REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2021 PUCV, Chile 54(107) 942-970

Aplicaciones de inteligencia artificial para la clasificación automatizada de propósitos comunicativos en informes de ingeniería[1]

Applications of artificial intelligence to automatic classification of communicative purposes in engineering reports

René Venegas
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
rene.venegas@pucv.cl

Recibido: 16-VII-2021 / **Aceptado:** 25-X-2021
DOI: 10.4067/S0718-09342021000300942

Resumen

La tarea de reconocer los patrones discursivos, por medio de los cuales los miembros de una comunidad discursiva académica logran cumplir los propósitos comunicativos de los géneros académicos disciplinares, es relevante para los procesos de incorporación de nuevos miembros, a través de procesos de alfabetización académica. Las investigaciones en géneros académicos se han focalizado en los géneros expertos y, en menor medida, en géneros producidos por los estudiantes universitarios, especialmente en el área de ingeniería. El análisis de patrones discursivos del género y la inteligencia artificial (IA), a través del Procesamiento del Lenguaje Natural (PLN) han tenido un desarrollo complementario, gracias tanto a los algoritmos de clasificación automatizada y la mayor disponibilidad de grandes cantidades de datos textuales, lo que ha posibilitado la clasificación de géneros discursivos académicos. En esta línea, el objetivo de este artículo es clasificar de manera automática las macromovidas (MM) del mesogénero informe de experiencia práctica en ingeniería. Para ello, se consideraron siete algoritmos de clasificación tradicionales, el modelo de aprendizaje profundo para español denominado BETO y sus correspondientes configuraciones. Entre los hallazgos, puede destacarse el mejor rendimiento general de SVM_lineal. Así también, se destaca que SVM_lineal, BETO y KNN son más efectivos para clasificar las movidas de algunas de las MMs. Estos resultados sugieren que la combinación de algoritmos sería un procedimiento útil para clasificar de mejor manera los macropropósitos de este mesogénero. Se proyecta evaluar estos algoritmos en un herramienta para retroalimentar la producción escrita desde la perspectiva del género discursivo.

Palabras Clave: Género informe, clasificación automatizada, aprendizaje profundo, BETO, procesamiento del lenguaje natural.

Abstract

The task of recognizing the discursive patterns that enable the members of an academic discursive community to achieve the communicative purposes of the academic disciplinary genres is relevant for the processes of incorporation of new members through academic literacy processes. Research on academic genres has focused on expert genres and, to a lesser extent, on genres produced by university students, especially in engineering. The analysis of discursive genre patterns and artificial intelligence (AI), through Natural Language Processing (NLP) have had a complementary development, thanks to both automated classification algorithms and the increased availability of large amounts of textual data, which has enabled the classification of academic discursive genres. The aim of this paper is to automate classification of macromoves (MM) of the meso-genre report of practical experience in engineering. For this purpose, seven traditional classification algorithms, the deep learning model for Spanish called BETO and their corresponding configurations were considered. Among the findings, the best overall performance of SVM_lineal stands out. Also, SVM_linear, BETO and KNN are more effective for the classification of moves in different MMs. These results suggest that the combination of algorithms would be a useful procedure to better classify the macro-purposes of this meso-genre. It is planned to evaluate these algorithms in a tool for genre-based feedback of written production.

Key Words: Report genre, automated classification, deep learning, BETO, natural language processing.

INTRODUCCIÓN

El estudio de la escritura en educación superior en el ámbito de la ingeniería es un tema relevante desde finales de los años noventa (Windsor, 1996; McKenna, 1997). De esta época destaca la noción de género para la enseñanza de la escritura colaborativa de informes en ingeniería (Walker, 1999). En el ámbito hispánico, si bien las investigaciones se han realizado en entornos y géneros diversos, estas han estado orientadas predominantemente a géneros expertos y en menor medida a los géneros epistémicos producidos por estudiantes, puesto que se les ha considerado versiones imperfectas, artificiales o meramente preparatorias (Gallardo, 2005; Navarro, 2018). Por lo mismo, solo hace unos pocos años el interés por las características de los géneros escritos por los estudiantes universitarios ha ido creciendo (Cassany & López, 2010; Parodi & Burdiles, 2015; Navarro, 2017; Bosio, 2018; Sologuren, 2021), aunque los resultados sociocontextuales, genéricos, lingüísticos y aplicados aún no se desarrollan fuertemente.

En el contexto latinoamericano, Narváez-Cardona (2016) plantea que el estudio de la escritura en ingeniería se trata de un área de indagación relativamente reciente, cuyo principal interés se orienta a escribir para aprender. Chile, por ejemplo, no ha sido ajeno a esta orientación, destacándose investigaciones en las que el género informe universitario ha sido objeto de estudio, debido a su utilización frecuente en los procesos de formación universitaria (Tapia, Burdiles & Arancibia, 2003; Harvey, 2005; Espejo, 2006; Harvey & Muñoz, 2006; Muñoz, 2006; Tapia & Burdiles, 2009; Muñoz, 2019; Sologuren, 2020a, 2020b, 2021). Si bien algunas de estas investigaciones han abordado

la descripción de los informes en el contexto de su producción en ingeniería, aportando con ello en explicitar la práctica discursiva por medio de la cual los estudiantes de ingeniería acreditan su conocimiento disciplinar, se observa aún un vacío en la caracterización lingüístico-discursiva de los propósitos comunicativos más relevantes del género. En este sentido, el análisis de patrones lingüístico-discursivos del género y de la inteligencia artificial (IA), a través del Procesamiento del Lenguaje Natural (PLN) han tenido un desarrollo innovador y complementario, gracias tanto a los algoritmos de clasificación automatizada, como a la mayor disponibilidad de grandes cantidades de datos textuales. Algunos estudios de este tipo para géneros académicos en la lengua inglesa son destacables (McCarthy & McNamara, 2007; Fang & Cao, 2015; Cotos & Pendar, 2016), así como otros en lengua española (Zamora, 2014; Venegas, 2015, 2020; Lillo, 2016), aunque en ninguno de estos casos se ha considerado la clasificación automática de los propósitos comunicativos del género informe. Nuestro estudio se constituye en un inicio para llenar este vacío, por lo que el objetivo de este artículo es clasificar de manera automática las macromovidas (MM) del mesogénero informe de experiencia práctica en ingeniería.

Para el logro de este objetivo se recopiló inicialmente un corpus de 270 informes, producidos por estudiantes de ingeniería civil informática, de ingeniería civil eléctrica e ingeniería civil electrónica de la Pontificia Universidad Católica de Valparaíso (Chile). Estos informes han sido producidos con el propósito de consignar las experiencias de los estudiantes en procesos de investigación. Los informes fueron, analizados retórico-discursivamente con el fin de identificar su organización en términos de macropropósitos comunicativos, para luego representar lingüísticamente la información a través n-gramas de lexemas, lemas y categorías morfológicas. Esta información sirvió de entrada a los 7 algoritmos de clasificación tradicionales y el modelo de aprendizaje profundo para español denominado BETO. Los hallazgos son de utilidad en el reconocimiento de patrones lingüísticos para la identificación de funciones discursivas, esto es, propósitos comunicativos de los géneros; así como para considerar su utilización en una herramienta de apoyo a la retroalimentación basada en el género discursivo de la escritura, desarrollada en el proyecto Fondecyt 1190639.

A continuación, se revisan conceptos asociados al género académico, especialmente el género informe en ingeniería, para luego describir algoritmos de clasificación automatizada tradicionales, así como el de aprendizaje profundo. Además, se presentan investigaciones sobre clasificación automatizada de géneros discursivos. Todo ello para presentar luego las decisiones metodológicas, el corpus, las representaciones lingüísticas y la configuración de los algoritmos utilizados. Más adelante se presentan los resultados más relevantes y se cierra con las conclusiones.

1. Antecedentes conceptuales

1.1. Escritura y género académico en la universidad

Un género académico, entendido como un evento comunicativo organizado en movidas y pasos (Swales, 1990), se corresponde con una determinada práctica social recurrente y repetitiva que se desarrolla en un espacio institucional con características singulares (Swales, 1990, 2004). En cuanto a práctica social, los géneros están relacionados con las formas de hacer y decir de las comunidades discursivas académicas, es decir, con el "ámbito de los comportamientos formales y altamente regulados desde el punto de vista social" (Castro, Hernández & Sánchez, 2010: 52). Así, el conjunto de los géneros académicos da forma al discurso académico (Hyland, 2009; Hyland & Paltridge, 2011) como un registro unificado de una lengua (Bathia, 1993, 2002) y constituyen un medio comunicativo, que permite a los expertos de diferentes comunidades de especialidad, interactuar discursivamente y compartir conocimientos, tanto entre sí como con participantes legos y semilegos (Parodi, 2010; Nesi & Gardner, 2012; Bosio, 2018). De esta manera, Parodi, Boudon y Julio (2014: 158) puntualizan que:

> "el discurso académico comprende aquellos géneros orales y escritos que posibilitan la construcción de significados en contextos de comunicación especializada entre estudiantes y profesores".

En esta línea, Navarro (2018) distingue entre géneros expertos y géneros de formación. Los primeros son escritos por sujetos con experiencia con el propósito de construir aportes al conocimiento científico y, por tanto, son leídos por pares con conocimientos afines. Se trata de recursos genéricos que no están necesariamente relacionados con la formación. Los segundos, son producidos por estudiantes para ser revisados por miembros expertos, teniendo una finalidad pedagógica, formativa y evaluativa.

