos 18Octubre, la Tabla 43 consolida los resultados de clasificación con la métrica *accuracy*.

Tabla 43 Resultados consolidados 18Octubre / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
18Octubre	Base	56%	56%	56%	56%	63%
	BT	54%	56%	56%	56%	73%

5.7.3 Resultados para Agresividad

En las siguientes figuras se representan los resultados de clasificación luego de la aumentación mediante *back-translation* para el conjunto de datos de Agresividad.

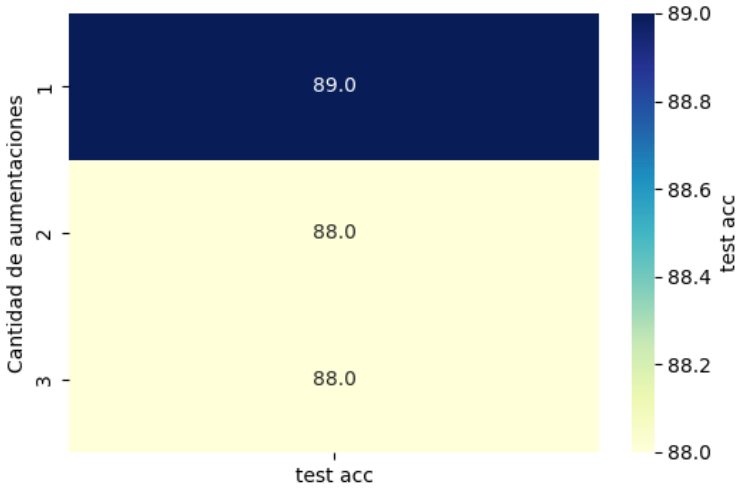

Fig. 72 Resultados Agresividad, BT, SVM

Los resultados para el clasificador SVM mostrados en la Fig. 72 indican que el mejor rendimiento de clasificación (89%) se obtiene al aumentar 1 vez el conjunto de datos. Mientras que, para el resto de las aumentaciones, el rendimiento de clasificación disminuye 1 punto en comparación al mejor resultado llegando a 88%.

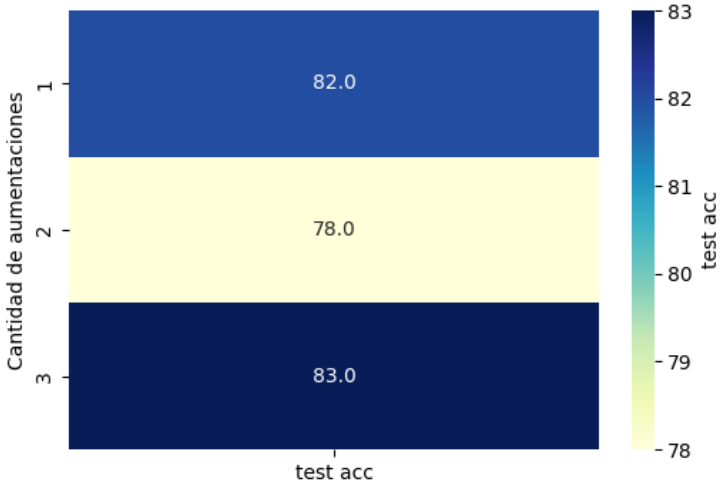

Fig. 73 Resultados Agresividad, BT, CNN

Los resultados de clasificación con el modelo CNN (Fig. 73) muestran una variación en los resultados de clasificación desde 78% en el punto más bajo, obtenido al aumentar el conjunto de datos 2 veces y 83% como mejor rendimiento de clasificación cuando el conjunto de datos es aumentado 3 veces.

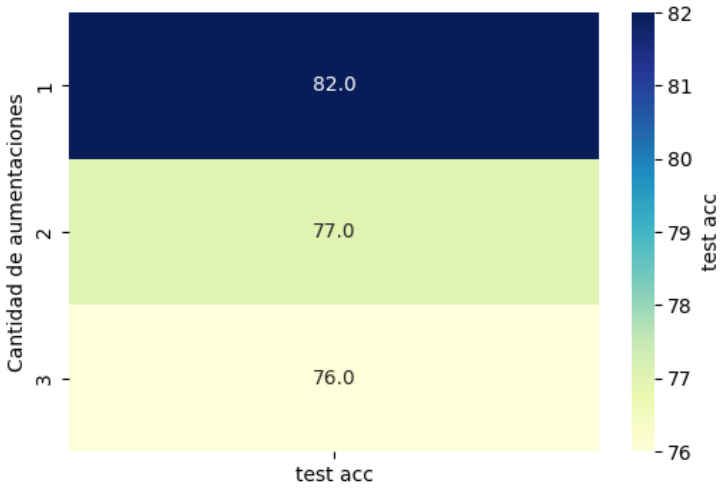

Fig. 74 Resultados Agresividad, BT, LSTM

Pasando al modelo de clasificación LSTM, los resultados mostrados en la Fig. 74 nos indican que el mejor rendimiento de clasificación es de 82% y es obtenido cuando el conjunto de datos es aumentado 1 vez. Por otra parte, el rendimiento de clasificación más bajo (76%) se obtiene cuando el conjunto de datos se aumenta 3 veces.

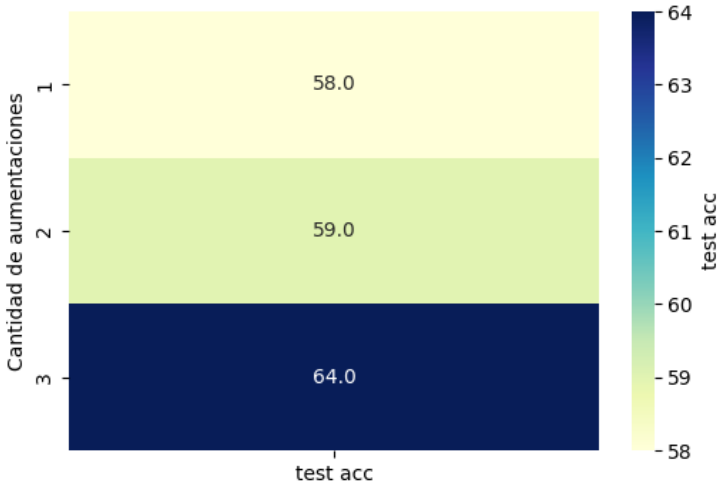

Fig. 75 Resultados Agresividad, BT, BiLSTM

Continuando con los resultados de clasificación, la Fig. 75 nos muestra que el mejor rendimiento de clasificación (64%) se logra al aumentar el conjunto de datos 3 veces. Mientras que, el rendimiento más bajo (58%) es obtenido cuando el conjunto de datos es aumentado 1 vez.

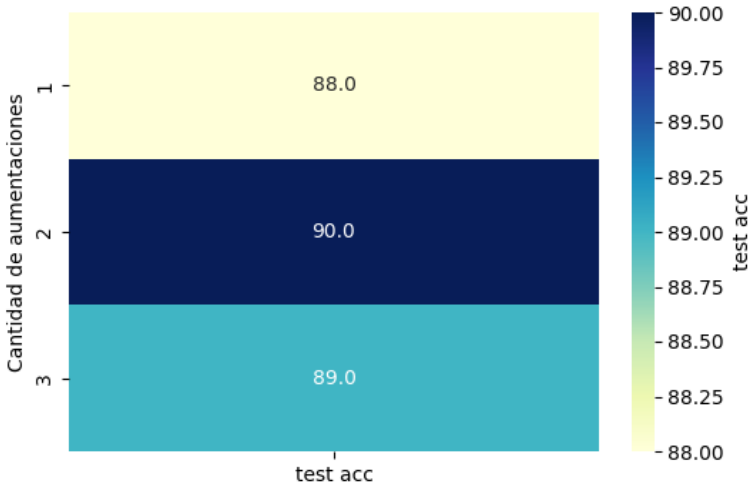

Fig. 76 Resultados Agresividad, BT, BERT

Los resultados de clasificación para el último modelo de clasificación (BERT), reflejados en la Fig. 76, nos indican que al aumentar 2 veces el conjunto de datos se obtiene el mejor rendimiento de clasificación (90%). Mientras que, el rendimiento más bajo (88%) es obtenido cuando el conjunto de datos es aumentado 1 vez.

Para finalizar con la revisión de resultados para el conjunto de datos Agresividad, la Tabla 44 consolida los resultados de clasificación y los compara con la línea base.

Tabla 44 Resultados consolidados Agresividad / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Agresividad	Base	90%	82%	84%	58%	90%
	BT	89%	83%	82%	64%	90%

5.7.4 Resultados para Emoji

A continuación, se presentan los resultados de clasificación luego de la aumentación mediante *back-translation* del conjunto de datos Emoji.

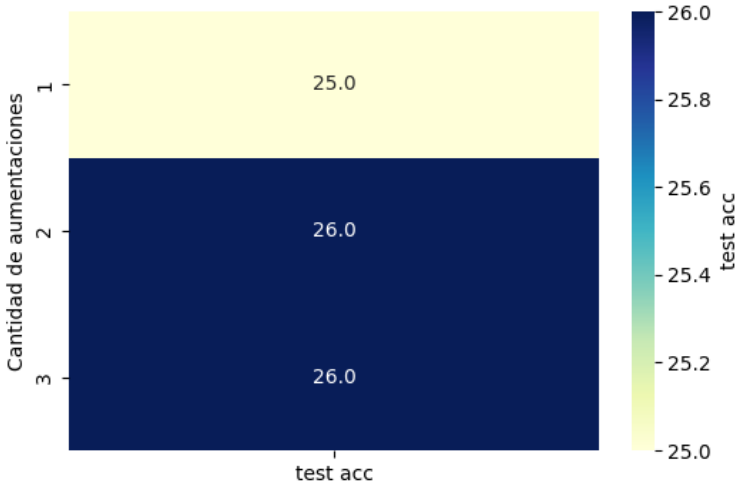

Fig. 77 Resultados Emoji, BT, SVM

Los resultados del clasificador SVM muestran que el mejor rendimiento (26%) se logra cuando el conjunto de datos es aumentado entre 2 y 3 veces, por el contrario, al aumentar el mismo conjunto de datos solamente 1 vez, se consigue el rendimiento más bajo de clasificación (25%). Lo anterior, es representado en la Fig. 77.

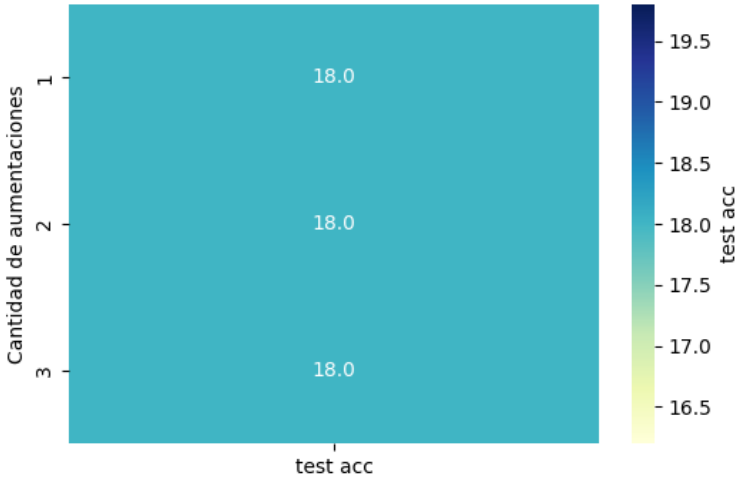

Fig. 78 Resultados Emoji, BT, CNN

Los resultados para el clasificador CNN plasmados en la Fig. 78 muestran que no existe variación en el porcentaje de clasificación con la métrica *accuracy* para el conjunto de datos Emoji aumentado con *back-translation*.

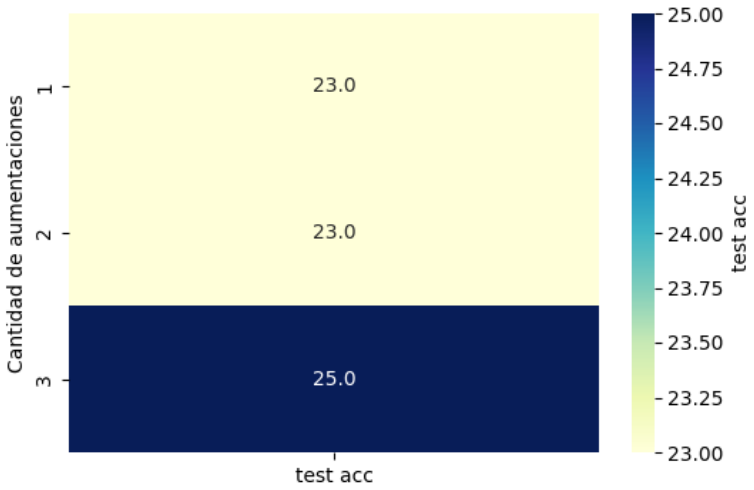

Fig. 79 Resultados Emoji, BT, LSTM

En cuanto a los resultados con el clasificador LSTM, la Fig. 79 nos indica que el mejor resultado de clasificación (25%) se da cuando el conjunto de datos es aumentado 3 veces. Por otra parte, cuando el mismo conjunto es aumentado 1 o 2 veces, el resultado de clasificación se mantiene en 23%.

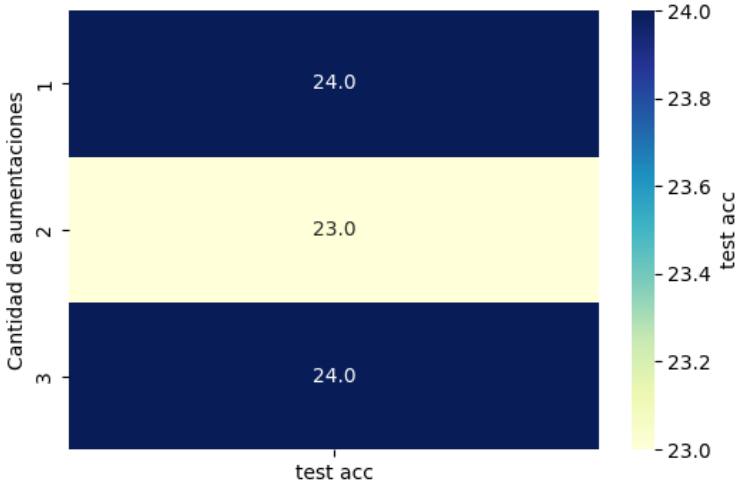

Fig. 80 Resultados Emoji, BT, BiLSTM

Los resultados de clasificación mostrados en la Fig. 80 corresponden al clasificador BiLSTM. En la figura se muestra que al aumentar el conjunto de datos 1 y 3 veces se obtienen

un rendimiento de 24% en accuracy. Mientras que, cuando el conjunto de datos es aumentado 2 veces el rendimiento conseguido es de 23%.

