PONTIFICIA UNIVERSIDAD
CATOLICA
DE VALPARAISO

INSTITUTO DE LITERATURA Y CIENCIAS DEL LENGUAJE

FACULTAD DE FILOSOFÍA Y EDUCACIÓN

# CLASIFICACIÓN SEMIAUTOMÁTICA DE MOVIDAS RETÓRICAS EN TRABAJOS FINALES DE GRADO A PARTIR DE LEMAS

Tesis para optar al grado de Licenciado en Lengua y Literatura Hispánica

**Alumno: Fernando Lillo Fuentes**

**Becario Proyecto FONDECYT 1140967**

**Profesor Guía: Dr. René Venegas Velásquez**

**Viña del Mar, Chile**
**2016**

# ÍNDICE

## ÍNDICE DE FIGURAS, TABLAS Y GRÁFICOS

### FIGURAS



### TABLAS

**GRÁFICOS**

# 1. INTRODUCCIÓN

A partir de los años 70, en el campo de la lingüística, se comenzaron a realizar variadas investigaciones que tuvieron como objeto de estudio al texto. Esto llevó a que diversos autores (Sandig, 1972; Werlich, 1975; Grosse, 1976; Longacre & Levinsohn, 1978**;** Bajtín, 1979; Isenberg, 1987; Biber, 1985, 1986; Posner & Gülich, 1986**;** Brinker, 1988; Adam, 1991, entre otros) indagaran en este campo y proporcionaran importantes contribuciones, como las denominadas tipologías textuales. En la actualidad, debido a la gran cantidad de documentos que circulan en formato digital y a la necesidad de organizarlos, se han realizado numerosos estudios que, a partir de ciertas tipologías textuales, han tenido como resultado una clasificación de textos más rápida y eficiente. La unión de dos disciplinas, informática y lingüística, ha permitido realizar tareas de clasificación automática de textos que permiten complementar los trabajos de tipologización manual, a partir del uso de una gran cantidad de información lingüística.

Si bien, en los últimos 10 años se han realizado variadas investigaciones en cuanto a la clasificación automática o semiautomática de textos, muy pocas de ellas se han hecho en español y la mayoría se ha centrado en tareas como Filtrado de correos electrónicos, clasificación de documentos digitales y recuperación de información, entre otras. A su vez, todos estos estudios se han realizado con textos específicos sin considerar los textos académicos a pesar de la importancia que han tenido en el último tiempo debido a su propósito; comunicar el conocimiento al interior de cada comunidad discursiva (Parodi, 2005). En este contexto, se han realizado escasas o nulas clasificaciones de los textos académicos, pues la mayoría de los estudios de estos textos se han orientado en la descripción de estos en términos estructurales y funcionales.

Gran parte de las indagaciones realizadas en torno a los textos académicos se han desarrollado en lengua inglesa y se han enfocado solo en algunos de los textos que componen el discurso académico, principalmente en los Artículos de Investigación Científica (AIC) (Gerbert, 1970; Salager-Meyer, 1994; Conrad, 1996; Bolivar, 2000; Hyland, 2002; Martín, 2003; Gotti, 2003) y en Tesis en niveles de Magíster y Doctorado (Hyland, 2008; Paltridge, 2002; Samraj, 2005; Thompson, 2005; Jara, 2009; Martínez, 2012; Zamora & Venegas, 2013).

Estos tres géneros no son los únicos que cumplen con el propósito mencionado, sino que existen otros, inscritos en el denominado macrogénero Trabajo Final de Grado (Venegas, 2010) que no han sido abordados en profundidad y que cumplen el mismo objetivo. Un ejemplo de estos, es el género Tesis de Licenciatura, el cual no ha sido descrito ni estudiado en profundidad (Venegas, 2014; Zamora, 2014) dado que la mayoría de las indagaciones en torno a la tesis se han enfocado en niveles de postgrado. Este género, entendido como un documento académico escrito que busca informar y acreditar méritos para la obtención de un título mediante los resultados y conclusiones de un trabajo de investigación (Moyano, 2000) ha obtenido gran importancia en el último tiempo debido a que es considerado un género de transición entre la vida académica y la profesional en diversas comunidades discursivas.

Una de las aristas en las que se ha investigado el discurso académico, en especial el género tesis, es en el análisis retórico-discursivo, entendido como aquel que sistematiza las funciones que se cumplen en los textos y el modo en que estas funciones se instancian en la superficie textual (Swales, 1990; Zamora, 2014; Venegas, Zamora & Galdames, 2016). Una de sus perspectivas más estudiadas es el Análisis de Género, el cual propone una forma de sistematización de los propósitos comunicativos que se cumplen en un texto, entendiéndolos como una serie de movimientos funcionales que se instancian a lo largo de un texto en términos de movidas y pasos retóricos (Swales, 1990; 2004). Respecto a estas unidades, algunos autores proponen que existe una unidad de mayor abstracción que la movida denominada macromovida, la cual recoge los macropropósitos generales que rigen la construcción de un texto (Parodi, 2010).

Tomando en cuenta el análisis retórico-discursivo, diversas investigación se han centrado en analizar y describir las macromovidas, movidas y pasos retóricos que caracterizan las tesis y sus apartados (Samraj, 2008; Jara, 2009; Silva, 2011; Soler, Carbonell & Gil, 2011; Martínez, 2012; Rodríguez, 2012; Tapia & Burdiles, 2013; León, 2014; Romero, 2015; Venegas, Zamora & Galdames, 2016). Sin embargo, a pesar de estas investigaciones, muy pocos estudios se han centrado en estudiar los rasgos léxicos que se presentan en las tesis y aún menos en la clasificación de estos géneros a partir de estos rasgos. Debido a lo anterior, el foco de esta investigación se centrará en responder a la pregunta: ¿Cuáles son los rasgos léxicos que permiten clasificar semiautomáticamente las movidas retóricas de las tesis de Licenciatura?

Tradicionalmente las clasificaciones de textos se han realizado a partir de las palabras o conjuntos de palabras, sin embargo, en la presente investigación se utilizarán los lemas para realizar dicha tarea. Al respecto, Sebastiani (2002) menciona que la lematización permite eliminar todo tipo de distractor que posean los documentos, ya que solo deja la forma canónica o lema de cada palabra, eliminando algunos de sus afijos asociados. Tomando en cuenta esta propuesta, en la presente investigación, se tendrá por objetivo clasificar semiautomáticamente las movidas retóricas de las tesis en el ámbito de Lingüística de la Licenciatura en Lengua y Literatura de la PUCV a partir de los lemas presentes en ellas. Para esto, se utilizará el subcorpus llamado TLing compuesto por 20 TFG de esta disciplina. Este corpus pertenece al proyecto FONDECYT 1140967 y corresponde a una muestra representativa de las Tesis realizadas entre los años 2009 y 2012 por los estudiantes de esta licenciatura.

Para cumplir con el objetivo descrito, esta investigación ha sido estructurada en cinco apartados distribuidos de la siguiente manera. En el apartado 2, se presentarán los conceptos y nociones básicas que sustentan esta tesis. En el número 3, se expondrán los sustentos metodológicos que permiten llevar a cabo este trabajo. En el apartado 4, se darán a conocer los resultados de la investigación, algunas comparaciones relevantes y una breve discusión a raíz de los resultados obtenidos. Finalmente, en el apartado 5, se dan a conocer las conclusiones que han surgido a partir del desarrollo de esta tesis, sus hallazgos, limitaciones y algunas de las proyecciones propuestas con el fin de cerrar discursivamente la presente investigación.

## 2. MARCO TEÓRICO

En este capítulo, se presentarán algunos conceptos teóricos fundamentales para comprender nuestra investigación. De este modo, en primer lugar abordaremos el concepto de discurso académico. Asociado a este, los conceptos de comunidad discursiva y género académico. Luego, se presentará la conceptualización de macrogénero trabajo final de grado y género tesis de licenciatura. Finalmente, se abordará la clasificación automática de textos y se presentarán dos algoritmos de clasificación de textos que serán utilizados en este estudio.

### 2.1.Discurso académico

Antes de definir discurso académico, debemos referirnos al discurso especializado, pues desde la perspectiva de Parodi (2005) entenderemos que el discurso académico es considerado un hipónimo del discurso especializado. Para Parodi (2005) este discurso está compuesto por un *continuum* de textos que van desde unos altamente especializados hasta otros más divulgativos y generales. Este autor, tomando como base los estudios de las lenguas para propósitos específicos (Schröder, 1991) define al discurso especializado de la siguiente manera:

> "(…) por una parte, un conjunto de textos que se distinguen y se agrupan por una co-ocurrencia sistemática de rasgos lingüísticos particulares en torno a temáticas específicas no cotidianas en los cuales se exige experiencia previa disciplinar de sus participantes (formación especializada dentro de un dominio conceptual particular de la ciencia y de la tecnología); por otra, son textos que revelan predominantemente una función comunicativa referencial y circulan en contextos situacionales particulares; todo ello implica que sus múltiples rasgos se articulan en singulares sistemas semióticos complejos y no de manera aislada y simple." (Parodi, 2005: 26).

Para Gotti (2008) el discurso especializado es el conjunto de textos que son reflejo de una práctica social que utiliza una lengua especializada, típica de una comunidad perteneciente al mundo académico, profesional y/o ocupacional. Desde la conceptualización de este autor, se desprende que uno de los ámbitos en los que se encuentra el discurso especializado es el contexto académico. Si bien, el inicio de las investigaciones del discurso académico se halla en

los estudios del inglés para propósitos específicos (Bhatia, 2002; Silver, 2006), el deslinde de Gotti (2008) proporciona una importante contribución a estos estudios.

Espejo (2006) establece que el discurso académico corresponde a todas las producciones lingüísticas elaboradas por los distintos miembros de ciertas comunidades discursivas con el fin de generar, transmitir o reproducir conocimiento científico. Por su parte, Silver (2006) expone que el discurso académico se constituye como tal debido a todas las prácticas discursivas que circulan en el ámbito académico, a saber: manuales, trabajos, exámenes, memorias, informes, tesis, etc. Al respecto, Cassany, Luna y Sanz (2000) mencionan que todo discurso académico posee un léxico preciso y específico y utiliza un registro formal de la lengua. En palabras de Zamora (2014) "el discurso académico se instancia en el uso real del lenguaje por medio del desarrollo de textos académicos, los cuales, según sus características, se agrupan en géneros que evidencian los rasgos compartidos en cada contexto" (Zamora, 2014:19).

En síntesis, el discurso académico circula en contextos académicos y universitarios, en los diferentes niveles y áreas del conocimiento (Bhatia, 2002; Espejo, 2006; Silver, 2006). Este discurso se realiza en el uso real del lenguaje en textos académicos, los que se agruparán según sus características en géneros discursivos que comparten rasgos en cada contexto. Según Bathia (2002) estos rasgos contextuales posibilitarán la variación de los textos de acuerdo a las funciones específicas que cumplirán, las cuales son determinadas por las disciplinas o comunidades académicas en las que circulan.

### 2.2.Comunidad discursiva

El término comunidad discursiva se relaciona directamente con discurso académico y especializado, puesto que cada comunidad comunica sus conocimientos de una manera determinada y específica. En esta investigación, se entenderá comunidad discursiva desde la perspectiva retórico-discursiva de Swales (1990), es decir, como un grupo socio-retórico de individuos que tienen propósitos particulares y que utilizan el lenguaje para concretar dichos propósitos. Según la propuesta de este autor, en cada comunidad existe un uso del lenguaje propio que permite distinguir a las comunidades, pues utilizarán un léxico especializado y estrategias retóricas y discursivas que caracterizan a cada una de ellas. A su vez, cada uno de

los integrantes de estas comunidades deberá dominar estos usos particulares para ingresar a ellas y convertirse en miembro.

Swales (1990: 24-25) señala que este término posee seis características, a saber:

> "1. Has a broadly agreed set of common public goals.
> 2. Has mechanisms of intercommunication among its members.
> 3. Uses its participatory mechanisms primarily to provide information and feedback.
> 4. Utilizes and hence possesses one or more genres in the communicative furtherance of its aims.
> 5. In addition to owning genres, a discourse community has acquired some specific lexis.
> 6. Has a threshold level of members with a suitable degree of relevant content and discourse expertise."

De estas características se puede desprender que cada miembro de la comunidad hará uso del lenguaje con el objetivo de construir y comunicar el conocimiento. Tanto los usos como los recursos que utilizan para cumplir con los propósitos establecidos son particulares de cada comunidad y permiten distinguirlas entre sí. Los objetivos que poseen estas comunidades son comunes y están previamente acordados, por lo que son intercomunicados entre sus participantes.

Si bien, la noción propuesta por Swales (1990) es la más utilizada y reconocida, existen otros autores que se han aproximado al concepto. Al respecto, Cassany (2009) propone que una comunidad discursiva es un grupo humano que comparte unas prácticas comunicativas particulares usando textos específicos para conseguir finalidades especificas entre autores y lectores con roles predeterminados. A través de los usos ya mencionados y de los textos propios que circulan en cada comunidad, los miembros de ellas comparten conocimientos específicos, habilidades cognitivas para procesarlos y una perspectiva de la realidad que es exclusiva del grupo del que forman parte. En síntesis, ser miembro de cada comunidad significa conocer los mecanismos de producción, transmisión y recepción del conocimiento, así como también los géneros y los usos que se hacen de estos en cada comunidad (Swales, 1990; Cassany, 2009).

Por su parte Parodi, Venegas, Ibáñez y Gutiérrez (2008: 48) plantean que una comunidad discursiva se compone "por un grupo de personas que comparten un conjunto de conocimientos, así como también las convenciones necesarias para interactuar discursivamente y compartir tales conocimientos". A su vez, exponen que a partir de la interacción entre los miembros de la comunidad, los sujetos que ingresan como aprendices se transforman en expertos asegurando la estabilidad de ella.

### 2.3.Géneros académicos

El concepto de género ha sido estudiado en diversos campos y ha tenido múltiples significados, por lo que se vuelve complejo obtener una única definición de él. Así, su origen deviene de la antigua Grecia, pues en dicha época Aristóteles (1998), en su Retórica, ya exponía que los discursos podían ser clasificados en tres grupos; los forenses o jurídicos, los deliberativos o políticos y los epidícticos o de ocasión. Para realizar esta clasificación, Aristóteles (1998) centra su atención en la práctica discursiva de su tiempo, tomando en cuenta sus finalidades, actantes, temas y formas verbales y no verbales de cada uno.

Con especto a la propuesta de Aristóteles (1998), se conocen algunos estudios que retomaron sus aportes (Quintiliano, 1942; Perelman y Olbrecht Tyteca, 2009), sin embargo, esta propuesta permaneció sin muchas variaciones hasta la irrupción de la Lingüística textual en los años setenta (Ciapuscio, 1994). Paralelamente, en esta época, el trabajo Bajtín (1999) comenzó a centrar su interés en los géneros discursivos al plantear que estos no se habían estudiado hasta entonces, ya que los estudios existentes solo se habían enfocado en ámbitos literarios y artísticos y no en determinados enunciados que se distinguen de otros debido a su naturaleza verbal común.

Desde este punto de vista, la teoría bajtiniana constituye la base de todos los estudios de géneros discursivos que se han desarrollado posteriormente. Así, Bajtín (1999) establece una relación directa entre las distintas esferas de la actividad humana y el uso de la lengua que se expresa en los diferentes enunciados, ya sean orales o escritos. De este modo, cada esfera del uso de la lengua elabora sus tipos relativamente estables de enunciados llamados géneros discursivos. A partir de lo anterior, se puede desprender que la riqueza de estos enunciados es enorme, pues

las esferas de la actividad humana no se agotan e incluso los géneros crecen a medida que estas se desarrollan (Bajtín, 1999).

A raíz de lo anterior, los géneros discursivos constituyen manifestaciones de carácter diverso y heterogéneo, cuya producción está directamente relacionada con el desarrollo de las esferas de la praxis o comunidades discursivas. Para Bajtín (1999) los géneros pueden separarse en primarios y secundarios, según su grado de complejidad. Los géneros primarios corresponden a aquellos que surgen de la esfera de la vida cotidiana, es decir, son enunciados reales que surgen de la comunicación inmediata, por ejemplo, una conversación. Mientras que los géneros secundarios, surgen en condiciones de comunicación cultural más compleja, más desarrollada y organizada, generalmente escrita, por ejemplo, investigaciones científicas.

Desde del Inglés con Propósitos Específicos, Swales (1990) lo define como el vehículo que permite el logro de los propósitos comunicativos. Menciona que involucra una comunidad de hablantes, pues es un evento comunicativo que posee propósitos comunicativos identificados y definidos por los miembros de la comunidad, quienes construyen sus géneros a partir de convenciones, formas y contenidos particulares de la comunidad de la que forman parte. En concordancia con Bajtín (1999), Swales (1990) expone que los géneros se identifican a partir de su forma, estructura, contenido y expectativas de la audiencia. Así, Swales (1990) afirma al respecto:

> "A genre comprises a class of communicative events, the members of wich share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rethorical action. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience." (Swales, 1990: 58).

Más tarde, desde una perspectiva psicosociodiscursiva Parodi (2008) entiende el género en función de tres dimensiones; Social, Lingüística y Cognitiva. Estas tres dimensiones resultan esenciales y están relacionadas, aunque el autor otorga especial relevancia a la dimensión cognitiva, pues menciona que ha sido descuidada en los estudios recientes. Desde este enfoque, Parodi (2008) define género discursivo de la siguiente manera:

> "(…) constituye una constelación de potencialidades de convenciones discursivas, sustentada por lo conocimientos previos de los hablantes/escritores y oyentes /lectores (almacenados en la memoria de cada sujeto), a partir de constricciones y parámetros contextuales, sociales y cognitivos…En su manifestación concreta, los géneros son variedades de una lengua que operan a través de un conjunto de rasgos lingüísticos-textuales co-ocurrentes sistemáticamente a través de las tramas de un texto y que se circunscriben lingüísticamente en virtud de propósitos comunicativos, participantes implicados (escritores y comprendedores), contextos de producción, ámbitos de uso, modos de organización discursiva, soportes y medios, etc." (Parodi, 2008:26).

En la definición anterior se entiende el género a partir de una concepción multidimensional e integral caracterizando el género desde tres ámbitos.

A partir de las teorías de género y la noción de discurso académico, surge el concepto de género académico (Bhatia, 2002a; 2002b; Parodi, 2008), entendido como el género o conjunto de géneros que pertenecen y caracterizan al discurso académico, en cada área disciplinar y comunidad discursiva en la que se desarrollen. Para Parodi (2008) los géneros del ámbito académico son presentados a través de un *continuum* por el cual el sujeto va avanzando desde los géneros menos especializados, como los escolares, hasta los más especializados, como los profesionales. Bhatia (2002a) relaciona el concepto de género académico con la variación del género según la disciplina en la cual circule y la comunidad discursiva en la que se desarrolle, ya que estas van a determinar las variaciones que se produzcan en el género y determinar su carácter académico.

## 2.4.Macrogénero trabajo final de grado

Para la obtención de un grado académico, ya sea de licenciatura, magister o doctorado, todo estudiante debe cumplir con el requisito de producir un texto que le permita obtener su grado académico y acceder a la comunidad discursiva a la que desea pertenecer. Al respecto, Arnoux (2006) señala que el trabajo final de grado (TFG) es un texto que permite a su autor, en este caso un estudiante, la entrada a una determinada comunidad discursiva demostrando, a través de una investigación científica, que cumplió con los requerimientos para ser miembro de ella. En esta misma línea, Venegas (2010) menciona que el TFG corresponde a una práctica discursiva clave en el paso de la vida estudiantil universitaria a la académico-científica, ya que se ha transformado en un rito de iniciación para el novato que ingresa a la nueva comunidad.

Desde la perspectiva de Venegas (2010), el TFG es entendido como un trabajo de investigación escrito de carácter evaluativo acreditativo, presentado por los estudiantes universitarios al término de sus estudios, como requisito para la obtención de un grado académico. Para Venegas (2010) el TFG se configura como un macrogénero discursivo, pues adquiere diversas formas textuales y también distintos nombres, según las diferentes comunidades discursivas en las que circule y se desarrolle. Con respecto al concepto de macrogénero, existen múltiples definiciones que permiten clasificar al TFG como uno de ellos. Dentro de las definiciones que destacan, encontramos la de Warta (1996), quien concibe el macrogénero como una unidad genérica de mayor jerarquía compuesta por otros géneros que se insertan o incrustan en su interior. Para Ibáñez (2008) el macrogénero es entendido como una unidad textual mayor, en cuyo interior se encuentran distintos textos que lo hacen funcionar, como por ejemplo, el género cuento en los textos escolares.

Respecto a este mismo concepto, García (2009), desde el marco del discurso profesional, señala que el macrogénero es una categoría abstracta que organiza el mapa de los géneros de un ámbito determinado. A su vez, este concepto agrupa un conjunto de géneros en función de los ámbitos de su uso particular dentro de la especialidad y su finalidad básica (García, 2009). Según la autora, lo que permite agrupar a un conjunto de géneros bajo un macrogénero es su propósito comunicativo compartido. Es interesante como esta última propuesta de la autora se condice

con la noción de colonias de géneros propuesta por Bhatia (2002a, 2002b, 2004), pues para este autor, colonias de género se define como "Una agrupación de géneros estrechamente relacionados que sirven a propósitos comunicativos ampliamente similares" (Bhatia, 2004: 59).

Respecto a esta última idea y las definiciones de macrogénero, Venegas, Zamora y Galdames (2016) proponen al "macrogénero trabajo final de grado (MGTFG) como una etiqueta genérica de un nivel de abstracción intermedio entre el género y la colonia de géneros. Este nivel genérico incluye la realización de distintos géneros (tesis, tesina, artículo, memoria, trabajo final, ensayo), orientados por un macropropósito evaluativo-acreditativo para la obtención de un grado académico determinado" (Venegas, Zamora & Galdames, 2016: 7).

### 2.5.Género Tesis de Licenciatura

Dentro de los géneros que circulan en el discurso académico, el género tesis es considerado uno de los más importantes, pues en el ámbito universitario, este género se ha vuelto una especie de ritual que el novato debe cumplir para acceder a su comunidad discursiva. Este género adquiere importancia en las diferentes comunidades, ya que se considera clave para demostrar el conocimiento que posee un estudiante al realizar una investigación y comunicarla por escrito (Koutsantoni, 2006). Su finalidad es informar y acreditar los méritos que posee el escritor a partir de los resultados de un trabajo de investigación (Moyano, 2000). Lo anterior se realiza con el fin de persuadir a una comunidad académica para recibir a un nuevo integrante. Al respecto, Meza (2013) propone que una característica central de este género es que los estudiantes deben ser capaces de glosar el discurso de otro, transformarlo y apropiarse de los modos, pues el nuevo integrante debe comunicar el conocimiento y ha de hacerlo de acuerdo a las normas o cánones establecidos por su disciplina.

Para Moyano (2000) la tesis se define como un documento académico que contiene los resultados de un trabajo de investigación generalmente complejo. Presenta la estructura de un artículo científico, aunque su revisión bibliográfica suele ser mucho más extensa y profunda. Sus destinatarios suelen ser investigadores en roles de docentes y su publicación es, generalmente, reducida, pues circula dentro de la institución en la que se desarrolló. Por su

parte, Sánchez (2012) menciona que la tesis es un texto de gran envergadura en el que el autor ha de formular problemas e hipótesis, contrastar estudios precedentes con su investigación, hilar los elementos expositivos con los argumentativos y componer la redacción final del propio texto de acuerdo a la comunidad académica en la que se realice.