Cabe señalar que, a pesar de la relevancia de los géneros en la formación universitaria y profesional, a menudo los estudiantes y docentes no son conscientes de las convenciones genéricas propias de la comunidad disciplinar que los acoge, aún cuando el conocimiento de los géneros es central en una alfabetización académica avanzada (Wennerstrom, 2003). Así, el cumplimiento de las convenciones genéricas de los textos se alcanza fundamentalmente por medio del ensayo y error o a través de pares que han logrado adecuarse a las necesidades del género, sin retroalimentación explícita o pertinente. En los últimos años, el estudio de los géneros académicos en español se ha focalizado preferentemente en los géneros expertos (Navarro, 2014, 2018). Solo hace unos pocos años el interés por las características de los géneros escritos por los estudiantes universitarios ha ido creciendo (Cassany & López, 2010; Parodi & Burdiles, 2015; Navarro, 2017; Bosio, 2018; Sologuren, 2021), aunque la integración sistemática

de los resultados sociocontextuales, genéricos, lingüísticos y aplicados aún no se desarrollan fuertemente.

1.2. El género informe en la formación de los ingenieros

En el ámbito de la formación en ingeniería, la producción textual se ha abordado desde la comunicación técnica, principalmente aquella desarrollada en Estados Unidos (McKenna 1997; Reave, 2004; Poe, Lerner & Craig, 2010). Destaca en este ámbito el trabajo de Walker (1999), quien asume la noción de género para enseñar a estudiantes de ingeniería a escribir informes colaborativamente. Para el contexto latinoamericano, Narváez-Cardona (2016), en relación con las iniciativas y estudios de escritura en ingeniería, plantea que se trata de un área de indagación reciente y que, en contraste con el foco de comunicación técnica y estandarización de Estados Unidos, en latinoamericana se presenta un interés por los dispositivos pedagógicos para su enseñanza.

En Chile, por ejemplo, se han producido trabajos que reflejan esta perspectiva, especialmente aquellos relacionados con el género informe universitario, el que ha sido objeto de estudio desde comienzos del siglo XXI (Tapia et al., 2003; Harvey, 2005; Espejo, 2006; Harvey & Muñoz, 2006; Muñoz, 2006; Tapia & Burdiles, 2009; Ávila & Cortés, 2017; Sologuren & Castillo, 2019; Sologuren, 2020a, 2020b, 2021; Venegas & Valdés, 2021). Esto porque el informe es, sin duda, uno de los géneros académicos más solicitados en las universidades chilenas y latinoamericanas, debido a que se utiliza como estrategia evaluativa en diversas disciplinas (Sologuren, 2021).

Para Harvey y Muñoz (2006) los informes son textos expositivos breves en los que se desarrolla un tema determinado, cuyos rasgos más comunes son la estructura, los aspectos formales, la coherencia, la brevedad, la claridad, el punto de vista personal y el desarrollo de un problema. En lo que concierne a los informes elaborados en disciplinas científicas, estos tienen como objetivo resolver una problemática planteada por el docente (Harvey & Muñoz, 2006).

En el caso de las ingenierías, para Amieva (2001: 1) un informe puede ser definido como una "exposición escrita relativa a un tema, problema o actividad con propósitos formales de comunicación". En efecto, en el ámbito de la enseñanza de la ingeniería los informes se posicionan como herramientas habituales en tareas académicas y profesionales para desarrollar competencias laborales como el diseño de situaciones específicas, la implementación de propuestas y el desarrollo de habilidades no técnicas para un desempeño adecuado en la empresa a nivel personal e interpersonal (Amieva, 2001). De acuerdo a ello, los informes en el ámbito profesional comunican resultados, análisis situacionales, resúmenes de actividades, exposición de planes o propuestas, etc. (Amieva, 2001). En relación a la alfabetización académica, la elaboración de informes se vincula con la realización de trabajos prácticos de campo o laboratorio, la resolución de problemas, el desarrollo de actividades de investigación, el registro de una actividad

técnica o profesional como las realizadas en pasantías, etc., con el propósito de regular las prácticas de los alumnos, o medir sus conocimientos (Amieva, 2001). En este sentido, Amieva (2001) explica que el informe permite el dominio de un saber y de un saber decir. El primero, se expresa por medio del contenido, el cual involucra procesos como interpretaciones, análisis, relaciones con las situaciones y la búsqueda de aplicaciones. El segundo, se refiere a la competencia discursiva, relacionada con la expresión y comunicación del conocimiento que se tiene sobre contenidos particulares de los diferentes ámbitos de la ingeniería.

En cuanto a su estructura, Harvey y Muñoz (2006) señalan que está compuesto por un inicio, desarrollo y conclusión. Según Reuter (2000), en la sección inicial, debe presentarse y justificar el tema a tratar y exponer los antecedentes conceptuales. El desarrollo se ocupa de la metodología para abordar la problemática, al igual que de la aplicación de la información presentada en el inicio. Por último, la conclusión implica la exposición de comentarios finales y, frecuentemente, comentarios personales.

Los aportes de Harvey y Muñoz (2006) son muy relevantes, aunque estos focalizan un punto de vista estructural muy general y no exclusivo para ingeniería, por lo que una complementación desde la perspectiva funcional para el análisis de los informes, así como la identificación y clasificación de propósitos comunicativos presentes en el género en esta disciplina son necesarias. Destaca en este sentido, el análisis de los informes de Martin y Rose (2008), quienes caracterizan a este género como una explicación procedural. Gardner (2012), por su parte, explora el contexto social de la escritura estudiantil con fines evaluativos en dos familias de géneros: Reporte de investigación y Recuento metodológico. A partir del análisis del registro y del género de los textos que presentan una estructura IMRD canónica (Introducción, métodos, resultados y discusión), la autora, distingue características lingüísticas desplegadas en diversos contextos disciplinares y en diferentes niveles de estudio.

En el campo del discurso profesional en español, Montolío (2010) y Montolío y López (2010) analizan informes de consultoría producidos por expertos del ámbito tecnológico y caracterizan la recomendación experta y el asesoramiento profesional como operaciones textuales muy significativas en la producción escrita cotidiana de un número creciente de profesionales. Así mismo junto con caracterizar los mecanismos lingüísticos y discursivos de este género, asociado con una organización discursiva argumentativa, relevan las dificultades que los profesionales manifiestan para su elaboración.

Un aporte importante en el estudio del género informe se ha venido desarrollando, desde el año 2008 en la Escuela Lingüística de Valparaíso (Parodi, 2008). Así, por ejemplo, el informe es definido por Parodi, Ibáñez y Venegas (2009: 90) como:

> "Género discursivo cuyo macropropósito es consignar situaciones, procedimientos y/o problemas. Idealmente, su contexto de circulación es el ámbito laboral y la relación entre los participantes es entre escritor experto y lector experto. Suele ser monomodal y presentar un modo de organización discursiva que es descriptiva".

En disciplinas específicas como la economía, el informe técnico ha sido reconocido como un macrogénero y se ha abordado desde el estudio del discurso profesional, diferenciando los géneros por las temáticas que tratan, las instituciones que los emiten, las audiencias originales declaradas y los grados de especialización (Parodi, Julio & Vásquez-Rocca, 2015).

Por otra parte, Muñoz (2019) caracteriza retórico-discursivamente informes académicos de ingeniería civil informática distinguiendo tres MMs, a saber, a) Introducir al lector en el tema o problema del informe; b) Exponer sustento conceptual de las herramientas, problemáticas o soluciones tratadas en el informe y c) Finalizar discursivamente la investigación presentada en el informe. Este análisis basado en una única asignatura del plan de estudios permite contar con una aproximación inicial al modelamiento de la práctica discursiva formativa en esta subdisciplina de la ingeniería civil. Asimismo, en esta disciplina, Sologuren (2020b) describió tipológica y topológicamente los géneros de formación en la etapa *capstone o* final del currículum en tres universidades chilenas. Entre sus hallazgos más relevantes se destaca el relevamiento de 7 familias de géneros o macrogéneros discursivos, a saber: a) Informe técnico, b) Planes, c) Requerimientos, d) Modelos, e) Metodologías, f) Ejercicio didácticos y g) Trabajo Final de Grado. Para la familia genérica Informe técnico (INT), cuyo macropropósito es consignar situaciones, procedimientos y/o problemas, Sologuren (2020b) identifica y describe tres mesogéneros: 1) Informe de laboratorio (ILAB), asociado a procesos de investigación que se presentan a través de géneros como informe de laboratorio, informe de algoritmo, informe de investigación e informe de terreno. Su organización retórica se presenta en la figura 1, 2) Informe de caso (ICAS), asociado a procesos curriculares entre los que se distinguen el informe de caso, reporte reflexivo e informe de práctica profesional y 3) Informe profesional (INP), asociado a procesos más orientados a la formación profesional, identificándose el informe de proyecto, informe de software, informe de evaluación, informe de diagnóstico, entre otros. Esta propuesta de análisis empírico mixto de los géneros relacionados con el currículum, basada en datos etnográficos y en corpus, es muy innovadora y rigurosa. Ello porque los resultados aportan un conocimiento amplio respecto a los géneros escritos, su rol formativo y su relevancia en los distintos momentos curriculares de la formación en ingeniería informática.

