risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 09/08/2019
Aceitação/Acceptance: 27/09/2019

Heurísticas para Data Augmentation en NLP: Aplicación a Revisiones de Artículos Científicos

Rubén Sánchez Acosta, Claudio Meneses Villegas, Brian Keith Norambuena

ruben.sanchez01@ucn.cl, cmeneses@ucn.cl, brian.keith@ucn.cl

Universidad Católica del Norte, Av. Angamos 0610, 1240000, Antofagasta, Chile.

DOI: 10.17013/risti.34.44–53

Resumen: Las técnicas de *data augmentation* son esenciales para entrenar algoritmos de *machine learning*, donde el conjunto de datos inicial es más pequeño que lo requerido debido a la complejidad del modelo. En modelos de aprendizaje automático, la robustez del proceso de entrenamiento depende altamente de grandes volúmenes de datos etiquetados, los cuales son costosos de producir. Un enfoque eficaz para tratar con este problema es generar automáticamente nuevos ejemplos etiquetados usando técnicas de *data augmentation*. En el procesamiento del lenguaje natural, en particular en el idioma español, hay una falta de técnicas bien definidas que permitan incrementar un conjunto de datos. En este artículo, se proponen un conjunto de heurísticas para *data augmentation* en NLP, las cuales son aplicadas en el dominio de las revisiones de artículos científicos.

Palabras-clave: Data Augmentation; NLP; Revisiones de Artículos.

Heuristics for Data Augmentation in NLP: Application to scientific paper reviews

Abstract: Data augmentation techniques are essential for training machine learning algorithms, where the initial data set is smaller than required due to the model complexity. In machine learning models, the robustness of the training process is highly dependent on large volumes of labeled data, which are expensive to produce. An effective approach to deal with this problem is to automatically generate new tagged examples using data augmentation techniques. In the processing of natural language, particularly in the Spanish language, there is a lack of well-defined techniques that allow increasing a set of data. In this article, we propose a set of heuristics for data augmentation in NLP, which are applied to the domain of reviews of scientific articles.

Keywords: Data Augmentation; NLP; Paper Reviews.

1. Introducción

En el ámbito de la minería de datos y el aprendizaje automático, el rápido aumento en la cantidad de datos disponibles ha facilitado el desarrollo de técnicas tales como *deep learning*,

que dada la complejidad de sus modelos requieren una enorme cantidad de datos para obtener un buen rendimiento. No obstante, no todos los dominios han visto este aumento en la disponibilidad de datos, en muchos casos debido al alto costo, ya sea económico o técnico, que puede significar la obtención de muestras rotuladas con un valor del atributo clase. Considerando estos antecedentes, varios de estos dominios se verían beneficiados de técnicas que permitan obtener datos adicionales con un menor costo asociado, en este contexto es que se utilizan las técnicas de *data augmentation*. En particular, este trabajo se centra en la aplicación de técnicas de *data augmentation* en el contexto de análisis de sentimientos aplicado al dominio de revisiones de artículos científicos, las cuales en su mayor parte corresponden a texto plano rotulados con la evaluación del revisor.

El problema base que se desea mejorar consiste en predecir si un artículo debería ser aceptado o rechazado en función de los comentarios aportados por los revisores. Otro aspecto importante en este dominio es poder cuantificar la consistencia entre los comentarios de un revisor con respecto a su evaluación, usualmente expresada en una escala multiclase. En este trabajo se aborda la primera parte de la investigación, que consiste en ampliar el conjunto de datos aplicando técnicas de *data augmentation*, en una segunda etapa de la investigación se utilizará el enfoque de transferencia de conocimientos, para entrenar un modelo basado en *deep learning* utilizando el conjunto de datos generados en el presente artículo.