Desde la Escuela Lingüística de Valparaíso (ELV) la tesis ha sido definida como un género discursivo inscrito en el denominado Macrogénero Trabajo Final de Grado cuyo macropropósito comunicativo es persuadir a un receptor acerca de un planteamiento teórico o ideológico (Parodi, Venegas, Ibáñez & Gutiérrez, 2008). Estas tesis pueden circular en diferentes niveles académicos, ya sea magister, doctorado o licenciatura. Que el estudiante cumpla con la escritura de su tesis y, los requisitos expuestos, es trascendental para su inserción en la comunidad, pues como se ha mencionado en párrafos anteriores, la escritura de la tesis se ha trasformado en un ritual que todo estudiante debe cumplir para obtener su grado académico (Venegas, 2010). Sin embargo, la escritura de la tesis no es solo un ritual con el que se debe cumplir, sino que también es "el camino para iniciarse en la cultura de la investigación" (Carlino, 2003:7).

En esta investigación, entenderemos el concepto de tesis de licenciatura desde la perspectiva de Tamola (2005) quien sostiene que es un trabajo escrito que cumple con la función de informar acerca del proceso y resultado de una investigación teórica o empírica con el fin de obtener el grado académico de licenciado. Este escrito será dirigido por un profesor guía, quien orientará, guiará y revisará el trabajo realizado por el alumno que tiene a su cargo, en este caso, un tesista. Esta misma autora menciona que las tesis variarán su grado de originalidad y complejidad dependiendo del contexto disciplinar y el grado académico en el que se desarrollen. A su vez, estas tesinas -en nuestro caso, tesis de licenciatura- no exigen al estudiante originalidad en sus conclusiones, sin embargo, requieren que la producción contenga la cantidad de información necesaria para ser sustentada, un modo de transmisión adecuado, adaptaciones al género y finalmente, el cumplimiento de las expectativas propias de la comunidad discursiva en la que se enmarca (Tamola, 2005).

En cuanto a la estructura del género tesis de licenciatura, desde la perspectiva de Paltridge (2002) se puede mencionar que la organización de las tesis en el ámbito de la Lingüística se asemejan a una de tipo tradicional simple, pues presentan todos los apartados que le corresponden a este tipo de estructura (Zamora & Venegas, 2013). Ahora bien, desde una perspectiva más funcionalista del lenguaje, estos trabajos pueden segmentarse mediante macromovidas (Zamora, 2014). Estas categorías funcionales relacionadas directamente con los propósitos comunicativos que se cumplen en las tesis, las entenderemos desde la perspectiva de Martínez (2012) quien entiende a la macrovida como una unidad retórica mayor a la movida y que permite realizar nuevas posibilidades de análisis retórico-funcionales. Así, en la presente investigación se clasificarán semiautomáticamente las movidas retóricas pertenecientes a las macromovidas Introducir al lector en la investigación (MM1) y Finalizar discursivamente la investigación (MM5).

### 2.6.Clasificación automática de textos

La necesidad de clasificar todo tipo de cosas, ordenar, jerarquizar y establecer tipos de objetos es intrínseca al ser humano (Ciaspucio, 1994) y se remonta a la antigua Grecia, pues en dicha época ya Aristóteles (1998) planteaba en su retórica una clasificación de los discursos que circulaban en sus tiempos. Para esto, el filósofo proponía atender a la naturaleza social de la práctica discursiva, distinguiendo criterios a partir del ámbito en el que se producen los discursos. En este sentido, mencionaba que para clasificarlos se debían tener en cuenta sus finalidades, actores, temas y formas verbales y no verbales propias de cada uno. Para Aristóteles (1998) la clasificación de estos discursos se podía realizar en 3 tipos, a saber: forenses o jurídicos, deliberativos o políticos y epidícticos o de ocasión.

Continuando con las ideas propuestas por Aristóteles (1998) más tarde, diversos autores (Sandig, 1972; Werlich, 1975; Grosse, 1976; Longacre & Levinsohn, 1978**,** 1987; Biber, 1985, 1986; Adam, 1991) continuaron con algunos de los postulados de él y contribuyeron significativamente en proponer diversas tipologías textuales. A partir de estas investigaciones, surgen diferentes propuestas de tipologizaciones, las que, generalmente, reflejan las concepciones lingüísticas dominantes de los momentos históricos en los que se originaron (Ciapuscio, 1994: 27). De esta manera, en los años setenta, Sandig (1972) propone una de las

primeras tipologizaciones de textos. En ella, tomaba en cuenta el esquema de acción socialmente normado de los textos no ficcionales, los grupos de usuarios de las clases textuales y las situaciones de utilización típicas para diferenciar unos textos de otros. Así, la autora utilizaba las oposiciones (+/-) de una serie de rasgos (hablado, espontáneo, monológico, tema preestablecido, primera persona, segunda persona, signos no verbales, etc.) para distinguir una clase textual de otra.

Más tarde, Grosse (1976) propuso una tipología de clases textuales que tuvo como base la tipologización homogénea. En este sentido, la propuesta de este autor fue una de las más serias (Ciapuscio, 1994) y tomó como parte de su tipología el concepto de función textual y función comunicativa de los textos para clasificarlos en ocho clases, a saber: Textos normativos, textos de contacto, textos que indican grupos, textos poéticos, textos en los que predomina la automanifestación, textos predominantemente exhortativos, clase de transición y textos en los que predomina la información. Según Grosse (1976) lo que determina la tipología de un texto no es solamente la función de él, sino, su función predominante.

Posteriormente, Werlich (1975) planteó una tipología de textos en el marco de la gramática textual del inglés. La base de esta tipología es la distinción de cinco textos básicos: descriptivos, narrativos, expositivos, argumentativos e instructivos. Esta tipología es considerada una de las mejores y la más fundamentada dentro de todas las propuestas realizadas en los años setenta (Ciapuscio, 1994). Más tarde, Longacre y Levinsohn (1978) proponen una tipología de textos basadas en el binarismo de dos ámbitos, el encadenamiento cronológico y la orientación hacia el agente. La combinación de estas características permitían distinguir entre cuatro tipos básicos de discursos, a saber: discurso narrativo, discurso procedural, discurso de conducta y discurso expositivo.

Por su parte, Biber (1985) menciona que para la investigación de la variación textual se deben tener en cuenta los análisis microscópicos y los macroscópicos. El primero de ellos, según el autor, brinda una descripción detallada de las funciones comunicativas de rasgos lingüísticos de textos determinados y el segundo, define dimensiones generales en la variación de un conjunto de textos. Para Biber (1985) las investigaciones lingüísticas centradas en el texto deben llevar a cabo estudios empíricos que identifiquen el conjunto de dimensiones textuales subyacentes que definen las similitudes y diferencias de los tipos de textos.

A partir de la propuesta de este último autor, podemos evidenciar que existe un paso desde la tipologización a la clasificación, pues los intentos realizados anteriormente para diferenciar y clasificar los textos no se concretan hasta que las investigaciones comienzan a tomar en cuenta los rasgos distintivos de cada texto para clasificarlo. Si bien, como se ha mostrado en párrafos anteriores, existen diversas propuestas de tipologizaciones y clasificaciones, la gran mayoría son realizadas por seres humanos, lo que significa un gasto de tiempo importante, además de imprimir en ellas algún nivel de subjetividad.

Debido a lo anterior y a los avances tecnológicos, en los últimos años se han comenzado a realizar clasificaciones de textos que a través de varios algoritmos computacionales permiten realizar esta tarea de manera más rápida y objetiva. Desde esta perspectiva, la clasificación automática de textos es un área que, a pesar del desarrollo que ha tenido la informática, aún no ha podido ser resuelta, pues posee una alta dimensionalidad y desbalance en las categorías que la hace mucho más compleja que, por ejemplo, una clasificación de imágenes (Cárdenas, Olivares & Alfaro, 2014). A partir de lo anterior, se hace necesaria la realización de una clasificación automática e interdisciplinar, pues utilizar los avances tecnológicos de la Informática y el conocimiento textual de la Lingüística puede aportar a realizar este proceso de manera automática y eficaz.

Generalmente, la clasificación automática ha estado ligada al desarrollo de las máquinas de aprendizaje, la inteligencia artificial y la inteligencia computacional, sin embargo, cada vez son más los ámbitos que utilizan estas clasificaciones con diversos fines. Desde la perspectiva de Sebastiani (2002) la clasificación automática de textos se entiende como la acción ejecutada por un sistema artificial sobre un conjunto N de elementos para ordenarlos en clases o categorías preestablecidas. Según Baeza y Ribeiro (1996) la clasificación automática de textos es un proceso supervisado, pues requiere de un conjunto de documentos previamente clasificados por expertos humanos para el entrenamiento del sistema.

Por su parte, Cárdenas, Olivares y Alfaro (2013) exponen que los clasificadores de textos son programas que utilizan algoritmos de clasificación para discriminar un texto de otro. Lo cual se realiza a partir variables previamente establecidas que permiten diferenciarlos unos de otros. En esta línea, Sebastiani (2002) sostiene que la clasificación automática de textos surge a partir de un preclasificado realizado por un conjunto de expertos que será utilizado como input para

generar un proceso más autónomo, menos subjetivo y con menos trabajo que el realizado por humanos. Al respecto Venegas (2007) concibe la clasificación automatizada como un proceso de aprendizaje matemático y estadístico, durante el cual un algoritmo implementado computacionalmente capta las características que distinguen cada categoría o clase de documentos de las demás, es decir, aquellas que deben poseer los documentos para pertenecer a esa categoría.

De esta manera, una clasificación se definirá como la acción ejecutada por un sistema artificial sobre un conjunto de elementos que serán ordenados en clases o categorías ya definidas. Un clasificador de textos utilizará conceptos aprendidos previamente para clasificar nuevos textos, es decir, agrupará N documentos en K grupos diferentes existentes. Los conceptos aprendidos previamente o características no indicarán de forma absoluta e inequívoca la pertenencia del texto a una clase o categoría en particular, sino más bien lo harán en función de una escala, graduación, probabilidad o porcentaje (Sebastiani, 2002; Venegas, 2007; Cárdenas, Olivares & Alfaro, 2014).

En cuanto a la ejecución de la tarea de clasificación, esta implica la recopilación de textos y la clasificación manual de ellos por un grupo de expertos. Luego, los documentos analizados deben ser extrapolados a una representación adecuada para aplicarles distintos tipos de algoritmos de clasificación que permitan obtener el clasificador. Para realizar la representación, generalmente, se suelen utilizar los modelos vectoriales para estas tareas, pues otorgan grandes beneficios y efectividad. Sin embargo, dependiendo de la tarea y el número de textos a clasificar, en algunas ocasiones se pueden utilizar clasificadores probabilísticos, debido a su óptimo funcionamiento con corpus numerosos (Baeza & Ribeiro, 1996; Sebastiani, 2002; Cárdenas, Olivares & Alfaro, 2014). En lo que sigue, profundizaremos en dos de estos algoritmos de clasificación, uno de tipo probabilístico, Bayesiano ingenuo, y otro de tipo vectorial llamado Máquina de soporte vectorial.

## 2.7.Clasificador Bayes Ingenuo

Este clasificador de tipo probabilístico estadístico es y ha sido ampliamente usado en la clasificación de textos debido a su sencillez, fácil uso y buenos resultados (Zhang, 2004). Puede predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra pertenezca a una clase particular (Malagón, 2003). Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados a partir de ejemplos clasificados previamente por grupos de expertos a modo de input (Zang, 2004; Malagón, 2013; Cárdenas, Olivares & Alfaro, 2014).

Algunos investigadores (Molina & García, 2004; Cerviño et al., 2004; Bordignon, Peri, Tolosa, Villa & Paoletti, 2004) han demostrado que este clasificador posee una alta exactitud y velocidad cuando es aplicado a una gran cantidad de datos textuales. A pesar de estos atributos, Caruana y Niculescu-Mizil (2006) mencionan que su desempeño es inferior al de otros métodos, como por ejemplo el de Máquina de Soporte Vectorial (MSV). A pesar de esta opinión, los autores sostienen que los resultados alcanzados por este clasificador son buenos, pues requieren solo un pequeño número de textos para su entrenamiento y su manejo es relativamente sencillo. Debido a este y otros factores, este clasificador se ha vuelto una opción ampliamente utilizada por los investigadores para realizar clasificaciones automatizadas de textos (Cárdenas, Olivares & Alfaro, 2014). Al respecto, Venegas (2007) expone que el objetivo de este método de aprendizaje matemático-estadístico es determinar cuál es la mejor hipótesis dado un conjunto de datos pre-existentes.

En cuanto al funcionamiento del clasificador Bayesiano Ingenuo, este se puede explicar de la siguiente manera. Si se denota P (D) como la probabilidad *a priori* de los datos y P(D|h) como la probabilidad de los datos dada una hipótesis, lo que se pretende estimar es P(h|D), es decir, la probabilidad posterior de h dado ciertos datos conocidos, esta probabilidad es llamada *a posteriori* o probabilidad condicional. Lo anterior se puede estimar con el siguiente teorema, llamado *Teorema de Bayes*:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Para estimar la hipótesis más probable se busca el mayor P(h|D), tal como se muestra en la siguiente imagen:

$$h_{MAP} = \arg\max_{h \in H}(P(h \mid D))$$
$$= \arg\max_{h \in H}\left(\frac{P(D \mid h)P(h)}{P(D)}\right)$$
$$= \arg\max_{h \in H}(P(D \mid h)P(h))$$

Ahora bien, como P(D) es una constante independiente de h, se asume que todas las hipótesis son igualmente probables, esto permite entonces concebir la hipótesis de máxima verosimilitud (ML, [maximum likelihood]) expresada en la siguiente ecuación :

$$h_{ML} = \arg\max_{h \in H}(P(D \mid h))$$

El clasificador bayesiano ingenuo se utiliza cuando se quiere clasificar un ejemplo descrito por un conjunto de atributos en un conjunto finito de clases (V), en nuestro caso, movidas retóricas. Esto es, clasificar un nuevo ejemplo de acuerdo con el valor más probable dado los valores de sus atributos. Al aplicar la ecuación relativa al valor más probable ($h_{máx)}$ al proceso de la clasificación se obtiene la siguiente ecuación:

$$v_{MAP} = \arg\max_{v_j \in V}(P(v_j \mid a_1, \ldots, a_n))$$
$$= \arg\max_{v_j \in V}\left(\frac{P(a_1, \ldots, a_n \mid v_j)P(v_j)}{P(a_1, \ldots, a_n)}\right)$$
$$= \arg\max_{v_j \in V}(P(a_1, \ldots, a_n \mid v_j)P(v_j))$$

Además, el clasificador Bayesiano ingenuo asume que los valores de los atributos son condicionalmente independientes dado el valor de la clase, por lo que se hacen ciertas las siguientes ecuaciones:

$$P(a_1, \ldots, a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$
$$\therefore P(v_j \mid a_1, \ldots, a_n) = P(v_j) \times \prod_i P(a_n \mid v_j)$$

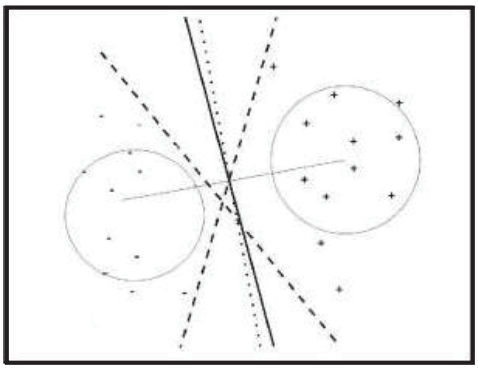
En este sentido, en los clasificadores bayesianos ingenuos o Naïve Bayes se asume que el efecto de un valor del atributo en una clase dada es independiente de los valores de los otros atributos. Esta suposición se llama independencia condicional de clase (Jurafsky & Martin, 2000; Molina & García, 2004; Bordignon et al., 2004). Ella permite simplificar los cálculos involucrados, siendo por esto que se le considera ingenuo (*naïve*) al método y, por lo mismo, sus resultados deben ser entendidos como una simplificación de la realidad. Cabe destacar que en este apartado solo se ha revisado el clasificador bayesiano ingenuo, a pesar de que existen otros que poseen similares características, a saber: simple Bayes (Gammerman & Thatcher, 1991) y Bayes independiente (Todd & Stamper, 1994).

## 2.8.Máquina de Soporte Vectorial

Otro de los algoritmos utilizados en la clasificación automática de textos son las Máquinas de Soporte Vectorial (MSV) o MSV, por sus siglas en inglés. Estas máquinas de aprendizaje toman distintas características de los elementos que quieren clasificar y los llevan a un espacio multidimensional. En este espacio, el algoritmo identifica un hiperplano que separa a los vectores de una clase a otra. Estas máquinas son bastante populares debido a su funcionamiento relativamente sencillo y sus buenos resultados. Además, algunos investigadores (Chan & Lin, 2003; Baldi, Fresconi & Smyth, 2003; Téllez, 2005; Betancourt, 2005) reconocen que el método minimiza los errores de los clasificadores sobre nuevos documentos a trabajar, por lo que su utilización se ha masificado en los últimos años. Este clasificador se basa en el modelo vectorial, definido inicialmente por Salton (1968) y ampliamente usado en operaciones de recuperación de información, así como también en operaciones de categorización automática, filtrado de información, clasificación de textos, etc. (Zazo et al., 2002; Manning & Schütze, 2003). El objetivo de la MSV es producir un modelo que permita predecir los valores de clasificación en la etapa de prueba conociendo solo los atributos (Baldi et al., 2003).

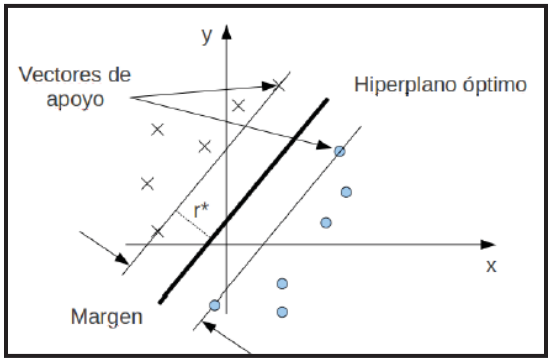
El problema que resuelve la MSV es identificar una frontera de decisión lineal entre dos grupos, a través de una línea que los separe maximizando el espacio del hiperplano. No obstante, la MSV no solo permitirá separar datos lineales, sino que a través del "truco de Kernel" separará

datos no lineales representándolos en una dimensión mayor a la que pertenecen. De esta manera, en un espacio multidimensional las posibilidades de separación de las clases serán múltiples, tal como se evidencian en la siguiente imagen.



Figuras 1: Alternativas de solución para un problema de clasificación binario

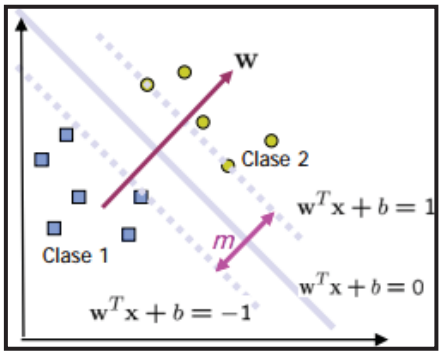
Sin embargo, se necesitará una separación óptima del hiperplano, sustentada por los vectores de soporte.




Figuras 2: Hiperplano sustentado con vectores de soporte

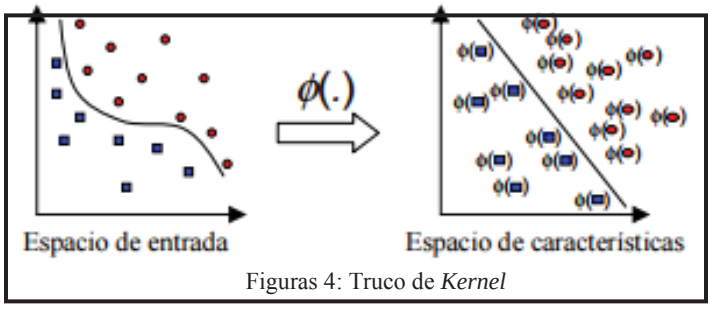
La frontera de separación óptima será aquella que minimice la probabilidad de que un nuevo punto sea mal clasificado. Para asegurar que el hiperplano sea el correcto, es decir, no sea azaroso y no aporte a generalizaciones de datos erróneas, Vapnick (2000) propone un hiperplano de separación óptima el cual tiene dos propiedades importantes. En primer lugar es único para cada grupo de datos separables linealmente, y el riesgo asociado de sobreestimación es más reducido que para cualquier otro hiperplano de separación. En segundo lugar, el margen de separación M del clasificador debe ser la distancia entre el hiperplano de separación y el

ejemplo de entrenamiento más cercano. De este modo, el hiperplano de separación óptimo es aquel que tenga el máximo margen.



Figuras 3: Hiperplano separado con distancia máxima de los datos

La característica más relevante de la MSV es la representación de datos no linealmente separables en otra dimensión a través del uso de las funciones kernel (lineal, polinominal, función de base radial, sigmoidal, etc.) (Betancourt, 2005). Esto se hace traspasando los datos desde el espacio de entrada *X* a un amplio espacio de características **X** mediante una función **O**, y resolviendo el problema de aprendizaje lineal en **X**(**0**: *X*—>**X**). (Christianini & Shaw-Taylor, 2002; Bautista, Guzmán & Figueroa, 2004).



Figuras 4: Truco de *Kernel*

### 2.9.Lema y lematización

Los conceptos de lema y lematización son neologismos que provienen del campo de la informática y que debido a que en los últimos años estos han sido adoptados en estudios lingüísticos se han incorporado como términos recurrentes en algunas áreas de estudio

(Peinado, 2003). En teoría lexicográfica, se define lema como la forma canónica a la que se reduce un paradigma flexivo y que se toma como representación de todas las variantes morfológicas de la palabra (Alcaraz & Martínez, 1993). Respecto a este concepto, Sabaj (2003) menciona que los lemas corresponden a la abstracción de un conjunto de formas paradigmáticamente relacionadas. El autor plantea que la abstracción será un consenso que representará a los verbos en infinitivo, sustantivos en singular y adjetivos en masculino singular. Para Goded, Ibáñez y Hoste (2015) lema es la forma canónica, forma de un diccionario o forma de cita que representa un conjunto de palabras. Los autores mencionan que el lema es la forma particular, elegida por convención, para representar un lexema.

Por otra parte, Peinado (2003) menciona que el lema o también llamado Stem, en inglés, es la porción o raíz de la palabra después de eliminar todos sus afijos asociados. Como se puede evidenciar a partir de esta última definición, existe una clara diferencia entre lo expuesto por este autor y las definiciones anteriores, pues Peinado (2003) iguala lema a raíz. Tomando en cuenta las definiciones propuestas por Gómez-Macker y Peronard (2005) discrepamos de la igualdad propuesta entre lema y raíz, puesto que raíz es definida por los autores como la unidad léxica que denota la idea nuclear de una palabra a la cual se le agregan los afijos derivativos, constituyéndola como la base de una familia de palabras (Gómez-Macker & Peronard, 2005:177) y lema es definido como una forma canónica a la que se reduce un paradigma flexivo por convención (Alcaraz & Martínez , 1993; Sabaj, 2003).

Tomando en cuenta estas definiciones, podemos mencionar que no es correcto igualar estos conceptos, pues refieren a unidades léxicas distintas. A modo de ejemplo, la palabra *Planetas* tendrá como raíz "plan", pues su raíz corresponde a la unidad léxica nuclear sin ningún afijo derivativo. Mientras que su lema será planeta, pues corresponde a la forma canónica determinada por una convención, en este caso el uso singular.