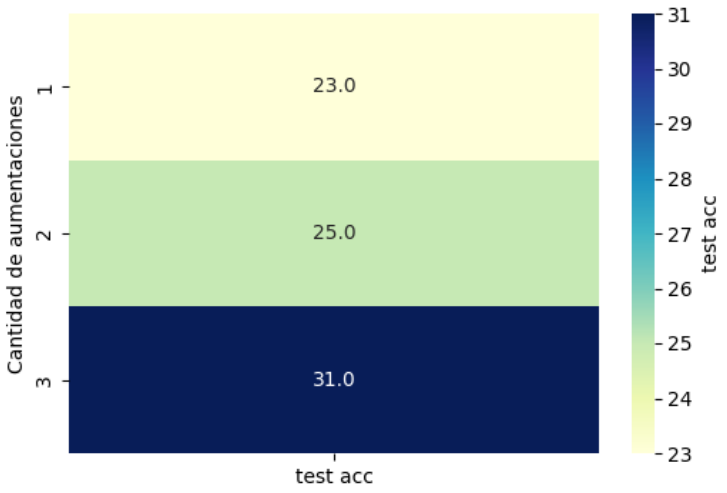

Fig. 81 Resultados Emoji, BT, BERT

La Fig. 81 muestra los resultados de clasificación con BERT, los resultados de clasificación indican que el mejor rendimiento (31%) es obtenido cuando el conjunto de datos es aumentado 3 veces. Mientras que, el rendimiento más bajo (23%) se registra cuando el mismo conjunto de datos es aumentado 1 vez.

Para finalizar con la revisión de resultados para el conjunto de datos Emoji, se presenta la Tabla 45 que muestra los resultados de clasificación por modelo y la comparación con la línea base.

Tabla 45 Resultados consolidados Emoji / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Emoji	Base	29%	22%	21%	23%	28%
	BT	26%	18%	25%	24%	31%

5.7.5 Resultados para Encuesta Docente Afecto

En este apartado se presentan los resultados de clasificación luego de aumentar el conjunto de datos Encuesta Docente Afecto.

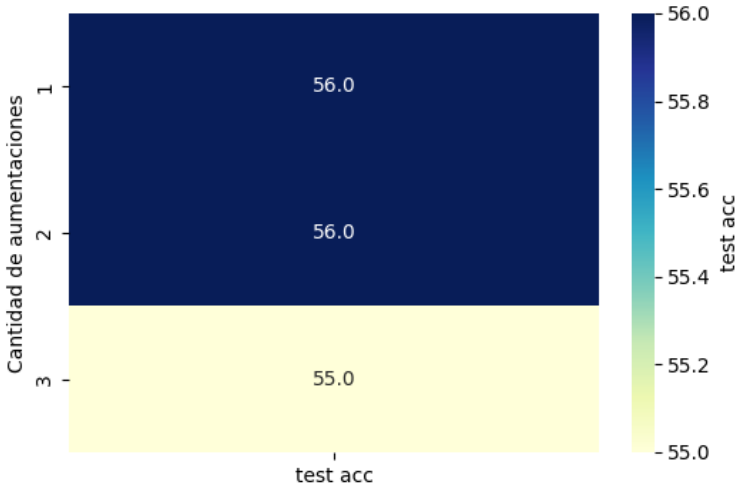

Fig. 82 Resultados Encuesta Docente Afecto, BT, SVM

La Fig. 82 muestra los resultados de clasificación con el modelo SVM, los resultados indican que al aumentar el conjunto de datos 1 o 2 veces produce un rendimiento de clasificación de 56%. Por el contrario, si el conjunto de datos es aumentado 3 veces el rendimiento de clasificación llega a 55% en la métrica *accuracy*.

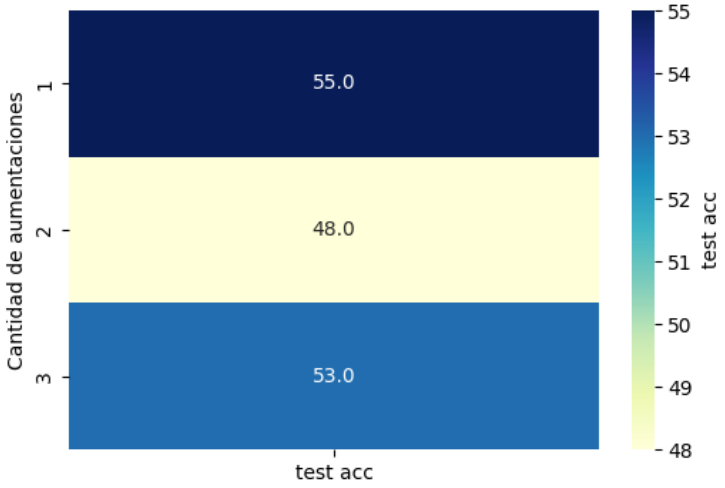

Fig. 83 Resultados Encuesta Docente Afecto, BT, CNN

Los resultados de clasificación con el clasificador CNN mostrados en la Fig. 83 indican que el rendimiento de clasificación más alto (55%) se producen cuando el conjunto de datos es aumentado 1 vez. Mientras que, al aumentar el conjunto de datos 2 veces produce el rendimiento de clasificación más bajo, llegando a 48%.

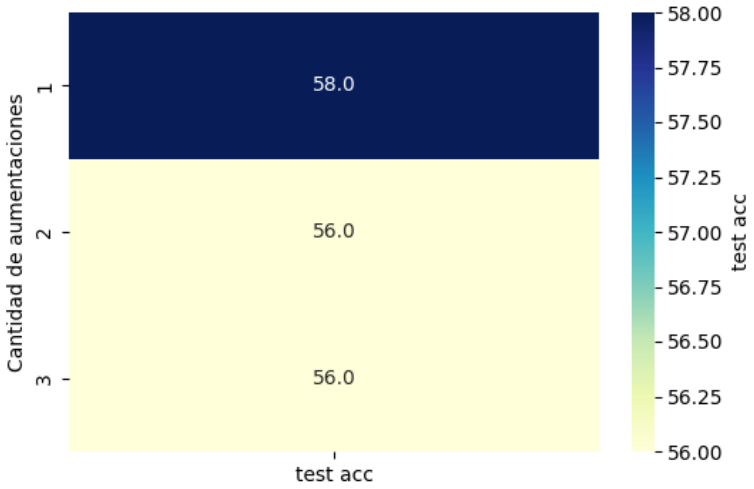

Fig. 84 Resultados Encuesta Docente Afecto, BT, LSTM

Los resultados de clasificación con el modelo LSTM mostrados en la Fig. 84, registran que al aumentar el conjunto de datos 1 vez se logra el mejor rendimiento de clasificación, llegando a 58%. En contraparte, cuando el conjunto de datos es aumentado 2 o 3 veces, el rendimiento de clasificación alcanza el 56% en ambos casos.

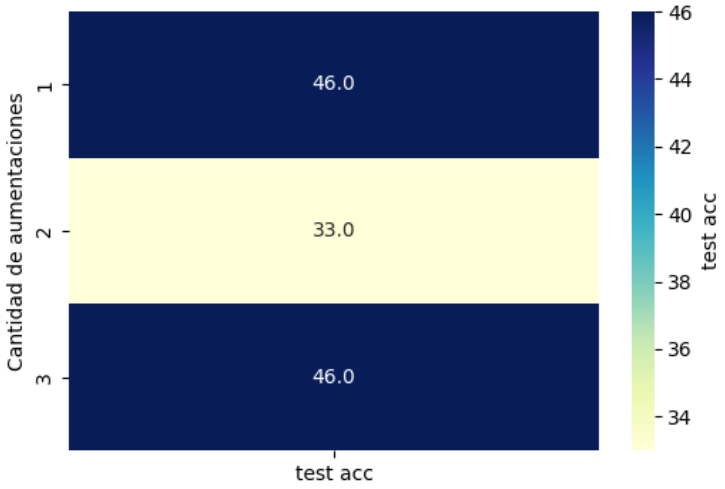

Fig. 85 Resultados Encuesta Docente Afecto, BT, BiLSTM

Los resultados de clasificación mostrados en la Fig. 85 nos indican que al aumentar 1 o 3 veces se obtiene el mismo rendimiento de clasificación (46%). Mientras que, al aumentar el conjunto de datos 2 veces se logra un rendimiento de 33%, 13% por debajo de los otros experimentos.

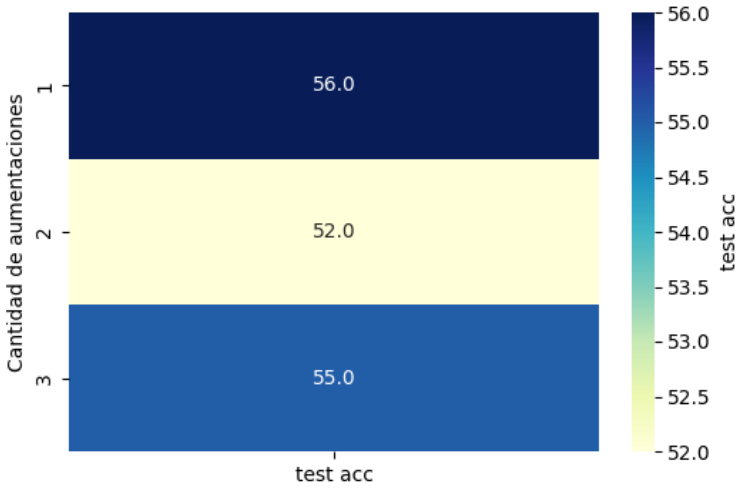

Fig. 86 Resultados Encuesta Docente Afecto, BT, BERT

La Fig. 86 registra los resultados de clasificación para el modelo BERT. Puede apreciarse que al aumentar el conjunto de datos 1 vez produce el mejor rendimiento, llegando a 56%. Por el contrario, al aumentar el conjunto de datos 2 veces produce el menor rendimiento de clasificación llegando a 52%.

Para finalizar con la revisión de los experimentos con el conjunto de datos Encuesta Docente Afecto aumentado con *back-translation*, se presenta la Tabla 46 que muestra los mejores resultados y los compara con la línea base.

Tabla 46 Consolidado resultados Encuesta Docente Afecto / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	BT	56%	55%	58%	46%	56%

5.7.6 Resultados para Encuesta Docente Agresividad

Los resultados de clasificación para el conjunto de datos Encuesta Docente Agresividad son presentados a continuación.

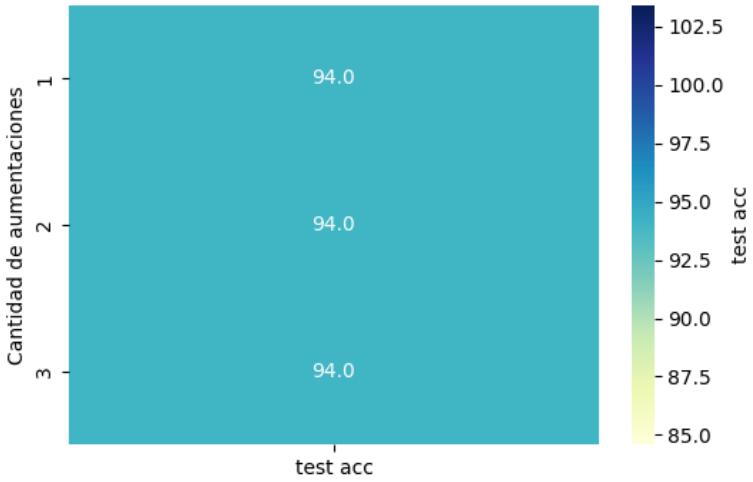

Fig. 87 Resultados Encuesta Docente Agresividad, BT, SVM

La Fig. 87 muestra que el rendimiento de clasificación con el clasificador SVM para el conjunto de datos aumentado hasta 3 veces con *back-translation* obtiene los mismos resultados 94%.

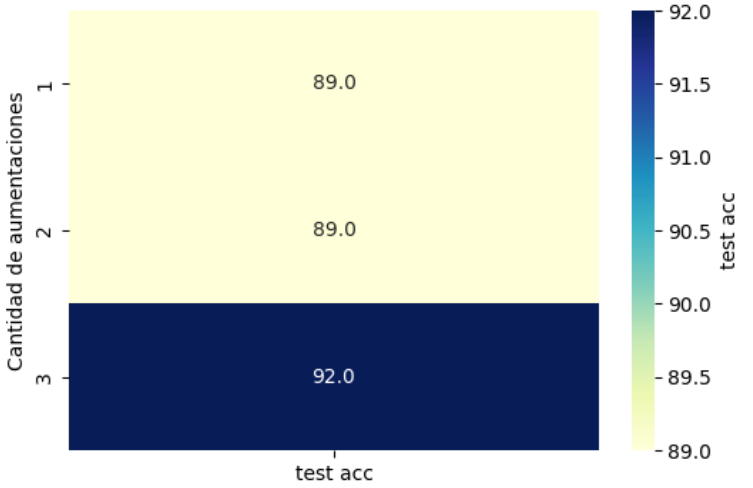

Fig. 88 Resultados Encuesta Docente Agresividad, BT, CNN

En cuanto a los resultados de clasificación con CNN, la Fig. 88 muestra que al aumentar el conjunto de datos 3 veces se registra el mejor rendimiento de clasificación (92%). Mientras que, cuando el conjunto de datos es aumentado 1 o 2 veces, el rendimiento de clasificación se mantiene en 89%.

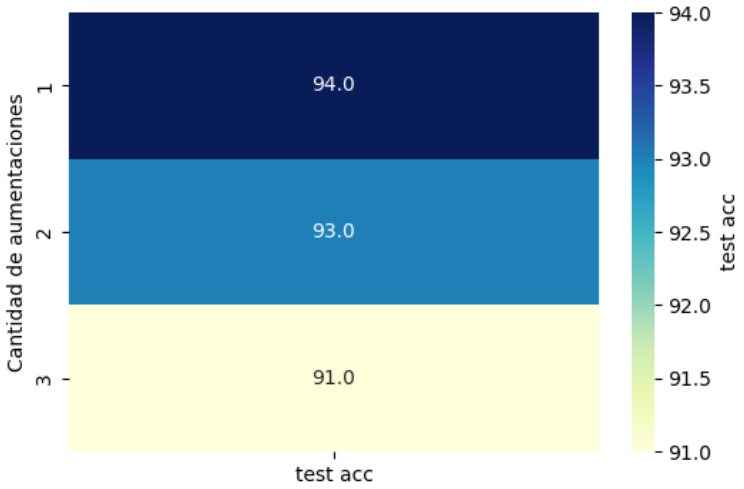

Fig. 89 Resultados Encuesta Docente Agresividad, BT, LSTM

Continuando con la revisión de resultados de clasificación luego de aumentar el conjunto de datos con *back-translation*, la Fig. 89 muestra que a medida que se generan más aumentaciones el porcentaje de clasificación baja desde 94% cuando el conjunto de datos se aumenta 1 vez y baja a 91% cuando es aumentado 3 veces.