Para utilizar técnicas de *data augmentation* de manera efectiva se debe hallar una función que tome un registro como entrada y lo convierta en otro que sea válido para el dominio establecido, manteniendo la clase con que se etiquetó el dato inicial (van Dyk & Meng, 2001). En el dominio específico estudiado en el presente trabajo, se deben encontrar técnicas para convertir una revisión de artículo científico en otra que, a pesar de ser generada artificialmente, su texto tenga sentido como revisión de artículo y mantenga el sentimiento positivo o negativo, en concordancia con el texto original. Para evaluar la efectividad de las técnicas utilizadas se asumirá el siguiente supuesto de concordancia semántica: "Si los cambios realizados al conjunto de datos fueran semánticamente incorrectos, entonces el rendimiento del clasificador disminuirá, en cambio, es lógico suponer que si el rendimiento aumenta entonces no hubo un cambio semántico significativo".

Las técnicas de *data augmentation* usualmente son definidas en base al dominio donde se utilizarán (Kobayashi, 2018). Algunas de las estrategias abordadas en el presente trabajo, a pesar de poder ser aplicadas a otros dominios, fueron definidas en función de las características del dominio de revisiones de artículos científicos y de la tarea de aprendizaje, es decir el análisis de sentimientos.

Las técnicas de *data augmentation* son ampliamente estudiadas y utilizadas en el dominio de la Visión por Computador (McLaughlin, Martinez, & Miller, 2015), (Fawzi, Samolowitz, Turaga, & Frossard, 2016). En este ámbito es menos complejo su uso debido a que las imágenes son expresadas como vectores, donde se representan las componentes RGB en cada píxel con un número entre 0 y 255 por cada componente. Esta representación tiene la ventaja de que los vectores cubren un espacio continuo, por lo tanto, si se cambian algunos valores en las componentes de los píxeles, la imagen sigue siendo un registro válido para el dominio que cubre. Además, una imagen que se rota, se traslada y se transforman de diferentes maneras, sigue siendo un registro válido muchas veces.

En el dominio del procesamiento de lenguaje natural (NLP, por sus siglas en inglés) (Grosz, Jones, & Webber, 1986), la situación es más compleja, pues cambiar una palabra por otra o alterar caracteres de manera aleatoria puede variar el significado del texto o directamente hacer que pierda todo significado. Por esta razón, cualquier función que se aplique a los datos debe ser elegida de manera cuidadosa para no llegar a corromperlos. En el lenguaje español se dificulta aún más la labor, pues muchas palabras tienen género, número y persona, los cuales al ser alterados pueden originar errores de concordancia en el texto.

Debido a la dificultad que conlleva adquirir datos para el dominio estudiado, ya que por lo general son confidenciales, es necesario recurrir a técnicas de *data augmentation* para mitigar el sobreajuste al realizar tareas de aprendizaje automático como análisis de sentimientos (Keith, Fuentes, & Meneses, 2019). El conjunto de datos utilizado se generó a partir del presentado en (Keith, Fuentes, & Meneses, 2017), cuenta con un total de 300 revisiones, 3 por cada artículo, donde 150 son positivas y 150 son negativas, para evitar un desbalance entre las clases.

Mediante las técnicas de *data augmentation* propuestas el conjunto de datos podrá ser incrementado, de forma tal que puedan realizarse técnicas avanzadas de análisis de sentimientos, como es el caso de *deep learning*, modelando el dominio de mejor manera y evitando el sobreajuste a los datos iniciales. Usualmente el aumento de los datos incrementa la dimensión del problema pues aparecen palabras que anteriormente no eran parte del corpus, por lo que se debe verificar que este incremento no perjudica el aprendizaje sobre el conjunto de datos aumentados.

2. Materiales y Métodos

En esta sección se enumeran primeramente las técnicas de data augmentation utilizadas, luego se explica de qué manera se llevó a cabo la experimentación. El proceso completo llevado a cabo se muestra en la Figura 1.