Con respecto a la lematización, Alcaraz y Martínez (1993) la entiende como el proceso de reducción de diferentes formas flexivas de una palabra a la forma canónica que se selecciona como lema. En esta misma línea, Gómez (2005) menciona que lematización es un neologismo que se aplica al proceso de eliminación automática de partes no esenciales de los términos

(sufijos, prefijos) para reducirlos a su parte esencial (*lema*) y facilitar la eficacia de la indización y su consiguiente recuperación. Al respecto, Peinado (2003) menciona que lematización es una técnica usada en la recuperación de datos para reducir las variantes morfológicas y mejorar la habilidad de los motores de búsqueda. Con respecto a este punto, algunos autores (Sebastiani, 2002; Peinado, 2003; Cárdenas, Olivares & Alfaro, 2014) mencionan que en la clasificación de textos se deben utilizar estrategias que permitan eliminar distractores y mejorar el proceso. Frente a esto, Sebastiani (2002) menciona que la lematización permite mejorar la clasificación de textos, pues eliminará todo tipo de distractores de los documentos, quedando solo el lema de la palabra y eliminando algunos de sus afijos asociados. Lo anterior se realiza con el fin de que las palabras que tienen el mismo significado conceptual sean representadas por su forma canónica para así simplificar la tarea del clasificador.

## 3. MARCO METODOLÓGICO

En el presente apartado, se dará cuenta de los aspectos metodológicos que sustentan esta investigación. De esta manera, en primer lugar se expondrá el enfoque, alcance y diseño de la investigación. Luego, se formularán las preguntas de investigación que serán respondidas en el presente estudio. En tercer lugar, se expondrá el objetivo general y los objetivos específicos de esta investigación. Posteriormente, se presentará el corpus trabajado en el estudio y, finalmente, la herramienta de recolección de datos y el detalle de los análisis.

### 3.1.Tipo de investigación

La investigación planteada en esta tesis se adscribe al enfoque metodológico de tipo cuantitativo, pues se buscaron, recolectaron y cuantificaron las categorías léxicas en las tesis del ámbito de la Lingüística en la Licenciatura en Lengua y Literatura Hispánica. Posteriormente se sistematizaron estas categorías, se establecieron generalizaciones y finalmente algunas predicciones a partir de los resultados obtenidos de ellas. De esta forma, se cuantificaron los datos de tipo léxico y se utilizaron técnicas cuantitativas de análisis de datos para clasificar y predecir las movidas retóricas según sus tri-gramas de lemas.

El alcance de nuestra investigación es de tipo descriptivo, ya que busca describir las variables léxicas, en este caso, lemas o tri-gramas de lemas que se presentan en las movidas retóricas de las macromovidas Introducir al lector en la investigación y Finalizar discursivamente la investigación en los Trabajos Finales de Grado en el ámbito de Lingüística de la Licenciatura en Lengua y Literatura de la PUCV.

En cuanto al diseño de la investigación, este será de tipo no experimental, ya que no se manipulará ninguna variable y se clasificarán tesis que fueron escritas en su contexto natural y sin ningún tipo de intervención. Respecto a este tipo de investigación, Hernández, Fernández y Baptista (2000) mencionan que “no se construye ninguna situación, sino que se observan situaciones ya existentes, no provocadas intencionalmente por el investigador”.

En esta investigación, las clases que se utilizarán para clasificar serán las movidas retóricas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5). Estas clases, se obtendrán a partir del modelo retórico-discursivo del Macrogénero

Trabajo Final de Grado en tesis de licenciatura (Venegas, Zamora & Galdames, 2016) desarrollado en los proyectos Fondecyt 1101039 y 1140967. En cuanto a los atributos, estos se procesaran mediante una herramienta semiautomática, ANMOP. Esta es una herramienta, desarrollada por el proyecto Fondecyt 11409967, que permite realizar un análisis léxico-gramatical y retórico-discursivo de un corpus que se almacena en la misma herramienta. En esta interfaz, los textos cargados pueden ser analizados a partir del modelo retórico-discursivo planteado por los proyectos Fondecyt 1101039 y 1140967 o bien, se pueden utilizar otras herramientas de análisis de corpus como Conexor. En nuestro caso, la herramienta antes mencionada permitirá lematizar las movidas seleccionadas con el fin de obtener tri-gramas de lemas que permitan más tarde clasificar los documentos en clases predefinidas.

### 3.2.Preguntas de investigación

Las preguntas que se pretenden responder con la presente investigación son las siguientes:

- ¿Cuáles son los rasgos léxicos que permiten clasificar semiautomáticamente las movidas retóricas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación en las tesis en el ámbito de la Lingüística en Licenciatura en Lengua y Literatura Hispanoamericana?

- ¿Cuáles son los atributos léxicos más representativos de las movidas retóricas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación que permiten clasificarlas semiautomáticamente?

- ¿Qué algoritmo de clasificación permite clasificar semiautomáticamente mejor las movidas retóricas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación?

## 3.3.Objetivos

La presente investigación tuvo los siguientes objetivos:

### 3.3.1. Objetivo general

- Clasificar semiautomáticamente las movidas retóricas de las tesis en el ámbito de la Lingüística de la Licenciatura en Lengua y Literatura Hispánica de la Pontificia Universidad Católica de Valparaíso a partir de sus lemas.

### 3.3.2. Objetivos específicos

- Identificar los atributos que caracterizan las movidas retóricas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación en las tesis en el ámbito de la Lingüística de Licenciatura en Lengua y Literatura Hispanoamericana de la PUCV.

- Identificar los tri-gramas de lemas que permitan clasificar semiautomáticamente de mejor manera las movidas retóricas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación en las tesis en el ámbito de la Lingüística de Licenciatura en Lengua y Literatura Hispanoamericana de la PUCV.

- Comparar el rendimiento de los algoritmos Bayesiano Ingenuo y Maquina de soporte vectorial en la clasificación de las movidas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación en las tesis en el ámbito de la Lingüística de Licenciatura en Lengua y Literatura Hispanoamericana de la PUCV.

### 3.4.Corpus

La muestra de la presente investigación está compuesta por 20 tesis de pregrado del ámbito de la Lingüística de la Licenciatura en Lengua y Literatura Hispánica de la PUCV realizadas entre los años 2009 y 2012. Esta muestra fue seleccionada a partir del subcorpus de tesis (TLING) del proyecto Fondecyt 1140967. Este subcorpus está compuesto por 33 trabajos finales de grado (TFG) desarrollados en este ámbito en los años ya mencionados. Sin embargo, en la presente investigación, por motivos fundaentalmente temporales, se utilizarán 20 (61%) tesis escogidas de manera aleatoria para conformar la muestra que será utilizada en el presente estudio. Así, dado que cada tesis está segementada por movidas, en nuestro estudio el total de segmentos textuales asociados a las movidas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5) corresponderá a 120 segmentos.

Los títulos de la muestra de tesis de lingüística con sus respectivos códigos se pueden apreciar en la siguiente tabla:

| Código | Título de la tesis |
|---|---|
| 2009_1 | Desarrollo del pensamiento analógico en adultos con escolaridad tardía y en proceso de alfabetización. |
| 2009_3 | La escritura en Licenciatura en Matemáticas. Un estudio exploratorio en torno a las representaciones sociales de profesores y alumnos. |
| 2009_4 | Influencia de la idiomaticidad y el contexto en la comprensión de locuciones idiomáticas en estudiantes de E/L2. |
| 2009_7 | La persona ausente y la inscripción del yo en las tesis de Licenciatura en Historia y Ciencias Sociales. |
| 2009_8 | Patrones retóricos en los géneros académicos: Un estudio fundado en los datos de la Introducción y Conclusiones de las tesis en Bioquímica. |
| 2009_9 | La escritura en la formación académica en humanidades: un estudio exploratorio acerca de las representaciones sociales de los profesores de la Licenciatura en Historia. |
| 2009_12 | Macroestructuras semánticas y punto de vista en dieciocho editoriales de El Mercurio. |
| 2010_1 | Hacia la configuración temática de los modos de organización discursivos argumentativo y narrativo. |
| 2010_2 | La influencia del bilingüismo en el desarrollo sintáctico: un estudio exploratorio en escolares de segundo año medio. |
| 2010_3 | Aproximación a una descripción multiregistro del español a través de la función de tema desde la lingüística sistémico funcional en textos argumentativos y descriptivos. |

| | |
|---|---|
| 2010_4 | La escritura en la universidad: una aproximación al texto expositivo desde el área humanista y científica. |
| 2009_2 | La escritura en la formación académica en ciencias. Un estudio exploratorio acerca de las representaciones sociales de los profesores de la licenciatura en bioquímica. |
| 2010_6 | Organización retórica del apartado discusión: un estudio exploratorio en tesis de licenciatura en historia. |
| 2010_7 | El modo visual gráfico y el verbal de un género académico en el sub-corpus de Químida correspondiente al Corpus PUCV 2010. |
| 2010_8 | Género tesis de licenciatura entre las disciplinas de piscología y Licenciatura en Literatura Hispánica a partir de los rasgos positivos y negativos de la dimensión informacional. |
| 2012_1 | La estructura del texto explicativo: su dominio a lo largo de la educación escolar. |
| 2012_2 | La complejidad sintáctica en los textos narrativos: su evolución en la edad escolar. |
| 2012_3 | Caracterización de trabajos finales de grado de licenciatura en filosofía, lingüística, literatura y psicología de la PUCV. |
| 2012_5 | Caracterización del posicionamiento en conclusiones de tesis de licenciatura, magíster y doctorado en filosofía y lingüística. |
| 2010_5 | Atribución del conocimiento en las tesis de licenciatura en las disciplinas de literatura hispánica y psicología. |

Tabla 1: Muestra del corpus

### 3.5.Procedimientos

Para la recolección del corpus escrito entre los años 2009 y 2012, se acudió a la biblioteca Mayor de Educación de la PUCV y se extrajeron todos los ejemplares escritos en estos años. Posteriormente, se procedió a fotocopiar cada tesis y luego se comenzó la etapa de digitalización. En este paso, en primer lugar, se escaneó cada uno de los ejemplares, luego, mediante el programa Omnipage profesional 15, se guardaron los documentos en tres archivos, a saber: txt, word y pdf. Posteriormente, los archivos en formatos txt fueron subidos y almacenados en la plataforma del proyecto Fondecyt 1140967 llamada ANMOP (Análisis de Movidas y Pasos).

El análisis retórico-discursivo de estos documentos se realizó de forma computacional a través de esta plataforma. Este análisis fue realizado de forma simultánea por tres analistas, los cuales leían el mismo TGF en la plataforma y, paralelamente, etiquetaban los pasos, movidas y macromovidas con base en el modelo retórico-discursivo ya mencionado (Ver Anexo 1). Una

vez que los tres analistas concluían el análisis del documento, el sistema otorgaba un porcentaje de acuerdo entre los participantes y luego, de forma aleatoria, seleccionaba un documento entre los dos analistas que habían obtenido el mejor porcentaje de acuerdo. Este documento era almacenado en la plataforma y quedaba en ella para ser sometido a todos los análisis disponibles en ANMOP. En la siguiente tabla se muestran los porcentajes de acuerdo de las tesis seleccionadas en el corpus de nuestra investigación:

| Código | Porcentaje de acuerdo entre analistas |
|---|---|
| 2009_1 | 97% |
| 2009_3 | 97% |
| 2009_4 | 100% |
| 2009_7 | 100% |
| 2009_8 | 100% |
| 2009_9 | 95% |
| 2009_12 | 93% |
| 2010_1 | 93% |
| 2010_2 | 99% |
| 2010_3 | 99% |
| 2010_4 | 98% |
| 2009_2 | 90% |
| 2010_6 | 95% |
| 2010_7 | 100% |
| 2010_8 | 95% |
| 2012_1 | 93% |
| 2012_2 | 98% |
| 2012_3 | 100% |
| 2012_5 | 98% |
| 2010_5 | 95% |

Tabla 2: Porcentajes de acuerdo entre analistas

La selección de documentos que tuvieran un porcentaje de acuerdo entre los analistas de más de un 90% aseguró que los análisis realizados por los sujetos fueran similares y tuvieran criterios unificados. En el caso que las tesis no tuvieran un porcentaje de acuerdo superior o igual a 90%, los analistas volvían a realizar el proceso de análisis (re análisis) y se reunían para establecer consenso. Este proceso se realizaba hasta que el porcentaje de acuerdo superara el 90%.

### 3.6.Herramientas de recolección y análisis de datos

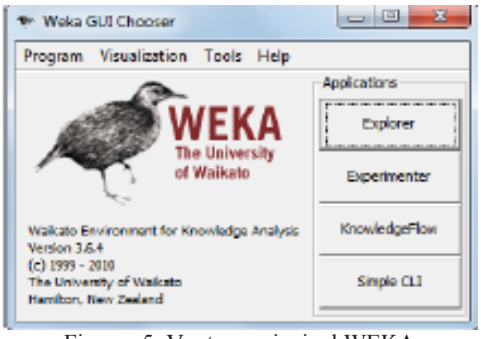
Para la recolección de datos léxicos se utilizó el lematizador de conexor, integrado en la plataforma ANMOP desarrollada por el proyecto Fondecyt 1140967. Los datos, en nuestro caso tri-gramas de lemas, fueron consignados en grillas realizadas con Excel que permitieron ordenar y sistematizar la información dispuesta en cada movida de las dos macromovidas seleccionadas de cada una de las 20 tesis del corpus. Así, se realizó una planilla de Excel que contenía todos los tri-gramas de lemas que aparecían en las movidas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación y la frecuencia relativa con la que aparecía cada trigrama de lemas de cada movida en cada tesis. De esta manera, si el trigrama de lema aparecía en un TFG, se insertaba la frecuencia relativa asignada a este dato, en el caso que no apareciera el trigrama en esa tesis en particular, se asignaba 0 a dicho valor.

Una vez recolectados, ordenados y sistematizados los datos, se procedió a trabajar con los dos algoritmos de clasificación mencionados en el apartado anterior, a saber: Bayesiano Ingenuo y Máquina de soporte vectorial (MSV). El proceso de clasificación se llevó a cabo con la plataforma de aprendizaje automático y minería de texto llamada WEKA, la cual incluye estos dos algoritmos clasificadores. A su vez, para validar los procesos de clasificación de ambos algoritmos, se utilizó una técnica de validación cruzada de 10 Fold. Esta técnica utilizada en análisis estadístico consiste en repetir 10 veces la clasificación y calcular la medida aritmética obtenida de las medidas de evaluación sobre diferentes participaciones.

En cuanto a la forma en la que se realizó el proceso de clasificación, en primer lugar, se procedió con el clasificador probabilístico Bayesiano Ingenuo, todos sus atributos y todas las movidas. Luego, se clasificó con este mismo algoritmo, pero con los mejores atributos, es decir, los mejores tri-gramas de lemas. Posteriormente, se procedió con el algoritmo Máquina de soporte vectorial (MSV) de la misma forma que con el anterior, es decir, primero con todos los atributos y luego solo con los mejores.

Cabe destacar que en el caso de MSV, se se utilizó un kernel Polinomial y un algortimo de optimización llamado secuential minimal optimation (SMO) disponible en WEKA. Este algoritmo desarrollado por Platt (1998) fue utilizado en nuestra investigación, puesto que permite resolver el problema de segmentación del MSV 1000 veces más rápido, puesto que

reemplaza, globalmente, todos los valores faltantes y transforma los atributos a valores binarios con el fin de aminorar el tiempo de clasificación.



Figuras 5: Ventana principal WEKA

Es importante mencionar que la selección de los mejores atributos, ya sea para el algoritmo Bayesiano ingenuo o Maquina de soporte vectorial, se realizó de manera inductiva utilizando el algoritmo Correlation-based Feature Subset Selection (CfsSubsetEval) propuesto por Hall (1998). Este algoritmo evalúa el valor de un subconjunto de atributos considerando su capacidad predictiva y grado de redundancia entre todos los atributos. De esta manera, el algoritmo selecciona como mejores atributos a los subconjuntos de características que están altamente relacionadas con la clase, en nuestro caso movidas, y que tienen una intercorrelación baja.

Finalmente, luego de la obtención de los datos de análisis, se llevó a cabo la interpretación y evaluación de los resultados teniendo en cuenta el coeficiente de fuerza de concordancia (Kappa) y el equilibrio entre la precisión y la exhaustividad del clasificador (F- Measure). En cuanto a los valores del coeficiente de fuerza de concordancia (kappa), se utilizó la propuesta de Landis y Koch (1997) para determinar si este era aceptable o no. A continuación, se presentan los índices de valoración:

| Valoración del coeficiente de Kappa | |
|---|---|
| 0,00 | Pobre |
| 0,01 – 0,20 | Leve |
| 0,21 – 0,40 | Aceptable |
| 0,41 – 0,60 | Moderado |
| 0,61 – 0,80 | Considerable |
| 0,81 – 1,00 | Casi perfecto |

Tabla 3: Valoración del coeficiente Kappa

Con respecto a los valores de F- Measure, estos corresponden a una medida común entre los trabajos de clasificación de textos, que permite unificar los resultados de precisión y exhaustividad (recall) siendo un equilibrio entre ellas (Venegas, 2015). Esta medida según Manning y Schütze (2003) es conocida como F, F1 o Fβ y expresa la media armónica entre precisión y recall, tal como se muestra en la siguiente ecuación:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

### 3.7.Definición de variables

Las variables o aspectos del objeto de estudio a investigar son las siguientes:

**Disciplina**

La variable disciplina académica alude a las tesis desarrolladas en el ámbito de la Lingüística en la Licenciatura en Lengua y Literatura Hispánica impartida en el Instituto de Literatura y Ciencias del Lenguaje de la PUCV. Estos trabajos finales de grado son desarrollados en el décimo semestre de la carrera y, generalmente, suelen desarrollarse en el marco de algún proyecto de investigación.

**Tri-gramas de lemas**

En cuanto a los tri-gramas, estos son n-grams compuestos por tres unidades léxicas que dan cuenta de la ocurrencia de la combinación prototípica de los elementos en los textos del corpus. En el caso de la presente investigación, estos elementos serán los lemas obtenidos con la interfaz conexor en ANMOP, entendidos como la forma canónica en la que se representa un paradigma flexivo, es decir, la representación determinada por una convención de una palabra sin sus afijos flexivos asociados (Alcaraz & Martínez, 1993; Sabaj, 2003; Goded, Ibáñez & Hoste, 2015).

**Modelo retórico-discursivo**

Este modelo retórico-discursivo es producto de los estudios realizados en los proyectos Fondecyt 1101039 y 1140967. Corresponde a un modelo de organización retórico-discursiva compuesto por 5 macromovidas, 18 movidas y 65 pasos (Ver Anexo 1). La relación entre cada uno de estos conceptos está dada por los niveles de abstracción de cada uno de ellos, siendo la macromovida la más abstracta y los pasos la realización de las movidas (Venegas, Núñez, Zamora & Santana, 2015).

# 4. RESULTADOS

En el presente apartado, se presentarán, interpretarán y compararán los resultados obtenidos a partir de la investigación realizada. De esta manera, en cada uno de los subapartados se expondrá la clasificación realizada por cada algoritmo y luego se realizará una comparación entre los resultados obtenidos por ellos. La presentación de los datos se realizará en seis subaparatados que serán descritos a continuación. En el primero, se expondrán los resultados de las clasificaciones de todas las movidas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5) realizadas con el algoritmo probabilístico Bayesiano Ingenuo (BI) y Máquina de soporte vectorial (MSV/SMO) con todos los atributos. En la siguiente sección, se presentaran los resultados obtenidos con los algoritmos BI y MSV/SMO de la clasificación de las movidas de las MM1 y MM5 con una selección de los mejores atributos. En el tercer acápite, se desplegarán los resultados de la clasificación de las movidas de la macromovida Introducir al lector (MM1) realizadas por los dos algoritmos con todos sus atributos.

En la sección cuatro, se darán a conocer los resultados de la clasificación de las movidas de la MM5 realizada por BI y MSV/SMO con todos sus atributos. En el siguiente capítulo, se expondrán los resultados de la clasificación de las movidas de la MM1 con una selección de los mejores atributos. Finalmente, se desplegarán los resultados de la tarea de clasificación de las movidas de la macromovida MM5 con ambos algoritmos y los mejores atributos.

## 4.1.Clasificación de todas las movidas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5) con todos sus atributos

En esta sección, se presentarán los resultados de las clasificaciones de todas las movidas de las MM1 y MM5 realizadas por los dos algoritmos de clasificación, a saber: Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV/SMO) con todos los atributos (14.739). De esta manera, se expondrán los resultados referentes a la precisión de la clasificación, la exhaustividad o recall, el porcentaje de clasificación alcanzado por el algoritmo, el coeficiente de fuerza de concordancia (Kappa) y la medida que unifica precisión y exhaustividad (F-Measure). Posteriormente, se desplegará la matriz de confusión que permite observar tanto la clasificación correcta realizada por el algoritmo como los errores que cometió en comparación a la clasificación humana.

### 4.1.1. Clasificación con el algoritmo Bayesiano Ingenuo (BI)

Los resultados de la clasificación de todas las movidas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5) realizada por este algoritmo se presentan en la siguiente tabla.

| | **MM1M1** | **MM1M2** | **MM1M3** | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|---|---|---|
| **Precisión** | 0,182 | 0,357 | 0,344 | 0,286 | 0,417 | 0,308 |
| **Recall** | 0,1 | 0,25 | 0,55 | 0,2 | 0,75 | 0,2 |
| **F- Measure** | 0,129 | 0,294 | 0,423 | 0,235 | 0,536 | 0,242 |
| **Kappa** | 0,21 | | | | | |
| **% clasificación correcta** | 34,1667 % | | | | | |

Tabla 4: Resumen clasificación BI con todas las movidas y atributos

A partir de los resultados expuestos en la Tabla 4, es posible evidenciar que la clasificación de todas las movidas realizada por BI con todos los atributos (14.739) no es satisfactoria, pues este algoritmo solo pudo clasificar de forma correcta un 34,1667% del corpus total, lo que se

considera muy bajo. A su vez, el coeficiente de fuerza de concordancia que alcanzó fue solo de 0,21, considerado leve, siendo que para que sea significativo se espera un Kappa sobre el 0,60. Con respecto a los valores alcanzados por movida, se evidencia que la movida MM1M1 es la que menos se pudo clasificar, pues la precisión con la que se clasifica esta movida es de 0,182, muy bajo, y la medida F es de solo un 0,129, lo que no permite su clasificación automática. En cuanto a la movida que mejor clasificó el algortitmo, esta fue la MM5M2, sin embargo, a pesar de obtener los resultados más altos de la clasificación, estos continúan sin ser significativos, pues solo alcanza un equilibrio entre precisión y exhaustividad de un 0,536, considerado bajo.

El detalle de la clasificación en estas movidas realizada por BI se presenta en la siguiente tabla.

| | **MM1M1** | **MM1M2** | **MM1M3** | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|---|---|---|
| MM1M1 | 2 | 3 | 8 | 3 | 3 | 1 |
| MM1M2 | 2 | 5 | 2 | 3 | 3 | 5 |
| MM1M3 | 2 | 0 | 11 | 0 | 7 | 0 |
| MM5M1 | 2 | 2 | 7 | 4 | 3 | 2 |
| MM5M2 | 0 | 1 | 3 | 0 | 15 | 1 |
| MM5M3 | 3 | 3 | 1 | 4 | 5 | 4 |

Tabla 5: Matriz de confusión de BI con todos los atributos y movidas

Como se desprende de la Tabla 5, en general la calificación del algoritmo es errónea, pues suele clasificar las categorías, movidas, en otras diferentes a las que identificaron los humanos. En este sentido, en la matriz de confusión, podemos evidenciar que el algoritmo tiende a clasificar la mayoría de los documentos en categorías que no pertenecen, recuperando muchos documentos erróneamente en la MM5M2. Lo anterior podría ser explicado desde dos puntos de vista, por medio de los propósitos comunicativos y por medio del algoritmo de clasificación. El primero de estos factores podría deberse a los recursos léxicos que se utilizan en las movidas MM5M2 y MM1M3, movidas que más confunde, para cumplir con los propósitos comunicativos. De esta forma, se puede evidenciar que muchos de los tri-gramas de lemas que se usan en la MM5M2 para cumplir con el propósito, también se utilizan en la movida MM1M3, puesto que en ambas movidas se hace alusión a los objetivos, preguntas de investigación,

resumen los hallazgos, entre otros. El segundo punto de vista está ligado a lo anterior, pues al compartir algunos tri-gramas y ser clasificados por un algoritmo probabilístico, este tiende a clasificar las movidas con base en probabilidades, por lo que si un documento reúne un porcentaje de características significativas, aunque no las reúna todas, igual podrá formar parte de esa movida.