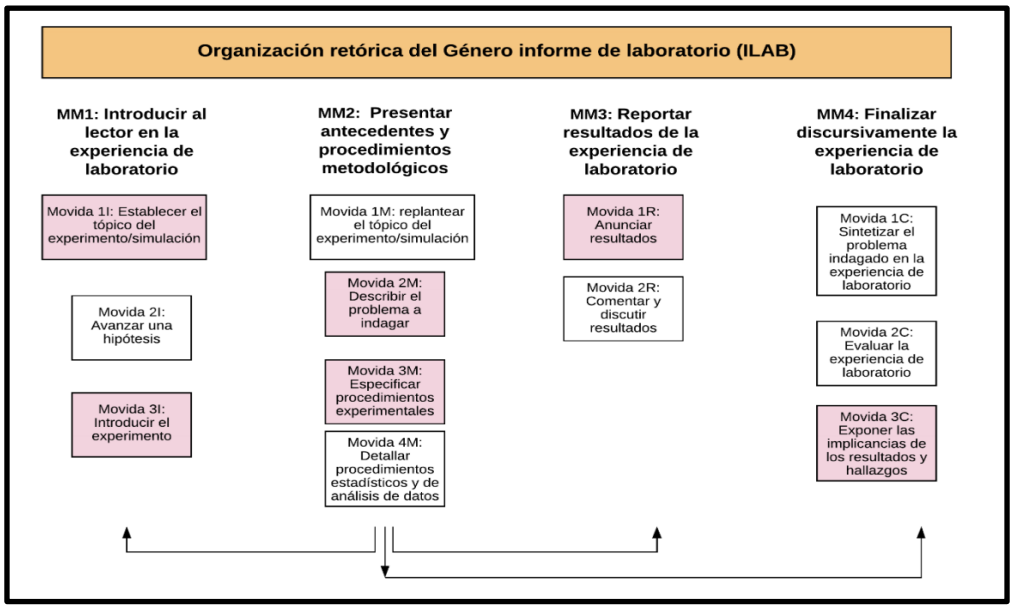

Figura 1. Organización retórica del mesogénero ILAB (Sologuren, 2020b: 335).

1.3. Géneros discursivos y clasificación automatizada

Como ya mencionamos, en los géneros discursivos existe una estrecha vinculación entre las formas lingüísticas y discursivas con la práctica social. Así la tarea de identificar, clasificar y explicitar los patrones discursivos por medio de los cuales los miembros de una comunidad discursiva logran cumplir sus propósitos (Kessler, Numberg & Schütze, 1997), se constituye en la base para sustentar los procesos de incorporación de nuevos miembros y, por lo mismo, ha sido útil en los procesos de alfabetización académica, así como para su análisis computacional. En relación con esto último, la vinculación entre el análisis lingüístico-discursivo y la inteligencia artificial (IA), a través del Procesamiento del Lenguaje Natural (PLN) ha dado un nuevo impulso al análisis de los géneros. Ello debido a que, el PLN es una subdisciplina aplicada que se centra en investigar y formular soluciones computacionales que faciliten la interrelación hombre-máquina y que permitan la automatización de procesos relacionados con la comunicación humana (Zhang & Lu, 2021). Entre sus tareas se encuentra la clasificación automatizada, mediante máquinas de aprendizaje; esto es, algoritmos de discriminación de atributos que permiten descubrir variables discriminatorias entre los textos de clases preexistentes (Jurafsky & Martin, 2000; Figuerola, Zazo & Berrocal, 2000).

1.3.1. Algoritmos de aprendizaje tradicional y de aprendizaje profundo para clasificación automatizada

Se entiende por clasificación automatizada al proceso de aprendizaje matemático estadístico, por medio del cual un algoritmo computacional identifica las características

que distinguen categorías o clases de documentos de las demás. Esta puede ser llevada a cabo utilizando distintas aproximaciones, tales como clasificación supervisada, no supervisada o semi-supervisada (Figuerola et al., 2000; Sebastiani, 2002). En particular, los algoritmos de clasificación supervisada son entrenados en un grupo de documentos, categorizados y etiquetados manualmente, acorde a algún criterio particular (tema, materia, origen, género, propósitos, etc.), conformando una clase. De esta manera, el objetivo de estos clasificadores es decidir en qué categoría debe ir cada texto nuevo, partiendo de un conjunto de atributos (caracteres, lexemas, categorías gramaticales, semánticas o una combinación de ellos) para conformar una representación, que permita al algoritmo asignar los documentos al esquema de clasificación previo (Alfaro & Allende, 2020; Figuerola et al., 2000). Algunos algoritmos de clasificación considerados tradicionales son los siguientes:

Los algoritmos probabilísticos (Naive Bayes): estos se basan en el teorema de Bayes, el cual permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. El algoritmo más conocido, y también el más simple, es el denominado Naive Bayes (o Bayes ingenuo), que estima la probabilidad de que un documento pertenezca a una categoría (Manning, Raghavan & Schütze, 2008; García, 2017). Dicha pertenencia depende de una serie de características, de las cuales se conoce la probabilidad de aparición en los documentos que pertenecen a la categoría en cuestión. Tales características son los atributos que conforman los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada categoría pueden obtenerse a partir de las frecuencias de aparición en el corpus de entrenamiento. Con dichas probabilidades, se puede estimar la probabilidad de que un nuevo documento, dado que contiene un conjunto determinado de atributos, pertenezca a cada una de las categorías. Entre los clasificadores de este tipo destacan: *Multinomial Naive Bayes* (MNB), variante que utiliza la frecuencia de variables discretas (por ejemplo, frecuencias de palabras); *Bernoulli Naive Bayes* (BNB), que utiliza valores booleanos y *Complement Naive Bayes* (COMP NB), variante que en vez de calcular la probabilidad del atributo para una clase específica, lo hace considerando la probabilidad del atributo en todas las clases (Manning, et al., 2008).

Algoritmo del vecino más próximo: La idea básica es que si se calcula la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento, aquél que sea más parecido indicará a qué clase o categoría se debe asignar el documento que se desea clasificar. Una de las variantes más conocidas de este algoritmo es la del *k-nearest neighbour* (KNN), que consiste en tomar los *k* documentos más parecidos, en lugar de solo el primero. Como en esos *k* documentos habrá, presumiblemente, de varias categorías se suman los coeficientes de cada una de ellas. Así, la categoría que más puntos acumule será la candidata idónea (Pérez, 2017). KNN es eficaz cuando el número de categorías posibles es alto, y cuando los documentos son heterogéneos y

difusos. Para inferir la categoría de un ejemplo desconocido, el algoritmo compara ese ejemplo con todos los ejemplos de entrenamiento, calculando la distancia euclidiana entre ellos, luego la clase mayoritaria de entre los *k* ejemplos más similares al de entrada es la categoría inferida (Pérez, 2017).

Árboles de Decisión o Clasificación: Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir de un conjunto numeroso de prototipos. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación, que puede modelarse mediante una estructura de tipo árbol (Jerez, 2018). Cada nodo interior contiene una pregunta sobre un atributo concreto y cada nodo hoja se refiere a una decisión de clasificación (Rokach & Maimon, 2008; Jerez, 2018). La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezando por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

Máquinas de Soporte Vectorial (SVM): Las SVM corresponden a un conjunto de algoritmos de aprendizaje supervisado (Vapnick, 2000). Para Pérez (2017: 14): "intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio multidimensional, separando las clases por un espacio lo más amplio posible". Así cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas en alguna de las clases predeterminadas. El uso de SVM es particularmente apropiado para trabajar con datos multidimensionales, tales como representaciones de vectores en un espacio de documentos textuales (Venegas, 2007). En términos geométricos, el problema que resuelve SVM es identificar una frontera de decisión lineal entre dos grupos, a través de una línea que los separe, maximizando el espacio del hiperplano (Sharma, 1996; Hair, Anderson, Tatham & Black, 1999). SVM incluye una operación llamada truco de kernel (estos son lineal, función de base radial, polinominal, sigmoidal, etc), lo que le permite realizar separaciones de los datos, optimizando la clasificación de los mismos. De este modo, el hiperplano de separación óptimo será aquel que tenga el máximo margen de separación (Guyon, Boser & Vapnik, 1993; Smola & Schölkopf, 2004).

A continuación presentamos una aproximación basada en aprendizaje profundo denominada BERT (*Bidirectional Encoder Representations from Transformers*), desarrollada inicialmente para el inglés (Devlin, Chang, Lee & Toutanova, 2018) y su versión en español, BETO (Cañete, Chaperon, Fuentes, Ho, Kang & Pérez, 2020).

Aprendizaje profundo BERT y BETO: El aprendizaje profundo o redes neuronales profundas, es un aspecto de la inteligencia artificial (AI) que se ocupa de emular el enfoque de aprendizaje que los seres humanos utilizan para obtener ciertos tipos de

conocimiento. En su forma más simple, el aprendizaje profundo puede considerarse como una forma de automatizar el análisis predictivo (TechTarget, 2021). De este modo, mientras que los algoritmos tradicionales de aprendizaje automático son lineales, los algoritmos de aprendizaje profundo se apilan en una jerarquía de creciente complejidad y abstracción, a través de múltiples capas de neuronas en las que distintas características o parte de ellas van siendo procesadas (ver Figura 2). Cada algoritmo en la jerarquía hace un a combinación lineal de la capa anterior para luego aplicar una transformación no lineal en su salida y utilizar lo que aprende, para crear un modelo estadístico como salida, hasta alcanzar un nivel de precisión aceptable. El número de capas de procesamiento a través de las cuales los datos deben pasar es lo que inspiró la etiqueta de profundidad o *deep*. Entre sus ventajas se destaca que el programa construye el conjunto de características por sí mismo sin supervisión.