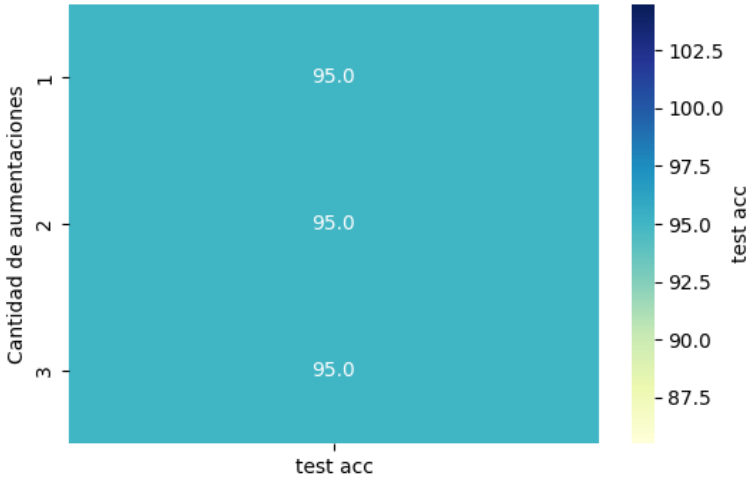

Fig. 90 Resultados Encuesta Docente Agresividad, BT, BiLSTM

La Fig. 90 muestra los resultados de clasificación con el modelo BiLSTM, los resultados indican que no existen diferencias en el porcentaje de clasificación alcanzado con este modelo al aumentar el conjunto de datos entre 1 y 3 veces.

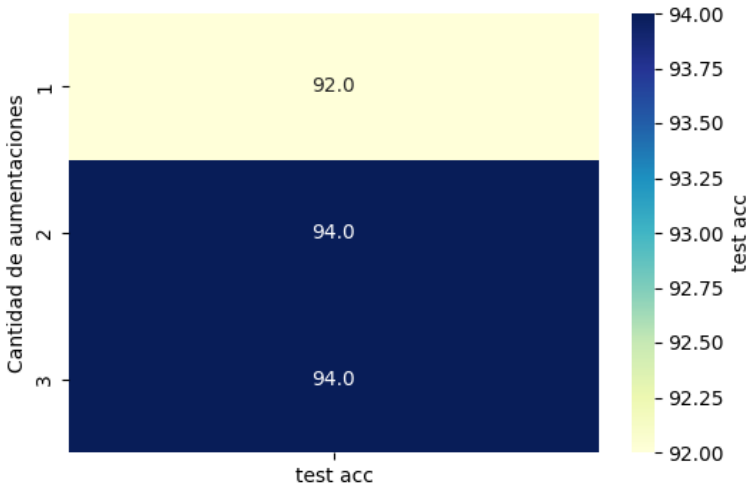

Fig. 91 Resultados Encuesta Docente Agresividad, BT, BERT

Por otra parte, los resultados de clasificación con el modelo BERT representados en la Fig. 91 indican que el mejor resultado (94%) se obtiene cuando el conjunto de datos es aumentado entre 2 y 3 veces. Mientras que, al aumentar 1 vez, el rendimiento llega a 92%.

Para finalizar con la revisión de resultados para el conjunto de datos Encuesta Docente Agresividad aumentado, la Tabla 47 muestra los mejores resultados por clasificador y los compara con la línea base.

Tabla 47 Consolidado resultados Encuesta Docente Agresividad / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	BT	94%	92%	94%	95%	94%

5.7.7 Resultados para Encuesta Docente Polaridad

En este apartado se muestran los resultados de clasificación sobre los conjuntos de datos aumentados de Encuesta Docente Polaridad mediante *back-translation.*

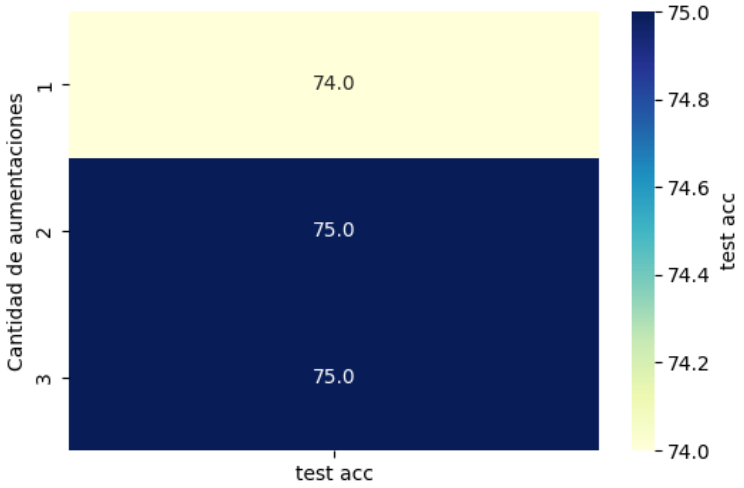

Fig. 92 Resultados Encuesta Docente Polaridad, BT, SVM

Los resultados en la Fig. 92 muestran que al aumentar el conjunto de datos en 2 o 3 veces se logra un rendimiento de clasificación de 75%. Mientras que, cuando el conjunto de datos es aumentado 1 vez, el rendimiento de clasificación llega al 74%.

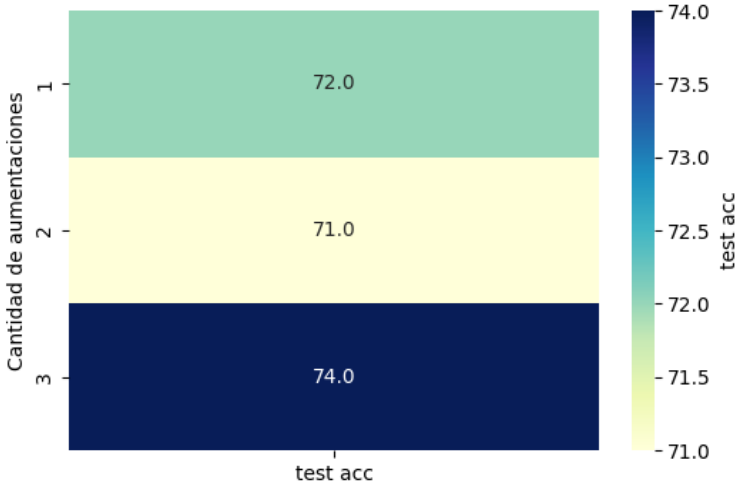

Fig. 93 Resultados Encuesta Docente Polaridad, BT, CNN

Los resultados de clasificación con CNN (Fig. 93) indican que al aumentar 3 veces el conjunto de datos se consigue un 74% de rendimiento. Mientras que, el rendimiento más bajo se produce cuando el conjunto de datos es aumentado 2 veces.

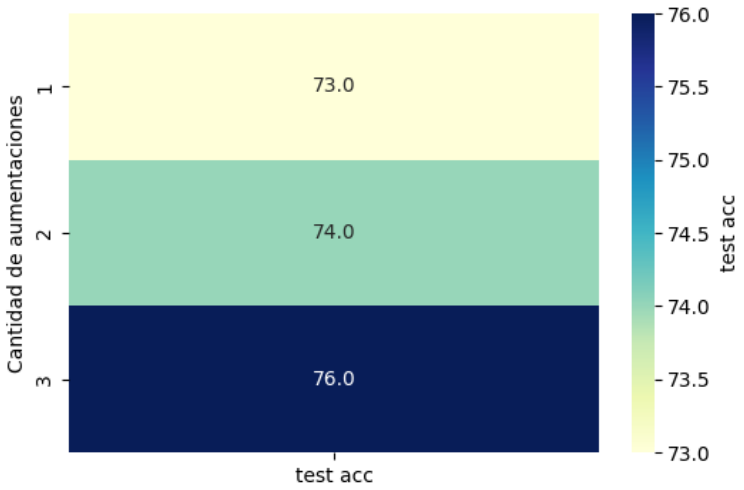

Fig. 94 Resultados Encuesta Docente Polaridad, BT, LSTM

En el caso del clasificador LSTM, los resultados de la Fig. 94 indican que a medida que se incrementa el número de aumentaciones, el rendimiento también aumenta. Llegando a un máximo de 76%. Por otra parte, el rendimiento más bajo alcanza el 73%.

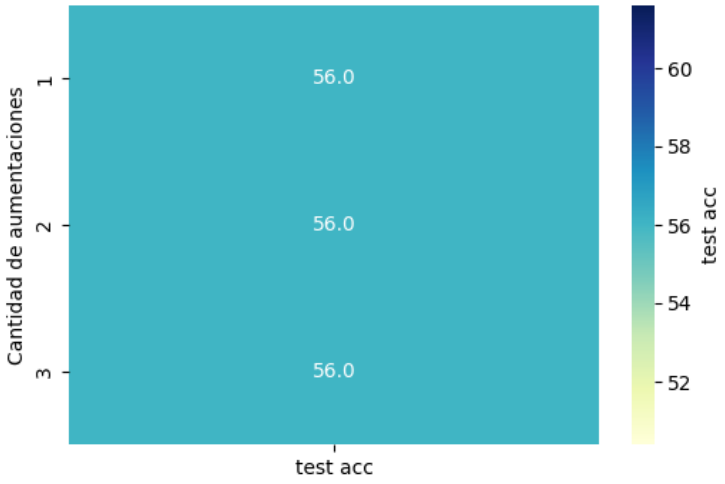

Fig. 95 Resultados Encuesta Docente Polaridad, BT, BiLSTM

Continuando con los resultados de clasificación, la Fig. 95 muestra el rendimiento sobre el conjunto de datos aumentado del clasificador BiLSTM. El rendimiento registrado no varía a medida que se aumenta el conjunto de datos, obteniendo un 56% en cada aumentación.

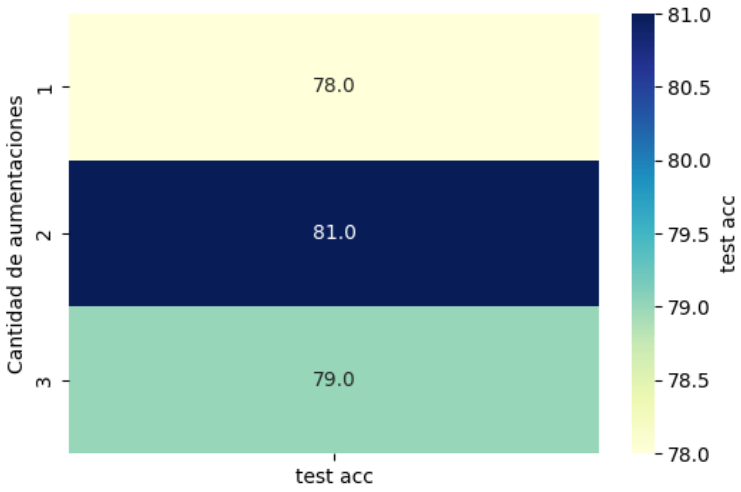

Fig. 96 Resultados Encuesta Docente Polaridad, BT, BERT

Los resultados para el clasificador BERT (Fig. 96) muestran que el mejor resultado de clasificación se produce cuando el conjunto de datos es aumentado 2 veces. Mientras que, el resultado más bajo se obtiene cuando el conjunto es aumentado 1 vez.

Para finalizar con la revisión de los resultados de clasificación para el conjunto de datos Encuesta Docente Polaridad aumentado, se muestra la Tabla 48 que resume los mejores resultados y los compara con la línea base obtenida en 5.4.

Tabla 48 Consolidado resultados Encuesta Docente Polaridad / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Polaridad	Base	71%	75%	72%	56%	80%
	BT	75%	74%	76%	56%	81%

5.7.8 Resultados para Encuesta Docente Seriedad

A continuación, se presentan los resultados de clasificación posterior a la aumentación con back translation para el conjunto de datos Encuesta Docente Seriedad.

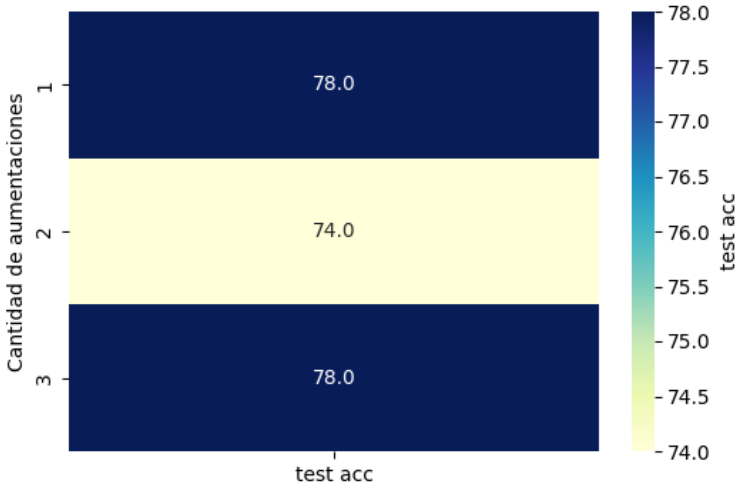

Fig. 97 Resultados Encuesta Docente Seriedad, BT, SVM

Al clasificar el conjunto de datos aumentado con el clasificador SVM se obtiene un máximo rendimiento de 78% cuando el conjunto de datos es aumentado 1 y 3 veces. Mientras que, al aumentar el conjunto 2 veces se logra un rendimiento de 74%. Lo anteriormente descrito es mostrado en la Fig. 97.

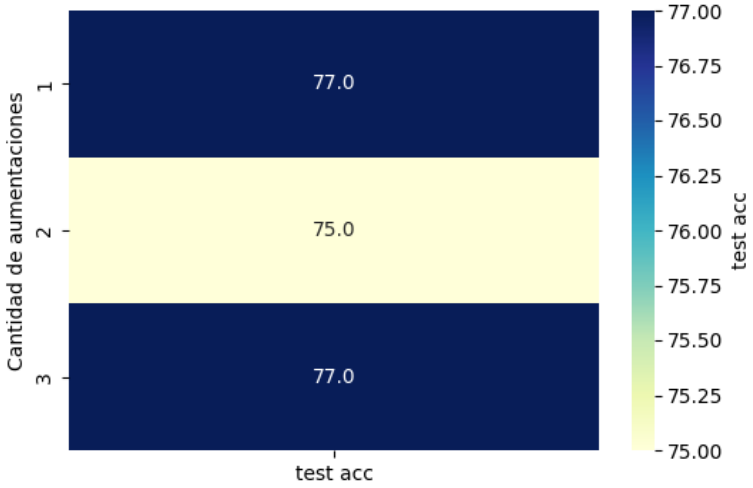

Fig. 98 Resultados Encuesta Docente Seriedad, BT, CNN

La Fig. 98 muestra los resultados de clasificación con el modelo CNN. En ella, se registra que el menor rendimiento de clasificación (75%) se produce al aumentar el conjunto de datos 2 veces, Para el resto de las aumentaciones (1 y 3) el rendimiento de clasificación alcanza un 77%.