### 4.1.2. Clasificación con Máquina de Soporte Vectorial (MSV/SMO)

A continuación, en la Tabla 6, se muestran los resultados obtenidos por el algoritmo Máquina de soporte vectorial (MSV) en la clasificación de las movidas retóricas de las macromovidas MM1 y MM5 con los 14.739 atributos totales. Antes de mostrar los resultados, se debe aclarar que para mejorar los logros de este clasificador (MSV) se utilizó un algortimo de optimización llamado secuential minimal optimation (SMO).

| | **MM1M1** | **MM1M2** | **MM1M3** | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|---|---|---|
| **Precisión** | 1 | 1 | 0,222 | 0 | 0,072 | 0,132 |
| **Recall** | 0,05 | 0,1 | 0,1 | 0 | 0,25 | 0,25 |
| **F- Measure** | 0,095 | 0,182 | 0,138 | 0 | 0,112 | 0,172 |
| **Kappa** | -0,05 | | | | | |
| **% clasificación correcta** | 12,5 % | | | | | |

Tabla 6: Resumen de clasificación MSV con todas las movidas y atributos

Tal como se muestra en la Tabla 6, esta clasificación, al igual que la realizada con BI, tampoco logra ser significativa, pues, en este caso, los porcentajes alcanzados por este clasificador son aún más bajos que los mencionados en la sección anterior. En este sentido, podemos observar que el algoritmo solo logra clasificar de forma correcta un 12,5% del corpus total y su Kappa es muy bajo, ya que ni siquiera alcanza el valor mínimo 0. Resulta interesante de la tabla anterior que la MM5M1 no logra ser clasificada, pues de todos los documentos, solo logra recuperar uno correctamente en esta movida.

En la Tabla número 7 se muestra la matriz de confusión de la clasificación realizada en las movidas de las macromovidas MM5 y MM1.

| | **MM1M1** | **MM1M2** | **MM1M3** | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|---|---|---|
| MM1M1 | 1 | 0 | 5 | 0 | 3 | 11 |
| MM1M2 | 0 | 2 | 0 | 0 | 15 | 3 |
| MM1M3 | 0 | 0 | 2 | 1 | 15 | 2 |
| MM5M1 | 0 | 0 | 0 | 0 | 16 | 4 |
| MM5M2 | 0 | 0 | 2 | 0 | 5 | 13 |
| MM5M3 | 0 | 0 | 0 | 0 | 15 | 5 |

Tabla 7: Matriz de confusión de MSV con todas las movidas y atributos

En estos resultados, podemos observar que por un lado, las mejores clasificaciones realizadas por este algoritmo alcanzan solo el 25% de identificación correcta, por lo que la tarea de clasificación realizada por MSV con todas las movidas y atributos es considerada insatisfactoria. Con respecto a las movidas MM5M2 y MM5M3, resulta significativo que la mayoría de las recuperaciones de documentos estén en estas dos categorías y que casi no existan documentos asociados a las MM1M1 y MM1M2. Lo anterior se podría deber a una de las características del algoritmo SMO utilizado en MSV, pues al ser un algoritmo binario con una representación muldimensional, clasifica las movidas de manera on-off. Lo anterior, en palabras simples significa que para MSV un documento pertenece o no pertenece a la clase, por lo que si el documento posee atributos compartidos con otro documento, será clasificado en el lugar donde el hiperplano se establezca a partir los vectores de soporte. A su vez, nos parece lógico que el clasificador tienda a recuperar la mayoría de los documentos en las movida MM5M2 y MM5M3, pues al tener como propósito comunicativo evaluar los hallazgos de la investigación y proveer una explicación crítica de los resultados y hallazgos, comparten muchos de los lemas utilizados en movidas anteriores (MM1M2; MM1M3), como por ejemplo, metodología, resultados, otras investigaciones, etc.

Respecto a la MM1M1, al tener como propósito destacar la importancia del tema y tener una generalización del tema en su interior, los atributos que posee esta movida son pocos

representativos, pues la mayoría de ellos se encuentran en todos las demás clases, por lo que resulta complejo clasificar esta movida con todos los atributos y este tipo de algoritmo.

### 4.1.3. Comparación de la clasificación entre BI y MSV

A continuación se presenta un Gráfico que contiene la comparación entre las tareas de clasificación de todas las movidas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación realizadas con los algoritmos Bayesiano ingenuo (BI) y Maquina de soporte vectorial (MSV) con todos sus atributos (14.739).

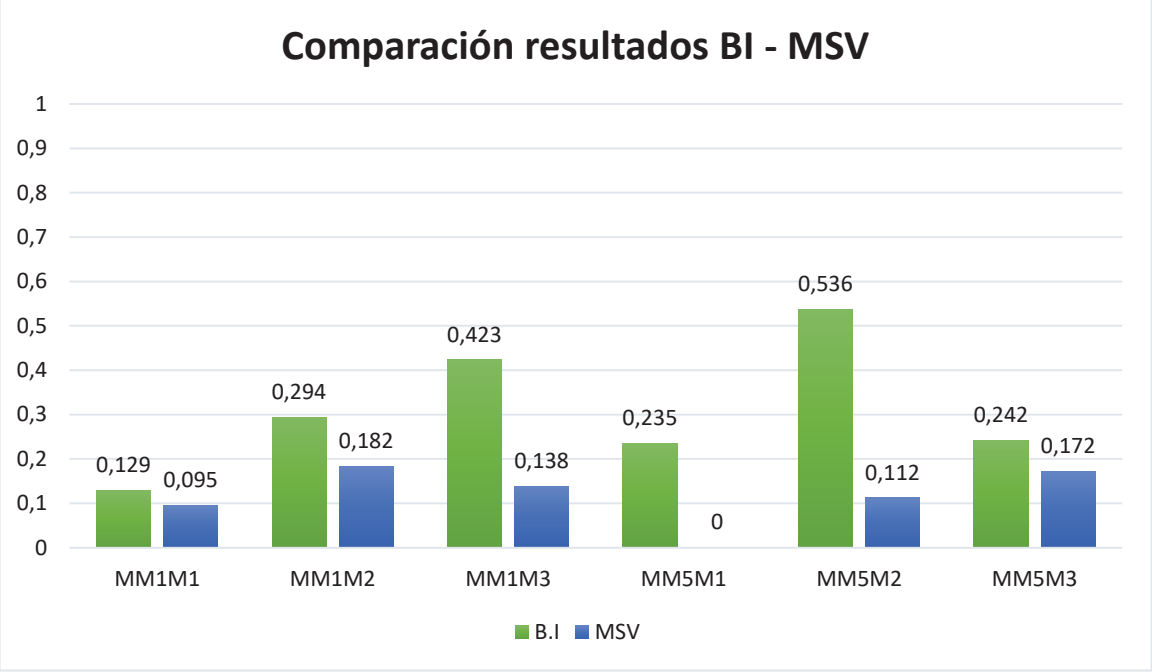



Gráfico 1: Comparación resultados BI y MSV con todas las movidas y atributos

Como se desprende del Gráfico anterior, si bien ninguno de los dos algoritmos, en términos generales, alcanzan un F considerado o un Kappa aceptable, existe una movida que es clasificada de forma satisfactoria, aunque baja (MM5M2). A partir de nuestra interpretación, podemos plantear que esta movida es la mejor clasificada, dado que es la única de las clases que tiene algunos atributos o tri-gramas de lemas con altas frecuencias relativas. Con respecto

a esta interpretación, se debe mencionar que este no puede ser el único factor, pues de lo contrario, MSV también debería clasificar a esta movida de forma correcta, sin embargo solo lo hace con un equilibrio entre precisión y exhaustividad de 0,112. Por lo anterior, sostenemos que la clasificación de la MM5M2 realizada por BI se ve beneficiada por las características o naturaleza probabilística de este algoritmo, pues BI clasifica en una clase X a todas las movidas que tienen una alta probabilidad de pertenecer a ella, aunque no posean todos los atributos para pertenecer, mientras que MSV clasifica de manera on-off, es decir, si está dentro del hiperplano pertencece y si no lo está, no pertenece.

## 4.2.Clasificación de todas las movidas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5) con los mejores atributos

Al no obtener los resultados esperados en la clasificación realizada con todos los atributos (14.739), se realizó una con ambos algoritmos en la que se tomaron solo los mejores atributos (23) de las movidas de las macromovidas MM1 y MM5 a través de un algoritmo llamado Correlation-based Feature Subset Selection (CfsSubsetEval). Lo anterior se realizó con dos fines, el primero, eliminar tri-gramas que no aportan a la clasificación y que solo suman más tiempo a la clasificación. El segundo, puesto que los mejores atributos permitirán caracterizar mejor la clase y mejorar la relación entre ellos y el propósito comunicativo de la movida. A continuación, se presentan las clasificaciones de todas las movidas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación realizadas con los algoritmos Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV) con los mejores atributos (Ver Anexo 2).

### 4.2.1. Clasificación con Bayesiano ingenuo (BI)

Los resultados más relevantes de la clasificación de todas las movidas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación realizada con el algoritmo probabilístico Bayesiano ingenuo se presentan en la Tabla 8.

| | MM1M1 | MM1M2 | MM1M3 | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|---|---|---|
| **Precisión** | 0,933 | 0,356 | 0,737 | 0,25 | 0,895 | 0,333 |
| **Recall** | 0,7 | 0,8 | 0,7 | 0,2 | 0,85 | 0,1 |
| **F- Measure** | 0,8 | 0,492 | 0,718 | 0,222 | 0,872 | 0,154 |
| **Kappa** | 0,47 | | | | | |
| **% clasificación correcta** | 55,8333 % | | | | | |

Tabla 8: Resumen clasificación BI todas las movidas con los mejores atributos

A partir de los resultados expuestos en la tabla anterior, se puede mencionar que la clasificación mejora sustancialmente en comparación a la realizada con todos los atributos, pues el Kappa alcanzado en esta oportunidad es de un 0,47, siendo considerado moderado. A su vez, el porcentaje de clasificación correcta asciende a un 55,83%, por lo que más del cincuenta por ciento del corpus de textos es clasificado de forma correcta.

De la tabla anterior desprendemos que la movida mejor clasificada fue nuevamente la MM5M2, pues alcanzó un equilibrio entre precisión y exhaustividad del 0,872, considerado alto y una recuperación de los documentos del 85%.

Otro resultado relevante es el alcanzado en las movidas MM1M1 y MM1M3, pues en ambos casos F supero el 0,70, lo cual es considerado alto. A pesar de estos resultados, las movidas MM5M1 y MM5M3 continúan sin alcanzar logros óptimos. Los resultados de estas clasificaciones podrían atribuirse a los valores alcanzados por sus tri-gramas de lemas en las frecuencias relativas y las cualidades del algoritmo probabilístico, pues este tiende a clasificar los documentos pertenecientes a estas movidas en la MM1M2 debido a las similitudes entre los atributos de MM5M1 y MM5M3 con ella. A su vez, las bajas frecuencias relativas de los atributos de las movidas de MM5 en comparación a los valores de MM1M2 provocan que la probabilidad de clasificación en MM1M2 aumente.

En la siguiente Tabla se detalla la matriz de confusión del algoritmo de clasificación BI para las movidas mencionadas.

| | MM1M1 | MM1M2 | MM1M3 | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|---|---|---|
| MM1M1 | 14 | 4 | 1 | 1 | 0 | 0 |
| MM1M2 | 0 | 16 | 1 | 2 | 0 | 1 |
| MM1M3 | 0 | 1 | 14 | 3 | 0 | 2 |
| MM5M1 | 0 | 13 | 2 | 4 | 0 | 1 |
| MM5M2 | 1 | 2 | 0 | 0 | 17 | 0 |
| MM5M3 | 0 | 9 | 1 | 6 | 2 | 2 |

Tabla 9: Matriz de confusión de BI con todas las movidas y mejores atributos

Se desprende de la Tabla anterior que, en general, la clasificación realizada por el algoritmo se efectúa de forma óptima, pues de un total de 20 documentos que deberían ser clasificados en cada movida, el algoritmo BI clasifica de la MM1M1 14 de forma correcta, de la MM1M2 16, de la MM1M3 14 y de la MM5M2 17. Con respecto a esta matriz, resulta interesante que el algoritmo clasifica muy bien las movidas MM1M1 y MM5M2, pues casi no incluye documentos erróneos en estas movidas y clasifica más del 75% de sus textos de forma correcta.

El problema se produce especialmente con la movida MM5M3, pues en este caso, el clasificador tiende a identificar los textos pertenecientes a esta movida en todos las demás movidas exceptuando la MM1M1. Este último punto nos parece relevante, pues que no clasifique los textos en esta movida puede deberse a los recursos léxicos que se utilizan para cumplir los propósitos comunicativos de ella, es decir, solo en la MM1M1 se presentan aspectos generales de la investigación y estos no se encontrarán en la MM5M3 debido a los propósitos que se persiguen en esta movida (asociados a las proyecciones, hipótesis de resultados y críticas de la investigación).

### 4.2.2. Clasificación con Máquina de Soporte Vectorial (MSV/SMO)

A continuación, se presentan los resultados de la clasificación de las movidas de las macromovidas MM1 y MM5 con el algoritmo vectorial MSV.

| | MM1M1 | MM1M2 | MM1M3 | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|---|---|---|
| **Precisión** | 1 | 0,357 | 0,786 | 0,444 | 0,938 | 0,259 |
| **Recall** | 0,65 | 0,25 | 0,55 | 0,2 | 0,75 | 0,7 |
| **F- Measure** | 0,788 | 0,294 | 0,647 | 0,276 | 0,833 | 0,378 |
| **Kappa** | 0,42 | | | | | |
| **% clasificación correcta** | 51,6667 % | | | | | |

Tabla 10: Resumen clasificación MSV todas las movidas con los mejores atributos

De los resultados de la tabla anterior se desprende que la clasificación correcta realizada por el algoritmo Maquina de soporte vectorial (MSV) con los mejores 23 atributos (Ver Anexo 2) es de solo un 51,66% con un Kappa de 0,42, considerado aceptable. Los mejores resultados de clasificación se alcanzan en la MM5M2 con un F de 0,833 y en MM1M1 con un F de 0,788. Este último dato es significativo, pues en las anteriores clasificaciones realizadas con todos los atributos, la movida MM1M1 con ambos clasificadores alcanzaba un F máximo de 0,129, por lo que la clasificación con los mejores atributos mejoró significativamente. También es importante mencionar que si bien, no se logra una clasificación significativa de MM5M1, esta si mejora en comparación a la realizada con el mismo algoritmo y todos sus atributos. Lo anterior se puede atribuir a las características del algoritmo y el conjunto de atributos sin los mejores, ya que al tener una alta dispersión de los atributos, el algoritmo tiende a clasificar de manera errónea las clases, pero al seleccionar solo los mejores, disminuye su dispersión y, por ende, se mejora la clasificación.

El detalle de la clasificación y los resultados obtenidos a través de la matriz de confusión de este clasificador se ilustran en la siguiente Tabla.

| | MM1M1 | MM1M2 | MM1M3 | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|---|---|---|
| MM1M1 | 13 | 0 | 0 | 1 | 0 | 6 |
| MM1M2 | 0 | 5 | 0 | 3 | 0 | 12 |
| MM1M3 | 0 | 2 | 11 | 0 | 0 | 7 |
| MM5M1 | 0 | 4 | 1 | 4 | 0 | 11 |
| MM5M2 | 0 | 1 | 0 | 0 | 15 | 4 |
| MM5M3 | 0 | 2 | 2 | 1 | 1 | 14 |

Tabla 11: Matriz de confusión de MSV con todas las movidas y mejores atributos

En cuanto a los resultados de la matriz de confusión, es posible evidenciar que las movidas MM1M1, MM1M3, MM5M2 Y MM5M3 son las mejores clasificadas, mientras que las movidas MM1M2 Y MM5M1 son las que obtienen menores porcentajes de textos clasificados de forma satisfactoria. Con respecto a los resultados, resultan interesantes los obtenidos en MM1M1, pues en este caso, el algoritmo clasifica 13 textos de forma correcta y 7 de forma incorrecta, lo cual es significativo si se compara con la clasificación realizada con el mismo algoritmo y todos sus atributos (14.739).

En cuanto a los textos que el algoritmo clasifica de forma errónea en otras movidas, llama la atención que en su mayoría sean clasificados en la MM5M3. Lo anterior puede ser atribuido a los tri-gramas de lemas que aparecen en esta movida, pues suelen compartirse con movidas como la MM1M2 y MM5M1. A este último punto se debe agregar que no solo los atributos podrían influir en la clasificación, sino que las características de este algoritmo vectorial, también podrían explicar el resultado de estas clasificaciones, pues si los atributos no son representativos de la movida y sus frecuencias relativas no son significativas, el algoritmo clasificará de forma binaria (pertenece o no pertenece) con base en los vectores de soporte trazados.

### 4.2.3. Comparación de la clasificación entre BI y MSV

A continuación, se presenta un Gráfico que contiene la comparación entre las tareas de clasificación de todas las movidas de las macromovidas Introducir al lector (MM1) y Finalizar

discursivamente la investigación (MM5) realizadas con los algoritmos Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV) con sus 23 mejores atributos.

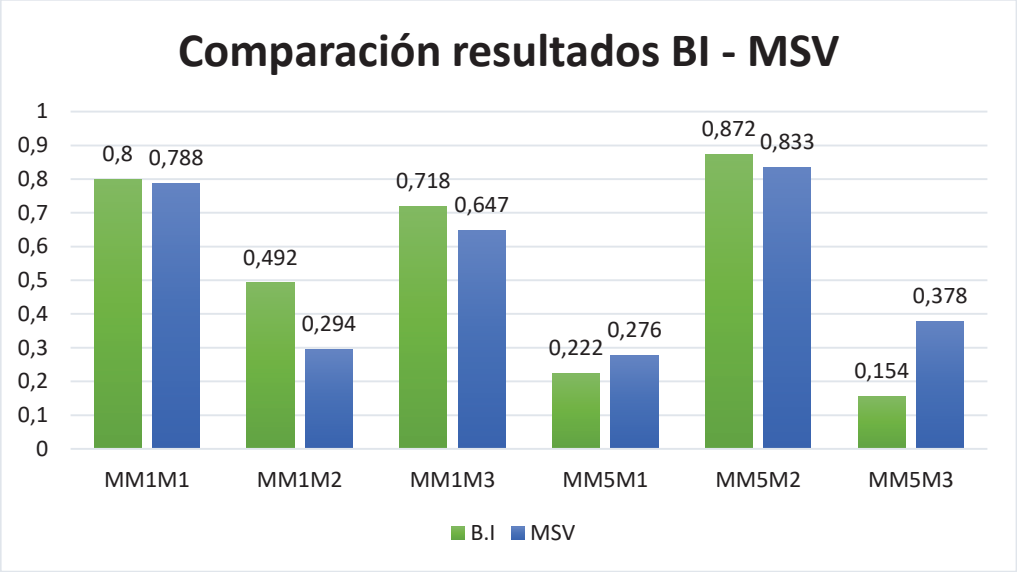



Gráfico 2: Comparación resultados BI y MSV con todas las movidas y mejores atributos

En el Gráfico anterior, se muestran las clasificaciones realizadas por ambos algoritmos con las movidas de las macromovidas Introducir al lector y Finalizar discursivamente la investigación y los mejores atributos. En comparación al Gráfico 1, se puede mencionar que las clasificaciones realizadas por ambos algoritmos mejoraron significativamente, pues sus Kappas aumentaron y sus F también lo hicieron, con excepción de las movidas MM5M1 y MM5M3.

De los datos obtenidos, se desprende que las clasificaciones de MM1M1, MM1M3 y MM5M2 son realizadas de forma más satisfactoria por el algoritmo Bayesiano ingenuo, aunque las diferencias entre ambos clasificadores es mínima. Con respecto a las peores clasificaciones, estas son las realizadas por ambos algortimos en MM1M2, MM5M1 y MM5M3. En esta última movida surge un dato interesante, pues MSV logra clasificarla con mejores resultados que BI, obteniendo un F, que si bien no es bueno, supera al alcanzado por Bayes. Este resultado se puede atribuir directamente a la naturaleza de ambos clasificadores, pues, al tener atributos compartidos con otras clases, el algoritmo BI toma la decisión de la clasificación con base en la probabilidad que el documento pertenezca a la clase, por lo que debido a las frecuencias relativas tiende a clasificar los documentos de la MM5M3 en las MM1M2. Por su parte, el

clasificador MSV tomará su decisión con base en la distribución que tengan los atributos respecto de los vectores, por lo que sí están dentro del corte para MM5M3 serán considerados como tal, a pesar de que por probabilidad pudiesen pertenecer a otra clase, pues este algoritmo al ser vectorial no tomará en cuenta el factor probabilístico.

## 4.3. Clasificación de las movidas de la macromovida Introducir al lector (MM1) con todos los atributos

En esta sección, se presentarán los resultados obtenidos por los algoritmos BI y MSV en la clasificación de las movidas de la macromovida introducir al lector (MM1) con todos sus atributos (14.739). De esta manera, primero se expondrán los resultados obtenidos por el primer algoritmo BI, luego se presentarán los datos obtenidos por MSV y, finalmente se realizará una comparación entre la clasificación de BI y MSV.

### 4.3.1. Clasificación con Bayesiano Ingenuo (BI)

Luego de realizar clasificaciones con todas las movidas de las dos macromovidas abordadas en esta investigación, se realizaron clasificaciones en las que solo se tomó una macromovida con el fin de reducir los atributos entregados al algoritmo. Los resultados obtenidos de la clasificación de todas las movidas de MM1 realizada por el algoritmo probabilístico BI con todos los atributos se ilustran en la Tabla número 12.

| | **MM1M1** | **MM1M2** | **MM1M3** |
|---|---|---|---|
| **Precisión** | 0,105 | 0,615 | 0,357 |
| **Recall** | 0,1 | 0,4 | 0,5 |
| **F- Measure** | 0,103 | 0,485 | 0,417 |
| **Kappa** | 0 | | |
| **% clasificación correcta** | 33,3333 % | | |

Tabla 12: Resumen clasificación BI con las movidas de MM1 y todos sus atributos

De la Tabla anterior se desprende que la clasificación realizada por Bayesiano ingenuo con todas las movidas de la macromovida Introducir al lector MM1 no es satisfactoria, puesto que el porcentaje de clasificación correcto de estas movidas alcanza solo un 33,33% con un Kappa de 0, obteniendo el nivel de kappa más bajo. En cuanto a las clasificaciones de cada movida, se puede apreciar que los equilibrios entre precisión y exhaustividad son realmente bajos, pues en la MM1M1 este valor alcanza solo un 0,103, siendo la clasificación más baja de las tres movidas que componen la macromovida Introducir al lector.