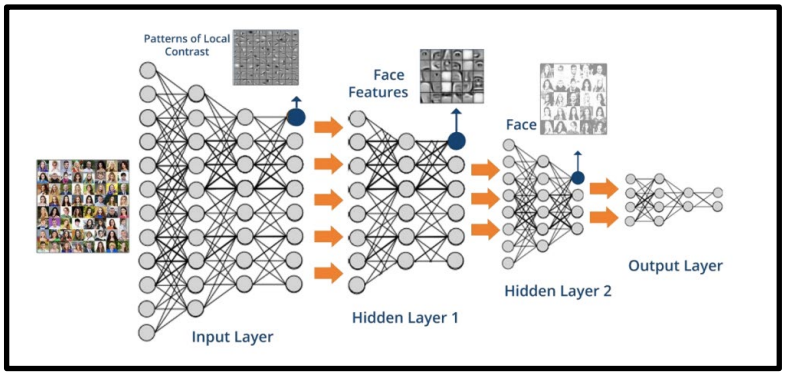

Figura 2. Datos y capas de entrada, capas ocultas y capas de salida en una red neuronal profunda para clasificación de rostros (Buigas, 2017).

De acuerdo con Cañete, et al. (2020), el campo del PLN ha hecho un progreso increíble en los últimos años. De este modo, dos de las características decisivas han sido la arquitectura *Transformer* (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, 2017) y la introducción de métodos de preentrenamiento no supervisado (Devlin et al., 2018), los que aprovechan enormes cantidades de corpus de texto no etiquetados. Basándose en ello, Devlin et al. (2018) y sus colegas de Google, propusieron la arquitectura BERT, la que permite extraer características de nivel intermedio como sintaxis e incrustaciones de palabras (normalmente afijos), útiles para muchas tareas de clasificación. Las arquitecturas de transformadores como BERT permiten a los usuarios aplicar la misma red previamente entrenada a nuevos problemas y lograr un rendimiento significativamente mayor. BERT fue preentrenado inicialmente utilizando un corpus en inglés y luego se probó con un corpus en chino (Devlin et al., 2018). Más tarde se creó una versión multilingüe de BERT (mBERT), preentrenada simultáneamente sobre un corpus que incluye más de 100 idiomas diferentes (Turc, Chang, Lee & Toutanova, 2019). El modelo mBERT ha mostrado un alto rendimiento cuando se ajusta a tareas específicas de cada idioma y ha logrado resultados de vanguardia en diferentes tareas

multilingües (Wu & Dredze, 2019). El primer modelo BERT preentrenado para el idioma español, BETO, fue propuesto por Cañete et al. (2020).

Según sus autores, BETO es un modelo que tiene 12 capas de autoatención con 16 nodos de atención cada una, utilizando 1024 nodos como tamaño oculto. En total el modelo tiene 110 millones de parámetros. Para el entrenamiento del modelo se utilizó un corpus de unos 3.000 millones de palabras en español. Además, se integró la técnica de *Dynamic Data Masking* (DDM), la que se refiere al uso de diferentes máscaras para la misma frase en el corpus, ello permite ocultar los datos y representarlos de manera artificial para su mejor automatización. El DDM utilizado fue de 10x, lo que significa que cada frase tuvo 10 máscaras diferentes. Cabe señalar, que se entrenaron dos modelos (considerando o no las mayúsculas) a través de 2 millones de pasos, con una tasa de aprendizaje de 0,0001. Estos modelos así como todos los parámetros necesarios están disponibles en https://github.com/dccuchile/beto. La conformación de este modelo ha permitido obtener mejores resultados en comparación con mBERT (Cañete et al., 2020), aportando así una aproximación innovadora y muy útil para aplicaciones de PNL en español.

1.4. Clasificación automatizada y géneros discursivos

Una de las primeras reflexiones relativas a la clasificación automatizada de los géneros fue la planteada por Ikonomakis, Kotsiantis y Tampakas (2005), sin embargo, tal reflexión no se acompañó de una investigación empírica. En oposición, McCarthy y MacNamara (2007), por una parte, clasificaron textos narrativos, de la historia y científicos. Los autores demostraron empíricamente que las siete primeras palabras eran atributos de clasificación adecuados. No obstante, la noción de género difiere de la utilizada en este artículo, pues la narración es un modo de organización discursiva (Parodi, Ibáñez & Venegas, 2010), mientras que los textos de la historia y científicos corresponden a registros discursivos (Halliday, 2004).

Por su parte, Fang y Cao (2015) clasificaron automáticamente 32 géneros del International Corpus of English (Greenbaum, 1991). Para la clasificación, los autores utilizaron un corpus de 500 textos, representados a partir de categorías morfológicas del tipo N_com_sing para sustantivo_común_singular, y dos máquinas de clasificación: *Naive Bayes* (NB) y *Multinomial Naive Bayes* (MNB). Esta última fue la más eficiente al obtener un valor F1= 0,794, entendido este valor como la media armónica entre los valores de precisión y exhaustividad de la clasificación. Los resultados muestran que a menor cantidad de clases de género, la representación del texto puede ser más concreta (solo palabras); mientras que a más cantidad, una mayor abstracción lingüística y especificidad (lemas, *POS*) es de utilidad para la clasificación.

Una aproximación empírica diferente es la que plantean Cotos y Pendar (2016), quienes clasifican los propósitos comunicativos de las introducciones de 1020 artículos

de investigación (20 artículos de 51 disciplinas), utilizando como representación de las oraciones unigramas y trigramas de lexemas y SVM en cascada para clasificar movidas y pasos de cada movida. Los resultados para las movidas son aceptables (F1 promedio 63,9%). Para los pasos, se obtienen valores variables (F1= 30,8% a 85,9%). El mérito de este trabajo es aportar en la clasificación automática de los propósitos comunicativos al interior de las introducciones de los textos que conforman el género académico Artículo de investigación en inglés.

En español, Venegas (2014, 2015) clasifica de manera automática las MMs que constituyen el trabajo final de grado, vinculando los niveles de análisis léxico-gramatical, léxico-semántico y retórico-discursivo. Como atributos se consideraron los rasgos léxico-gramaticales (lexemas, lemas y categorías gramaticales) y léxico-semánticos, operacionalizados a partir de dos indicadores: solapamiento de palabras de contenido (SPC) y similitud semántica (SS). El corpus se encuentra constituido por 16 tesis de pregrado, realizadas entre los años 2009 y 2013 en la disciplinas de ingeniería civil informática, y se utilizó *Naive Bayes*. Para una mejor representación se utilizó el método denominado *Correlation-based Feature Subset Selection* (CbFSS, Hall, 1999). Así, con la representación de trigramas de categorías gramaticales se obtuvo un F1 promedio =0,764 en tanto que con la combinación de atributos SPC+Adjetivo+Preposición+Pronombre se obtuvo para las F1 promedio =0,58. Cabe destacar que la selección de atributos, si bien aportó mejores índices de clasificación con trigramas de categorías gramaticales, su uso implicó un alto costo computacional en tiempo. En este misma línea, más tarde, Venegas (2020) clasificó las MMs 'introducir al lector en la investigación' y 'finalizar discursivamente la investigación' en un corpus de 190 tesis de licenciatura de agronomía, kinesiología, informática, lingüística y psicología. Los textos se segmentaron como fragmento y como oración. Los atributos - lexemas, lemas y categorías morfológicas- se cuantificaron en términos de 3-gramas. Para la clasificación se utilizaron 6 máquinas de aprendizaje: *Naive Bayes*, *Multinomial Naive Bayes*, SVM, NuSVM; Árboles de decisión y Redes neuronales artificiales. El resultado más destacado para clasificar las MMs en todas las disciplinas fue usando SVM con kernel lineal (F1 promedio = 0,781). Este resultado concuerda con lo presentado por Cotos y Pendar (2016) y mejora los datos obtenidos en Venegas (2015). En cuanto a la segmentación, las oraciones fueron más efectivas con trigramas de lema como de lexema. El aporte de Venegas (2020) fue identificar la mejor configuración para clasificar los propósitos comunicativos considerados en ambas MMs.

2. Procedimientos metodológicos

2.1. Corpus

Esta investigación tiene por objetivo clasificar de manera automática las macromovidas del mesogénero informe de experiencia práctica en ingeniería. Para el logro del mismo se utilizó inicialmente un corpus de 270 informes producidos, entre los

años 2017 y 2019, por estudiantes de ingeniería civil informática (ICI), de ingeniería civil eléctrica (ICE) e ingeniería civil electrónica (ICEL) de la Pontificia Universidad Católica de Valparaíso. Estos informes han sido producidos en contextos de enseñanza-aprendizaje en los que los estudiantes debían consignar sus experiencias en procesos de investigación, a través de géneros específicos como informe de laboratorio, informe de algoritmo, informe de investigación e informe de terreno.

Estos informes fueron analizados a nivel retórico-discursivo por etiquetadores entrenados, en el marco del proyecto Fondecyt 1190639, quienes identificaron su organización en términos de movidas (propósitos) y macromovidas (macropropósitos). Esta tarea fue desarrollada con el apoyo de la herramienta HERMES, (http://www.redilegra.com/index.php/hermes/). El análisis de movidas y macromovidas identificadas fue validado por dos magísteres en lingüística aplicada. De este proceso se seleccionaron 57 informes (ICI=9, ICE=20, IECL=28) para la etapa de clasificación, dado el mayor consenso entre los etiquetadores y la homogeneidad genérica de los informes. Cabe señalar, que para esta investigación no se establecieron diferencias disciplinares, dado el interés en los propósitos del género. En la Tabla 1 se muestra la organización retórico-discursiva identificada para este mesogénero de experiencia práctica (MGIEP).