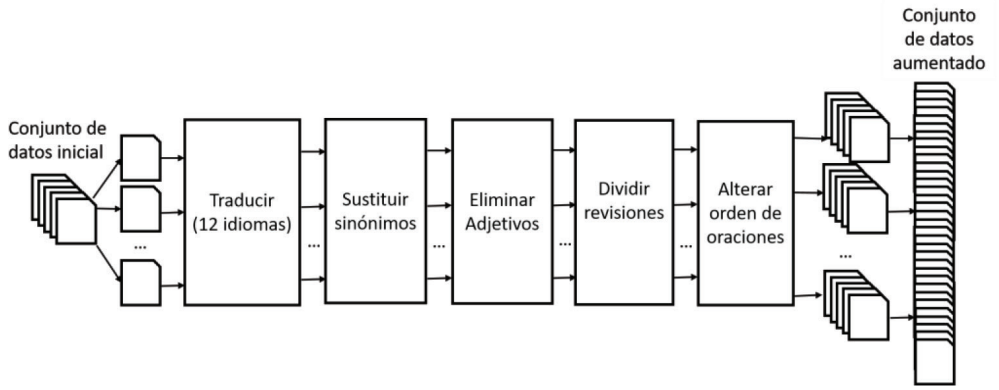

Figura 1 – Proceso de *data augmentation* realizado a los datos

2.1. Técnicas de data augmentation

Se utilizaron 5 heurísticas para aumentar los datos: 1) traducir el texto a idiomas extranjeros para luego regresarlo al español; 2) intercambiar palabras por sinónimos; 3) eliminar adjetivos; 4) alterar el orden de las oraciones; y 5) transformar revisiones extensas en varias revisiones pequeñas. Estas técnicas no son utilizadas de manera independiente, sino de forma sucesiva lo que hace que la cantidad de variaciones que se pueden aplicar al texto sea mucho mayor.

Debido a que cada etapa va aumentando la cantidad de datos, se decidió ordenar las heurísticas de más a menos complejas computacionalmente, de esta forma las técnicas más complejas trabajarán con menos datos, por lo que se logra disminuir el tiempo de generación de forma considerable. Utilizar las heurísticas en otro orden podría incrementar el costo computacional hasta el punto que dejaría de ser factible generar un conjunto de datos aumentado de manera automática. Con la presente investigación no se pretende encontrar el mejor orden para la aplicación de las técnicas, solamente se busca encontrar un orden que incremente los datos de manera consistente y que pueda mejorar el rendimiento del modelo. A continuación, se explica con más detalle cada una de estas heurísticas de manera individual.

Traducción a idiomas extranjeros

La traducción a otros idiomas es uno de los campos en que *deep learning* ha marcado la diferencia y de alguna manera se puede sacar partido a la eficacia de los algoritmos actuales. Google Translate es capaz de traducir textos a más de 120 idiomas diferentes, los cuales, al ser traducidos nuevamente al español, varían con respecto al texto inicial en muchos sentidos, desde palabras hasta estructuras gramaticales completas.

Se debe tomar en consideración que, si los lenguajes son muy similares gramaticalmente al idioma español, al ser traducido el texto de vuelta se obtiene nuevamente la frase inicial y si tienen grandes diferencias gramaticales con el castellano, el significado de la frase que se genera de vuelta puede no tener mucha relación con el texto inicial. Por esta razón, se debe encontrar un balance y elegir idiomas con los cuales se obtengan frases lo más distintas posible al texto inicial manteniendo su significado.

Se revisó manualmente la utilización de varios idiomas, fueron seleccionados aquellos que generaban las revisiones más diferentes a la original, pero que mantenían su significado. En el algoritmo propuesto se eligieron 12 idiomas con los que se obtuvieron resultados satisfactorios: afrikáans, búlgaro, inglés, francés, alemán, griego, japonés, portugués, eslovaco, sueco, turco y vietnamita.

Sustitución por sinónimos

Los sinónimos son palabras diferentes que comparten un mismo significado, por lo tanto, si alguna palabra es sustituida por su sinónimo, la frase mantendría exactamente el mismo significado y por consiguiente el sentimiento expresado. Esta técnica es utilizada en trabajos anteriores (Zhang, Zhao, & LeCun, 2015). Las partes de la oración que comúnmente tienen sinónimos asociados son los sustantivos, adjetivos y verbos, por esta razón en el presente artículo solamente se seleccionan este tipo de palabras para ser sustituidas por su sinónimo.