Las clasificaciones realizadas con las otras dos clases, también son bajas, ya que ninguna de ellas alcanza un nivel sobre el 0,5. Si bien, la mejor clasificación realizada por el algoritmo es la hecha de la MM1M2, esta no alcanza un nivel satisfactorio, pues solo recupera un 40% de los documentos de esta movida con una precisión de 0,615.

En la Tabla 13 se detallan las clasificaciones correctas y erróneas realizadas por este algoritmo probabilístico.

| | **MM1M1** | **MM1M2** | **MM1M3** |
|---|---|---|---|
| MM1M1 | 2 | 5 | 13 |
| MM1M2 | 7 | 8 | 5 |
| MM1M3 | 10 | 0 | 10 |

Tabla 13: Matriz de confusión de BI con las movidas de MM1 y mejores atributos

A partir de los resultados de la matriz de confusión, se puede observar que el clasificador BI tiende a clasificar los documentos en todas las clases sin discriminar. Si bien, la mejor recuperación de documentos es la realizada con MM1M3, esta solo alcanza un 50% de clasificación correcta, por lo que no es considerada satisfactoria. Una posible explicación a este resultado podría ser atribuida a la utilización de todos los atributos (14.739) de las movidas de la macromovida Introducir al lector, ya que al ser muy parecidos entre ellos, poseer frecuencias relativas bajas y ser demasiados, el clasificador puede obtener resultados confusos. A este punto se debe agregar la naturaleza de Bayes, dado que al ser un algoritmo probabilístico y no tener atributos representativos de cada movida, tiende a clasificar los documentos de forma azarosa.

Se debe agregar a los datos ya expuestos los valores de F, pues en la matriz de confusión apreciamos que la movida mejor clasificada es la MM1M3, sin embargo, al observar F, nos

percatamos que este valor es bajo, por lo que la clasificación realizada de esta movida puede ser azarosa, ya que la precisión de la clasificación es de 0,357.

### 4.3.2. Clasificación con Máquina de Soporte Vectorial (MSV/SMO)

A continuación, se presentan los resultados obtenidos por MSV en la clasificación realizada de las movidas de la macromovida introducir al lector MM1M1 con todos los atributos.

| | **MM1M1** | **MM1M2** | **MM1M3** |
|---|---|---|---|
| **Precisión** | 0,667 | 0,37 | 0,667 |
| **Recall** | 0,1 | 1 | 0,1 |
| **F- Measure** | 0,174 | 0,541 | 0,174 |
| **Kappa** | 0,1 | | |
| **% clasificación correcta** | 40% | | |

Tabla 14: Resumen clasificación MSV con las movidas de MM1 y todos sus atributos

Como se presenta en la Tabla anterior, los resultados de la clasificación realizada por Maquina de soporte vectorial alcanzan un 40% con un Kappa de 0,1, lo que es considerado insatisfactorio, pues el algoritmo no alcanza a clasificar ni siquiera la mitad del corpus de forma correcta. En cuanto a la clasificación de las movidas, podemos notar que los mejores valores son los alcanzados en MM1M2, pues recupera el 100% de los documentos, aunque con una precisión de clasificación muy baja, ya que solo alcanza un 0,37.

En cuanto a la clasificación realizada con las otras dos movidas, se evidencia que la recuperación es muy baja en comparación a la de MM1M2, sin embargo, la precisión de clasificación en estas dos movidas es significativamente superior.

En la siguiente Tabla se ilustran los detalles de la clasificación de las movidas de la macromovida Introducir al lector MM1M1 realizada por MSV con sus 14.739 atributos.

| | MM1M1 | MM1M2 | MM1M3 |
|---|---|---|---|
| MM1M1 | 2 | 17 | 1 |
| MM1M2 | 0 | 20 | 0 |
| MM1M3 | 1 | 17 | 2 |

Tabla 15: Matriz de confusión de MSV con las movidas de MM1 y mejores atributos

De la matriz anterior, se puede desprender que el algoritmo MSV tiende a clasificar muy mal las movidas MM1M1 y MM1M3, pues solo logra recuperar el 10% de los documentos pertenecientes a estas movidas. Resulta significativo que la mayoría de los documentos asociados a la MM1M1 sean clasificados en la MM1M2. Algo similar a esto es lo que ocurre con MM1M3, dado que clasifica 17 de los documentos en MM1M2. Los resultados obtenidos en estas movidas llaman la atención, ya que solo logra clasificar bien MM1M2 y la mayoría de los documentos los atribuye a ella. Una posible explicación a este fenómeno se puede deber a los rasgos léxicos utilizados en MM1M1 y MM1M3 para instanciar los propósitos comunicativos. En este sentido, por ejemplo, la MM1M1 posee tri-gramas de lemas que tienen frecuencias relativas muy bajas en comparación a los valores de MM1M2, a su vez, los atributos de esta movida suelen aparecer en otras movidas o clases, por lo que el algoritmo al representar multidimensionalmente estos datos y tener una alta dispersión, tenderá a clasificar los documentos con base en los vectores, pero debido a la homogeneidad de los atributos podría clasificarlos algunos de forma errónea.

### 4.3.3. Comparación de la clasificación entre BI y MSV

A continuación, se presenta un Gráfico que contiene la comparación entre las tareas de clasificación de todas las movidas de la macromovida Introducir al lector (MM1) realizadas con los algoritmos Bayesiano ingenuo (BI) y Maquina de soporte vectorial (MSV) con todos sus atributos.

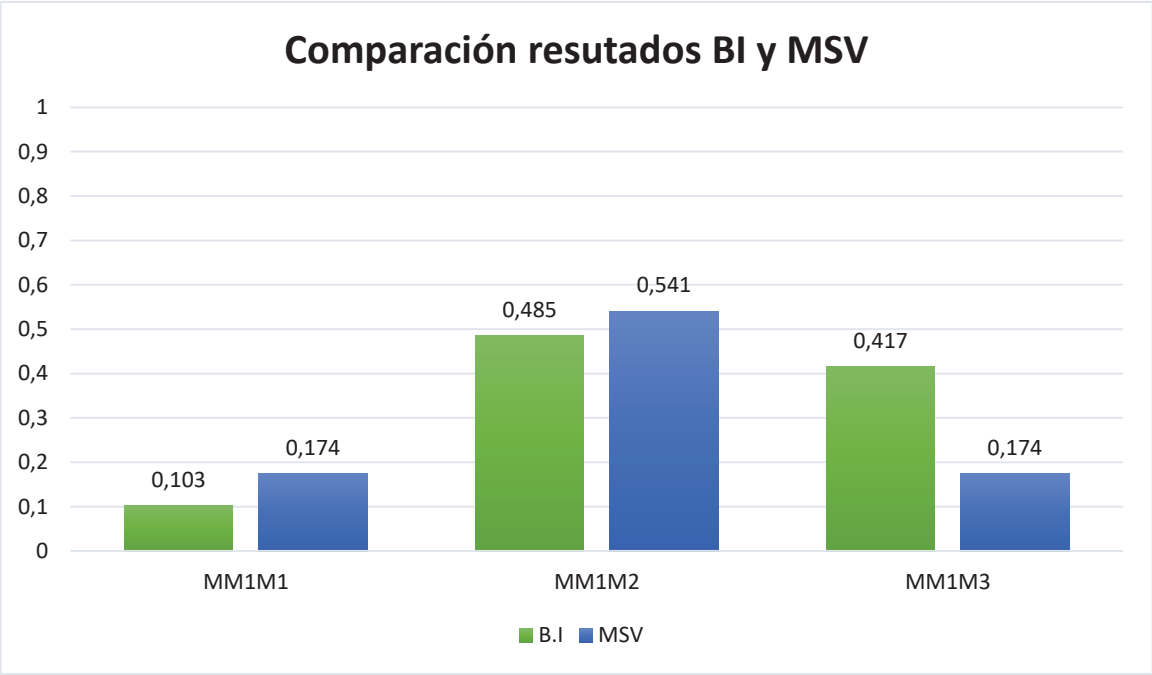



Gráfico 3: Comparación resultados BI y MSV con las movidas de MM1 y todos sus atributos

A partir del Gráfico anterior, es posible desprender que existen algunas diferencias entre las tareas de clasificación realizadas por BI y MSV, pues en las tres movidas de la macromovida Introducir al lector se alcanzan diferencias significativas entre ambos algoritmos. Si bien, ninguna de las dos clasificaciones logra un nivel satisfactorio, llama la atención que en MM1M1 y MM1M2 la mejor clasificación la realice MSV, pues desde la literatura se tiende a sostener que BI realiza tareas de clasificación con mejores resultados que MSV. Desde una interpretación de los resultados, se podría atribuir esta baja clasificación a los tri-gramas de lemas utilizados y a sus frecuencias relativas, pues al observar detenidamente la matriz de frecuencias utilizada para la clasificación, se evidencia una alta cantidad de tri-gramas de lemas con frecuencias relativas bajas y tri-gramas compartidos entre muchas movidas.

## 4.4. Clasificación de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) con todos sus atributos

En la presente sección, se expondrán los resultados obtenidos por los algoritmos Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV) en la clasificación de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) con sus 14.739 atributos. Con el fin de presentar los datos, primero se desplegará la clasificación realizada por BI, luego la ejecutada por MSV y, finalmente, una comparación entre las tareas realizadas por ambos algoritmos.

### 4.4.1. Clasificación con Bayesiano Ingenuo (BI)

En la Tabla 16 se presentan los resultados obtenidos por el algoritmo BI en la clasificación de las movidas retóricas de la macromovida Finalizar discursivamente la investigación (MM5).

| | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|
| **Precisión** | 0,438 | 0,515 | 0,455 |
| **Recall** | 0,35 | 0,85 | 0,25 |
| **F- Measure** | 0,389 | 0,642 | 0,323 |
| **Kappa** | 0,225 | | |
| **% clasificación correcta** | 48,33% | | |

Tabla 16: Resumen clasificación BI con las movidas de MM5 y todos los atributos

De los datos obtenidos, se desprende que la clasificación de las movidas de la MM5 realizada por el BI alcanza un porcentaje de clasificación correcta de un 48,33% con un Kappa de 0,225. A partir de estos datos, se establece que la tarea ejecutada por BI no alcanza un nivel satisfactorio, pues ni siquiera logra clasificar correctamente el 50% del corpus total y su Kappa obtenido es considerado entre leve y aceptable.

De la clasificación por movidas, se puede mencionar que los mejores resultados son los obtenidos en la MM5M2, pues en este caso, BI obtiene un F de 0,642, recuperando un 85% de

los textos pertenecientes a esta movida con una precisión del 0,515. Si bien, este resultado es relativamente bueno, la precisión alcanzada por este algoritmo en las tres movidas es baja, lo cual no asegura que la tarea de clasificación se pueda realizar de forma confiable.

| | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|
| MM5M1 | 7 | 9 | 4 |
| MM5M2 | 1 | 17 | 2 |
| MM5M3 | 8 | 7 | 5 |

Tabla 17: Matriz de confusión de BI con las movidas de MM5 y todos los atributos

En la matriz anterior se observa que, a nivel general, existe una clasificación errónea de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) con los 14.739 atributos. Con respecto a las clasificaciones de MM5M1, se evidencia que la mayoría de los documentos (9) son clasificados en la movida MM5M2 y unos pocos (4) los identifica como MM5M3. En el caso de MM5M2, la mayoría de los textos son recuperados de forma correcta, aunque la precisión con la que se realiza la clasificación es considerada baja. En cuanto a la movida MM5M3, esta tiene una clasificación de textos aún más baja que la de MM5M1, pues Bayes solo clasifica de forma correcta un 25% de los textos con una precisión de 0,45. Sumado a lo anterior, el algoritmo BI tiende a identificar los textos de esta movida como MM5M1 y MM5M2. Este último dato resulta interesante, pues la mayoría de los textos clasificados de forma errónea por este algoritmo son identificados como parte de MM5M2, lo cual se puede deber, en gran parte, a las bajas frecuencias de los atributos y los tri-gramas de lemas compartidos entre las tres movidas clasificadas en este acápite.

### 4.4.2. Clasificación con Máquina de Soporte Vectorial (MSV/SMO)

A continuación, se presentan los datos obtenidos por el algoritmo MSV en la clasificación de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) con todos sus atributos.

| | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|
| **Precisión** | 1 | 0,455 | 0,362 |
| **Recall** | 0,1 | 0,25 | 0,85 |
| **F- Measure** | 0,182 | 0,323 | 0,507 |
| **Kappa** | 0,1 | | |
| **% clasificación correcta** | 40% | | |

Tabla 18: Resumen clasificación MSV con las movidas de MM5 y todos sus atributos

A partir de la Tabla número 18, se puede deducir que la clasificación realizada por Máquina de soporte vectorial es insatisfactoria, pues solo clasifica correctamente un 40% de los textos del corpus con un Kappa de 0,1, lo cual es considerado leve y no significativo, ya que el Kappa deseado debe ser de un 0,6 en adelante. En cuanto a las clasificaciones de cada movida, se observa que los mejores valores son los alcanzados en la de MM5M3, pues en este caso, el equilibrio entre precisión y exhaustividad es de 0,57, por lo que el algoritmo recupera el 85% de los textos con una precisión de 0,362. Si bien, la clasificación es la más alta de las movidas de la macromovida Finalizar discursivamente la investigación (MM5), la precisión de clasificación es muy baja, por lo que la tarea continúa siendo insatisfactoria. Respecto de la movida MM5M1, el algoritmo solo recupera el 10% de los documentos, pero lo realiza con una muy buena precisión. En el caso de MM5M2, solo es recuperado el 25% de los documentos y la precisión es de solo 0,455, por lo que a diferencia de lo que ha ocurrido en otras clasificaciones realizadas, el algoritmo MSV no clasifica de forma satisfactoria la movida MM5M2.

En la Tabla 19 se despliega la matriz de confusión que contiene los resultados de la clasificación de las movidas de la MM5 con el algoritmo Máquina de soporte vectorial con todos sus atributos.

| | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|
| MM5M1 | 2 | 3 | 15 |
| MM5M2 | 0 | 5 | 15 |
| MM5M3 | 0 | 3 | 17 |

Tabla 19: Matriz de confusión de MSV con las movidas de MM5 y todos sus atributos

En la tabla anterior se observa que la clasificación realizada por (MSV), en general es errónea, a excepción de la MM5M3 que logra clasificar 17 textos de forma correcta. Con respecto a la movida MM5M1, en ella es donde el algoritmo logra recuperar menos textos, pues de un total de 20, solo logra clasificar 2, es decir el 10%. A pesar de este mal resultado, resulta interesante que MSV, en este caso particular, no incluye textos pertenecientes a otras movidas en ella.

De los resultados obtenidos, resulta significativo que el algoritmo clasifique los textos de las movida MM5M1 y MM5M2 en la MM5M3, pues la gran mayoría de los documentos son recuperados en esta movida de forma errónea. Lo anterior, podría ser explicado debido a, como se ha mencionado en acápites anteriores, la naturaleza del clasificador vectorial y los tri-gramas de lemas de esta movida con sus respectivas frecuencias relativas, pues estos suelen compartirse con movidas como la MM5M1 y MM5M2 y en dichas movidas sus frecuencias suelen ser muy bajas.

### 4.4.3. Comparación de la clasificación entre BI y MSV

A continuación se presenta un Gráfico que contiene la comparación entre las tareas de clasificación de todas las movidas de la macromovida Finalizar discursivamente la investigación (MM5) realizadas con los algoritmos BI y MSV, con los 14.739 atributos.

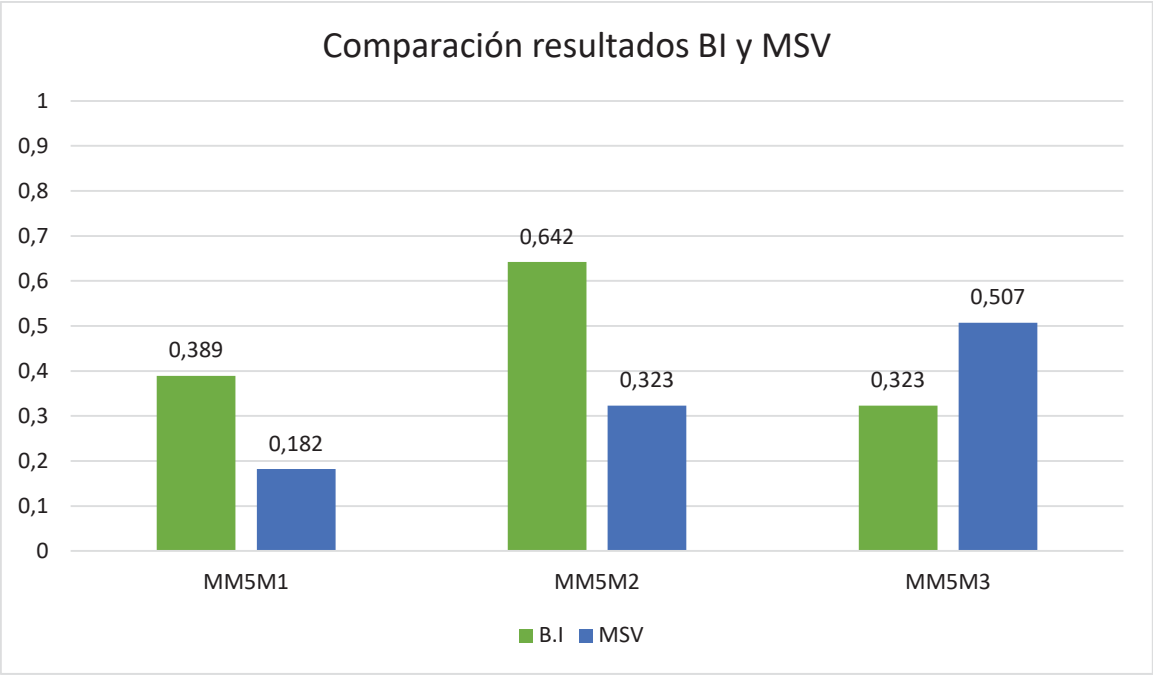



Gráfico 4: Comparación resultados BI y MSV con las movidas de MM5 y todos sus atributos

A partir del Gráfico anterior, se evidencia un comportamiento significativamente diferente entre ambos clasificadores, pues obtienen resultados que en cada movida varían. En la movida MM5M1, el clasificador probabilístico alcanza un F mayor que el del clasificador vectorial, aunque en ninguno de los casos resulta una clasificación satisfactoria, pues sus F son muy bajos. Con respecto a la MM5M2, en ella surge la primera diferencia notoria entre ambos clasificadores, ya que BI logra un F de 0,64, lo que es considerado bueno, pues recupera 17 textos de 20 con una precisión de 0,515. Mientras que Máquina de soporte vectorial alcanza un F de 0,323, recuperando solo 5 textos de forma correcta con una precisión de 0,455.

En cuanto a los resultados de la clasificación de la movida MM5M3, resulta interesante que, en este caso, el algoritmo MSV alcanza un valor de F mayor al de BI, por lo que recupera 17 textos de forma correcta con una precisión de 0,362. Si bien, esta precisión es baja, la recuperación de textos es alta y su F es significativo, por lo que en esta movida en particular, el algoritmo Máquina de soporte vectorial (MSV) realiza una clasificación mejor que Bayesiano Ingenuo (BI).

Tomando en cuenta los resultados ya expuestos y el Kappa obtenido en cada clasificación, en esta comparación se ha demostrado que el clasificador Bayesiano ingenuo (BI), en general,

obtiene mejores resultados de clasificación de las movidas de la macromovida Finalizar discursivamente la investigación cuando se toman en consideración todos los atributos para realizar esta tarea.

## 4.5. Clasificación de las movidas de la macromovida Introducir al lector (MM1) con los mejores atributos

En la presente sección, se expondrán las clasificaciones de las movidas de la macromovida Introducir al lector (MM1) realizadas con BI y MSV con la selección de sus 26 mejores atributos (Ver Anexo 3). Para presentar los resultados, en primer lugar, se presentaran los datos obtenidos de la clasificación con el algoritmo probabilístico (BI). Luego, se desplegarán los resultados logrados por Máquina soporte vectorial (MSV) y, finalmente, se realizará una comparación entre la clasificación efectuada por ambos algoritmos.

### 4.5.1. Clasificación con Bayesiano Ingenuo (BI)

En la Tabla 20, se presentan los resultados obtenidos por el algoritmo Bayesiano ingenuo (BI) en la clasificación de las movidas retóricas de la macromovida Introducir al lector (MM1) con sus 26 mejores atributos.

| | **MM1M1** | **MM1M2** | **MM1M3** |
|---|---|---|---|
| **Precisión** | 1 | 0,826 | 0,909 |
| **Recall** | 0,75 | 0,95 | 1 |
| **F- Measure** | 0,857 | 0,884 | 0,952 |
| **Kappa** | 0,85 | | |
| **% clasificación correcta** | 90% | | |

Tabla 20: Resumen clasificación BI con las movidas de MM1 y sus mejores atributos

A partir de los datos expuestos en la Tabla anterior, es posible señalar que la clasificación de las movidas de la macromovida Introducir al lector realizada con BI con sus mejores atributos alcanza un 90% de recuperación de textos con un Kappa de 0,85. Lo anterior es significativo, pues por primera vez en las clasificaciones realizadas se obtiene un Kappa considerado casi perfecto y una clasificación correcta de las clases tan alta. En cuanto a la clasificación de cada movida de MM1, se observa que MM1M3 obtiene los mejores resultados, ya que alcanza un equilibrio entre precisión y exhaustividad de 0,952, recuperando el 100% de los documentos con una precisión de 0,909.

Con respecto a los MM1M2, su clasificación también es alta, dado que logra un F de 0,884, clasificando el 95% de los textos con una precisión de 0,826. En cuanto a la clasificación de MM1M1, esta es la más baja de las hechas en esta movida, pues alcanza un equilibrio entre precisión y exhaustividad de 0,857, recuperando el 75% de los textos con una precisión de 1, lo que permite que no incluya textos de otras movidas en ella.

En la siguiente Tabla, se despliega el detalle de las clasificaciones realizadas por Bayesiano ingenuo en la clasificación de las movidas de la macromovida Introducir al lector.

| | **MM1M1** | **MM1M2** | **MM1M3** |
|---|---|---|---|
| MM1M1 | 15 | 4 | 1 |
| MM1M2 | 0 | 19 | 1 |
| MM1M3 | 0 | 0 | 20 |

Tabla 21: Matriz de confusión de BI con las movidas de MM1 y sus mejores atributos

Los datos expresados en la Tabla 21 permiten señalar que, en general, la clasificación realizada por el algoritmo Bayesiano ingenuo es muy buena, dado que, en el caso de las movidas MM1M2 y MM1M3 la mayoría de los documentos son recuperados de forma correcta, a excepción de un documento de MM1M2 que es clasificado en MM1M3. Con respecto a la clasificación de la movida MM1M1, esta, si bien es bastante buena, el algoritmo BI clasifica algunos documentos en la MM1M2, lo cual se podría deber a la semejanza que existe entre los atributos de las movidas y la naturaleza del clasificador probabilístico.

### 4.5.2. Clasificación con Máquina de Soporte Vectorial (MSV/SMO)

Los resultados concernientes a la clasificación de las movidas de la macromovida Introducir al lector MM1 realizada por el algoritmo Máquina de soporte vectorial con sus mejores 26 atributos (Ver Anexo 3) son presentados en la siguiente Tabla.

| | MM1M1 | MM1M2 | MM1M3 |
|---|---|---|---|
| **Precisión** | 1 | 0,714 | 1 |
| **Recall** | 0,75 | 1 | 0,85 |
| **F- Measure** | 0,857 | 0,833 | 0,919 |
| **Kappa** | 0,8 | | |
| **% clasificación correcta** | 86,66% | | |

Tabla 22: Resumen clasificación MSV con las movidas de MM1 y sus mejores atributos

A partir de la lectura de la Tabla número 22, se puede establecer que la clasificación de las movidas de la macromovida Introducir al lector realizada por el algoritmo Máquina de soporte vectorial es buena, ya que alcanza un porcentaje de clasificación correcta de un 86,66% con un Kappa de 0,8.