Tabla 1. Modelo retórico-discursivo del MGIEP.

MODELO MESO-GÉNERO INFORME DE EXPERIENCIA PRÁCTICA		
CÓDIGO	**FUNCIÓN**	**DESCRIPCIÓN**
MM1	**Macromovida 1: Sintetizar el contenido del informe**	**Anunciar los principales puntos que se abordarán en el informe**
MM1_M1	Movida 1: Describir de la experiencia realizada	Describir el tema y problema que sustentan la experiencia informada.
MM1_M2	Movida 2: Describir los procedimientos	Mencionar los procedimientos realizados para el desarrollo de la experiencia.
MM1_M3	Movida 3: Anticipar resultados	Mencionar sintéticamente los principales resultados de la experiencia.
MM1_M4	Movida 4: Especificar palabras claves	Mencionar los términos relevantes en la experiencia informada.
MM2	**Macromovida 2: Introducir al lector a la problemática de la experiencia práctica**	**Orientar al lector en relación con el tema, presentar los supuestos conceptuales más relevantes que guían la investigación, indicar su propósito y justificar la relevancia de realizar la investigación.**
MM2_M1	Movida 1: Establecer el territorio	Destacar la importancia del tema, desde lo general a lo específico, mostrando la necesidad de su investigación.
MM2_M2	Movida 2: Delimitar el problema a indagar	Establecer el problema específico que abordará la experiencia práctica.
MM3	**Macromovida 3: Presentar antecedentes y procedimientos**	**Desarrollar los elementos que sustentan la experiencia a nivel teórico, metodológico y práctico.**
MM3_M1	Movida 1: Establecer territorio temático	Situar temáticamente la investigación, justificando su relevancia y dando cuenta de investigaciones y conceptos relevantes en el área.
MM3_M2	Movida 2:Describir el problema a indagar	Destacar el aspecto específico del tema que se trabajará en la experiencia práctica.
MM3_M3	Movida 3: Especificar procedimientos metodológicos	Explicar los pasos procedimentales realizados para el desarrollo de la experiencia práctica
MM4	**Macromovida 4: Reportar resultados**	**Presentar los resultados obtenidos en la experiencia práctica.**
MM4_M1	Movida 1: Anunciar resultados	Destacar los hallazgos obtenidos y representarlos a través de artefactos semióticos matemáticos variados (tablas, ecuaciones, gráficos,etc.)
MM4_M2	Movida 2: Comentar e interpretar resultados	Entregar una explicación coherente de los resultados obtenidos
MM5	**Macromovida 5: Finalizar discursivamente la experiencia práctica**	**Concluir la declaración de la experiencia práctica, recordando al lector sus aspectos más relevantes, evaluando los hallazgos obtenidos y presentando proyecciones.**
MM5_M1	Movida 1: Sintetizar el problema indagado en la experiencia práctica	Sintetizar el tema indagado, su puesta en práctica y los resultados obtenidos
MM5_M2	Movida 2: Evaluar la experiencia práctica	Valorar la experiencia en el contexto de la formación académica y/o profesional.
MM5_M3	Movida 3: Exponer las implicaciones de los resultados y hallazgos	Identificar aspectos derivados de la experiencia para potenciales aplicaciones o futuras investigaciones.

2.2. Clases, representación y atributos

Para la clasificación automática hemos considerado como clases las cinco Macromovidas (MM) del MGIEP conformadas por las movidas de cada MM. La Tabla 2 muestra la distribución del total de oraciones para cada MM. Cabe señalar, que en el

proceso de limpieza de datos se eliminaron oraciones repetidas, aquellas menores a 15 palabras y aquellas con más de 80 palabras. La clasificación de cada una de ellas se realizó con algoritmos tradicionales (ver 1.3.1) y aprendizaje profundo: K Vecinos Cercanos (KNN), Árboles de clasificación y regresión (CART), *Naive Bayes Multinomial* (MNB), *Naive Bayes Bernoulli* (BNB), *Complement Naive Bayes* (COMP NB), Máquina de Soporte Vectorial (SVM con kernel lineal) y BETO.

Tabla 2. Total de oraciones por Macromovidas en el corpus.

Clases (MM)	Oraciones
MM1: Sintetizar el contenido del informe	170
MM2: Introducir al lector a la problemática de la experiencia práctica	426
MM3: Presentar antecedentes, procedimientos de la experiencia práctica	2913
MM4: Reportar resultados	4481
MM5: Finalizar discursivamente la experiencia práctica	335
Total de oraciones	**8325**

Con el fin de representar cuantitativamente la información lingüística contenida en las oraciones, se consideró la frecuencia de tres tipos de atributos lingüísticos: lexemas, lemas y categorías gramaticales. Cada una de estas representaciones se vectorizaron, usando técnicas frecuentes en PLN, esto es: *Bag of Words vectorizer* (BOW), *Term frequency – Inverse document frequency vectorizer* (TFIDF) *Word hashing vectorizer* (Hash), así como cada uno de los anteriores vectorizadores, pero descartando las *Stop Words* (+SW). Además, se incluyó en la representación distintos tamaños de n-gramas y combinaciones de ellos. Así se utilizaron, unigramas (n11), unigramas y bigramas (n12), unigramas, bigramas y trigramas(n13), bigramas (n22) y trigramas(n33). Para todas las clasificaciones se consideró una distribución del 70% de los datos para entrenamiento y un 30% para la prueba del modelo. Además, se utilizó una validación cruzada de tamaño 5 (Kfold 5), así los valores que se reportan como resultado es el promedio de 5 clasificaciones con distintos ejemplos tomados al azar para cada algoritmo y tarea. Por último, se balancearon los datos, para ajustar las diferencias y evitar el sobreaprendizaje de los algoritmos de clases con más oraciones.

Para la clasificación con BETO se utilizó el código disponible en Github (https://github.com/dccuchile/beto). En su ejecución se utilizaron los siguientes hiperparámetros para hacer el ajuste fino: *learning rate*=0.001 y *dropout* en la penúltima capa de la red = 0.3. Además, se utilizó el optimizador Adam.

2.3. Herramientas y medidas de evaluación

Para la representación de los lexemas a lemas y categorías gramaticales se utilizó el etiquetador de *POS*, basado en aprendizaje profundo *SpaCy* (www.spacy.io). Se programaron cada una de las máquinas de clasificación y sus variaciones, utilizando los algoritmos disponibles en *Scikit-learn* (http://scikit-learn.org/stable/supervised_learning.html#supervised-learning). Toda la programación se hizo en *Python*, quedando disponible en *Google Colab*

https://colab.research.google.com/drive/1R4TskQA3QpVXYgV04qOtPbQJri8UIx76#scrollTo=b1gc_4Td-muE. La evaluación de los resultados realizó con las siguientes medidas:

$$a = \frac{TP+TN}{TP+FP+TN+FN} \quad p = \frac{TP}{TP+FP} \quad r = \frac{TP}{TP+FN} \quad F1 = 2\,\frac{pr}{p+r}$$

Figura 3. Medidas de evaluación a= *Accuracy* (Exactitud), p= *Precision* (Precisión) y r = *Recall* (Exhaustividad). TP=*True positives* (verdaderos positivos), TN=*True Negatives* (verdaderos negativos), FP= *False Positives* (falsos positivos) y FN= *False Negatives* (Falsos negativos).

Cabe destacar que la medida F1 (o Valor F) se concibe como la media armónica entre precisión, el valor que considera los ejemplos correctos identificados (TP) y los ejemplos identificados que corresponden a otra clase (FP); y exhaustividad, el valor que considera los ejemplos correctos identificados (TP) y aquellos que no se lograron identificar para la clase respectiva. Ello permite combinar en un solo valor ambas medidas, lo que es muy práctico para comparar el rendimiento combinado de la precisión y la exhaustividad entre varias resultados de clasificación. Además, se debe señalar que los valores de F1 que se reportan corresponden a macro-F1, es decir, al valor promedio de las 5 clasificaciones realizadas en la validación cruzada (Kfold=5).

3. Resultados

A continuación, en la Tabla 3, se presentan las cinco mejores configuraciones de algoritmos tradicionales para las tareas de clasificación de las 5 MMs. Entre los resultados observamos que independiente del tipo de clasificador, el mejor atributo para representar la información lingüística son los unigramas y bigramas de lemas. Esto concuerda con la clasificación de MMs en otros géneros académicos (Cotos & Pendar, 2016; Venegas, 2020).

Tabla 3. Mejores resultados para clasificación de MMs del MGIEP.

Algoritmo	Representación	n-gramas	vectorización	Exactitud	Desv.est.
MNB	lema	n12	BOW+SW	74,48%	1,30%
COMP NB	lema	n12	BOW	74,38%	1,05%
COMP NB	lema	n12	BOW+SW	74,29%	0,95%
SVM_lineal	**lema**	**n12**	**TFIDF+SW**	**75,09%**	**0,91%**
SVM_lineal	lema	n12	TFIDF	74,26%	0,96%

Además, es destacable que el uso de las técnicas de vectorización BOW ('bolsa de palabras') y TFIDF, usadas como métodos de normalización de las frecuencias del atributo son las más predictoras. En ambos casos sin considerar las *Stop Words* (palabras

funcionales del tipo conectores, conjunciones, artículos, etc.). Por lo tanto son las palabras de contenido semántico las que mejores resultados pueden aportar a la clasificación.