En español la sustitución por un sinónimo no es trivial, pues este debe coincidir en género, número y persona con la palabra inicial para mantener la concordancia de la frase. No podemos simplemente sustituir "casa" por "hogar" pues su diferencia de género crearía un error de concordancia en el texto. Para evitar este problema, una vez que se sustituye la palabra por su sinónimo el texto es traducido al inglés y luego regresado al español. Al realizar este procedimiento el texto regresa sin errores de concordancia y con el mismo significado.

En el algoritmo propuesto se les asigna una probabilidad a las partes de la oración y por cada palabra del texto, si es sustantivo, adjetivo o verbo, se sustituye, con una probabilidad igual a la asignada, por su sinónimo. Las probabilidades de sustitución establecidas por defecto son de 0.2 para cada una de las partes de la oración que intervienen (sustantivo, adjetivo y verbo). Luego de tener el texto final, se traduce al inglés y se regresa al español para eliminar errores de concordancia. Se utiliza la librería "Stanford POS Tagger" para identificar las partes de la oración y el API Google Translate de la plataforma Google Cloud para realizar las traducciones.

Eliminación de adjetivos

Los adjetivos son palabras que complementan a los sustantivos y en muchas ocasiones sucede que una frase es totalmente comprensible si se eliminan varios de estos. Al eliminar algunos de los adjetivos, por lo general, el texto mantiene el sentimiento expresado. Una vez son eliminados algunos adjetivos, la revisión se traduce al inglés y se regresa al español para corregir, en la medida de lo posible, incongruencias o faltas de concordancia.

Esta heurística está inspirada en la técnica dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) utilizada en redes neuronales para evitar el sobreajuste. El clasificador se robustece con esta eliminación, pues debe ser capaz de identificar el sentimiento de una frase, a pesar de que falten algunos adjetivos.

En el algoritmo propuesto se identifican los adjetivos utilizando la librería "Stanford POS Tagger" (Manning et al., 2014) y cada uno es eliminado con una probabilidad de 0.2. La librería "Standford POS Tagger" es ampliamente utilizada en la literatura e identifica las partes de la oración con un rendimiento sobre el 96% de *accuracy* (Toutanova, Klein, Manning, & Singer, 2003).

Alteración del orden de las oraciones

Para cualquier revisión, alterar el orden de las oraciones mantiene el sentimiento de la opinión inicial y conserva a grandes rasgos la estructura del dominio en que se definen los datos. Esta estrategia añade robustez al modelo generado, pues el algoritmo de aprendizaje no condiciona sus decisiones al orden de las oraciones. En el presente trabajo se realiza una permutación aleatoria de las cláusulas y se crea una revisión con el nuevo orden.

Reducción de oraciones extensas

Cuando la revisión del experto es extensa, un humano puede identificar si la opinión es positiva o negativa al leer solamente un subconjunto de las oraciones del texto. Si es una crítica positiva, la mayoría de las oraciones deben ser positivas, y si es negativa pasaría lo contrario.

Por lo tanto, si se escogen subconjuntos de las oraciones, eligiéndolas de manera aleatoria, es muy probable que las nuevas revisiones se mantengan expresando el mismo sentimiento que la revisión original. Los casos en que la nueva revisión varíe de sentimiento, los cuales existirán en pequeña proporción, son un tipo de ruido beneficioso para el aprendizaje, pues lo hace más robusto debido a que el clasificador puede inferir que algunas de las oraciones del texto, a pesar de ser negativas, no afectan en gran medida el sentimiento general.

En el dominio de revisiones de artículos científicos se ha observado que comúnmente las opiniones positivas se expresan en una misma región del texto y en otra las negativas, por lo que las cláusulas que manifiestan un mismo sentimiento suelen ser cercanas, lo que hace que escoger oraciones contiguas no sea una buena idea. Esta es la razón por la cual las oraciones se eligen de manera aleatoria.