En cuanto a la clasificación de las movidas, se observa que la MM1M3 es la mejor clasificada por este algoritmo, dado que logra un equilibrio entre precisión y exhaustividad de 0,919, recuperando el 85% de los documentos con una precisión de 1. En cuanto a la MM1M1, el valor que alcanza F es de 0,857, lo que significa que recupera el 75% de los textos pertenecientes a esta movida con una casi perfecta precisión. Con respecto a la MM1M2, si bien, obtiene un valor F inferior a las demás movidas, 0,833, este continúa siendo alto, pues recupera el 100% de los documentos con una precisión de 0,714.

El detalle de las clasificaciones ejecutadas por el algoritmo vectorial (MSV) se presenta en la Tabla 23.

| | MM1M1 | MM1M2 | MM1M3 |
|---|---|---|---|
| MM1M1 | 15 | 5 | 0 |
| MM1M2 | 0 | 20 | 0 |
| MM1M3 | 0 | 3 | 17 |

Tabla 23: Matriz de confusión de MSV con las movidas de MM1 y sus mejores atributos

De la matriz de confusión anterior se desprende que la MM1M2 es la movida que mejor recuperación de textos obtiene, sin embargo, en ella se clasifican algunos textos que pertenecen a las otras movidas de esta macromovida. Con respecto a los documentos clasificados en MM1M3, de un total de 20 textos recupera 17 y clasifica erróneamente 3 en la MM1M2. En el caso de a MM1M1, esta movida alcanza la menor recuperación de textos de las movidas de MM1, pues 5 de sus documentos son clasificados en la movida MM1M2. Al respecto, resulta interesante que este clasificador tienda a recuperar erróneamente algunos documentos en la movida MM1M2, lo cual se relaciona directamente con el valor de precisión de esta movida. Desde la interpretación, podríamos mencionar que lo anterior se atribuye a la existencia de algunos tri-gramas de lemas compartidos entre las tres movidas de la macromovida Introducir al lector. Si al dato anterior, se suma la naturaleza del algoritmo y las bajas frecuencias relativas de algunos de los atributos que comparten con la MM1M2, es esperable que el MSV clasifique mal los documentos que presentan las características mencionadas, pues desde la literatura se menciona que este algoritmo necesita atributos significativos y con altas frecuencias para realizar clasificaciones satisfactorias (Cárdenas, Olivares & Alfaro, 2014).

### 4.5.3. Comparación de la clasificación entre BI y MSV

A continuación se presenta el Gráfico 5 que contiene la comparación entre las tareas de clasificación de todas las movidas de las macromovidas Introducir al lector (MM1) realizadas con los algoritmos Bayesiano ingenuo (BI) y Maquina de soporte vectorial (MSV) con sus mejores 26 atributos (Ver Anexo 3).

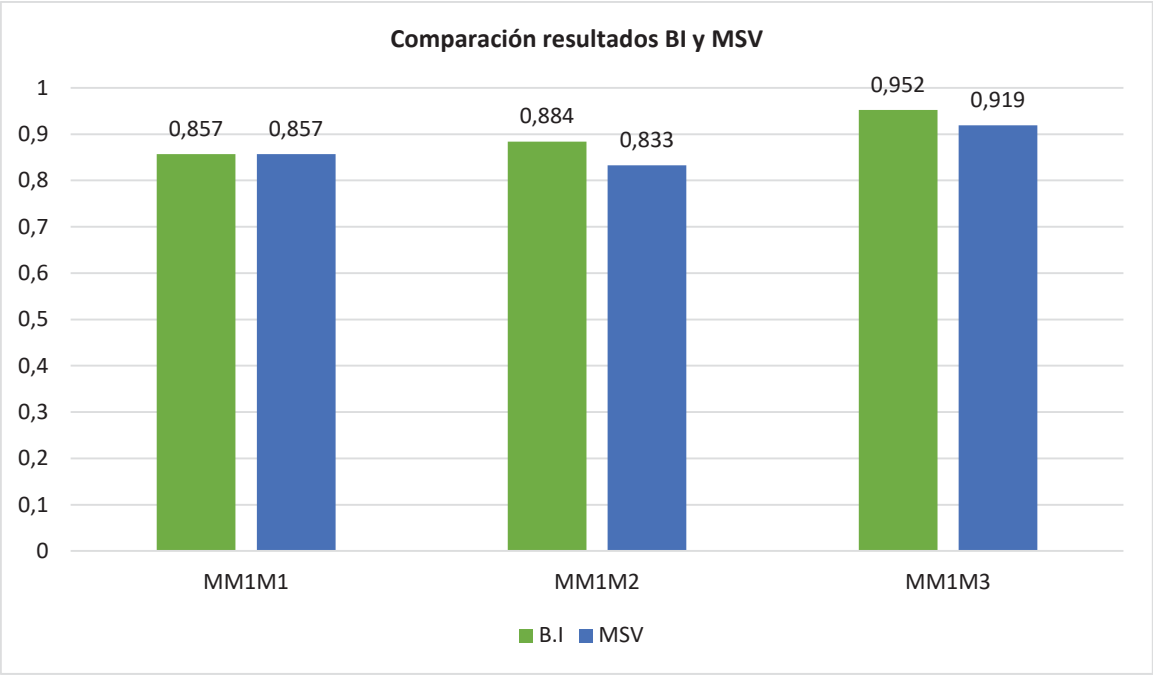



Gráfico 5: Comparación resultados BI y MSV con las movidas de MM1 y sus mejores atributos

Como se puede desprender del gráfico 5, las clasificaciones de las movidas de la macromovida Introducir al lector (MM1) realizadas por los algoritmos BI y MSV con sus mejores 26 atributos son consideradas buenas, pues todas ellas alcanzan un equilibrio entre precisión y exhaustividad mayor o igual a 0,833.

En cuanto a las diferencias entre los dos clasificadores, en la MM1M1 ambos algoritmos tienen un desempeño idéntico, dado que alcanzan el mismo valor de F. Con respecto a MM1M2, en este caso si existe una pequeña diferencia entre Bayesiano ingenuo y Máquina de soporte vectorial, ya que BI obtiene un resultado sutilmente superior al de MSV. Parecida es la clasificación de MM1M3, pues, en este caso, ambos clasificadores alcanzan valores similares, aunque BI, nuevamente obtiene un valor levemente superior al de MSV.

En conclusión, ambos algoritmos realizan una clasificación satisfactoria de las movidas de la macromovida Introducir al lector con sus mejores atributos, a pesar de que los resultados indican que los logros de BI son levemente superiores a los alcanzados por MSV.

## 4.6. Clasificación de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) con los mejores atributos

En la presente sección, se expondrán los resultados obtenidos en la clasificación de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) realizada con los algoritmos Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV) con los mejores 24 atributos (Ver Anexo 4).

### 4.6.1. Clasificación con Bayesiano Ingenuo (BI)

Los resultados de la clasificación de las movidas retóricas de la macromovida Finalizar discursivamente la investigación (MM5) realizada con el algoritmo probabilístico Bayesiano ingenuo (BI) se presentan en la Tabla 24.

| | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|
| **Precisión** | 1 | 0,952 | 0,594 |
| **Recall** | 0,35 | 1 | 0,95 |
| **F- Measure** | 0,519 | 0,976 | 0,731 |
| **Kappa** | 0,65 | | |
| **% clasificación correcta** | 76,66% | | |

Tabla 24: Resumen clasificación BI con las movidas de MM5 y sus mejores atributos

Como se desprende de la Tabla número 24, los resultados de la clasificación de las movidas retóricas de la macromovida Finalizar discursivamente la investigación obtienen un porcentaje de clasificación de 76,66% con un Kappa de 0,65, lo cual es valorado como considerable. Con respecto a la clasificación de cada movida, la MM5M2 es la que obtiene mejores resultados de clasificación, pues su equilibrio entre precisión y exhaustividad es de 0,976, recuperando el 100% de los documentos pertenecientes a esta movida con una precisión de 0,952. En el caso de MM5M3, el valor de F es de un 0,731, lo que significa que clasifica el 95% de los textos de forma correcta y lo hace con una precisión de 0,594. Si bien, la recuperación de los documentos de esta movida es muy buena, su precisión indica que la clasificación de este algoritmo es alta,

pero puede presentar algunos errores de clasificación. Con respecto a MM5M1, en este caso, el valor de F es de 0,519, considerado bajo, por lo que la tarea de clasificación automática de esta movida es realizada de forma compleja. Lo anterior concuerda con los resultados obtenidos, pues el algoritmo en esta movida solo recupera un 35% de los textos, aunque la precisión con que lo realiza es muy buena, pues obtiene un valor máximo.

A continuación se despliega la matriz de confusión de la clasificación ya mencionada.

| | **MM5M1** | **MM5M2** | **MM5M3** |
|---|---|---|---|
| MM5M1 | 7 | 0 | 13 |
| MM5M2 | 0 | 20 | 0 |
| MM5M3 | 0 | 1 | 19 |

Tabla 25: Matriz de confusión de BI con las movidas de MM5 y sus mejores atributos

A partir de los resultados de la Tabla 25, es posible observar que la movida que mejor es clasificada por el algoritmo BI es la MM5M2, dado que todos los documentos de esta movida son recuperados de forma correcta. En cuanto a MM5M3, en ella surgen dos datos interesantes, pues por una parte, recupera el 95% de los textos de esta movida y, por otra parte, el algoritmo clasifica erróneamente 13 documentos que no pertenecen a esta movida. Estos falsos positivos pueden atribuirse a los rasgos léxicos utilizados para cumplir con los propósitos comunicativos de las movidas MM5M3 y MM5M1 y por ende, en la similitud de sus tri-gramas de lemas. Sumado a lo anterior, los atributos de la MM5M1 tienen frecuencias relativas muy bajas, por lo que en algunas ocasiones el algoritmo probabilístico puede clasificar sus documentos en otras clases donde debido a los valores de las frecuencias la probabilidad de pertenecer sea mayor.

### 4.6.2. Clasificación con Máquina de Soporte Vectorial (MSV/SMO)

En la Tabla número 26, se presentan los resultados de la clasificación de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) con el algoritmo Máquina de soporte vectorial (MSV) y los mejores 24 atributos de las movidas (Ver Anexo 4).

| | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|
| **Precisión** | 1 | 1 | 0,606 |
| **Recall** | 0,45 | 0,9 | 1 |
| **F- Measure** | 0,621 | 0,947 | 0,755 |
| **Kappa** | 0,675 | | |
| **% clasificación correcta** | 78,33% | | |

Tabla 26: Resumen clasificación MSV con las movidas de MM5 y sus mejores atributos

De la Tabla anterior, se desprende que la clasificación correcta de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) alcanza un porcentaje de 78,33% con un Kappa de 0,675, lo que es considerable para tareas de clasificación. Respecto de las clasificaciones de cada movida, la mejor clasificación es la realizada en la MM5M2, pues en este caso, el algoritmo obtiene un F de 0,974, lo que significa que recupera el 90% de los documentos de esta movida con una precisión muy alta. En el caso de la MM5M3, el F alcanzado es de 0,755, donde si bien recupera el 100% de los documentos, la precisión con que lo hace solo alcanza un 0,606. La movida que alcanza la peor clasificación de las movidas de MM5 es MM5M1, pues en este caso, el valor de F alcanzado es de 0,621, lo que se traduce en una recuperación de documentos del 45% con una precisión de 1. Si bien el valor de la precisión es muy bueno, la recuperación de documentos es muy baja, por lo que se considera insatisfactoria la clasificación de esta movida.

En la Tabla 27 se detallan los datos de la clasificación de las movidas de la macromovida MM5 realizada por el algoritmo MSV.

| | MM5M1 | MM5M2 | MM5M3 |
|---|---|---|---|
| MM5M1 | 9 | 0 | 11 |
| MM5M2 | 0 | 18 | 2 |
| MM5M3 | 0 | 0 | 20 |

Tabla 27: Matriz de confusión de MSV con las movidas de MM5 y sus mejores atributos

A partir de la matriz de confusión anterior, se observa que la movida que recupera el menor porcentaje de documentos es la MM5M1, pues clasifica 9 textos de los 20 que pertenecen a esta movida. Con respecto a MM5M2, la clasificación mejora bastante, pues clasifica 18 documentos bien y solo deja 2 documentos en otra movida. En el caso de MM5M3, el algoritmo clasifica bien los 20 textos pertenecientes a esta movida, sin embargo, clasifica otros documentos en ella que no le pertenecen. Lo anterior podría atribuirse a la dispersión que existe en los atributos de MM5M1, las bajas frecuencias relativas de ellos y el hecho que estos atributos se comparten en algunos documentos de esta movida y la MM5M3.

### 4.6.3. Comparación de la clasificación entre BI y MSV

A continuación, se presenta un Gráfico que contiene la comparación entre las tareas de clasificación de todas las movidas de la macromovida Finalizar discursivamente la investigación (MM5) realizadas con los algoritmos Bayesiano ingenuo (BI) y Maquina de soporte vectorial (MSV) con sus mejores 24 atributos.

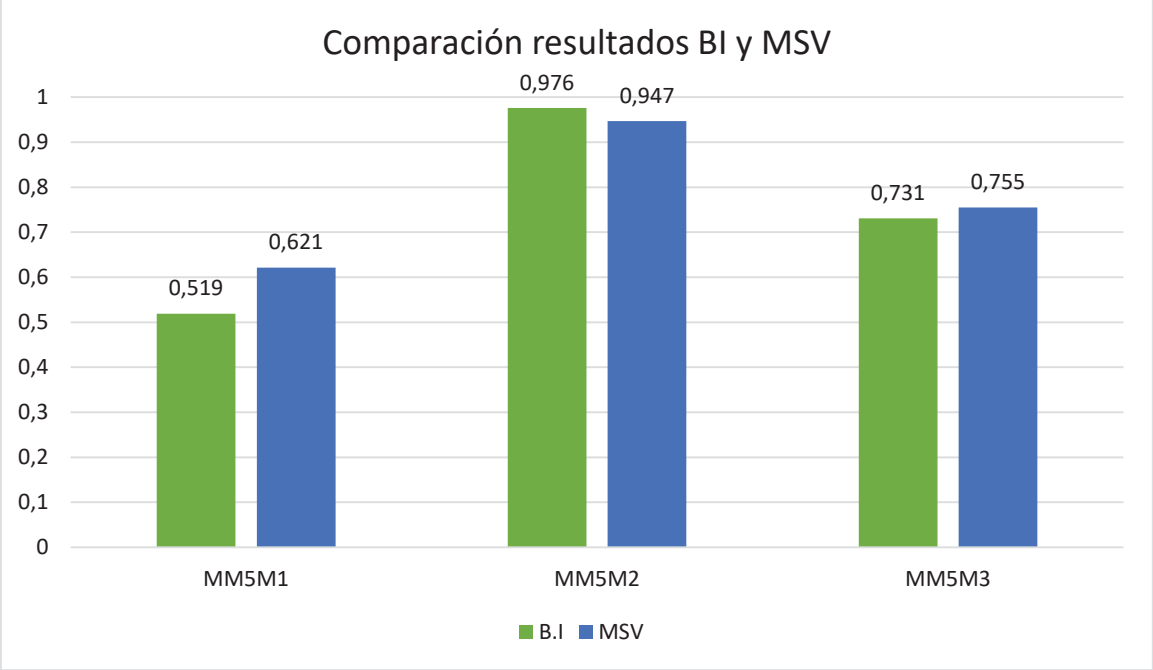



Gráfico 6: Comparación resultados BI y MSV con las movidas de MM5 y sus mejores atributos

A partir del Gráfico anterior, es posible desprender que existen algunas diferencias entre las clasificaciones de las movidas de la macromovida Finalizar discursivamente la investigación realizadas por los algoritmos Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV). Si bien, estas diferencias no son significativas, en la MM5M1 se aprecia que el algoritmo que mejor clasifica esta movida es el vectorial, superando en 0,102 el valor de F obtenido por Bayes. En el caso de MM5M2, los valores de F se invierten, dado que, en este caso, el valor de equilibrio entre precisión y exhaustividad alcanzado por Bayesiano ingenuo es sutilmente superior al del algoritmo vectorial. Con respecto a la movida MM5M3, los valores de F son muy parecidos, aunque MSV supera en 0,024 el valor de F alcanzado por el algoritmo BI. Finalmente, si bien, las clasificaciones de las movidas de la macromovida Finalizar discursivamente la investigación (MM5) son buenas, la realizada en la MM5M1 por ambos algoritmos es significativamente más baja. Lo anterior se puede interpretar debido a las bajas frecuencias relativas que tienen los atributos de estas movidas, pues si fuesen diferencias debido a la naturaleza de los algoritmos, no serían ambos resultados tan bajos.

### 4.7.Discusión

A partir de los resultados analizados en el apartado anterior, se puede establecer que los algoritmos Bayesiano ingenuo (BI) y Máquina de soporte vectorial (MSV) no realizan clasificaciones semiautomáticas satisfactorias cuando tienen como clases a todas las movidas de MM1 y MM5 y todos los atributos de estas movidas. Con respecto a estos resultados, sostenemos que se deben en gran parte a la cantidad de atributos -tri-gramas de lemas- que se obtienen de todas estas movidas, dado que los algoritmos reciben 14.739 atributos con frecuencias relativas muy bajas y, en algunos casos, muchos atributos compartidos entre varias movidas.

La representatividad que los atributos hacen de las clases resulta trascendental al momento de realizar clasificaciones semiautomáticas, ya que si los atributos no son representativos de las clases, cualquiera de los dos algoritmos realizará clasificaciones erróneas. En este sentido,

debido a su naturaleza probabilística, por un lado, Bayes, al tener como atributos tri-gramas de lemas compartidos entre las clases y con frecuencias relativas muy bajas, es decir, atributos poco representativos, tenderá a realizar las clasificaciones de forma azarosa y poco precisas.

Por su parte, el clasificador vectorial (MSV), también realizará clasificaciones inexactas si los atributos no son representativos, ya que si estos no "caracterizan" o representan a la clase y sus frecuencias relativas son bajas, realizará una representación multidimensional homogénea, con alta dispersión y poca representatividad de los datos. Una representación multidimensional con estas características solo dificultará la tarea de clasificación, ya que al realizarla con un algoritmo que clasifica los documentos a través de un hiperplano, a pesar de poder utilizar el truco de Kernel para mejorar las clasificaciones, MSV no podrá separar los documentos de forma satisfactoria cuando estén superpuestos, homogeneizados o muy dispersos.

De los resultados obtenidos también destaca la relación que existe entre el propósito comunicativo de cada movida retórica y los rasgos léxicos que se utilizan para cumplir con él. Lo anterior quedó evidenciado al momento de seleccionar los tri-gramas de lemas de cada movida, puesto que, si bien existen varios que se comparten entre ellas, hay otros que son específicos de cada una y reflejan su propósito comunicativo. Debido a lo anterior, sostenemos que los atributos que se entreguen al clasificador deben reflejar el propósito de la movida, dado que de esta forma, se mejorará la representación de ellas, caracterizando mejor las clases y por ende, mejorando la clasificación de los TFG.

En cuanto a la clasificación de todas las movidas retóricas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5) con sus mejores 26 atributos (Ver Anexo 2), los resultados mejoraron bastante en comparación con los logrados tomando todos sus atributos. Sin embargo, los porcentajes de clasificación correcta, el recuperado de textos pertenecientes a cada clase y la precisión de las clasificaciones continuaron siendo bajas e insatisfactorias, por lo que sostenemos que existen otros factores, ligados a la representación de los atributos, que afectan las tareas de clasificación de los dos algoritmos. En este sentido, atribuimos estos resultados a la posible relación que existe entre algunos de los propósitos comunicativos de las movidas de una macromovida y otra.

Si bien, sabemos que los macropropósitos de ambas macromovidas (MM) no son los mismos, ambos tienen cuestiones en común y en algunos aspectos se asemejan. Lo que sucede a nivel de macromovida, se replica en las movidas, por lo que varios de sus rasgos léxicos también serán similares. Las relaciones que existen entre ambas MM refieren a que en MM5 se retoman algunos de los aspectos que se exponen y esbozan en MM1. A modo de ejemplo, en la macromovida Introducir al lector (MM1) se presentan tres movidas en las que se dan a conocer los aspectos generales del tema de investigación, los supuestos conceptuales, el vacío, el objetivo de investigación, las preguntas e hipótesis, el resumen de los métodos, el anuncio de principales productos, entre otros. Y muchos de estos temas se retoman en las movidas de Finalizar discursivamente la investigación (MM5), pero con otro foco (evaluar, explicar, valorar, etc.). Debido a lo anterior, muchos tri-gramas que aparecen en MM1 también aparecerán en MM5, pues al relacionarse sus propósitos, también se relacionan sus rasgos léxicos.

Una posible solución a lo anterior es la clasificación por macromovidas, pues como se observó en los resultados, al realizar las clasificaciones con las movidas de cada macromovida por separado, los logros mejoraban significativamente. La mejora de estos datos se adjudica a la disminución de la dispersión de los atributos y la representación de las características de cada clase, puesto que al disminuir las movidas, se disminuyen las clases y por ende, algunos de los tri-gramas que compartían las clases se vuelven únicos y representativos, ya que solo caracterizan a la categoría que quedó. A partir de lo anterior, el algoritmo tendrá un menor número de clases para clasificar y atributos representativos que mejorarán la tarea de clasificación.

Otro aspecto que resulta interesante de los resultados es que algunas movidas si presentan tri-gramas de lemas o atributos que son representativos de ellas. Este es el caso de MM5M2, donde sus tri-gramas son específicos y sus frecuencias relativas son altas. Estas características y los resultados de su clasificación los adjudicamos al propósito comunicativo de MM5M2 y los rasgos léxicos asociados a él, puesto que, en ninguna otra movida se puede referir a indicar fortalezas o limitaciones del estudio realizado, responder las preguntas de investigación y

confrontar los resultados con otras investigaciones. El hecho que este propósito no se pueda realizar en otra movida y que exista una relación entre los rasgos léxicos y los propósitos comunicativos, hacen que la clasificación de esta movida sea la mejor de todas las realizadas en el presente estudio.

En síntesis, si bien, los tri-gramas de lemas con sus mejores atributos permiten clasificar las movidas retóricas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5), los logros obtenidos no son satisfactorios. Sin embargo, se cumple muy bien la clasificación cuando se realizan las clasificaciones que toman las movidas de solo una macromovida. Tomando en cuenta lo expuesto en todo este apartado, llegamos a la conclusión de que uno de los aspectos más relevantes para lograr una clasificación semiautomática satisfactoria es la representatividad que los atributos hacen de una clase o categoría, en nuestro caso la movida. Como hemos mencionado anteriormente, la naturaleza de los clasificadores no se puede modificar, ni tampoco la cantidad de clases, pues deseamos que la clasificación sea lo más semiautomática posible, pero sí podemos mejorar la calidad y representatividad de los atributos para obtener una mejor clasificación. Con base en lo anterior y con el fin de mejorar la representatividad que hacen los atributos de cada clase, proponemos que la inclusión de la técnica de clasificación por redes de palabras puede contribuir a mejorar la representatividad de los atributos, pues esta técnica considera la dependencia, relación y coocurrencia entre las palabras (Cárdenas, Olivares & Alfaro, 2014), lo que permite mejorar la representatividad de los atributos que utiliza el algoritmo para clasificar. Lo anterior, desde nuestro punto de vista y en la línea de los autores mencionados, ayudaría a disminuir los errores típicos de la clasificación de textos debido al parecido que puede existir entre los atributos de un texto y otro, pues no solo tomaría los lemas o trigramas de lemas, sino que tendría en consideración la relación entre los lemas, su dependencia y aparición en el texto.