Se observa que los algoritmos probabilísticos y vectoriales conforman las mejores configuraciones entre los algoritmos tradicionales. Se destaca SVM_lineal, como el más apropiado para trabajar con estos datos textuales. La exactitud indica que en general el modelo ha sido bien entrenado y que permite reconocer la mayoría de las clases. En este caso, destaca SVM_lineal con unigramas y bigramas de lema con una vectorización TFIDF sin considerar las *Stop Words* (75,09%). En la Tabla 4, se presenta el detalle de los valores correspondientes a las medidas de evaluación para este clasificador. Como se observa, la precisión promedio del modelo (0,86) es muy buena para este tipo de tarea, considerando que es una tarea altamente compleja, dada la gran cantidad de variaciones posibles en el cumplimiento de los propósitos comunicativos, el desbalance de los datos y la relativamente simple representación de las oraciones.

Tabla 4. Medidas de evaluación con SVM_lineal.

Clase	Precisión	Exhaustividad	F1	Exactitud
MM1	0,94	0,31	0,47	
MM2	0,88	0,26	0,4	
MM3	0,75	0,67	0,7	
MM4	0,73	0,9	0,81	
MM5	1	0,27	0,43	
Macro promedio	**0,86**	**0,48**	**0,56**	**75,09%**

Esto implica que la dispersión de los datos es baja, lográndose una buena concentración de ejemplos que efectivamente corresponden a las MMs. Por otra parte, la exhaustividad, la cual evalúa los ejemplos positivos que efectivamente predecimos correctamente, tiene un buen rendimiento. En general se espera por azar un 20% suponiendo que las clases estén balanceadas, en este caso se obtiene más del doble (0,48). No obstante, el modelo aún incluye en cada clase un conjunto de ejemplos que en realidad corresponden a otra MM, particularmente en MM2, MM5 y M1.

En síntesis, el valor macro-F1 nos permite establecer que si queremos clasificar los macropropósitos, a partir de los unigramas y bigramas de lemas, obtendremos un éxito del 56% total. Este resultado es considerado bueno, aunque mejorable. Particularmente, es difícil para este modelo identificar correctamente las oraciones que corresponden a MM2, MM5 y MM1, esto debido a que son clases en las que naturalmente hay menos cantidad de oraciones y porque discursivamente estas MMs tienen una función de macrosemantización, a través de la repetición o parafraseo, de parte de la información que aparece en las MM3 y MM4.

Para el caso de la clasificación con BETO, como se mencionó previamente, se utilizó como ajuste fino un *learning rate* = 0,001, un *dropout* en la penúltima capa de la red = 0,3

y se utilizó el optimizador Adam. Para la representación del texto se utilizó el identificador de *tokens* (*wordpiece*) incluido en el algoritmo.

Como se observa en la Tabla 5, el resultado de exactitud alcanzó un 59%, lo que indica un buen rendimiento, aunque menor en un 16% a SVM_lineal.

Tabla 5. Medidas de evaluación para la clasificación de MMs con BETO.

Clase	Precisión	Exhaustividad	F1	Exactitud
MM1	0,12	0,9	0,21	
MM2	0,24	0,83	0,38	
MM3	0,8	0,63	0,71	
MM4	0,86	0,54	0,66	
MM5	0,17	0,75	0,27	
Macro promedio	**0,43**	**0,71**	**0,43**	**59,00%**

Como se observa, la precisión promedio alcanzada con este algoritmo (0,43) es un 50% menor que con SVM_lineal, observándose una mayor dispersión. Por otra parte, la exhaustividad es mucho mayor (0,71). Esto indica que el algoritmo distingue mejor que SVM_lineal aquellos ejemplos que efectivamente pertenecen a la categoría y no incluye los que pertenecen a otra. No obstante lo anterior, en general el valor promedio de macro-F1 es menor al obtenido con SVM_lineal. MM3 es la única en la que se observa un mayor valor de F1 (0,71), pero la diferencia es mínima como para hacer alguna distinción relevante.

3.1. Mejores algoritmos para cada MM

A continuación, se sintetiza la comparación del rendimiento de cada clasificador tradicional con BETO para cada una de las MMs. Esto con el fin de determinar si al separar el problema por subclases los clasificadores pueden entregar mejores resultados. De este modo, para la clasificación se consideran las oraciones de cada movida (subclase) en cada una de las macromovidas (clase). Así, por ejemplo, para MM1 se identifican 4 movidas, entre las que se distribuyen 170 oraciones. Esto supone que la clase es conocida y que las subclases y la cantidad de oraciones variará para cada tarea de clasificación.

Los datos indican que para este tipo de clasificación BETO es capaz de clasificar mejor las movidas que los algoritmos tradicionales en tres de las cinco MMs (MM1, MM2 y MM3), esto es indicativo de la mejor capacidad que tiene el aprendizaje profundo para identificar patrones, usando los vectores de palabras (*words embbedings*). No obstante, MM4 y MM5 son mejor clasificadas utilizando MNB y KNN, respectivamente.

Tabla 6. Mejor configuración de clasificadores tradicionales o de aprendizaje profundo.

Clase	Exactitud	Precisión	Exhaustividad	F1
Movidas--> MM1				
SVM	0,33	0,08	0,25	0,12
Beto	**0,38**	**0,36**	**0,51**	**0,38**
Movidas--> MM2				
KNN	0,48	0,5	0,39	0,49
Beto	**0,55**	**0,55**	**0,58**	**0,56**
Movidas-->MM3				
MNB	**0,6**	0,51	0,47	0,47
Beto	0,59	**0,57**	**0,59**	**0,57**
Movida-->MM4				
MNB	**0,47**	**0,59**	0,4	**0,42**
Beto	0,39	0,37	**0,49**	0,37
Movidas-->MM5				
KNN	**0,52**	**0,55**	0,42	**0,44**
Beto	0,38	0,37	**0,46**	0,36

Es interesante destacar, al menos en este caso, que SVM_lineal con menor cantidad de información deja de ser tan efectivo frente a los otros algoritmos más simples de implementar e incluso frente a BETO. Con todo lo anterior, los valores de F1, en general, son menores a los obtenidos previamente, excepto para MM2 con BETO. De este modo, surge una configuración de clasificación que podría resumirse del siguiente modo (ver Tabla 7), de acuerdo a los mejores valores de F1. Así, la mejor representación para la mayoría de los algoritmos de clasificación es a través de unigramas y bigramas de lemas, así como trigramas en el caso de MM5. También es recurrente la vectorización de TFIDF sin considerar *stop words*, permitiendo la mejor normalización de las ocurrencias de las agrupaciones probabilísticas de lemas.

Tabla 7. Mejores valores de F1 para macromovida.

Clase	Algoritmo	F1	Representación	n-gramas	Vectorizador	tipo tarea
MM1	SVM_lineal	0,47	lema	n12	tfidf+sw	MM juntas
MM2	BETO	0,56	wordpiece	N/A	vector de palabras	MM separadas
MM3	SVM_lineal	0,71	lema	n12	tfidf+sw	MM juntas
MM4	SVM_lineal	0,81	lema	n12	tfidf+sw	MM juntas
MM5	KNN	0,44	lema	n13	tfidf+sw	MM separadas

En cuanto a los algoritmos, se observa que SVM_lineal es el que presenta mejor rendimiento para tres de las cinco MMs cuando se consideran todas las MMs en la tarea. Los valores de F1 para MM4 y MM3 son muy destacables. Esto es relevante, pues son las clases que mayor cantidad de oraciones tiene, por lo que el algoritmo tiene la posibilidad de aprender a distinguirlas mejor. Asimismo, son MMs que se distinguen claramente por su contenido y función retórico-discursiva. En este sentido, cabe destacar que 'Presentar antecedentes y procedimientos' y 'Reportar resultados' constituyen los propósitos centrales de este tipo de mesogénero, por lo que la efectividad del algoritmo es un buen sustento para afirmar que es posible vincular la representación computacional del texto con su función discursiva nuclear. Por otra parte, KNN resultó ser el mejor algoritmo para la clasificación de las movidas de la MM5 'Finalizar discursivamente la experiencia práctica'. Lo mismo ocurrió para la clasificación de las movidas de MM2 con BETO. Cabe destacar, que este resultado es muy bueno comparado con experiencias similares (Venegas, 2015, 2020). De ello se desprende que si no se conocen las macromovidas del texto a clasificar la configuración con SVM_lineal es la más adecuada. Ahora si el caso es que sí se sabe de qué macromovida se trata BETO y KNN son mejores clasificadores para distinguir las movidas de MM2 y MM5, respectivamente.

Por último, con BETO se obtienen en general mejores resultados de exhaustividad, 23% de mejora sobre SVM_lineal en la recuperación de ejemplos relevantes, esto es, que efectivamente corresponden a las clases indagadas.