La técnica empleada en el presente trabajo, consiste en generar una cantidad de revisiones pequeñas proporcional al tamaño de la revisión original. Cada revisión pequeña se forma seleccionando de manera aleatoria 5 oraciones del texto inicial. Una revisión de 5 oraciones es lo suficientemente extensa como para expresar un sentimiento bien definido. Los párrafos básicos con los que se enseñan a desarrollar ideas en edades tempranas cuentan con 5 oraciones, esta es la longitud utilizada por varios educadores (Seo, 2007), (Bormuth, Carr, Manning, & Pearson, 1970), (van Gilst & Villalobos, 1996). Debido al planteamiento anterior, se asume que, al formar revisiones pequeñas de al menos 5 oraciones, estarán expresando una opinión concreta y bien definida. Además, al contener un número impar de oraciones se previene, en cierta medida, la situación donde el sentimiento positivo y el negativo tengan la misma relevancia.

2.2. Ejemplo del proceso

En esta sección se muestra un ejemplo para ilustrar todo el proceso que sigue una revisión desde su estado inicial, pasando por cada etapa y llegando a convertirse en un texto diferente. Como ejemplo, se utiliza una revisión artificial de 3 oraciones, debido a que es un texto corto la etapa de dividirla en párrafos de 5 oraciones no se aplica. Todo el proceso se ilustra en la Figura 2, como se puede apreciar se obtiene una revisión distinta que conservan el sentimiento de la opinión inicial.

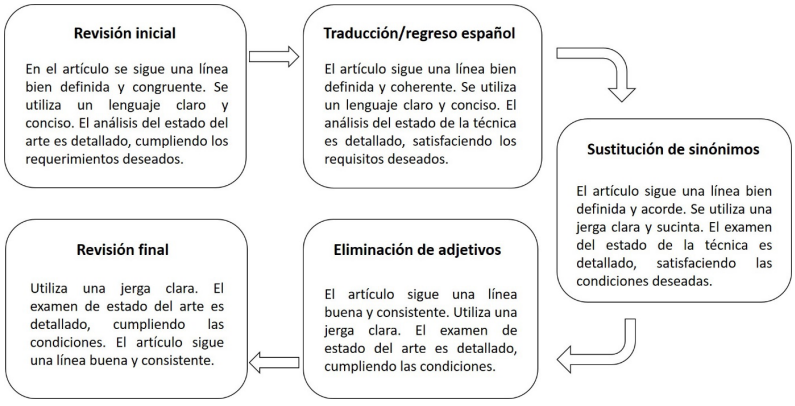

Figura 2 – Proceso de transformación que sufre una revisión al aumentar el conjunto de datos

2.3. Marco del Experimento

El método utilizado para mostrar la efectividad de las técnicas propuestas consiste en el entrenamiento de un modelo para la clasificación de sentimiento, mediante regresión logística, con el 80% de los datos del conjunto inicial y el cálculo de su *accuracy* con el 20% restante. Luego, se aplican las técnicas de *data augmentation* al 80% de entrenamiento y se entrena una vez más el modelo. El resultado esperado es mejorar el *accuracy* al clasificar el 20% que se mantuvo sin cambios. El proceso de validación utilizado se ilustra en el diagrama de la Figura 3. En esta primera etapa de la investigación se contempló generar un modelo simple a modo de prueba de concepto, por lo que se decidió entrenar un modelo de regresión logística para la tarea de análisis de sentimiento.

En esta etapa solo se desea obtener el rendimiento de la clasificación como prueba de viabilidad de los métodos de *data augmentation*, por lo que no se aplicaron técnicas más avanzadas de validación. Esto se realizará en una próxima etapa, donde se empleará la técnica de k-folds para sustentar los experimentos.