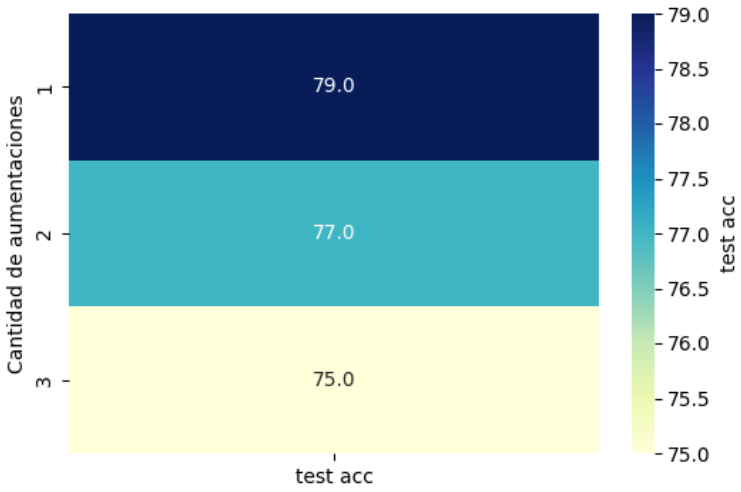

Fig. 99 Resultados Encuesta Docente Seriedad, BT, LSTM

Los resultados para el clasificador LSTM representados en la Fig. 99 indican que al aumentar el número de aumentaciones el rendimiento de clasificación disminuye. Cuando el conjunto es aumentado 1 vez el rendimiento alcanza un 79%. Mientras que, cuando el conjunto es aumentado 3 veces, el rendimiento llega a 75%.

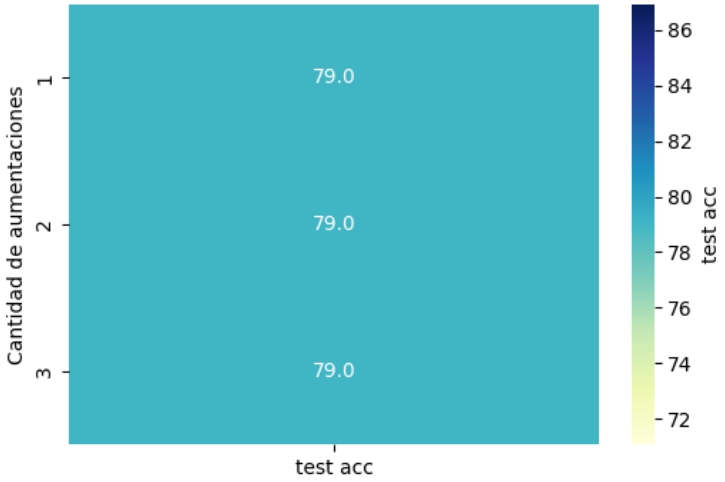

Fig. 100 Resultados Encuesta Docente Seriedad, BT, BiLSTM

La Fig. 100 muestra que el rendimiento de clasificación con el clasificador BiLSTM es el mismo no importando el número de veces que es aumentado el conjunto de datos, alcanzando un 79% de rendimiento en la métrica accuracy.

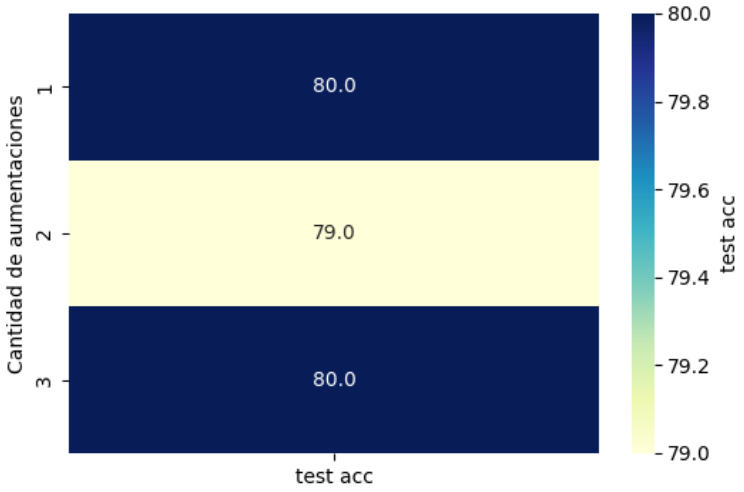

Fig. 101 Resultados Encuesta Docente Seriedad, BT, BERT

La Fig. 101 representa los resultados para el clasificador BERT. Puede observarse en la figura que el rendimiento de clasificación más bajo para los conjuntos de datos aumentados se produce cuando existen 2 aumentaciones. Mientras que, para 1 y 3 aumentaciones, el rendimiento alcanza el 80%.

Para finalizar con la presentación de resultados, la Tabla 49 muestra los mejores resultados obtenidos al clasificar los conjuntos de datos aumentados.

Tabla 49 Consolidado resultados Encuesta Docente Seriedad / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	BT	78%	77%	79%	79%	80%

5.7.9 Resultados para Titulares de Diario

A continuación, se presentan los resultados de clasificación para el conjunto de datos Titulares de Diario aumentado con *back-translation*.

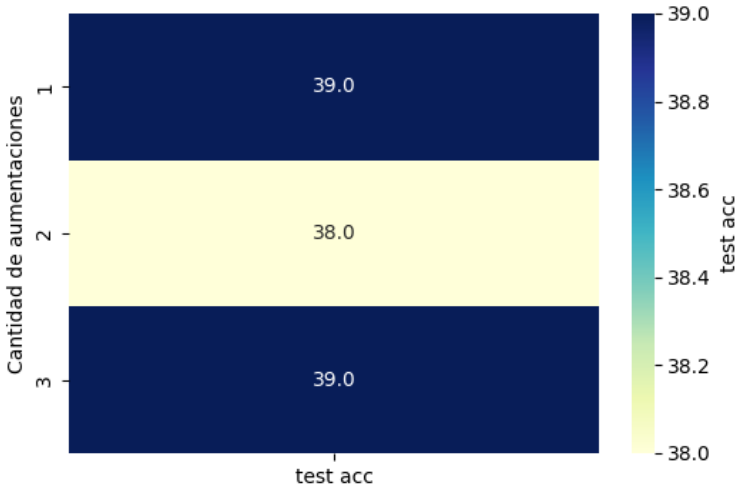

Fig. 102 Resultados Titulares Diarios, BT, SVM

Los resultados de clasificación con SVM son mostrados en la Fig. 102 indican que cuando el conjunto de datos es aumentado 2 veces se obtiene el menor rendimiento de clasificación con 38%. Por otro lado, al aumentar el conjunto de datos 1 y 3 veces el rendimiento de clasificación aumenta a 39%.

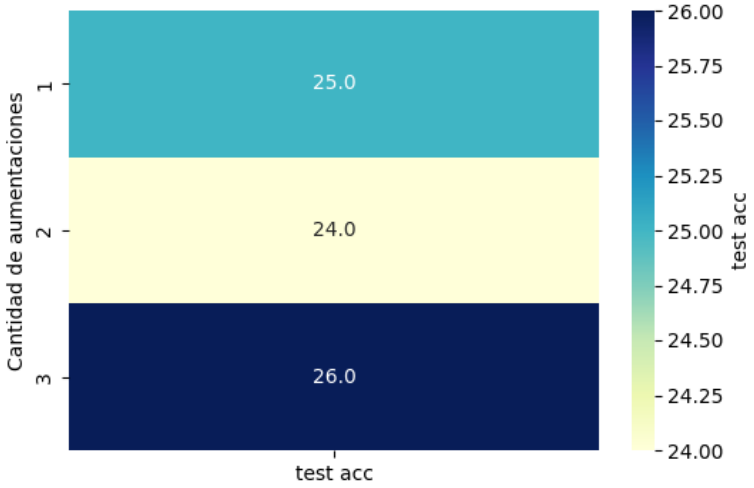

Fig. 103 Resultados Titulares Diarios, BT, CNN

En el caso del clasificador CNN, los resultados representados en la Fig. 103 indican que al aumentar 3 veces el conjunto de datos el rendimiento de clasificación alcanza un 26%. Por otra parte, el rendimiento más bajo se obtiene cuando el conjunto de datos es aumentado 2 veces, llegando a 24%.

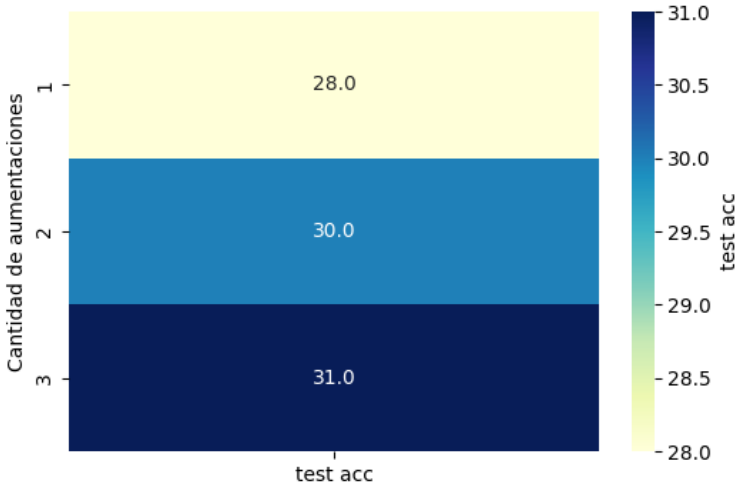

Fig. 104 Resultados Titulares Diarios, BT, LSTM

Los resultados de clasificación con LSTM mostrados en la Fig. 104 indican que a medida que se incrementa el número de aumentaciones, el rendimiento también se incrementa. Al aumentar 1 vez el conjunto de datos se registra un 28%. Mientras que, cuando el conjunto es aumentado 3 veces se alcanza un 31%.

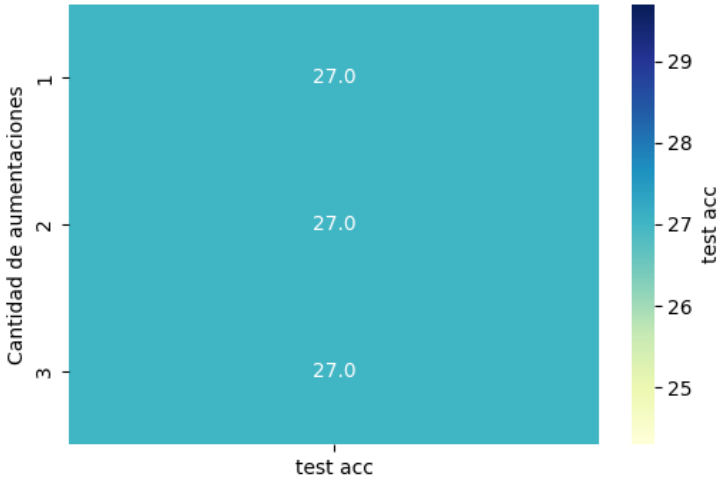

Fig. 105 Resultados Titulares Diarios, BT, BiLSTM

La Fig. 105 muestra que cuando el conjunto de datos aumentado es clasificado con BiLSTM el rendimiento es igual para todos los experimentos realizados, alcanzando un 27%.

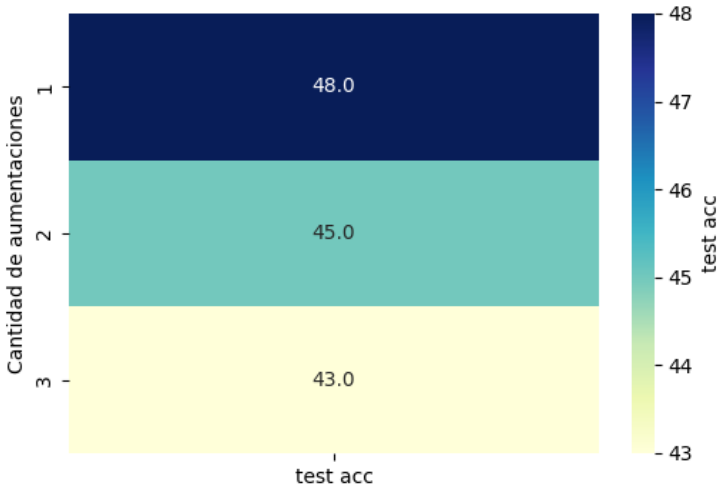

Fig. 106 Resultados Titulares Diarios, BT, BERT

Los resultados de clasificación con BERT (Fig. 106) muestran que a medida que se incrementa el número de aumentaciones el rendimiento de clasificación baja. Esto queda reflejado con el rendimiento al aumentar el conjunto de datos 1 vez (48%) que supera en 5 puntos al rendimiento con 3 aumentaciones (43%).

Para finalizar con la revisión de los resultados de clasificación para el conjunto de datos Titulares Diarios aumentado, se presenta la Tabla 50 que muestra los mejores resultados y los compara con la línea base.

Tabla 50 Consolidado resultados Titulares Diarios / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Titulares Diarios	Base	39%	26%	31%	27%	44%
	BT	39%	26%	31%	27%	48%

5.7.10 Resultados para Violencia Género

A continuación, se presentan los resultados de clasificación del conjunto de datos Violencia de Género aumentado mediante *back-translation*.

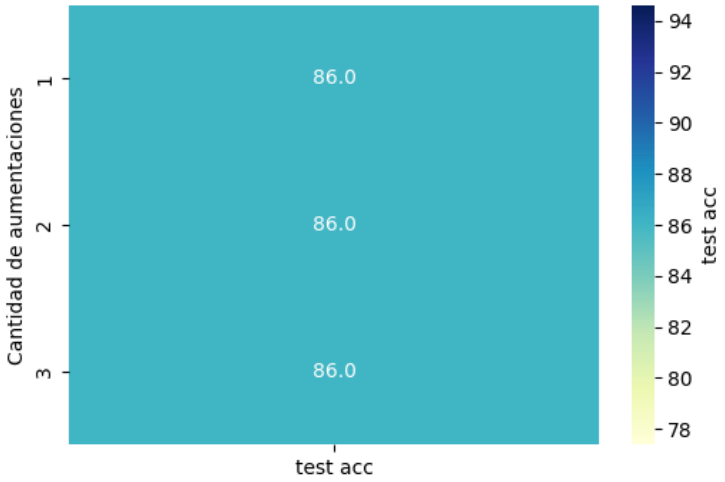

Fig. 107 Resultados Violencia Género, BT, SVM

Los resultados para el clasificador SVM (Fig. 107) indican que no importa la cantidad de veces que el conjunto de datos sea aumentado, el rendimiento de clasificación se mantiene en 86% en la métrica accuracy.