# 5. CONCLUSIONES

En el presente apartado, finalizaremos nuestra investigación evaluando la forma en la que el presente trabajo cumple el objetivo de investigación, expondremos los principales hallazgos, resultados, limitaciones y propondremos algunas proyecciones para futuras investigaciones. De esta manera, en primer lugar, se evaluará el cumplimiento del objetivo a partir de los principales resultados obtenidos. Luego, se darán a conocer las principales fortalezas y limitaciones de la investigación. Finalmente, se concluirá con proyecciones que permitan realizar futuras investigaciones.

La importancia que los textos académicos han tenido en los últimos años y, en especial la que ha tenido el género tesis para algunas comunidades discursivas, ha llevado a la necesidad de profundizar las indagaciones en él. La presente investigación cobra relevancia, pues buscó profundizar en estos estudios a través de las clasificaciones semiautomáticas de este género. A partir de lo anterior, creemos que nuestra investigación es una aproximación importante en estos estudios y puede contribuir al desarrollo de futuras clasificaciones que se realicen en torno a los textos académicos.

El objetivo general que guío el presente estudio fue "Clasificar semiautomáticamente las movidas retóricas de las tesis en el ámbito de la Lingüística de la Licenciatura en Lengua y Literatura Hispánica de la Pontificia Universidad Católica de Valparaíso a partir de sus lemas". Mediante el desarrollo de la presente investigación, podemos decir que se cumplió con este objetivo, puesto que se clasificaron las movidas retóricas de las macromovidas Introducir al lector (MM1) y Finalizar discursivamente la investigación (MM5). A su vez, no solo se realizó esta tarea de clasificación, sino que se ejecutaron varias tareas con el fin de probar cuáles eran los mejores atributos que permitían clasificar estas categorías. Para complementar aún más la investigación desarrollada, todas las clasificaciones (con todas las clases, selección de clases, todos los atributos, mejores atributos) se hicieron con dos de los algoritmos más utilizados en clasificación de textos. Lo anterior, se realizó con el fin de determinar cuál era el algoritmo de clasificación que permitía realizar la tarea de forma más satisfactoria.

Los resultados de nuestra investigación indican que los algoritmos Bayesiano ingenuo y Máquina de soporte vectorial no realizan una clasificación satisfactoria cuando tienen como

clases a todas las movidas de las macromovidas MM1 y MM5 y todos sus atributos. También se pudo observar que los logros de los algoritmos mejoran significativamente cuando se realizan las clasificaciones con las mismas clases, pero con una selección de sus mejores atributos. Sin embargo, la mejor clasificación se realiza cuando tienen como clases las movidas de solo una macromovida y sus atributos corresponden a la selección de los mejores tri-gramas. En este último caso, los resultados son satisfactorios, ya que, en forma general, alcanzan porcentajes de clasificación sobre el 90% con Kappas considerables.

Con respecto a la comparación del mejor algoritmo para realizar tareas de clasificación, los resultados indican que, en general, no existe mucha diferencia entre el logro de ambos, sin embargo, BI obtiene logros levemente superiores en la clasificación de la mayoría de las pruebas. A pesar de estos resultados, en algunos casos donde los atributos son representativos y tienen frecuencias relativas altas, las clasificaciones realizadas por MSV son ligeramente superiores a las de BI, teniendo precisiones de entre 0,8 y 1 con una recuperación de textos entre el 90% y el 100%.

A raíz de lo anterior surge nuestro primer hallazgo, pues el logro de las clasificaciones no dependerá totalmente de la naturaleza del algoritmo, sino de la representatividad que los atributos hagan de las clases, en este caso movidas. La calidad de los atributos, los valores de las frecuencias relativas y la no aparición de tri-gramas compartidos determinan el éxito de la clasificación, dado que los algoritmos son entrenados con estos tri-gramas y sus características para posteriormente clasificar nuevos documentos. Otro hallazgo que se liga al anterior es la relación que existe entre los propósitos comunicativos de las movidas y sus rasgos léxicos. De este modo, para asegurar el logro de una clasificación satisfactoria se deben tener como atributos los tri-gramas que reflejen el propósito comunicativo de la movida, pues de esa forma serán representativos de ella y permitirán mejorar la calidad de la clasificación.

En cuanto a las fortalezas o logros descritos anteriormente, estos permiten valorar nuestra investigación como una aproximación a las clasificaciones semiautomáticas de textos académicos a partir de sus rasgos léxicos. A su vez, nuestro estudio permite abrir un campo de investigación que ha sido muy poco estudiado, pues como ya hemos mencionado, las clasificaciones semiautomáticas de textos han solido estar relacionadas a ciertos tipos de textos y a tareas específicas que no incluyen los textos académicos.

La principal limitación de nuestra investigación se relacionó con el número de tesis que se utilizó como corpus para realizar las clasificaciones, pues pensamos que aumentar el corpus permitiría representar mejor las clases y, por ende, enriquecer los resultados obtenidos. Otra dificultad enfrentada fue la poca representatividad que algunos atributos hacían de las clases. Sin embargo, esta se transformó en uno de los aspectos más interesantes de nuestra investigación, pues desde la literatura encontramos que la mayoría de las diferencias entre la clasificación de BI y MSV se relacionarían con la naturaleza de los algoritmos, sin embargo, en la presente investigación, encontramos diferencias significativas entorno a la representatividad de los atributos de las clases en relación a sus propósitos comunicativos. Como hemos mencionado en apartados anteriores, la representatividad que los atributos permitan hacer de los propósitos comunicativos de cada movida es clave para las clasificaciones, puesto que estos rasgos léxicos dan cuenta de los propósitos. Estos rasgos se relacionan a los propósito debido a que ellos instancian los movimientos funcionales, por lo que un estudio que no haya elegido bien los atributos, en nuestro caso llemas, o bien estos no representen significativamente tendrá clasficaciones cuyos resultados serán insatisfactorios.

A partir de lo anterior, en trabajos posteriores, esperamos poder incluir en las clasificaciones de estos dos algoritmos las redes de palabras, puesto que con su uso se podría aumentar la representatividad que los atributos hacen de sus clases debido a la consideración de la dependencia, relación y coocurrencia entre los lemas. Otra proyección refiere a la inclusión de otras macromovidas de los TFG en las clasificaciones, ya que así se podría realizar una clasificación completa de todo este género. Finalmente, creemos que sería interesante clasificar trabajos finales de grado de otras disciplinas, pues de esta manera, se podrían contrastar los resultados obtenidos con los clasificadores en una disciplina y otra y establecer generalidades.

## REFERENCIAS BIBLIOGRÁFICAS


Adam, J. (1991). Les textes: types et protolypes. París: Nathan.

Alcaraz, E. & Martínez, M. (1993). *Diccionario de Lingüística Moderna*. Barcelona: Editorial Ariel S.A.

Aristóteles (1998). *La Retórica.* Madrid: Alianza.

Arnoux, E. (2006). Incidencia de la lectura de pares y expertos en la reescritura de tramos del trabajo de Tesis. *Revista de Lingüística Teórica y Aplicada, 44*(1), 95-118.

Bajtín, M. (1999). *Estética de la creación verbal*. México D.F.: Siglo Veintiuno.

Baeza, R. & Ribeiro, B. (1996). *Modern information retrieval: The concepts and technology behind search.* Reading, M.A.: Addison-Wesley.

Baldi, P., Frasconi, P. & Smyth, P. (2003). *Modeling the Internet and the web*. Chichester: John Wiley.

Bautista, E., Guzmán, E. & Figueroa, J. (2004). Predicción de múltiples puntos de series utilizando support vector machines. *Computación y sistemas*, *7*(3), 148-155.

Bhatia, V. (2002a). Applied genre analysis: A multi-perspective model. *Ibérica*, 4, 3-19.

Bhatia, V. (2002b). A generic view of academic discourse. En J. Flowerdew (Ed.), *Academic*

Bhatia, V. (2004). *World of written discourse. A genre-based view*. Londres: Continuum.

Betancourt, G. (2005). Las máquinas de soporte vectorial. *Scientia e Técnica*, *11*(27), 67-72.

Biber, D. (1985). Investigating Macroscopic Textual Variation Through Multifeature/Multidimentional Analyses. *Linguistic*, 23, 337-360.

Bordignon, F., Peri, J., Tolosa, G., Villa, D. & Paoletti, L. (2004). *Experimentos en clasificación automática de noticias en español utilizando el modelo bayesiano* [en línea]. Disponible en: http://www.unlu.edu.ar/~tyr/TYR-publica/paper-unlu-bayes-2004.doc

Carlino, P. (2003). La experiencia de escribir una tesis: Contextos que la vuelven más difícil. Ponencia presentada en el II Congreso Internacional Cátedra UNESCO Lectura y Escritura. Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile, 5-9 de mayo de 2003.

Cassany, D., Luna, M. & Sanz, G. (2007). *Enseñar Lengua*.Barcelona, España: Grao.

Cassany, D. (2009). Metodología para trabajar con géneros discursivos [En línea]. Disponible en: http://www.upf.edu/pdi/daniel_cassany/_pdf/b08/UPVEuskera08.pdf

Cárdenas, J., Olivares, G., & Alfaro, R. (2014). Clasificación automática de textos usando redes de palabras. *Revista signos*, *47*(86), 346-364.

Caruana, R. & Niculescu-mizil, A. (2006). *An empirical comparison of supervised learning algorithms*. Ponencia presentada en el International Conference on Machine learning, Pittsburgh, Estados Unidos.

Ciapuscio, G. (1994). *Tipos textuales*. Buenos Aires: Universidad de Buenos Aires.

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

García, I. (2009). *Divulgación médica y traducción: El género información para pacientes*. Berna: Peter Lang.

Gómez-Macker, L. & Peronard, M. (2005). *El lenguaje Humano: Léxico fundamental para la iniciación lingüística.* Valparaíso: Euvsa.

Gotti, M. (2008). *Investigating Specialized Discourse*. Bern: Peter Lang.

*Hernández* Sampieri, R., *Fernández*, C. y *Baptista*, P. (2010). *Metodología de la Investigación.* Madrid: Mc Graw Hill.

Ibáñez, R. (2008). El texto disciplinar y el acceso al conocimiento desde el análisis del género: ¿Regulación del conocimiento o persuasión? En G. Parodi (Ed.), *Géneros académicos y géneros profesionales: accesos discursivos para saber y hacer* (pp. 219-246). Valparaíso: Ediciones Universitarias de Valparaíso.

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

Koutsantoni, D. (2006). Rhetorical strategies in engineering research articles and research theses: Advanced academic literacy and relations of power. *Journal of English for Academic Purposes, 5*(1), 19-36.

Longacre, R. & Levinsohn, S. (1978). Currtent Trends in Textlinguistic (W. Dressier, Ed.) Berlin: Gruyter.

Malagón, C. (2003). Clasificadores Bayesianos: El algortimo Naïve Bayes. [En línea]. Disponible en: https://www.nebrija.es/~cmalagon/inco/Apuntes/bayesian_learning.pdf

Manning, C. & Schütze, H. (2003). *Foundations of statistical natural language processing.* Cambridge: MIT Press.

Martínez, J. D. (2012). Descripción y Variación Retórico-Funcional del Género *Tesis*. Doctoral: Un Análisis desde dos Disciplinas y de dos Comunidades Discursivas a partir del corpus Te-DICE 2010. Tesis doctoral, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Meza, P. (2013). La comunicación del conocimiento en las secciones de tesis de lingüística: Determinación de la variación entre grados académicos. Tesis de doctorado, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Molina, J. & García, J. (2004). *Técnicas de análisis de datos en aplicaciones prácticas utilizando Microsoft Excel y Weka* [en línea]. Disponible en: http://galahad.plg.inf.uc3m.es/~docweb/ad/transparencias/apuntesAnalisisDatos.pdf

Moyano, E. (2000). *Comunicar ciencia*. Buenos Aires: Secretaría de Investigaciones, Universidad Nacional de Lomas de Zamora.

Paltridge, B. (2002). Thesis and dissertation writing: an examination of published advice and actual practice. *English for Specific Purposes*. 21 (2), 125-143.

Parodi, G. (2005). *Discurso especializado e instituciones formadoras*. Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. (2008). Géneros académicos y géneros profesionales: acceso discursivo para saber y hacer. Valparaíso, Chile: Ediciones Universitarias

Parodi, G., Venegas, R., Ibáñez, R. & Gutiérrez, R. (2008). Géneros del discurso en el Corpus PUCV-2006: Criterios, definiciones y ejemplos. En G. Parodi (Ed.), *Géneros académicos y géneros profesionales: Accesos discursivos para saber y hacer* (pp. ). Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. (2010). The rhetorical organization of the textbook genre across disciplines: 'a colony in loops'? *Discourse Studies,* 12(2), 195-222.

Peinado, R. (2003). Lematización para palabras médicas complejas: Implementación de un algoritmo en LISP. *Aplicaciones de Procesamiento de Lenguaje Natural*, 87–96.

Perelman, CH & Olbrechts Tyteca, (2009). Tratado de la argumentacion: La nueva retórica. Madrid: Gredos.

Quintiliano, M. (1942). Instituciones oratorias. Madrid: Hernando.

Sabaj, O. (2003). El comportamiento de los verbos abstractos en el corpus pucv-2003. Tesis de doctorado, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Sánchez, D. (2012). La elaboración de la tesis doctoral en las universidades de habla hispana: dificultades y planteamientos de mejora. *Revista Iberoamericana de Educación*, 60(3), 1-12.

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

Silver, M. (2006). *Language across disciplines. Towards a critical reading of contemporary academic discourse*. Florida: Brown Walker Press.

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.

Tamola, D. (2005). La tesina de licenciatura. En L. Cubo de Severino (Coord.), *Los textos de la Ciencia. Principales clases del discurso académico-científico* (pp. 235-265). Córdoba: Comunicarte.

Téllez, A. (2005). *Extracción de información con algoritmos de clasificación* [en línea]. Disponible en: http://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-AlbertoTellez.pdf

Todd, B. & Stamper, R. (1993). The formal design and evaluation of a variety of medical diagnostic programs. Technical Monograph PRG-109. Oxford University Computing Laboratory.

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

Venegas, René. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista signos*, *40*(63), 239-271.

Venegas, R (2010). "Caracterización de géneros evaluativos como Trabajos Finales de Grado en licenciatura y magíster a través de cuatro disciplinas: Desde los patrones léxico-gramaticales y retórico-estructurales al andamiaje de la escritura académico disciplinar".Proyecto Fondecyt 1101039. Valparaíso, Chile.

Venegas, R. (2014). "Hacia un modelo de análisis semiautomatizado de la organización retórico-discursiva de los Trabajos Finales de Grado de licenciatura en ciencias y humanidades". Proyecto Fondecyt 1140967. Valparaíso, Chile.

Venegas, R., Zamora, S. & Galdames, A. (2016). Hacia un modelo retórico-discursivo del macrogénero Trabajo Final de Grado en Licenciatura. *Revista Signos. Estudios de Lingüística*, 49(S1), 247-279.

Warta, V. (1996). *Embedded case reports: A genre-analysis issue in teaching English for medical purposes.* Tesis de Magíster, Aston University, Aston, Estados Unidos.

Weinrich, H. (1975). Estructura y función de los tiempos en el lenguaje. Madrid: Gredos.

Zhang, H. (2004). *The optimality of Naive Bayes* [en línea]. Disponible en: http://www.citeulike.org/user/JoSeK/article/370404

Zamora, S. & Venegas, R. (2013). Estructura y propósitos comunicativos en tesis de licenciatura y magíster. *Revista Literatura y Lingüística*

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, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile.

Zazo, A., Figuerola, C., Alonso, J.L. & Gómez, R. (2002). *Recuperación de información utilizando el modelo vectorial* [en línea]. Disponible en: http://tejo.usal.es/inftec/2002/DP-TOIA-IT-2002-006.pdf

## ANEXOS

### Anexo N°1: Modelo de análisis retórico-discursivo

| **MACROMOVIDA 1: INTRODUCIR AL LECTOR EN LA INVESTIGACIÓN**<br>El macropropósito de esta macromovida es orientar al lector en relación con la temática, presentar los supuestos conceptuales más relevantes que guían la investigación, indicar su propósito y justificar la relevancia de realizar la investigación. | |
|---|---|
| **Movida 1: ESTABLECER EL TERRITORIO** | |
| El propósito de esta movida es destacar la importancia del tema, mostrando la necesidad de su investigación. | |
| **Paso 1:** Generalización del tópico con especificidad creciente | **DESCRIPCIÓN**:<br>Se identifica el tópico sobre el que versará la investigación. Se organiza desde la mención de aspectos generales a otros particulares relativos al tema de investigación. |
| **Paso 2**: Presentación biográfica del autor en estudio | **DESCRIPCIÓN:**<br>Se presenta brevemente a el (los) autores de la(s) obra(s) a investigar, incluyendo hitos relevantes de su historia personal, profesional o artística. |
| **Movida 2: ESTABLECER EL NICHO** | |
| Esta movida tiene el propósito de establecer el área específica de trabajo o vacío, indicando las limitaciones de las investigaciones previas o un área de interés novedosa poco abordada. Suele aparecer de manera relativamente explícita. | |
| **Paso 1a:** Indicación del vacío | **DESCRIPCIÓN:**<br>Se explicita un nicho dentro de un campo de estudio, que puede ser justificado a través de dos estrategias: a) limitaciones de trabajos anteriores; b) temática poco estudiada en un campo |
| **Paso 1b:** Presentación de información conocida | **DESCRIPCIÓN:**<br>Se evidencia la relevancia del nicho para el área disciplinaria a través de la mención de hallazgos relevantes en investigaciones previas relacionadas con el nicho. |
| **Paso 2:** Presentación de justificaciones positivas | **DESCRIPCIÓN:**<br>Se establece la importancia de investigar el nicho de la manera que se propone en la investigación. |
| **Movida 3: OCUPACIÓN DEL NICHO** | |
| El propósito de esta movida es explicitar los objetivos de la investigación, las preguntas o hipótesis, asociados al vacío identificado. Además se delinea la estructura organizativa de la tesis. | |
| **Paso 1:** Anuncio del objetivo de la investigación | **DESCRIPCIÓN:** |

| | |
|---|---|
| | Se presenta y contextualiza el objetivo de la investigación a desarrollar. |
| **Paso 2:** Presentación de algunas preguntas de investigación y/o hipótesis | **DESCRIPCIÓN:**<br>Se plantean las interrogantes o hipótesis que guían la investigación. Se encuentran elementos como preguntas directas o indirectas o indicación léxica de un problema. |
| **Paso 3:** Clarificación de algunas definiciones | **DESCRIPCIÓN:**<br>Se definen conceptos relevantes para el desarrollo de la investigación. |
| **Paso 4:** Resumen de los métodos | **DESCRIPCIÓN:**<br>Se exponen los aspectos generales de la metodología, utilizados para la realización de la investigación. |
| **Paso 5:** Anuncio de los principales productos | **DESCRIPCIÓN:**<br>Se adelantan los principales hallazgos obtenidos en la investigación. |
| **Paso 6:** Establecimiento del valor de la investigación | **DESCRIPCIÓN:**<br>Se explicita la importancia o aportes significativos de la investigación al campo disciplinar al que pertenece o sus proyecciones para temas relacionados. |
| **Paso 7:** Delineado de la estructura | **DESCRIPCIÓN:**<br>Se explicita la organización de la tesis enunciando su estructura y declarando, en algunas ocasiones, brevemente el propósito de cada uno de sus apartados.<br>Se utilizan expresiones como “primero”, “luego”, “a continuación”, “finalmente”, entre otros; además se utilizan tiempos en presente o futuro. |
| **MACROMOVIDA 2: PRESENTAR INVESTIGACIONES PREVIAS Y ANTECEDENTES CONCEPTUALES RELEVANTES**<br><br>El macropropósito de esta macromovida es andamiar conceptualmente la investigación a desarrollar, presentando investigaciones relacionadas o desarrollando un marco teórico o conceptual, que le permita interpretar los alcances de su investigación. | |
| **Movida 1: SÍNTETIZAR DE LA INFORMACIÓN A PRESENTAR** | |
| El propósito de esta movida es adelantar información, de modo sintético, que será desarrollada posteriormente en la presentación de investigaciones similares y antecedentes conceptuales relevantes. | |
| **Paso 1:** Presentación de la estructura y/o contenidos<br>a tratar | **DESCRIPCIÓN:**<br>Se explicita la estructura del apartado o contenidos que conforman la presentación |

| | |
|---|---|
| | de investigaciones previas y antecedentes conceptuales relevantes. |
| **Paso 2 :** Alusión a acciones o contenidos desarrollados en otro apartado | **DESCRIPCIÓN:**<br>Se alude a la relación existente entre los conceptos identificados de las investigaciones relevantes con los procedimientos de análisis que se llevarán a cabo. |
| **Movida 2: ESTABLECER EL TERRITORIO TEMÁTICO** | |
| Esta movida tiene como propósito situar temáticamente la investigación, justificando su relevancia y dando cuenta de investigaciones en el área. | |
| **Paso 1:** Presentación del estado actual del conocimiento y prácticas no investigadas | **DESCRIPCIÓN:**<br>Se hace revisión, teórica y terminológica, de los conceptos asociados a la investigación. |
| **Paso 2:** Afirmación de la centralidad del tema | **DESCRIPCIÓN:**<br>Se explicita la importancia del tema investigado, a partir de necesidades investigativas emanadas o explícitamente establecidas en la revisión bibliográfica. |
| **Paso 3:** Investigaciones similares | **DESCRIPCIÓN:**<br>Se mencionan investigaciones que, dentro del área temática, han considerado un objeto, objetivos, procedimientos, técnica de recolección de datos y/o hipótesis similares a las propuestas en la propia investigación. |
| **Movida 3: CREAR UN NICHO DE INVESTIGACIÓN** | |
| El propósito de esta movida es indicar un vacío en el área temática en que se desarrolla la investigación, a través de la revisión de la literatura existente. | |
| **Paso 1:** Contraargumentación | **DESCRIPCIÓN:**<br>Se revisan críticamente investigaciones similares a partir de argumentos propios o de otros autores. |
| **Paso 2:** Indicación del vacío | **DESCRIPCIÓN:**<br>Se delimita el área a investigar, a partir de las carencias o debilidades argumentadas anteriormente a nivel teórico y/o metodológico en el área de estudio. |
| **Paso 3:** Presentación de argumentos confirmatorios | **DESCRIPCIÓN:**<br>Se destaca el aporte de una investigación a través del comentario sobre el valor o fuerza o contribución de una cita. |
| **Paso 4:** Relevancia de los argumentos indagados para su propia investigación | **DESCRIPCIÓN:**<br>Se explicita la aplicabilidad o relevancia de otras investigaciones en favor de la propia. |
| **Paso 5:** Resumen de conocimientos para establecer una posición | **DESCRIPCIÓN:** |