CONCLUSIONES

En esta investigación, comparamos diversas configuraciones de algoritmos tradicionales y de aprendizaje profundo, con el objetivo de clasificar de manera automática las macromovidas y movidas del mesogénero informe de experiencia práctica en ingeniería. Así, combinamos la perspectiva del análisis del género y la IA a través del PLN, usando el aprendizaje de máquinas tradicionales, y una de las aproximaciones más actuales en el ámbito del aprendizaje profundo para el español, BETO. De este modo, el desafío se focalizó en identificar la mejor representación

lingüística y de vectorización para entrenar tales algoritmos e identificar los propósitos comunicativos del mesogénero MGIEP. Los resultados nos permiten establecer que entre los clasificadores tradicionales, los que mejor rendimiento ofrecen en términos de macro-F1 son los probabilísticos (MNB y COMP NB) y SVM_lineal. Este último destaca con la mejor exactitud 75,09% y macro-F1(0,56) para la clasificación de todas las MMs juntas (12568 oraciones). Este valor es muy satisfactorio. Si bien es menor al reportado para otros géneros (Venegas, 2015; Cotos & Pendar, 2016; Venegas, 2020) no es completamente comparable, dado que las lenguas, las clases y la cantidad de datos varía grandemente. Por su parte, con BETO no se pudo mejorar estos valores. Esto podría deberse a una heterogeneidad genérica en los informes. Con el fin de indagar más en el problema, se abordó este en términos de la clasificación de las movidas de cada MM separadamente. En esta situación, destacamos el buen rendimiento de BETO para la MM1, MM2 y MM3 y KNN para la MM5. Ello implica que si las macromovidas son analizadas separadamente de acuerdo con las movidas, estos clasificadores se convierten en una buena opción para asignar a cada clase las oraciones que corresponden.

En general, ninguno de los algoritmos es mejor que otros para clasificar todas las MMs, probablemente debido a las variabilidad introducida por los tipos de informes, la disciplina o el desbalance de las clases. Así, SVM_lineal se destaca como el mejor modelo de clasificación para la clasificación de las MMs, representando las oraciones como unigramas y bigramas de lemas vectorizados con TFIDF sin considerar las palabras funcionales, al obtener valores de precisión altos en general, aunque menor exhaustividad que BETO. Al respecto, BETO destaca por tener una alta exhaustividad general, ello asegura que las oraciones que se identifiquen, si bien no sean todas las que corresponden, estas serán las correctas. Además, BETO permite identificar las movidas de MM2, una de las más difíciles de clasificar, con el mayor valor de F1 (0,56). Finalmente, KNN se destaca por permitir un mejor valor de F1 (0,44) para la clasificación de las movidas de MM5, utilizando uni, bi y trigramas de lemas y TFIDF .

Todo lo anterior, sugiere que para clasificar una oración producida en un informe correspondiente al MGIEP, que permite la consecución de un propósito comunicativo determinado, la mejor solución es combinar algoritmos tradicionales y de aprendizaje profundo. Estos resultados son muy útiles, sobre todo por la identificación de la mejor representación lingüística (trigrama de lema), lo que confirma investigaciones previas similares (Venegas, 2020). No obstante, se observan desafíos para la implementación de estos clasificadores en un sistema real de retroalimentación, por ejemplo. Ello debido a que computacionalmente requieren un alto grado de parametrización, de recursos de procesamiento memoria y de tiempo de respuesta. Estas restricciones favorecerían a las variantes NB, dada su menor complejidad (Venegas, 2007).

A modo de proyección, nos proponemos evaluar en nuestro sistema de apoyo a la escritura Thot (http://www.redilegra.com/Frontend/#/) estos modelos con el fin de

proveer retroalimentación basada en el género para la escritura de estudiantes de ingeniería. Además, proyectamos, incorporar mayor cantidad de corpus anotado, así como información del contexto en el que ocurren las movidas. Asimismo, consideraremos para las movidas y macromovidas más difíciles de detectar la adopción de un conjunto de reglas para apoyar su reconocimiento automático.

REFERENCIAS BIBLIOGRÁFICAS

Alfaro R. & Allende, H. (2020). Clasificación de textos multi-etiquetados con modelo Bernoulli multi-variado y representación dependiente de la etiqueta. *Revista Signos. Estudios de Lingüística*, *53*(104), 549-567.

Amieva, R. L. (2001). Elaboración de informes en la enseñanza de la ingeniería. Facultad de Ingeniería. Gabinete de Asesoramiento Pedagógico. Universidad Nacional de Río Cuarto en línea]. Disponible en: http://www.ehu.eus/pat/compe/lanak/elaboracion_de_informes_en_la_ensennanza_de_la_ingenieria_informes_lengua%20y%20contenido.pdf

Ávila, N. & Cortés, A. M. (2017). El género informe de caso en la formación inicial docente: Una aproximación basada en la actividad. *Lenguas Modernas*, 50, 153-174

Bhatia, V. (1993). *Analysing genre: Language use in professional settings*. Londres: Longman.

Bhatia, V. (2002). Professional discourse: Towards a multidimensional approach and shared practice. En C. Candlin (Ed.), *Research and practice in professional discourse* (pp. 39-60). Hong Kong: City University of Hong Kong Press.

Bosio, V. (2018). ¿Podemos mejorar la calidad de la escritura en el posgrado? Algunas respuestas a partir de un proceso de investigación-acción. *Revista Brasileira de Lingüística Aplicada* [en línea]. Disponible en: https://dx.doi.org/10.1590/1984-6398201812959

Buigas, J. (2017). *Guía rápida de inteligencia artificial. Así funciona Deep Learning* [en línea]. Disponible en: https://puentesdigitales.com/2017/11/15/guia-rapida-de-inteligencia-artificial-asi-funciona-el-deep-learning/

Cañete, J., Chaperon, G., Fuentes, R. Ho, J-H, Kang, H. & Pérez, J. (2020). Spanish pre-trained Bert model and evaluation data. Ponencia presentada en el workshop PML4DC, ICLR 2020 [en línea]. Disponible en: https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf

Cassany, D. & López, C. (2010). De la Universidad al mundo laboral: Continuidad y contraste entre las prácticas letradas académicas y profesionales. En G. Parodi (Ed.), *Alfabetización académica y profesional en el siglo XXI: Leer y escribir desde las disciplinas* (pp. 347-374). Santiago: Ariel.

Castro, M. C., Hernández, L. A. & Sánchez, M. (2010). El ensayo como género académico: Una aproximación a las prácticas de escritura en la universidad pública mexicana. En G. Parodi (Ed.), *Alfabetización académica y profesional en el siglo XXI: Leer y escribir desde las disciplinas* (pp. 49-79). Santiago: Ariel.

Cotos, E. & Pendar, N. (2016). Discourse classification into rhetorical functions for AWE feedback. *Calico Journal*, *33*(1), 92-116.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding [en línea]. Disponible en: https://arxiv.org/pdf/1810.04805v2.pdf

Espejo, C. (2006). La movida concluyendo en torno al tema en informes de investigación elaborados por estudiantes universitarios. *Onomázein*, *13*(1), 35-54.

Fang, Ch. & Cao, J. (2015). *Text genres and registers: The computation of linguistic features*. Berlín: Springer.

Figuerola, C., Zazo, A. & Berrocal, J. (2000). *Categorización automática de documentos en español: Algunos resultados experimentales* [en línea]. Disponible en: http://imhotep.unizar.es/jbidi/jbidi2000/14_2000.pdf

Gallardo, S. (2005). La monografía universitaria como aprendizaje para la producción de artículos científicos. En G. Vázquez (Coord.), Español con fines académicos: de la comprensión a la producción de textos (pp. 13-28). Madrid: Edinumen.

García, A. (2017). *Inteligencia artificial. Fundamentios prácticas y aplicaciones*. Ciudad de México: Alfaomega.

Gardner, S. (2012). Genres and registers of student report writing: An SFL perspective on texts and practices. *Journal of English for Academic Purposes*, *11*(1), 52-63.

Greenbaum, S. (1991). ICE: The International Corpus of English. *English Today*, 7, 3-7.

Guyon, I., Boser, B. & Vapnik, V. (1993). Automatic capacity tuning of very large VC dimension classifiers. En S. J. Hanson, J. Cowan & C. L. Giles (Eds.), *Advances in Neural Information Processing Systems* (pp. 147-154). San Mateo CA: Morgan Kaufmann.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1999). *Análisis multivariante*. Madrid: Prentice Hall.

Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. Tesis de Doctorado, Universidad de Waikato, Hamilton, Nueva Zelanda [en línea]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.3848&rep=rep1&type=pdf

Halliday, M. A. K. (2004). *An introduction to functional grammar*. Londres: Arnold.

Harvey, A. (2005). La evaluación en el discurso de informes escritos por estudiantes universitarios chilenos. En M. Pilleux (Ed.), *Contextos del Discurso* (pp. 215-228). Valdivia: Universidad Austral de Chile.

Harvey, A. & Muñoz, D. (2006). El género informe y sus representaciones en el discurso de los académicos. *Estudios Filológicos*, 41, 95-114.

Hyland, K. (2009). *Academic discourse*. Londres: Continuum.

Hyland, K. & Paltridge, B. (2011). *Continuum companion to discourse analysis*. Londres: Continuum.

Ikonomakis, M., Kotsiantis, S. & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, *4*(8), 966-974.

Jarpa, M. (2016). La resolución de problema como género académico evaluativo: Organización retórica y uso de artefactos multimodales. *Revista Signos. Estudios de Lingüística*, *49*(2), 350-376.

Jerez, A. (2018). Análisis del PageRank como factor de peso en la clasificación automática de textos. Tesis Informe de Proyecto de Título Ingeniería Civil Informática. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Jurafsky, D. & Martin, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Nueva Jersey: Prentice-Hall.

Kessler, B., Numberg, G. & Schütze, H. (1997). Automatic detection of text genre. Actas del 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 32-38.

Lillo, F. (2016). Clasificación semiautomática de movidas retóricas en trabajos finales de grado a partir de lemas. Tesis de Licenciatura, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Manning, Ch., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Martin, J. & Rose, D. (2008). *Genre relations: Mapping culture*. Londres: Equinox.