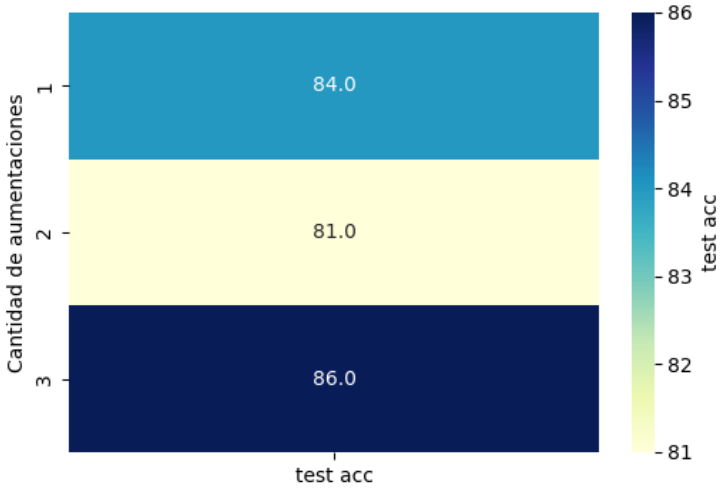

Fig. 108 Resultados Violencia Género, BT, CNN

Los resultados para el clasificador CNN (Fig. 108) muestran que el mejor rendimiento de clasificación se logra cuando el conjunto de datos es aumentado 3 veces, llegando a 86%. Mientras que, el rendimiento más bajo (81%) es obtenido cuando el mismo conjunto es aumentado 3 veces.

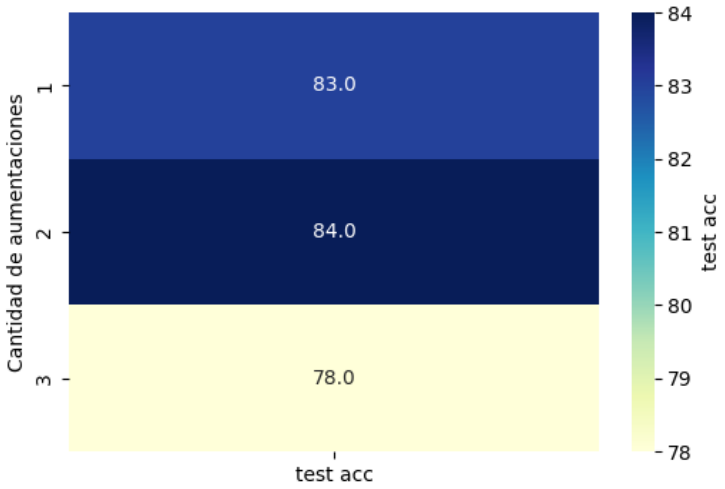

Fig. 109 Resultados Violencia Género, BT, LSTM

En el caso del clasificador LSTM, la Fig. 109 muestra que el mejor rendimiento de clasificación se logra cuando el conjunto de datos es aumentado 2 veces, llegando a 84%. Por otra parte, el rendimiento más bajo (78%) es obtenido cuando el conjunto de datos es aumentado 2 veces.

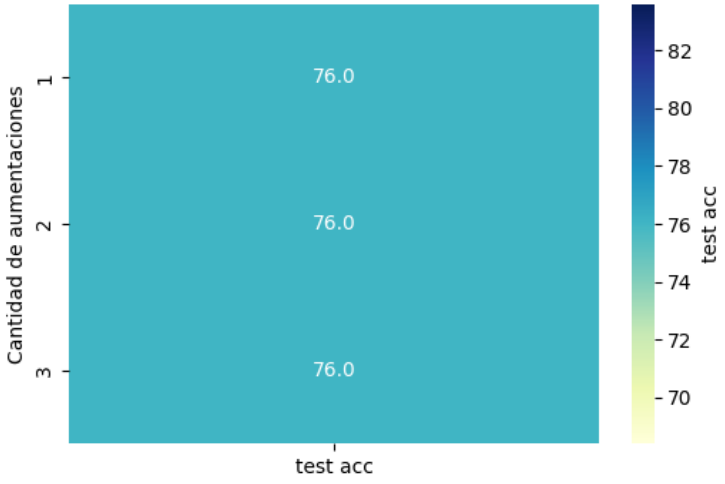

Fig. 110 Resultados Violencia Género, BT, BiLSTM

La Fig. 110 muestra que los resultados de clasificación con BiLSTM no importa la cantidad de aumentaciones que se realicen sobre el conjunto de datos, el rendimiento se mantiene constante en 76%.

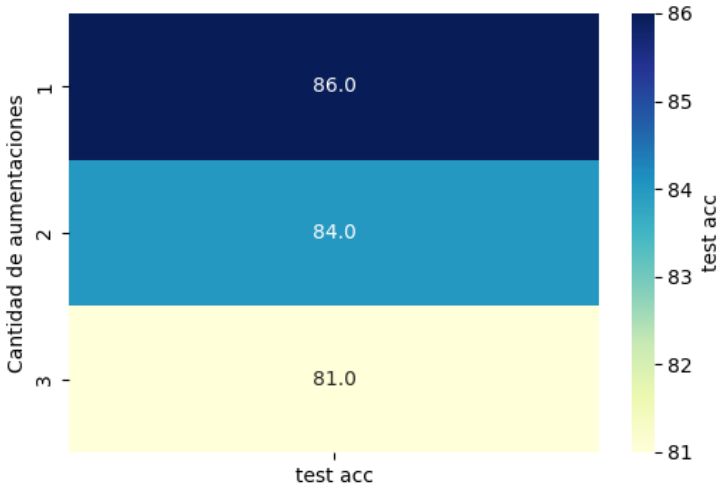

Fig. 111 Resultados Violencia Género, BT, BERT

Por otra parte, los resultados de clasificación con BERT (Fig. 111) indican que a medida que se incrementa la cantidad de aumentaciones el rendimiento de clasificación decrece. Esto queda reflejado en la diferencia entre aumentar el conjunto de datos 1 vez (86%) y 3 veces (81%) marcando una diferencia de 5%.

Para finalizar con la presentación de resultados de clasificación para el conjunto de datos Violencia de Género, la Tabla 51 refleja los mejores resultados y los compara con la línea base del punto 5.3.

Tabla 51 Consolidado resultados Violencia Género / BT

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
Violencia Género	Base	86%	76%	83%	76%	83%
	BT	86%	86%	84%	76%	86%

5.7.11 Resultados consolidados BT

Para finalizar con los resultados con la técnica de aumentación por parafraseo de oraciones, se presenta la Tabla 52 que consolida los resultados con BT (*back-translation*) para cada conjunto de datos. En la tabla se encuentran resaltados en rojo los rendimientos de

clasificación que están por debajo de la línea base y en azul el mejor resultado para el conjunto de datos en términos de puntos porcentuales por sobre la línea base.

Tabla 52 Resultados consolidados BT

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
18Octubre	Base	56%	56%	56%	56%	63%
	BT	**54%**	56%	56%	56%	**73%**
Agresividad	Base	90%	82%	84%	58%	90%
	BT	**89%**	83%	**82%**	**64%**	90%
Emoji	Base	29%	22%	21%	23%	28%
	BT	**26%**	**18%**	**25%**	24%	31%
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	BT	**56%**	55%	**58%**	**46%**	56%
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	BT	**94%**	**92%**	**94%**	95%	**94%**
Encuesta Docente Polaridad	Base	71%	75%	72%	56%	80%
	BT	**75%**	**74%**	**76%**	56%	81%
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	BT	**78%**	77%	79%	79%	**80%**
Titulares Diarios	Base	39%	26%	31%	27%	44%
	BT	39%	26%	31%	27%	**48%**
Violencia Género	Base	86%	76%	83%	76%	83%
	BT	86%	**86%**	84%	76%	86%

5.8 Resultados generales de la experimentación con aumentación.

La siguiente tabla (Tabla 53) registra los resultados finales de la aumentación con las 3 técnicas seleccionadas y realiza una comparación entre estas y la línea base.

Las técnicas de aumentación presentes en la tabla corresponden a:

- EDA: técnica de aumentación por transformación
- EDA-B: balanceo de clases utilizando EDA
- SentiGAN: técnica de aumentación por generación
- BT: técnica de aumentación por parafraseo *back-translation*

Tabla 53 Resultados finales de experimentación

Conjunto de datos	**Técnica DA**	**Clasificador**				
		SVM	**CNN**	**LSTM**	**BiLSTM**	**BERT**
18 Octubre	Base	56%	56%	56%	56%	63%
	EDA	**54%**	58%	56%	56%	67%
	EDA-B	N/A	N/A	N/A	N/A	N/A
	BT	**54%**	56%	56%	56%	**73%**
	SentiGAN	58%	44%	42%	17%	3%
Agresividad	Base	90%	82%	84%	58%	90%
	EDA	91%	83%	85%	67%	89%
	EDA-B	N/A	N/A	N/A	N/A	N/A
	BT	89%	83%	82%	64%	90%
	SentiGAN	99%	**100%**	99%	64%	94%
Emoji	Base	29%	22%	21%	23%	28%
	EDA	29%	19%	25%	25%	25%
	EDA-B	24%	8%	5%	**3%**	**3%**
	BT	26%	18%	25%	24%	**31%**
Encuesta Docente Afecto	Base	57%	52%	59%	33%	56%
	EDA	58%	54%	57%	**46%**	61%
	EDA-B	N/A	N/A	N/A	N/A	N/A
	BT	56%	55%	58%	**46%**	56%
	SentiGAN	43%	43%	25%	13%	4%
Encuesta Docente Agresividad	Base	95%	95%	95%	95%	96%
	EDA	95%	95%	95%	95%	96%
	EDA-B	93%	91%	90%	5%	**4%**

Conjunto de datos	Técnica DA	Clasificador				
		SVM	CNN	LSTM	BiLSTM	BERT
	BT	94%	92%	94%	95%	94%
	SentiGAN	90%	92%	95%	95%	96%
Encuesta Docente Polaridad	Base	71%	75%	72%	56%	80%
	EDA	75%	76%	**77%**	56%	82%
	EDA-B	75%	69%	71%	56%	78%
	BT	75%	74%	76%	56%	81%
	SentiGAN	73%	71%	66%	**31%**	72%
Encuesta Docente Seriedad	Base	79%	76%	79%	79%	77%
	EDA	79%	76%	79%	79%	**81%**
	EDA-B	76%	75%	75%	21%	**20%**
	BT	78%	77%	79%	79%	80%
	SentiGAN	79%	79%	79%	79%	80%
Titulares Diarios	Base	39%	26%	31%	27%	44%
	EDA	41%	27%	36%	30%	43%
	EDA-B	38%	**33%**	25%	**8%**	25%
	BT	39%	26%	31%	27%	48%
	SentiGAN	34%	26%	29%	22%	42%
Violencia de Género	Base	86%	76%	83%	76%	83%
	EDA	86%	**87%**	85%	76%	87%
	EDA-B	86%	85%	**73%**	76%	84%
	BT	86%	86%	84%	76%	86%
	SentiGAN	85%	85%	81%	50%	78%

Capítulo 6 Discusión de resultados

A continuación, se comentan los resultados en un contexto más crítico, no obstante, es necesario considerar que nuestras interpretaciones se basan o limitan según los experimentos desarrollados, considerando las restricciones técnicas y las implementaciones, así como las características de los corpus utilizados en este proyecto.

6.1 Técnicas de aumentación por transformación

Dos tipos de experimentos fueron llevados a cabo con EDA [31] con resultados muy disímiles. El primero de los experimentos consistió en aumentar los conjuntos de datos sin importar el balance de las clases presentes en estos. Los resultados de este experimento muestran que la mayor parte de los conjuntos de datos aumentados incrementó el rendimiento de los clasificadores en las tareas de análisis de sentimientos (SA) y análisis de emociones (EA). Mientras que los resultados del segundo experimento, que buscó balancear la cantidad de muestras por clase en cada conjunto de datos provocaron un decrecimiento en el rendimiento de clasificación de los modelos de DL en 3 de los conjuntos de datos que fueron balanceados con EDA (Emoji, Encuesta Docente Agresividad y Encuesta Docente Seriedad).

El primer experimento cuyos resultados son mostrados a través de la tabla Tabla 54, permite observar que no es especialmente relevante cuantas veces es aumentado el conjunto de datos. Mientras que, el porcentaje de modificación de palabras en las oraciones resulta tener un impacto positivo en la aumentación. Este comportamiento indica que más allá de la cantidad de muestras que pueda tener el conjunto de datos a clasificar, la variable que más influye es la diversidad de dichas muestras.

Tabla 54 Mejores rendimientos EDA

Conjunto de datos	**Rendimiento vs línea base**	**N° aumentaciones**	**% modificación de palabras**	**N° palabras por oración**	**Tarea de clasificación**
18 de Octubre	+5%	1	20%	17	EA
Emoji	+4%	1	20%	11	EA

Conjunto de datos	Rendimiento vs línea base	N° aumentaciones	% modificación de palabras	N° palabras por oración	Tarea de clasificación
Encuesta Docente Afecto	+13%	1	5%	5	EA
Titulares de Diarios	+5%	1	10%	12	EA
Agresividad	+9%	6	30%	22	SA
Encuesta Docente Agresividad	0%	1	5	5	SA
Encuesta Docente Polaridad	+5%	2	20%	5	SA
Encuesta Docente Seriedad	+4%	6	10	5	SA
Violencia de Género	+9%	2	30%	44	SA

Otro hallazgo relacionado con el primer experimento es que entre más palabras tenga una oración, es necesario aplicar un mayor porcentaje de modificación de las palabras presentes en ella. Teniendo en mente este comportamiento, es posible evitar el *overfitting* al momento de entrenar los clasificadores, ya que, a un mayor porcentaje de modificación de palabras, más diversas son las oraciones producidas.

6.2 Técnicas de aumentación por generación

Dada la forma de generación de oraciones utilizada por SentiGAN [58], puede decirse que los atributos más impactan en la generación de oraciones es el largo promedio de estas y la tarea de clasificación que se busca resolver con el conjunto de datos. Esto queda patente al revisar la tabla Tabla 55, en la que puede apreciarse que los conjuntos de datos que tienen más de 15 palabras por oración y que corresponden a la tarea de SA ven incrementado el rendimiento de clasificación con respecto a la línea base. En contraparte, los conjuntos aumentados generados a partir de oraciones con una cantidad promedio de palabras menor a 15 y que están enfocadas en EA ven perjudicado el rendimiento de clasificación con respecto a la línea base.