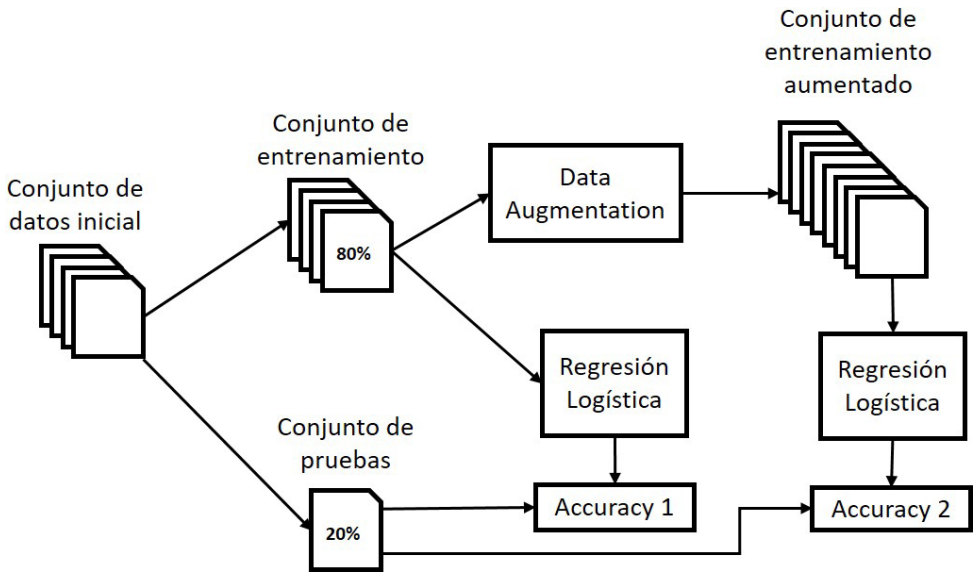

Figura 3 – Proceso de validación. Si el proceso de *data augmentation* fue beneficioso para el aprendizaje entonces el *accuracy 2* debería ser mayor que el *accuracy 1*

3. Resultados

Al aplicar las heurísticas mencionadas en la sección 2.1 y siguiendo el proceso mostrado en la Figura 3, se obtuvieron dos modelos de regresión logística, uno entrenado el 80% de los datos originales y el otro entrenado con los datos aumentados obtenidos desde el 80% de los datos originales. Ambos modelos fueron evaluados con el 20% de los datos originales. Los resultados obtenidos en términos de las métricas de rendimiento de un clasificador son resumidos en la Tabla 1.

Los resultados obtenidos con el conjunto de datos original mediante regresión logística son comparables a los reportados por Keith et al. en sus trabajos (Keith, Fuentes, & Meneses, 2017) (Keith, Fuentes, & Meneses, 2019), donde los autores obtienen como línea base 0.68 de *accuracy* y 0.64 de F1 para Naïve Bayes, 0.70 de *accuracy* y 0.69 de F1 para SVM. Por otra parte, la mejora obtenida con el conjunto de datos aumentado es comparable a la que Keith et al. obtienen al utilizar su algoritmo de Scoring y su propuesta híbrida en el caso de clasificación binaria. En particular, los autores de dicho trabajo obtuvieron 0.81 en ambas métricas para su propuesta de algoritmo de Scoring y 0.79 en ambas métricas para su propuesta híbrida. Por lo que de estos resultados se puede deducir que la aplicación de *data augmentation* provee de una mejora de rendimiento comparable, aunque marginalmente menor, a la obtenida al aplicar el método híbrido o el algoritmo de Scoring. Es posible incluso que con la combinación de *data augmentation* y estos dos métodos los rendimientos pudiesen ser mayores para este conjunto de datos en el caso de clasificación binaria.

Métricas	**Conjunto de datos original**		**Conjunto de datos aumentado**	
Cantidad de revisiones	300		14525	
Accuracy	0.67		0.74	
	Aceptado	**Rechazado**	**Aceptado**	**Rechazado**
Cantidad de revisiones	150	150	7235	7290
Precision	0.68	0.65	0.76	0.73
Recall	0.59	0.74	0.66	0.82
F1	0.63	0.69	0.71	0.77

Tabla 1 – Rendimiento obtenido con un modelo de regresión logística sobre el conjunto de datos original versus aumentado.

4. Discusión

Las heurísticas enunciadas lograron aumentar de manera significativa el volumen de datos, por un factor de 48 veces la cantidad de revisiones con respecto a los datos originales. Al clasificar un subconjunto de prueba del 20% de los datos originales con un modelo entrenado con los datos aumentados, se obtuvo un incremento en 7% del *accuracy* con respecto a un modelo entrenado solo con los datos originales. Esto muestra empíricamente el efecto positivo de aplicar las heurísticas de *data augmentation* en este dominio.