McCarthy, P. M. & McNamara, D. S. (2007). Are seven words all we need? Recognizing genre at the sub-sentential level. En D. S. McNamara & G. Trafton (Eds.), *Actas del 29th Annual Conference of the Cognitive Science Society* (pp. 1295-1300). Austin, TX: Cognitive Science Society.

McKenna, B. (1997). How engineers write: An empirical study of engineering report writing. *Applied Linguistics*, *18*(2), 189-211.

Montolío, E. (2010). Mejorar las recomendaciones contenidas en los informes elaborados por consultores. La optimización del discurso. *Onomazéin*, *21*(1), 237-253.

Montolío, E. & López, A. (2010). Especificidades discursivas de los textos profesionales frente a los textos académicos: El caso de la recomendación profesional. En G. Parodi (Ed.), *Alfabetización académica y profesional en el siglo XXI: Leer y escribir desde las disciplinas* (pp. 215-245). Santiago: Ariel.

Muñoz, D. (2006). Estructura y patrones léxicos en informes escritos de estudiantes universitarios. *Onomázein*, *13*(1), 55-71.

Muñoz, G. (2019). Caracterización retórico-discursiva de informes académicos de Ingeniería Civil Informática de la Pontificia Universidad Católica de Valparaíso. Tesis de Licenciatura, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Narváez-Cardona, E. (2016). Latin-American writing initiatives in Engineering from Spanish-speaking countries. *Ilha do Desterro*, *69*(3), 223-248.

Navarro, F. (2014). Géneros discursivos e ingreso a las culturas disciplinares. Aportes para una didáctica de la lectura y escritura en educación superior. En F. Navarro (Ed.), *Manual de escritura para carreras de humanidades* (pp. 29-52). Buenos Aires: Editorial de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires.

Navarro, F. (2017). De la alfabetización académica a la alfabetización disciplinar. En R. Ibáñez & C. González (Eds.), *Alfabetización disciplinar en formación inicial docente. Leer y escribir para aprender* (pp.7-15). Valparaíso: Ediciones Universitarias de Valparaíso.

Navarro, F. (2018). Más allá de la alfabetización académica: Las funciones de la escritura en educación superior. En M. Alves & V. Iensen Bortoluzzi (Eds.*), Formação de Professores: Ensino, linguagens e tecnologia*s (pp. 13-49). Porto Alegre: Editora Fi.

Nesi, H. & Gardner, S. (2012). Familie of genres of assesed writing. En H. Nesi & S. Gardner (Eds.), *Genres across the disciplines: Student writing in Higher education* (pp. 21-56). Cambridge: Cambridge University Press.

Parodi, G. (2008) (Ed.) *Géneros Académicos y géneros profesionales: Accesos discursivos para saber y hacer*. Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. (2010). Academic and professional genre variation across four disciplines: Exploring the PUCV 2006 Corpus of written Spanish. *Linguagem em (Dis)curso*, *10*(3), 535-567.

Parodi, G. & Burdiles, G. (2015). Encapsulación y tipos de coherencia referencial y relacional: El pronombre 'ello' como mecanismo encapsulador en el discurso escrito de la economía. *Onomázein, 33*(1), 107-129.

Parodi, G., Ibáñez, R. & Venegas, R. (2009). El Corpus PUCV-2006 del Español: Identificación y definición de los géneros discursivos académicos y profesionales. *Literatura y Lingüística*, 20, 71-101.

Parodi, G., Ibáñez, R. & Venegas, R. (2010). Discourse genres in PUCV-2006 Corpus of Academic and Professional Spanish: Criteria, definitions and examples. En G. Parodi (Ed.), *Discourse genres in Spanish: Academic and professional connections* (pp. 39-68). Ámsterdam: John Benjamins.

Parodi, G., Boudon, E. & Julio, C. (2014). El Manual de Economía: Género entre dos mundos disciplinares. En G. Parodi & G. Burdiles (Eds.), *Leer y escribir en contextos académicos y profesionales. Géneros, corpus y métodos*. Barcelona: Ariel.

Parodi, G., Julio, C. & Vásquez-Rocca, L. (2015). Los géneros del Corpus PUCV-UCSC-2013 del discurso académico de la economía: El caso del Informe de Política Monetaria. *Revista ALED*, *15*(3), 179-200.

Pérez, S. (2017). Análisis y clasificación de textos con técnicas semi supervisadas aplicado a área atención al cliente. Tesis de Ingeniería Civil Informática. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Poe, M., Lerner, N. & Craig, J. (2010) *Learning to communicate in Science and Engineering case studies from MIT*. Cambridge: MITY Press.

Reave, L. (2004). Technical communication instruction in engineering schools. A survey of top-ranked U.S. and Canadian programs. *Journal of Business and Technical Communication, 8*(4), 452-490.

Reuter, Y. (2000). *La description: Des théories à l'enseignement-apprentissage*. París: ESF éditeur.

Rokach, L. & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. Nueva Jersey: World Scientific.

Sebastiani, F. (2002). *Machine learning in automated text categorization. ACM Computing Surveys, 34*(1), 1-47.

Sharma, S. (1996). *Applied multivariate techniques*. Nueva York: Wiley.

Smola, A. & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199-222.

Sologuren, E. (2020a). Géneros de formación en el ciclo capstone de Ingeniería Civil Informática: Exploraciones al curriculum. *Revista de estudios y experiencias en educación*, *19*(41), 167-198.

Sologuren, E. (2020b). Prácticas de escritura en la universidad: géneros de formación académica en la carrera de ingeniería civil informática. Tesis de Doctorado en Lingüística, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Sologuren, E. (2021). Escritura académica en tres escuelas de ingeniería chilenas: La familia 'Informe técnico' como macrogénero discursivo en el área de Ingeniería Civil Informática. *Educatio Siglo XXI*, *39*(1), 107-130.

Sologuren, E. & Castillo, M. (2019). La construcción del Ethos en informes de laboratorio producidos por estudiantes universitarios: contrastes en el discurso académico en español. *Letras de Hoje*, *54*(3), 369-384.

Swales, J. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.

Tapia, M., Burdiles, G. & Arancibia, B. (2003). Aplicación de una pauta diseñada para evaluar informes académicos universitarios. *Revista Signos. Estudios de Lingüística*, *36*(54), 249-257.

Tapia, M. & Burdiles, G. (2009). Una caracterización del género informe escrito. *Letras*, *51*(78), 17-49.

TechTarget (2021). *Aprendizaje profundo (Deep learning)* [en línea]. Disponible en: *https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-profundo-deep-learning*

Turc,I. Chang, M.-W., Lee, K. & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv:1908.08962v2* [en línea]. Disponible en: https://arxiv.org/pdf/1908.08962.pdf

Vapnick, V. (2000). *The nature of statistical learning theory*. Nueva York: Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing* systems, 5998-6008. [en línea]. Disponible en: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Venegas, R. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista Signos. Estudios de Lingüística*, *40*(63), 239-271.

Venegas, R. (2014). Clasificación automatizada de macromovidas en el género TFG en base a patrones léxico- gramaticales y léxico- semánticos. Ponencia presentada en el congreso RITERM, Santiago, Chile.

Venegas, R. (2015). Clasificación automatizada de macrovidas en el género TFG a partir de patrones léxico-gramaticales y léxico-semánticos. En L. Ruiz, A. Muñoz, M. R. Álvarez, Y. Pérez & D. Jackson (Eds.), *Comunicación Social: Retos y perspectivas* (pp. 614- 619). Santiago de Cuba: Centro de Lingüística Aplicada.

Venegas, R. (2020). Clasificación automatizada de macromovidas discursivas en el género tesis: Escritura académica y aprendizaje de máquinas. En C. Zapata & B. Manríquez (Coords.), *Tecnologías del lenguaje Humano: Aplicaciones desde la lingüística computacional y de corpus* (pp. 93-105). Medellín: Editorial de la Universidad de Medellín.

Venegas, R. & Valdés, M. (2021). Evidencias léxico-gramaticales de inserción disciplinar en informes de Ingeniería Civil Informática. *Boletín de Filología*, 1097-1114.

Walker, K. (1999). Using genre theory to teach students engineering lab report writing: A collaborative approach. *IEEE transactions on professional communication*, *42*(1), 12-19.

Wennerstrom, A. (2003). *Discourse analysis in the classroom. Volume 2. Genres of writing*. Ann Arbor: University of Michigan Press.

Windsor, D. (1996). *Writing like and engineer: A rhetorical education*. Londres: Routledge.

Wu, Sh. & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv:1904.09077* [en línea]. Disponible en: https://arxiv.org/pdf/1904.09077.pdf

Zamora, S. (2014). Clasificación de las movidas retóricas de la macromovida Introducir al Lector en Trabajos Finales de Grado a partir de rasgos léxico-gramaticales y léxico-semánticos. Tesis de Magíster en Lingüística Aplicada, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Zhang, C. & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration,* 23 [en línea]. Disponible en: https://doi.org/10.1016/j.jii.20

[1]AGRADECIMIENTO

El autor expresa su agradecimiento a todos los asistentes técnicos y tesistas del proyecto Fondecyt 1190639, en particular, a Sebastián Rodríguez y Esteban Mohr por la colaboración que brindaron en la etapa de clasificación automática. Este artículo contó con el financiamiento parcial de FONDECYT REGULAR 1190639, el apoyo del Núcleo de Investigación en Procesamiento del Lenguaje Natural Aplicado #NiPLNA (https://bit.ly/3hOUKMn) y REDILEGRA (www.redilegra.com).