En cuanto a los clasificadores, el que tiene mejor rendimiento de clasificación con los conjuntos de datos aumentados con SentiGAN es CNN con una mejora de 18% en el conjunto de datos de Agresividad y con un 9% en el conjunto de datos de Violencia de Género. Este comportamiento se explica debido a que SentiGAN en el proceso adversarial (generador vs discriminador) utiliza una red CNN para determinar cuáles son las muestras que serán agregadas al conjunto de datos aumentado.

Tabla 55 Mejores rendimientos SentiGAN

Conjunto de datos	**Rendimiento vs línea base**	**Clasificador**	**Cantidad de palabras por oración**	**Tarea de clasificación**
18 de Octubre	+2%	SVM	17	EA
Encuesta Docente Afecto	-52%	BERT	5	EA
Titulares de Diarios	-5%	SVM / BiLSTM	12	EA
Agresividad	+18%	CNN	22	SA
Encuesta Docente Agresividad	-5%	SVM	5	SA
Encuesta Docente Polaridad	+2%	SVM	5	SA

Conjunto de datos	Rendimiento vs línea base	Clasificador	Cantidad de palabras por oración	Tarea de clasificación
Encuesta Docente Seriedad	+3%	CNN / BERT	5	SA
Violencia de Género	+9%	CNN	44	SA

6.3 Técnicas de aumentación por parafraseo

En el caso de la aumentación mediante *back-translation* los resultados obtenidos (

Tabla 56) indican que este tipo de aumentación tiene un impacto positivo en el rendimiento de clasificación en la mayor parte de los conjuntos de datos, especialmente cuando estos son clasificados con modelos DL. Si se analizan los resultados con base en la cantidad de aumentaciones realizadas y los idiomas utilizados, puede apreciarse que el uso de *back-translation* para aumentar los conjuntos de datos tiene una incidencia positiva a partir de la utilización de un idioma en los resultados de clasificación obtenidos posterior a la aumentación.

Al obtener buenos resultados transversales sobre los conjuntos de datos sin importar las características de estos, puede interpretarse que el elemento principal que impacta en la aumentación con estas técnicas es la herramienta de traducción utilizada, en este trabajo se utilizó el Traductor de Google que contiene diccionarios en múltiples idiomas, permitiendo la generación de oraciones lo suficientemente diferentes para generar un impacto positivo en la clasificación.

Tabla 56 Mejores rendimientos back-translation

Conjunto de datos	Rendimiento vs línea base	Clasificador	N° aumentaciones	Idiomas	Tarea de clasificación
18 de Octubre	+10%	BERT	2	EN/DE	EA
Emoji	+4%	LSTM	3	EN/DE/FR	EA
Encuesta Docente Afecto	+13%	BiLSTM	1	EN	EA
Titulares de Diarios	+4%	BERT	1	EN	EA
Agresividad	+6%	BiLSTM	3	EN/DE/FR	SA
Encuesta Docente Agresividad	0%	N/A	N/A	N/A	SA
Encuesta Docente Polaridad	+4%	SVM / LSTM	2 / 3	EN/DE EN/DE/FR	SA
Encuesta Docente Seriedad	+3%	BERT	1	EN	SA
Violencia de Género	+10%	CNN	3	EN/DE/FR	SA

6.4 Técnicas de aumentación en general

Al analizar el impacto de la aumentación en el rendimiento de clasificación, se puede observar lo siguiente:

- Uno de los factores determinantes en el impacto que tienen los conjuntos de datos aumentados sobre el rendimiento de clasificación es que tan distintas son las muestras creadas por la aumentación. En todos los experimentos realizados, la diversidad fue factor determinante al momento de incrementar el rendimiento de los clasificadores por sobre la cantidad de muestras artificiales creadas con la aumentación.
- Otro de los factores que inciden en la calidad de las muestras creadas artificialmente con las técnicas de aumentación es el largo promedio de la oración. Tal como pudo observarse en la discusión de las técnicas por transformación y generación, cuando un conjunto de datos contiene oraciones con un promedio de palabras menor a 15 palabras se registra un impacto menor o negativo en el rendimiento de clasificación.
- La técnica de aumentación por parafraseo, *back-translation,* tiene un impacto positivo en la mayor parte de los conjuntos de datos al agregar una oración similar a la existente pero lo suficientemente distinta sintácticamente para que no sea la misma pero que conversa la semántica, evitando con esto que los clasificadores sean entrenados con conjuntos de datos con muestras muy similares lo cual lleva a overfitting.

Otro tema relevante al momento de analizar el impacto de las técnicas de aumentación en el rendimiento de clasificación son los desafíos que se enfrentan cuando se aplican estas técnicas. Dentro de los desafíos que se encontraron al aplicar las técnicas de aumentación se encuentran:

- Las técnicas de aumentación utilizadas y revisadas no están entrenadas en el idioma español. Por lo que fue necesario modificarlas de forma que pudiesen aumentar los conjuntos de datos que se describieron en el Capítulo 4. En el caso de EDA, dicha modificación fue el cambio del diccionario léxico utilizado por la técnica y el cambio de las *stopwords* en inglés por stopwords en español. Por otra parte, SentiGAN fue entrenado con distintos conjuntos de datos en inglés con una mayor cantidad de muestras y con un promedio de palabras por oración superior al

de los conjuntos de datos utilizados en este trabajo, por ende, fue necesario reentrenar el modelo con los conjuntos de datos en español.

- Otro de los desafíos encontrados con SentiGAN fue que la documentación de la técnica que acompaña al código en el repositorio de Github[16] es escaza y no se condice con lo expresado en el trabajo de Wang et al. [58] que indica que SentiGAN se encuentra preparado para la generación de oraciones para múltiples clases. En la práctica, el código provisto en el repositorio se encuentra preparado para generar muestras para 1 sola clase. Esto repercutió negativamente en el tiempo necesario para la ejecución de los experimentos llevados a cabo en este trabajo.
- Un desafío adicional fue que las técnicas de aumentación generativas analizadas en el estado del arte se encuentran escritas en versiones de Python y dependencias que ya no se encuentran vigentes. Por ejemplo, SentiGAN fue escrita en Python 2.7 y utiliza TensorFlow 1.4. Estos requerimientos hicieron necesario crear ambientes virtuales que pudiesen ejecutar la técnica y que tuviesen el poder de cómputo suficiente para su ejecución.
- El conjunto de datos Emoji fue especialmente desafiante, la razón se encuentra en la versión del conjunto de datos utilizada que no incluye la información semántica de los emojis, esto provocó que el conjunto de datos perdiera información importante para determinar las emociones subyacentes llevando a malos resultados de clasificación en general por lo que no serán considerados en las recomendaciones entregadas en el Capítulo 7.

[16] https://github.com/Nrgeup/SentiGAN

Capítulo 7 Guía de selección

A partir de los resultados de los experimentos realizados sobre los conjuntos de textos en español descritos en el Capítulo 4 se generó las siguientes guías de selección de técnicas de aumentación en la forma de un árbol de decisión desde el punto de vista de la tarea de clasificación que aborda el conjunto de datos. Cabe destacar que este árbol se presenta a modo descriptivo y no predictivo dado el volumen se datos e instancias por cada clase del conjunto de datos de resultados.

El primer árbol de decisión (Fig. 112) representa la selección de la técnica de aumentación en función del impacto que tiene sobre el rendimiento de clasificación para la tarea de clasificación EA.

Por otra parte, el segundo árbol de decisión (Fig. 113) presenta los mejores caminos en la selección de una técnica de aumentación de acuerdo con el impacto en el rendimiento respecto a la línea base para los conjuntos de datos para la tarea de clasificación SA.

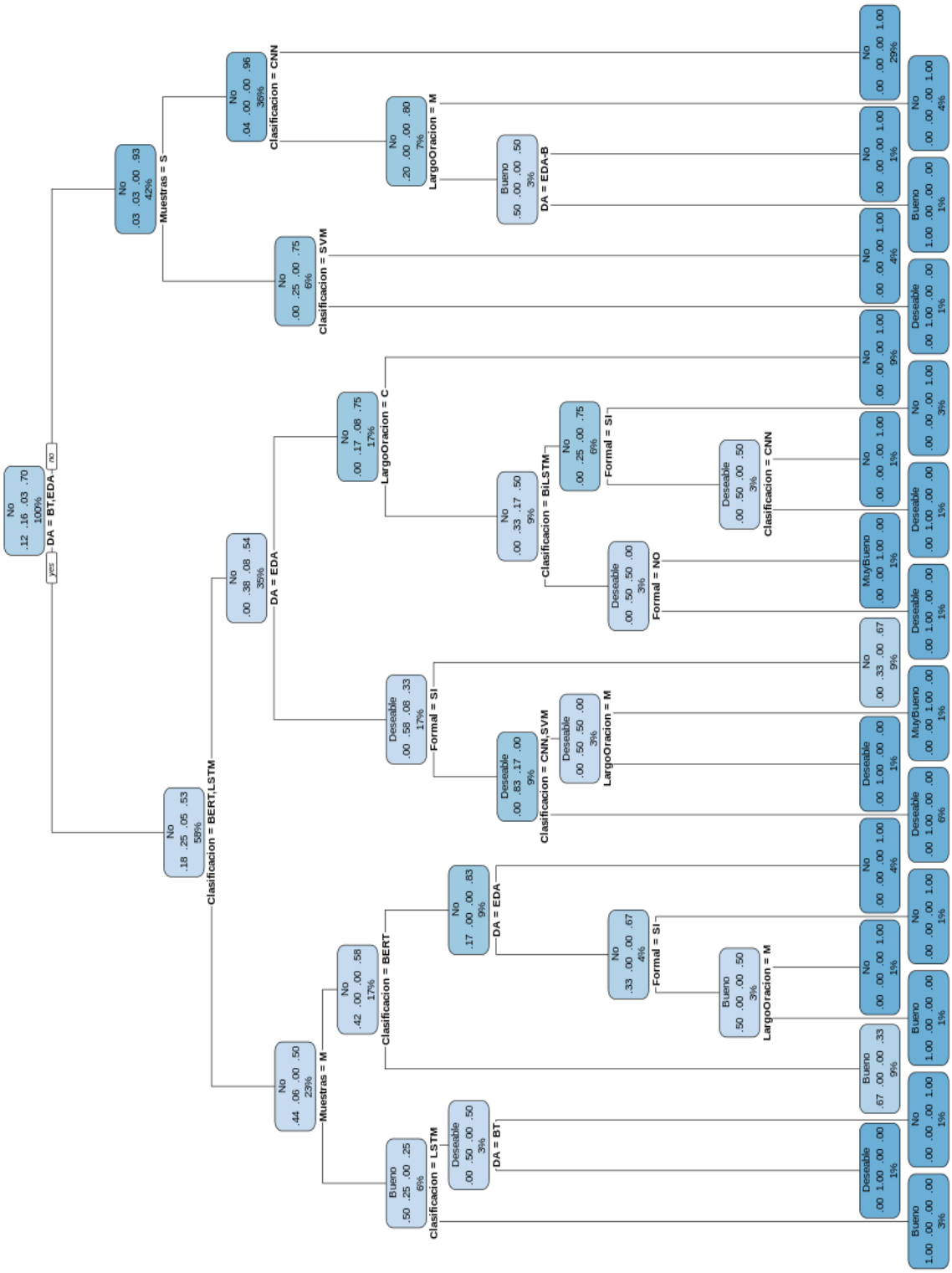

Fig. 112 Guía de selección para conjuntos EA

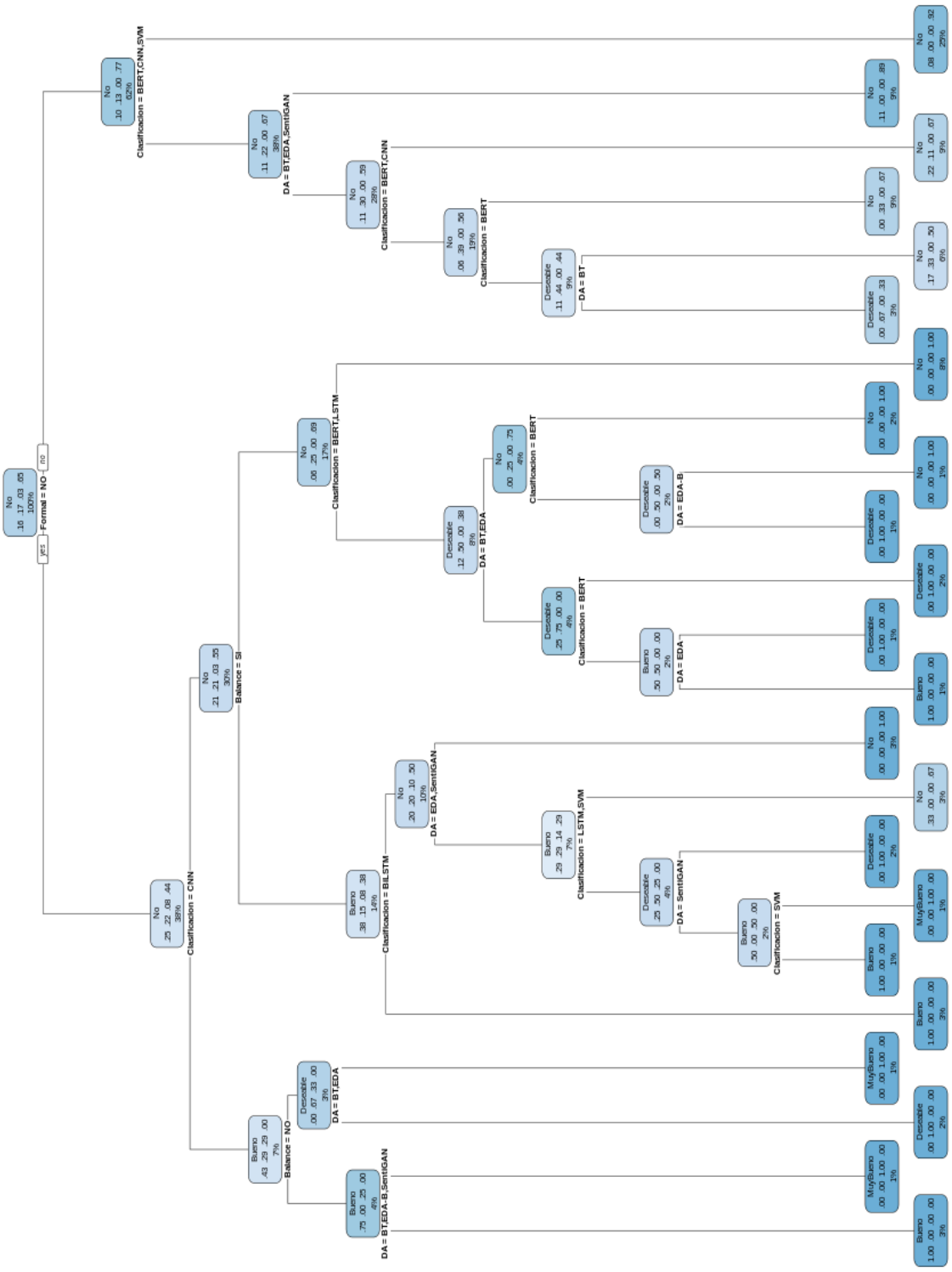

Fig. 113 Guía de selección para conjuntos SA

Cuando se realiza un análisis de Afectos, en los casos en que la aumentación reporta buenos resultados se obtienen reglas como las siguientes:

Tabla 57 Reglas de selección conjuntos EA (resultados positivos)

Bueno WHEN	DA IS	BT or EDA	&	Clasificacion IS	BERT	&	Muestras IS	L or S			
Bueno WHEN	DA IS	BT or EDA	&	Clasificacion IS	LSTM	&	Muestras IS	M			
Bueno WHEN	DA IS	EDA	&	Clasificacion IS	LSTM	&	Muestras IS	L	&	LargoOracion IS	M
Bueno WHEN	DA IS	EDA-B	&	Clasificacion IS	CNN	&	Muestras IS	L or M	&	LargoOracion IS	M
Deseable WHEN	DA IS	BT	&	Clasificacion IS	BERT	&	Muestras IS	M			
Deseable WHEN	DA IS	EDA	&	Clasificacion IS	CNN or SVM	&	Muestras IS	L			
Deseable WHEN	DA IS	EDA	&	Clasificacion IS	BiLSTM	&	Muestras IS	L	&	LargoOracion IS	M
Deseable WHEN	DA IS	BT	&	Clasificacion IS	BiLSTM	&	Muestras IS	M	&	LargoOracion IS	C
Deseable WHEN	DA IS	BT	&	Clasificacion IS	CNN	&	Muestras IS	L	&	LargoOracion IS	C
Deseable WHEN	DA IS	EDA-B or SentiGAN	&	Clasificacion IS	SVM	&	Muestras IS	S			
MuyBueno WHEN	DA IS	EDA	&	Clasificacion IS	BiLSTM	&	Muestras IS	L	&	LargoOracion IS	C
MuyBueno WHEN	DA IS	BT	&	Clasificacion IS	BiLSTM	&	Muestras IS	L	&	LargoOracion IS	C

Cuando se realiza un análisis de Afectos, en los casos en que la aumentación no reporta buenos resultados se obtienen reglas como las siguientes:

Tabla 58 Reglas de selección conjuntos EA (resultados negativos)

No WHEN	DA IS	EDA	&	Clasificacion IS	BiLSTM or CNN or SVM	&	Muestras IS	M or S			
No WHEN	DA IS	EDA	&	Clasificacion IS	BERT	&	Muestras IS	M			
No WHEN	DA IS	EDA	&	Clasificacion IS	LSTM	&	Muestras IS	L	&	LargoOracion IS	C
No WHEN	DA IS	EDA	&	Clasificacion IS	LSTM	&	Muestras IS	S			
No WHEN	DA IS	BT	&	Clasificacion IS	LSTM	&	Muestras IS	L or S			
No WHEN	DA IS	BT	&	Clasificacion IS	SVM	&	Muestras IS	L	&	LargoOracion IS	C
No WHEN	DA IS	BT	&	Clasificacion IS	CNN or SVM	&	Muestras IS	M	&	LargoOracion IS	C
No WHEN	DA IS	BT	&	Clasificacion IS	BiLSTM or CNN or SVM	&	LargoOracion IS	M			
No WHEN	DA IS	EDA-B or SentiGAN	&	Clasificacion IS	BiLSTM or CNN or LSTM	&	Muestras IS	S			
No WHEN	DA IS	SentiGAN	&	Clasificacion IS	CNN	&	Muestras IS	L or M	&	LargoOracion IS	M
No WHEN	DA IS	EDA-B or SentiGAN	&	Clasificacion IS	CNN	&	Muestras IS	L or M	&	LargoOracion IS	C
No WHEN	DA IS	EDA-B or SentiGAN	&	Clasificacion IS	BERT or BiLSTM or LSTM or SVM	&	Muestras IS	L or M			

Cuando se realiza un análisis de sentimientos, en los casos en que la aumentación reporta buenos resultados se obtienen reglas como las siguientes:

Tabla 59 Reglas de selección conjuntos SA (resultados positivos)

Bueno when	LargoOracion IS	L	&	Clasificacion IS	CNN	&	DA IS	BT or EDA-B or SentiGAN
Bueno when	LargoOracion IS	M	&	Clasificacion IS	BiLSTM			
Bueno when	LargoOracion IS	M	&	Clasificacion IS	SVM	&	DA IS	SentiGAN
Bueno when	LargoOracion IS	L	&	Clasificacion IS	BERT	&	DA IS	EDA
Deseable when	LargoOracion IS	C	&	Clasificacion IS	BERT	&	DA IS	BT
Deseable when	LargoOracion IS	M	&	Clasificacion IS	CNN	&	DA IS	BT or EDA
Deseable when	LargoOracion IS	M	&	Clasificacion IS	LSTM or SVM	&	DA IS	EDA
Deseable when	LargoOracion IS	L	&	Clasificacion IS	BERT	&	DA IS	BT
Deseable when	LargoOracion IS	L	&	Clasificacion IS	LSTM	&	DA IS	BT or EDA
Deseable when	LargoOracion IS	L	&	Clasificacion IS	BERT	&	DA IS	EDA-B
MuyBueno when	LargoOracion IS	L	&	Clasificacion IS	CNN	&	DA IS	EDA
MuyBueno when	LargoOracion IS	M	&	Clasificacion IS	CNN	&	DA IS	SentiGAN
MuyBueno when	LargoOracion IS	M	&	Clasificacion IS	LSTM	&	DA IS	SentiGAN

Cuando se realiza un análisis de sentimientos, en los casos en que la aumentación no reporta buenos resultados se obtienen reglas como las siguientes:

Tabla 60 Reglas de selección conjuntos SA (resultados negativos)

No when	LargoOracion IS	C	&	Clasificacion IS	BERT	&	DA IS	EDA or SentiGAN
No when	LargoOracion IS	M	&	Clasificacion IS	BERT	&	DA IS	EDA or SentiGAN
No when	LargoOracion IS	C	&	Clasificacion IS	CNN	&	DA IS	BT or EDA or SentiGAN
No when	LargoOracion IS	C	&	Clasificacion IS	SVM	&	DA IS	BT or EDA or SentiGAN
No when	LargoOracion IS	C	&	Clasificacion IS	BERT or CNN or SVM	&	DA IS	EDA-B
No when	LargoOracion IS	C	&	Clasificacion IS	BiLSTM or LSTM			
No when	LargoOracion IS	M	&	Clasificacion IS	BERT or LSTM or SVM	&	DA IS	BT
No when	LargoOracion IS	L	&	Clasificacion IS	BERT	&	DA IS	SentiGAN
No when	LargoOracion IS	L	&	Clasificacion IS	LSTM	&	DA IS	EDA-B or SentiGAN
No when	LargoOracion IS	L	&	Clasificacion IS	BiLSTM or SVM			

Las reglas presentadas en la guía de selección del Capítulo 7, entregan una idea de la importancia de contar con conocimiento que permita la creación de una guía que oriente acerca de la aplicación de técnicas de aumentación de datos. Estas reglas, obtenidas a partir del análisis de los resultados obtenidos en este trabajo, están clasificadas en 4 niveles de acuerdo con el grado de mejora en los resultados obtenidos a partir de:

- Las características del conjunto de datos (corpus) utilizado: tamaño del conjunto, largo promedio de las oraciones, formalidad del lenguaje.
- El tipo de técnica de aumentación a utilizar.
- El algoritmo de clasificación a aplicar.
- El enfoque del análisis, es decir, análisis de sentimientos (clasificación en dos clases) o análisis de emociones (clasificación de más de dos clases).

Por ejemplo, la regla presentada como ***"Bueno WHEN DA IS BT OR EDA & Clasificacion IS LSTM & Muestras IS M"*** describe que para un corpus de tamaño Mediano (M), sin importar la formalidad del texto o el largo promedio de las oraciones, si se utiliza el algoritmo de clasificación LSTM, se sugiere la utilización de aumentación de datos mediante las técnicas EDA o BT para obtener un aumento de entre 3% y 5% en el rendimiento de clasificación en comparación a no utilizar aumentación de datos. Lo anterior, para el caso de realizar Análisis de Emociones (EA).

También existen reglas que sugieren la no utilización de técnicas de aumentación de datos de acuerdo con ciertas condiciones. Por ejemplo, la regla de la Tabla 60 expresada como ***"No WHEN LargoOracion IS L & Clasificacion IS BiLSTM OR SVM"*** indica que cuando se tiene un conjunto de datos con un largo promedio de oraciones grande (L) y se desea utilizar como algoritmo de clasificación BiLSTM o SVM, no se sugiere la aplicación de técnicas de aumentación, puesto que su rendimiento se mantiene o empeora al comparar con el caso de no utilizar aumentación de datos. Todo lo anterior, para el caso de realizar Análisis de Sentimientos.

Capítulo 8 Conclusiones y trabajo futuro

8.1 Conclusiones

Este trabajo entrega un aporte al procesamiento del lenguaje natural por medio de un *framework* que guie la selección de una técnica de aumentación que permita incrementar artificialmente conjuntos de datos en español tomando como punto de partida la tarea de clasificación.

Para lograr este objetivo, fue necesario analizar el estado del arte de la aumentación de textos mediante una extensa revisión de literatura cuyo resultado permitió, en primer lugar, entender que la aumentación de textos consiste en una serie de métodos utilizados para incrementar artificialmente un conjunto de datos etiquetados. En segundo lugar, se obtuvo una clasificación de las técnicas de aumentación de textos de acuerdo con el tipo de manipulación que se realiza sobre el conjunto de datos mediante la taxonomía de Abonizio et al. [30] que condujo a la selección de las técnicas más representativas de la aumentación de textos.

Luego de la experimentación realizada con las técnicas de transformación, generación y parafraseo, puede concluirse que:

1. Dependiendo de la técnica de aumentación y el clasificador utilizado, es posible impactar positivamente el rendimiento de clasificación.
2. Cuando se trata de conjuntos de datos que tienen por objetivo el análisis de emociones (EA), una de las mejores alternativas es el uso de la técnica de parafraseo *back-translation* independiente de las características del conjunto de datos.
3. El parámetro más relevante a la hora de realizar aumentación de los conjuntos de datos con EDA fue el porcentaje de modificación de palabras en una oración, este parámetro se encarga de agregar la diversidad necesaria al conjunto de datos para que las oraciones creadas sean lo suficientemente distintas del conjunto de datos original. Sin embargo, vale la pena destacar que EDA no se preocupa de mantener la semántica de una oración, por lo que podría ser interesante explorar otras técnicas de aumentación basadas en EDA que se preocupen de mantener la semántica.
4. Al balancear los conjuntos de datos con EDA, el rendimiento de los clasificadores decae considerablemente, este comportamiento puede darse debido a la calidad de las oraciones y el bajo rendimiento de línea base de las oraciones originales

puede ser propagado al aumentar dichas oraciones profundizando el mal rendimiento de las clases en los modelos de DL y ML.

5. En cuanto a las técnicas generativas, los resultados obtenidos indican que su utilización en los conjuntos de datos estudiados produce buenos rendimientos de clasificación cuando el conjunto de datos corresponde a la tarea SA, contiene más de 1.000 muestras (tamaño L), las clases se encuentran balanceadas y se utiliza el clasificador CNN.

Las guías de selección propuestas, al ser construidas con base en los resultados obtenidos en este trabajo, se consideran un punto de partida en la selección de técnicas de aumentación de textos en español que puede seguir mejorando en la medida que se realicen mayores experimentos que permitan hacerlas extensibles a un mayor número de características presentes en los conjuntos de datos en español.

Para finalizar, se puede concluir con base en los experimentos realizados en este trabajo que la utilización de técnicas de aumentación para los conjuntos de textos en español tiene un impacto positivo en el rendimiento de clasificación con diversos modelos de DL y ML, a pesar de que los conjuntos estudiados pueden ser considerados pequeños en comparación con los conjuntos utilizados en inglés por la mayor parte de los trabajos presentes en la literatura revisada. Tanto el código como los resultados obtenidos se encuentran disponibles en el repositorio de Gitlab[17] https://gitlab.com/rgutierrezb/dataaugmentation.

[17] https://about.gitlab.com

8.2 Trabajo futuro

Dados los resultados obtenidos en el balanceo de clases con EDA, en el futuro es deseable explorar los efectos del balanceo mediante aumentación de los conjuntos de textos que aborden la tarea de análisis de emociones en español (EA) con el fin de mejorar los resultados obtenidos en este trabajo.

Las técnicas de aumentación utilizadas en este trabajo corresponden a las más representativas de las categorías transformación, generación y parafraseo de la taxonomía de Abonizio et al. [30], por lo que en futuros trabajos puede considerarse la experimentación con las técnicas restantes de esta taxonomía, especialmente los modelos de lenguaje[18] como GPT, que hoy por hoy han cobrado gran notoriedad gracias a la implementación de herramientas como ChatGPT[19].

En un sentido más práctico, en futuros trabajos se contempla la implementación de una aplicación web basada en el *framework* propuesto en este trabajo con el fin de entregar más herramientas para el procesamiento del lenguaje natural en español.