Con el uso de técnicas de *data augmentation* se buscaba no solo aumentar el volumen de datos y de esta forma mitigar un posible sobreajuste del modelo a los datos de entrenamiento, sino mantener la consistencia de los datos originales con respecto a los datos aumentados. Esto se logró ampliamente, dado el aumento evidenciado en *precision*, *recall* y medida *F1* obtenida tanto en la clase positiva (aceptado) como en la clase negativa (rechazado).

Estos resultados permitirán en una siguiente etapa aplicar a este dominio de manera más confiable técnicas que requieren ajustar una gran cantidad de parámetros durante su aprendizaje, tal como sucede con redes neuronales profundas. De manera adicional, se planea experimentar también con técnicas de *transfer learning*, es decir, redes neuronales profundas pre-entrenadas sobre reconocimiento de patrones en texto plano.

5. Conclusiones

Respecto a las heurísticas utilizadas, algunas de ellas fueron formuladas en base a trabajos relacionados, mientras que otras son propuestas teniendo en cuenta un criterio subjetivo inspirado en diferentes dominios, considerando un sentido eminentemente práctico. A pesar de ello, los resultados de su aplicación fueron positivos, en el sentido que se cumplió el objetivo de su utilización, validando la consistencia de los datos artificialmente generados, a través de la aplicación de un clasificador simple para estimar cuantitativamente su beneficio.

El dominio de revisiones de artículos científicos, además de su relevancia científica, presenta un sinnúmero de particularidades que requieren un cuidadoso pre-procesamiento sintáctico y semántico, a fin de obtener modelos consistentes con las características del dominio (e.g., amplio desbalance de clases, amplia variabilidad del tamaño de las revisiones, limitado número de revisiones, fuerte sesgo hacia los aspectos negativos). Esto lo hace un dominio desafiante para cualquier tarea de aprendizaje automático, que requiere una cuidadosa ingeniería de datos previa a la etapa de generación de modelos.

A pesar de las características del dominio y de la subjetividad de las heurísticas, los resultados obtenidos son alentadores, dado que presentan un beneficio cuantitativo de un aumento significativo en el tamaño de los datos y a la vez un mejor rendimiento de un clasificador basado en los datos aumentados en comparación a uno basado solo en los datos originales.

Referencias

Bormuth, JR, Carr, J, Manning, J, & Pearson, D. (1970). Children's Comprehension of Between and Within Sentence Syntactic Structures. *Journal of Educational Psychology*, *61*(5), 349–357.

Fawzi, A., Samolowitz, H., Turaga, D., & Frossard, P. (2016). Adaptive data augmentation for image classification. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Phoenix, Arizona, USA, 3688-3692. IEEE. DOI: 10.1109/ICIP.2016.7533048.

Grosz, B.J., Jones, K.S., & Webber, B.L. (1986). *Readings in natural language processing*. Los Altos, California, USA: Morgan Kaufmann.

Keith, B., Fuentes, E., & Meneses, C. (2019). Sentiment analysis and opinion mining applied to scientific paper reviews. *Intelligent Data Analysis*, *23*(1), 191–214.

Keith, B., Fuentes, E., & Meneses, C. (2017). A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews. In *Proceedings of WISDOM 2017 at ACM SIGKDD Conference*, Halifax, Nova Scotia, Canada. Retrieved from: https://sentic.net/wisdom2017fuentes.pdf

Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 452–457.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60. DOI: 10.3115/v1/P14-5010.

McLaughlin, N., Martinez, J., & Miller, P. (2015). Data-augmentation for reducing dataset bias in person re-identification. In *Proceedings of the 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Karlsruhe, Germany. 1–6. DOI: 10.1109/AVSS.2015.7301739.

Seo, B.I. (2007). Defending the five-paragraph essay. *English Journal*, *97*(2), 15–16.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Toutanova, K., Klein, D., Manning, C.D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 1, 173–180.

van Dyk, D., & Meng, X. (2001). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1-50.

van Gilst, L., & Villalobos, J. (1996). *The Five-Paragraph Essay: Legacy or Liability in English Writing Classrooms outside the US*. Retrieved from: https://eric.ed.gov/?id=ED401537

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, 649–657. Retrieved from: https://arxiv.org/abs/1509.01626