[18] https://aws.amazon.com/es/what-is/large-language-model/

[19] https://chat.openai.com/

Referencias

1. Tahayna, B.M.A., Ayyasamy, R.K., Akbar, R.: Automatic Sentiment Annotation of Idiomatic Expressions for Sentiment Analysis Task. IEEE Access. 10, 122234–122242 (2022). https://doi.org/10.1109/ACCESS.2022.3222233.
2. Iosifidis, V., Ntoutsi, E.: Sentiment analysis on big sparse data streams with limited labels. Knowl Inf Syst. 62, 1393–1432 (2020). https://doi.org/10.1007/S10115-019-01392-9/TABLES/17.
3. Chen, J., Luo, L., Ji, B., Zhao, S., Zhang, Y.: A Joint Learning Sentiment Analysis Method Incorporating Emoji-Augmentation. In: 2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS). pp. 348–354. IEEE (2022). https://doi.org/10.1109/CCIS57298.2022.10016405.
4. Sun, X., He, J.: A novel approach to generate a large scale of supervised data for short text sentiment analysis. Multimed Tools Appl. 79, 5439–5459 (2020). https://doi.org/10.1007/S11042-018-5748-4/FIGURES/3.
5. Pei, Y., Chen, S., Ke, Z., Silamu, W., Guo, Q.: AB-LaBSE: Uyghur Sentiment Analysis via the Pre-Training Model with BiLSTM. Applied Sciences 2022, Vol. 12, Page 1182. 12, 1182 (2022). https://doi.org/10.3390/APP12031182.
6. Abdul Qudar, M.M., Bhatia, P., Mago, V.: ONSET: Opinion and Aspect Extraction System from Unlabelled Data. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 733–738. IEEE (2021). https://doi.org/10.1109/SMC52423.2021.9658689.
7. Haralabopoulos, G., Torres, M.T., Anagnostopoulos, I., McAuley, D.: Text data augmentations: Permutation, antonyms and negation. Expert Syst Appl. 177, 114769 (2021). https://doi.org/10.1016/J.ESWA.2021.114769.
8. GB, S., Jacob, I.J.: A Semantic Approach for Computing Speech Emotion Text Classification Using Machine Learning Algorithms. In: 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). pp. 1–5. IEEE (2022). https://doi.org/10.1109/ICEEICT53079.2022.9768465.
9. Li, G., Wang, H., Ding, Y., Zhou, K., Yan, X.: Data augmentation for aspect-based sentiment analysis. International Journal of Machine Learning and Cybernetics. 14, 125–133 (2023). https://doi.org/10.1007/S13042-022-01535-5/TABLES/3.
10. Biolchini, J., Gomes Mian, P., Candida Cruz Natali, A., Horta Travassos, G.: Systematic Review in Software Engineering. (2005).
11. Tang, T., Tang, X., Yuan, T.: Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text. IEEE Access. 8, 193248–193256 (2020). https://doi.org/10.1109/ACCESS.2020.3030468.
12. Kumar, S., Khan, M.B., Hasanat, M.H.A., Saudagar, A.K.J., AlTameem, A., AlKhathami, M.: Sigmoidal Particle Swarm Optimization for Twitter Sentiment Analysis. Computers, Materials & Continua. 74, 897–914 (2022). https://doi.org/10.32604/CMC.2023.031867.
13. Jurafsky, D., Martin, J.H.: Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Third Edition draft Summary of Contents.
14. Bogoradnikova, D., Makhnytkina, O., Matveev, A., Zakharova, A., Akulov, A.: Multilingual Sentiment Analysis and Toxicity Detection for Text Messages in Russian. In: 2021 29th Conference of Open Innovations Association (FRUCT). pp. 55–64. IEEE (2021). https://doi.org/10.23919/FRUCT52173.2021.9435584.

15. Balakrishnan, V., Shi, Z., Law, C.L., Lim, R., Teh, L.L., Fan, Y.: A deep learning approach in predicting products' sentiment ratings: a comparative analysis. Journal of Supercomputing. 78, 7206–7226 (2022). https://doi.org/10.1007/S11227-021-04169-6/TABLES/9.
16. Yuan, H., Song, Y., Hu, J., Ma, Y.: Design of Festival Sentiment Classifier Based on Social Network. Comput Intell Neurosci. 2020, (2020). https://doi.org/10.1155/2020/8824009.
17. Shehu, H.A., Sharif, M.H., Sharif, M.H.U., Datta, R., Tokat, S., Uyaver, S., Kusetogullari, H., Ramadan, R.A.: Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data. IEEE Access. 9, 56836–56854 (2021). https://doi.org/10.1109/ACCESS.2021.3071393.
18. Rafi-Ur-Rashid, Md., Mahbub, M., Adnan, M.A.: Breaking the Curse of Class Imbalance: Bangla Text Classification. ACM Transactions on Asian and Low-Resource Language Information Processing. 21, 1–21 (2022). https://doi.org/10.1145/3511601.
19. Liu, S., Lee, K., Lee, I.: Document-level multi-topic sentiment classification of Email data with BiLSTM and data augmentation. Knowl Based Syst. 197, 105918 (2020). https://doi.org/10.1016/j.knosys.2020.105918.
20. Wang, L., Xu, X., Liu, C., Chen, Z.: M-DA: A Multifeature Text Data-Augmentation Model for Improving Accuracy of Chinese Sentiment Analysis. Sci Program. 2022, (2022). https://doi.org/10.1155/2022/3264378.
21. Tan, K.L., Lee, C.P., Lim, K.M.: RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. Applied Sciences 2023, Vol. 13, Page 3915. 13, 3915 (2023). https://doi.org/10.3390/APP13063915.
22. Devlin, J., Chang, M.-W., Lee, K., Google, K.T., Language, A.I.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
23. Dhiman, A., Toshniwal, D.: AI-based Twitter framework for assessing the involvement of government schemes in electoral campaigns. Expert Syst Appl. 203, 117338 (2022). https://doi.org/10.1016/j.eswa.2022.117338.
24. Srinivasarao, U., Sharaff, A.: Machine intelligence based hybrid classifier for spam detection and sentiment analysis of SMS messages. Multimed Tools Appl. 1–31 (2023). https://doi.org/10.1007/S11042-023-14641-5/TABLES/7.
25. Kelsingazin, Y., Akhmetov, I., Pak, A.: Sentiment Analysis of Kaspi Product Reviews. In: 2021 16th International Conference on Electronics Computer and Computation (ICECCO). pp. 1–5. IEEE (2021). https://doi.org/10.1109/ICECCO53203.2021.9663854.
26. Duwairi, R., Abushaqra, F.: Syntactic- and morphology-based text augmentation framework for Arabic sentiment analysis. PeerJ Comput Sci. 7, 1–25 (2021). https://doi.org/10.7717/PEERJ-CS.469/SUPP-4.
27. Santoso, N., Mendonça, I., Aritsugi, M.: Text Augmentation Based on Integrated Gradients Attribute Score for Aspect-based Sentiment Analysis. In: 2023 IEEE International Conference on Big Data and Smart Computing (BigComp). pp. 227–234. IEEE (2023). https://doi.org/10.1109/BigComp57234.2023.00044.
28. Wang, Q.: Learning From Other Labels: Leveraging Enhanced Mixup and Transfer Learning for Twitter Sentiment Analysis. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 336–343. IEEE (2021). https://doi.org/10.1109/ICTAI52525.2021.00055.
29. Bayer, M., Kaufhold, M.A., Reuter, C.: A Survey on Data Augmentation for Text Classification. ACM Comput Surv. 55, (2022). https://doi.org/10.1145/3544558.
30. Abonizio, H.Q., Paraiso, E.C., Barbon, S.: Toward Text Data Augmentation for Sentiment Analysis. IEEE Transactions on Artificial Intelligence. 3, 657–668 (2022). https://doi.org/10.1109/TAI.2021.3114390.

31. Wei, J., Zou, K.: EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.
32. Krishnan, J., Anastasopoulos, A., Purohit, H., Rangwala, H.: Cross-Lingual Text Classification of Transliterated Hindi and Malayalam. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 1850–1857. IEEE (2022). https://doi.org/10.1109/BigData55660.2022.10021079.
33. Shang, Y., Su, X., Xiao, Z., Chen, Z.: Campus Sentiment Analysis with GAN-based Data Augmentation. In: 2021 13th International Conference on Advanced Infocomm Technology (ICAIT). pp. 209–214. IEEE (2021). https://doi.org/10.1109/ICAIT52638.2021.9702068.
34. Gupta, R.: Data augmentation for low resource sentiment analysis using generative adversarial networks. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019-May, 7380–7384 (2019).
35. Carrasco, X.A., Elnagar, A., Lataifeh, M.: A Generative Adversarial Network for Data Augmentation: The Case of Arabic Regional Dialects. Procedia CIRP. 189, 92–99 (2021). https://doi.org/10.1016/J.PROCS.2021.05.072.
36. Luo, J., Bouazizi, M., Ohtsuki, T.: Data Augmentation for Sentiment Analysis Using Sentence Compression-Based SeqGAN with Data Screening. IEEE Access. 9, 99922–99931 (2021). https://doi.org/10.1109/ACCESS.2021.3094023.
37. Sun, T., Jing, L., Wei, Y., Song, X., Cheng, Z., Nie, L.: Dual Consistency-enhanced Semi-supervised Sentiment Analysis towards COVID-19 Tweets. IEEE Trans Knowl Data Eng. 1–13 (2023). https://doi.org/10.1109/TKDE.2023.3270940.
38. Kodiyala, V.S., Mercer, R.E.: Emotion Recognition and Sentiment Classification using BERT with Data Augmentation and Emotion Lexicon Enrichment. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 191–198. IEEE (2021). https://doi.org/10.1109/ICMLA52953.2021.00037.
39. Almasre, M.A.: Enhance the Aspect Category Detection in Arabic Language using AraBERT and Text Augmentation. In: 2022 Fifth National Conference of Saudi Computers Colleges (NCCC). pp. 1–4. IEEE (2022). https://doi.org/10.1109/NCCC57165.2022.10067648.
40. Ha, S., Grubert, E.: Hybridizing qualitative coding with natural language processing and deep learning to assess public comments: A case study of the clean power plan. Energy Res Soc Sci. 98, 2214–6296 (2023). https://doi.org/10.1016/j.erss.2023.103016.
41. Lee, J., Kim, J.: Improving Generation of Sentiment Commonsense by Bias Mitigation. In: 2023 IEEE International Conference on Big Data and Smart Computing (BigComp). pp. 308–311. IEEE (2023). https://doi.org/10.1109/BigComp57234.2023.00061.
42. Tahayna, B., Ayyasamy, R.K., Akbar, R., Subri, N.F.B., Sangodiah, A.: Lexicon-based Non-Compositional Multiword Augmentation Enriching Tweet Sentiment Analysis. In: 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS). pp. 19–24. IEEE (2022). https://doi.org/10.1109/AiDAS56890.2022.9918749.
43. Wang, X., Sheng, Y., Deng, H., Zhao, Z.: Information and Control ICIC International c ◯2019 ISSN. International Journal of Innovative Computing. 15, 227–246 (2019). https://doi.org/10.24507/ijicic.15.01.227.
44. Hu, L., Li, C., Wang, W., Pang, B., Shang, Y.: Performance Evaluation of Text Augmentation Methods with BERT on Small-sized, Imbalanced Datasets. In: 2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI). pp. 125–133. IEEE (2022). https://doi.org/10.1109/CogMI56440.2022.00027.
45. Tan, K.L., Lee, C.P., Anbananthen, K.S.M., Lim, K.M.: RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. IEEE Access. 10, 21517–21525 (2022). https://doi.org/10.1109/ACCESS.2022.3152828.

46. Al-Jamal, W.Q., Mustafa, A.M., Ali, M.Z.: Sarcasm Detection in Arabic Short Text using Deep Learning. In: 2022 13th International Conference on Information and Communication Systems (ICICS). pp. 362–366. IEEE (2022). https://doi.org/10.1109/ICICS55353.2022.9811153.
47. Kraus, M., Feuerriegel, S.: Sentiment analysis based on rhetorical structure theory:Learning deep neural networks from discourse trees. Expert Syst Appl. 118, 65–79 (2019). https://doi.org/10.1016/j.eswa.2018.10.002.
48. Feng, Z., Zhou, H., Zhu, Z., Mao, K.: Tailored text augmentation for sentiment analysis. Expert Syst Appl. 205, 117605 (2022). https://doi.org/10.1016/j.eswa.2022.117605.
49. Wei, S., Yu, D., Lv, C.: Text Editing for Augmented Distilled BERT. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 437–442. IEEE (2020). https://doi.org/10.1109/ICTAI50040.2020.00075.
50. Omran, T.M., Sharef, B.T., Grosan, C., Li, Y.: Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. Data Knowl Eng. 143, 102106 (2023). https://doi.org/10.1016/j.datak.2022.102106.
51. Body, T., Tao, X., Li, Y., Li, L., Zhong, N.: Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models. Expert Syst Appl. 178, 115033 (2021). https://doi.org/10.1016/j.eswa.2021.115033.
52. Liu, X., Zhong, Y., Wang, J., Li, P.: Data augmentation using Heuristic Masked Language Modeling. International Journal of Machine Learning and Cybernetics. 1–15 (2023). https://doi.org/10.1007/S13042-023-01784-Y/FIGURES/7.
53. Xu, N., Mao, W., Wei, P., Zeng, D.: MDA: Multimodal Data Augmentation Framework for Boosting Performance on Sentiment/Emotion Classification Tasks. IEEE Intell Syst. 36, 3–12 (2021). https://doi.org/10.1109/MIS.2020.3026715.
54. Pandey, S., Akhtar, Md.S., Chakraborty, T.: Syntactically Coherent Text Augmentation for Sequence Classification. IEEE Trans Comput Soc Syst. 8, 1323–1332 (2021). https://doi.org/10.1109/TCSS.2021.3075774.
55. Jiang, Q., Chen, L., Zhao, W., Yang, M.: Toward Aspect-Level Sentiment Modification Without Parallel Data. IEEE Intell Syst. 36, 75–81 (2021). https://doi.org/10.1109/MIS.2021.3052617.
56. Shyang, Y.K., Yan, J.L.S.: A Text Augmentation Approach using Similarity Measures based on Neural Sentence Embeddings for Emotion Classification on Microblogs. In: 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). pp. 1–6. IEEE (2020). https://doi.org/10.1109/IICAIET49801.2020.9257826.
57. Duong, H.-T., Nguyen-Thi, T.-A., Hoang, V.T.: Vietnamese Sentiment Analysis under Limited Training Data Based on Deep Neural Networks. (2022). https://doi.org/10.1155/2022/3188449.
58. Wang, K., Wan, X.: SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. (2018).
59. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. (2016).
60. Lepe-Faúndez, M., Segura-Navarrete, A., Vidal-Castro, C., Martínez-Araneda, C., Rubio-Manzano, C.: Detecting aggressiveness in tweets: A hybrid model for detecting cyberbullying in the Spanish language. Applied Sciences (Switzerland). 11, (2021). https://doi.org/10.3390/app112210706.
61. Martinez-Araneda, C., Segura, A., Vidal-Castro, C., Elgueta, J.: Is news really pessimistic? Sentiment Analysis ofChilean online newspaper headlines. Indian J Sci Technol. 11, 1–8 (2018). https://doi.org/10.17485/ijst/2018/v11i22/102251.

62. Martinez-Araneda, C., Calbullanca Viluñir, R., Segura, A., Vidal-Castro, C., Gomez-Meneses, P.: IMPROVING OF AUTOMATIC DETECTION OF GENDER VIOLENCE IN SPANISH SONGS LYRICS BY AUGMENTATION DATA AND UNDERSAMPLING.
63. WordNet, https://mitpress.mit.edu/9780262561167/, last accessed 2023/12/20.