| | |
|---|---|
| | Se resumen las posturas teórico-metodológicas revisadas, así como los argumentos relevados, para plantear la postura o perspectiva que guiará la investigación. |
| **Movida 4: OCUPAR EL NICHO A INVESTIGAR** | |
| El propósito de esta movida es introducir los aspectos teórico-metodológicos adoptados en la investigación. | |
| **Paso 1:** Objetivo de la investigación, pregunta de investigación, hipótesis | **DESCRIPCIÓN:** Se anuncia el objetivo, pregunta y/o hipótesis a raíz del nicho de investigación identificado en las movidas previas. |
| **Paso 2:** Postura crítica del Marco Teórico | **DESCRIPCIÓN:** Se indica una posición con respecto a otros autores. No se limita a la importancia o relevancia de estos a la investigación, sino que puede ser una posición contraria o incluso opinar sobre un tema no relacionado directamente con la investigación de la tesis. |
| **Paso 3:** Presentación del diseño de la investigación o procesos metodológicos relevantes | **DESCRIPCIÓN:** Se presentan aspectos y procedimientos metodológicos utilizados en la investigación. |
| **Paso 4:** Interpretación de la terminología utilizada en la tesis | **DESCRIPCIÓN:** Se explica un concepto o se lo reinterpreta, a través del parafraseo. |
| **Paso 5:** Justificación del tema escogido | **DESCRIPCIÓN:** Se entregan las razones que llevaron al autor a escoger el tema desarrollado en su tesis. |
| **MACROMOVIDA 3: EXPONER LOS PROCEDIMIENTOS METODOLÓGICOS**<br><br>El macropropósito de esta macromovida es sustentar metodológicamente la investigación, seleccionando un paradigma de investigación, de acuerdo con la disciplina en la que se desarrolla. | |
| **Movida 1: PRESENTAR INFORMACIÓN PREPARATORIA** | |
| El propósito de esta movida es dar a conocer todas aquellas informaciones relevantes (tipo de investigación, diseño, sujetos, corpus, procedimientos e instrumentos) relativa a los procesos metodológicos que se llevaron a cabo en la realización de la investigación. | |
| **Paso 1:** Presentación de la estructura o contenidos a tratar | **DESCRIPCIÓN:** Se presenta la estructura del apartado o contenidos que estarán presentes en esta sección de la tesis. |
| **Paso 2:** Articulación entre marco teórico, el problema de investigación y los aspectos metodológicos | **DESCRIPCIÓN:** Se plantean conexiones entre el objetivo de investigación, antecedentes conceptuales y |

| | |
|---|---|
| | los procedimientos metodológicos escogidos para abordar el problema de investigación. |
| **Movida 2: PRESENTAR LOS ASPECTOS METODOLÓGICOS DE LA INVESTIGACIÓN** | |
| El propósito de esta movida es dar a conocer todos aquellos aspectos metodológicos y procedimentales que sustentan la investigación. | |
| **Paso 1**: Presentación del enfoque, alcance y diseño de la investigación | **DESCRIPCIÓN:** Se describe y justifica el enfoque, alcance y diseño de la investigación. |
| **Paso 2:** Presentación de objetivos | **DESCRIPCIÓN:** Se realiza una declaración explícita de los objetivos generales y específicos de la investigación. |
| **Paso 3:** Presentación de las preguntas de investigación | **DESCRIPCIÓN:** Se realiza una declaración explícita de las preguntas que guían la investigación. |
| **Paso 4:** Presentación de las hipótesis | **DESCRIPCIÓN:** Se declara explícitamente la hipótesis del estudio. |
| **Paso 5:** Explicación de aspectos metodológicos | **DESCRIPCIÓN:** Se explica con mayor detalle algún aspecto de la metodología o procedimientos utilizados. |
| **Movida 3: DELIMITAR EL OBJETO DE ESTUDIO** | |
| El propósito de esta movida es caracterizar el objeto estudiado en la investigación. | |
| **Paso 1:** Presentación de los sujetos o materiales involucrados en el estudio | **DESCRIPCIÓN:** Se realiza una declaración de los sujetos o materiales seleccionados para ser estudiados durante la investigación. |
| **Paso 2:** Ubicación y características contextuales de la muestra | **DESCRIPCIÓN:** Se realiza una especificación y caracterización del contexto desde el cual se seleccionaron los sujetos o materiales para la obtención de los datos de investigación. |
| **Paso 3:** Presentación del instrumento de estudio | **DESCRIPCIÓN:** Se presenta el instrumento que posibilitará la recolección de datos para ser analizados en la investigación. |
| **Movida 4: ESPECIFICAR PROCEDIMIENTOS** | |
| El propósito de esta movida es mencionar específicamente todos los pasos procedimentales que se han realizado en el proceso de investigación. | |
| **Paso 1:** Detalle de los procedimientos de recolección de datos | **DESCRIPCIÓN:** Se explican los pasos procedimentales realizados para la recolección de los datos necesarios para la investigación. |

| | |
|---|---|
| **Paso 2:** Sustento conceptual de aspectos metodológicos | **DESCRIPCIÓN:**<br>Se define conceptualmente los procedimientos metodológicos que posibilitan el desarrollo de la investigación. |
| **Paso 3 :** Definición de variables | **DESCRIPCIÓN:**<br>Se definen conceptual y operacionalmente las variables estudiadas en la investigación. |
| **Paso 4 :** Evaluación de procedimientos | **DESCRIPCIÓN:**<br>Se evalúan los procedimientos metodológicos que se han llevado a cabo en la investigación. |
| **Paso 5 :** Justificación de metodología | **DESCRIPCIÓN:**<br>Se justifica la metodología utilizada en la tesis. |
| **MACROMOVIDA 4: DAR CUENTA DE LOS RESULTADOS Y SU INTERPRETACIÓN EN EL CONTEXTO DE LA INVESTIGACIÓN**<br><br>El macropropósito de esta macromovida es presentar los resultados y proponer una interpretación de ellos de acuerdo con el marco teórico o conceptual y las investigaciones previas relacionadas con la investigación. | |
| **Movida 1: PRESENTAR INFORMACIÓN PREPARATORIA** | |
| Esta movida tiene como propósito introducir al lector en la presentación de los resultados. | |
| **Paso 1:** Presentación del apartado a través de un epígrafe | **DESCRIPCIÓN:**<br>Se incluye la cita de un autor, relacionada con la temática de la investigación. Esta se integra antes de comenzar la exposición de los resultados. |
| **Paso 2:** Presentación de la estructura y/o contenidos a tratar | **DESCRIPCIÓN:**<br>Se explicita la estructura o contenidos que estarán presentes en la sección. |
| **Movida 2: REPORTAR RESULTADOS** | |
| Esta movida tiene como propósito presentar los resultados de una manera organizada, a través de recursos verbales y no verbales. Se pueden incluir datos estadísticos o ejemplos relevantes que ilustren los resultados. | |
| **Paso 1:** Presentación de los resultados por medio de recursos verbales y no verbales (tablas, gráficos, esquemas, etc.) | **DESCRIPCIÓN:**<br>Se presenta cada uno de los resultados, a través de comentarios sintéticos, o bien, tablas, cuadros, gráficos, esquemas, entre otros. |
| **Paso 2:** Presentación de ejemplos | **DESCRIPCIÓN:**<br>Se presentan ejemplos que respaldan los resultados. |
| **Paso 3 :** Referencia a anexos o descripción del contenido de los mismos | **DESCRIPCIÓN:**<br>Se presentan uno o varios anexos relacionados con el o los resultados, |

| | |
|---|---|
| | indicando su ubicación o describiendo su contenido. |
| **Movida 3: INTERPRETAR RESULTADOS** | |
| Esta movida tiene como propósito interpretar los datos obtenidos en relación con los objetivos e hipótesis de la investigación y con la literatura relevante. | |
| **Paso 1:** Interpretación de datos obtenidos | **DESCRIPCIÓN:** Se interpretan los datos obtenidos en función de su significatividad o pertinencia con respecto de las técnicas de análisis de datos utilizadas. |
| **Paso 2:** Comparación de resultados con la Literatura | **DESCRIPCIÓN:** Se contrastan los resultados obtenidos en la investigación con los obtenidos en otras investigaciones relacionadas. |
| **Paso 3:** Evaluación de resultados y hallazgos | **DESCRIPCIÓN:** Se comentan las limitaciones, confirmaciones o generalizaciones posibles de los resultados. Además se destacan hallazgos relevantes. |
| **Paso 4:** Justificación de resultados y hallazgos | **DESCRIPCIÓN:** Se entregan razones teóricas o metodológicas que permiten justificar los hallazgos desprendidos del análisis. |
| **Movida 4: EVALUAR EL ESTUDIO** | |
| Esta movida tiene como propósito evaluar los hallazgos de la investigación, dando cuenta de su aporte al campo de investigación. | |
| **Paso 1:** Presentación de información preparatoria a la evaluación | **DESCRIPCIÓN:** Se introducen los principales aspectos a considerar para la evaluación de los resultados. |
| **Paso 2:** Valoración de la investigación | **DESCRIPCIÓN:** Se explicita la importancia o relevancia de su propia investigación. |
| **MACROMOVIDA 5: FINALIZAR DISCURSIVAMENTE LA INVESTIGACIÓN**<br><br>El macropropósto de esta macromovida es concluir la investigación, recordando al lector los aspectos más relevantes de la investigación, proponiendo una interpretación lo más sistematizada posible de los hallazgos obtenidos y presentando proyecciones de la investigación realizada. | |
| **Movida 1: CONSOLIDAR LA OCUPACIÓN DEL NICHO** | |
| Esta movida tiene por propósito reconfirmar la novedad del tema o el tratamiento del problema investigado. Además, se recuerdan las razones que llevaron a ocupar el nicho de investigación. | |
| **Paso 1:** Presentación de la estructura o contenidos a tratar | **DESCRIPCIÓN:** Se mencionan y describen muy brevemente los principales subtemas que se abordarán |

| | |
|---|---|
| | en la conclusión, indicando su orden de aparición en el texto. |
| **Paso 2:** Establecimiento de la relevancia del problema o tema investigado | **DESCRIPCIÓN:**<br>Se destaca la importancia del tema investigado y se señalan los aportes de la investigación. |
| **Paso 3:** Resignificación hipótesis u objetivos en función de la investigación realizada | **DESCRIPCIÓN:**<br>Se reinsertan los objetivos y las hipótesis que guiaron la investigación. El tesista realiza una evaluación respecto del cumplimiento de los objetivos propuestos y del grado de confirmación de la hipótesis presentada. |
| **Paso 4:** Resumen general de la investigación | **DESCRIPCIÓN:**<br>Se sintetiza la temática y se menciona el objetivo general de la investigación. |
| **Movida 2: EVALUAR LA INVESTIGACIÓN** | |
| Esta movida tiene como propósito evaluar los hallazgos de la investigación, dando cuenta del aporte del estudio al campo de investigación en el que se inserta. | |
| **Paso 1:** Indicación de fortalezas y limitaciones del estudio | **DESCRIPCIÓN:**<br>Se indican los aportes teóricos o metodológicos de los hallazgos obtenidos en la investigación. También se mencionan las debilidades del proceso investigativo presentado en la tesis. |
| **Paso 2:** Evaluación de la metodología | **DESCRIPCIÓN:**<br>Especificar si la metodología fue adecuada o tuvo falencias en su aplicación. Si fue capaz de llevar a cabo los objetivos propuestos y qué cambios o mejoras le aplicarían. |
| **Paso 3:** Resumen de los resultados propios | **DESCRIPCIÓN:**<br>Se presentan de manera sintetizada los principales hallazgos de la investigación, explicándolos brevemente. |
| **Paso 4:** Confrontación de los resultados obtenidos con otras investigaciones | **DESCRIPCIÓN:**<br>Se valoran o justifican los resultados obtenidos en la investigación a partir de la confrontación con los resultados de otras investigaciones realizadas en la misma área. |
| **Paso 5:** Resolución de las preguntas de investigación y/o contrastación de la hipótesis | **DESCRIPCIÓN:**<br>Se relacionan explícitamente los resultados obtenidos con la hipótesis de la investigación o, en su defecto, da respuesta a las preguntas de investigación que guiaron la misma. |

| Movida 3: EXPONER LAS IMPLICACIONES DE LOS RESULTADOS Y HALLAZGOS | |
|---|---|
| Esta movida tiene como propósito proveer una explicación crítica de los resultados y los hallazgos, así como también proyectar futuras aplicaciones e investigaciones relacionadas con la investigación presentada. | |
| **Paso 1:** Hipótesis de una explicación de los resultados | **DESCRIPCIÓN:** Se postulan explicaciones para los resultados y hallazgos, en función de una perspectiva global del área investigada. |
| **Paso 2:** Presentación de una visión crítica de la investigación | **DESCRIPCIÓN:** Se delimita el alcance de los hallazgos de la investigación al proceso de investigación. Además se evalúa la metodología y procedimientos que llevaron a estos hallazgos. |
| **Paso 3:** Identificación de resultados o hallazgos clave para una futura investigación o aplicación | **DESCRIPCIÓN:** Se identifican aspectos pertinentes derivados de la tesis para potenciales aplicaciones o futuras investigaciones. |

**Anexo N°2: Mejores tri-gramas de lemas de todas las movidas de las macromovidas MM1 Y MM5**

```
acceso_y_la
anÃ¡lisis_de_los
conocimiento_ser_suficiente
de_cierre_en
de_educaciÃ³n_superior
de_este_investigaciÃ³n
de_imposiciÃ³n_de
determinado_tipo_de
el_anÃ¡lisis_de
el_concepto_de
el_desarrollo_de
el_marco_metodolÃ³gico
el_marco_teÃ³rico
en_el_marco
estar_mucho_orientado
este_investigaciÃ³n_ser
haber_presentar_este
principalmente_en_el
pues_se_deber
que_si_haber
se_presentar_el
se_presentar_los
```

**Anexo N°3: Mejores tri-gramas de lemas de las movidas de MM1**

2005_orientar_la
anÃ¡lisis_de_los
conocimiento_ser_suficiente
corpus_y_el
de_cierre_en
de_educaciÃ³n_superior
de_el_lenguaje
de_este_investigaciÃ³n
de_imposiciÃ³n_de
de_los_dato
de_los_resultado
dentro_de_este
el_anÃ¡lisis_de
el_desarrollo_de
el_marco_metodolÃ³gico
el_marco_teÃ³rico
el_presente_trabajo
en_primero_lugar
este_investigaciÃ³n_ser
estudiante_a_insertar
marco_teÃ³rico_que
se_presentar_el
se_presentar_los
travÃ©s_de_la
uno_de_los

**Anexo N°4: Mejores tri-gramas de lemas de las movidas de MM5**

acceso_y_la
complejo_que_se
con_el_uso
de_la_teorÃ-a
determinado_tipo_de
dicha_influenciar_los
el_trabajo_de
en_el_marco
en_espaÃ±a_de
estudiado_para_en
evaluativo_dentro_de
exigente_ser_el
gramÃ¡tica_asÃ-_como
haber_presentar_este
haber_trabajar_en
la_etapa_final
lo_que_se
por_lo_tanto
principalmente_en_el
pues_se_deber
someter_a_las
superior_con_este
texto_acadÃ©mico_y

**Anexo N°5: Resultados de clasificaciones: Bayesiano ingenuo con todas las movidas y atributos**

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     TRI_LEMAS
Instances:    120
Attributes:   14739
[list of attributes omitted]
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                                                  Class
Attribute                                         MM1M1   MM1M2   MM1M3
MM5M1   MM5M2   MM5M3
                                                (0.17)  (0.17)  (0.17)


Correctly Classified Instances          41               34.1667 %
Incorrectly Classified Instances        79               65.8333 %
Kappa statistic                          0.21
Mean absolute error                      0.2195
Root mean squared error                  0.4682
Relative absolute error                 79.0172 %
Root relative squared error            125.6292 %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.1       0.09       0.182     0.1       0.129
0.511    MM1M1
                 0.25      0.09       0.357     0.25      0.294
0.638    MM1M2
                 0.55      0.21       0.344     0.55      0.423
0.652    MM1M3
                 0.2       0.1        0.286     0.2       0.235
0.591    MM5M1
                 0.75      0.21       0.417     0.75      0.536
0.769    MM5M2
                 0.2       0.09       0.308     0.2       0.242
0.66     MM5M3
Weighted Avg.    0.342     0.132      0.315     0.342     0.31
0.637

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
  2  3  8  3  3  1 |  a = MM1M1
  2  5  2  3  3  5 |  b = MM1M2
```

```
  2  0 11  0  7  0 |  c = MM1M3
  2  2  7  4  3  2 |  d = MM5M1
  0  1  3  0 15  1 |  e = MM5M2
  3  3  1  4  5  4 |  f = MM5M3
```

- **Anexo N°6: Resultados de clasificaciones: Máquina de soporte vectorial con todas las movidas y atributos**

```
=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -
V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     TRI_LEMAS
Instances:    120
Attributes:   14739
[list of attributes omitted]
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

SMO

Correctly Classified Instances          15               12.5    %
Incorrectly Classified Instances       105               87.5    %
Kappa statistic                         -0.05
Mean absolute error                      0.2881
Root mean squared error                  0.4027
Relative absolute error                103.7333 %
Root relative squared error            108.0535 %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.05      0          1         0.05      0.095
0.139    MM1M1
                 0.1       0          1         0.1       0.182
0.552    MM1M2
                 0.1       0.07       0.222     0.1       0.138
0.39     MM1M3
                 0         0.01       0         0         0
0.515    MM5M1
                 0.25      0.64       0.072     0.25      0.112
0.194    MM5M2
                 0.25      0.33       0.132     0.25      0.172
0.462    MM5M3
Weighted Avg.    0.125     0.175      0.404     0.125     0.117
0.375

=== Confusion Matrix ===
```

```
  a  b  c  d  e  f   <-- classified as
  1  0  5  0  3 11 |  a = MM1M1
  0  2  0  0 15  3 |  b = MM1M2
  0  0  2  1 15  2 |  c = MM1M3
  0  0  0  0 16  4 |  d = MM5M1
  0  0  2  0  5 13 |  e = MM5M2
  0  0  0  0 15  5 |  f = MM5M3
```

- **Anexo N°7: Resultados de clasificaciones: Bayesiano ingenuo con todas las movidas y mejores atributos**

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes

Naive Bayes Classifier

                                 Class
Attribute                        MM1M1   MM1M2   MM1M3   MM5M1   MM5M2
MM5M3
                                (0.17)  (0.17)  (0.17)  (0.17)  (0.17)

Correctly Classified Instances          67               55.8333 %
Incorrectly Classified Instances        53               44.1667 %
Kappa statistic                          0.47
Mean absolute error                      0.1557
Root mean squared error                  0.3438
Relative absolute error                 56.0367 %
Root relative squared error             92.2438 %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.7       0.01       0.933     0.7       0.8
0.905    MM1M1
                 0.8       0.29       0.356     0.8       0.492
0.818    MM1M2
                 0.7       0.05       0.737     0.7       0.718
0.89     MM1M3
                 0.2       0.12       0.25      0.2       0.222
0.704    MM5M1
                 0.85      0.02       0.895     0.85      0.872
0.938    MM5M2
                 0.1       0.04       0.333     0.1       0.154
0.749    MM5M3
Weighted Avg.    0.558     0.088      0.584     0.558     0.543
0.834

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
```

```
 14  4  1  1  0  0 |  a = MM1M1
  0 16  1  2  0  1 |  b = MM1M2
  0  1 14  3  0  2 |  c = MM1M3
  0 13  2  4  0  1 |  d = MM5M1
  1  2  0  0 17  0 |  e = MM5M2
  0  9  1  6  2  2 |  f = MM5M3
```

- **Anexo N°8: Resultados de clasificaciones: Máquina de soporte vectorial con todas las movidas y mejores atributos**

```
Test mode:10-fold cross-validation

Correctly Classified Instances          62               51.6667 %
Incorrectly Classified Instances        58               48.3333 %
Kappa statistic                          0.42
Mean absolute error                      0.2454
Root mean squared error                  0.3454
Relative absolute error                 88.3333 %
Root relative squared error             92.693  %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.65      0          1         0.65      0.788
0.9      MM1M1
                 0.25      0.09       0.357     0.25      0.294
0.674    MM1M2
                 0.55      0.03       0.786     0.55      0.647
0.935    MM1M3
                 0.2       0.05       0.444     0.2       0.276
0.621    MM5M1
                 0.75      0.01       0.938     0.75      0.833
0.935    MM5M2
                 0.7       0.4        0.259     0.7       0.378
0.664    MM5M3
Weighted Avg.    0.517     0.097      0.631     0.517     0.536
0.788

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 13  0  0  1  0  6 |  a = MM1M1
  0  5  0  3  0 12 |  b = MM1M2
  0  2 11  0  0  7 |  c = MM1M3
  0  4  1  4  0 11 |  d = MM5M1
  0  1  0  0 15  4 |  e = MM5M2
  0  2  2  1  1 14 |  f = MM5M3
```

- **Anexo N°9: Resultados de clasificaciones: Bayesiano ingenuo con MM1 y todos los atributos**

```
Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          20               33.3333 %
Incorrectly Classified Instances        40               66.6667 %
Kappa statistic                          0
Mean absolute error                      0.4444
Root mean squared error                  0.6667
Relative absolute error                100      %
Root relative squared error            141.4214 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.1       0.425      0.105     0.1       0.103
0.281    MM1M1
                 0.4       0.125      0.615     0.4       0.485
0.669    MM1M2
                 0.5       0.45       0.357     0.5       0.417
0.5      MM1M3

Weighted Avg.    0.333     0.333      0.359     0.333     0.335
0.483

=== Confusion Matrix ===

  a  b  c   <-- classified as
  2  5 13 |  a = MM1M1
  7  8  5 |  b = MM1M2
 10  0 10 |  c = MM1M3
```

- **Anexo N°10: Resultados de clasificaciones: Máquina de soporte vectorial con MM1 y todos los atributos**

```
=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -
V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     TRI_LEMAS
Instances:    60
Attributes:   14739

Test mode:10-fold cross-validation

Time taken to build model: 0.17 seconds
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          24               40      %
Incorrectly Classified Instances        36               60      %
Kappa statistic                          0.1
Mean absolute error                      0.4704
Root mean squared error                  0.5676
Relative absolute error                105.8333 %
Root relative squared error            120.4159 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.1       0.025      0.667     0.1       0.174
0.217    MM1M1
                 1         0.85       0.37      1         0.541
0.575    MM1M2
                 0.1       0.025      0.667     0.1       0.174
0.429    MM1M3
Weighted Avg.    0.4       0.3        0.568     0.4       0.296
0.407

=== Confusion Matrix ===

  a  b  c   <-- classified as
  2 17  1 |  a = MM1M1
  0 20  0 |  b = MM1M2
  1 17  2 |  c = MM1M3
```

- **Anexo N°11: Resultados de clasificaciones: Bayesiano ingenuo con MM1 y los mejores atributos**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               90      %
Incorrectly Classified Instances         6               10      %
Kappa statistic                          0.85
Mean absolute error                      0.0713
Root mean squared error                  0.2449
Relative absolute error                 16.0411 %
Root relative squared error             51.9476 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
```

```
                 0.75      0          1         0.75      0.857
0.969    MM1M1
                 0.95      0.1        0.826     0.95      0.884
0.973    MM1M2
                 1         0.05       0.909     1         0.952
0.975    MM1M3
Weighted Avg.    0.9       0.05       0.912     0.9       0.898
0.972

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  4  1 |  a = MM1M1
  0 19  1 |  b = MM1M2
  0  0 20 |  c = MM1M3
```

- **Anexo N°12: Resultados de clasificaciones: Máquina de soporte vectorial con MM1 y los mejores atributos**

```
Test mode:10-fold cross-validation

Correctly Classified Instances          52               86.6667 %
Incorrectly Classified Instances         8               13.3333 %
Kappa statistic                          0.8
Mean absolute error                      0.2556
Root mean squared error                  0.3277
Relative absolute error                 57.5    %
Root relative squared error             69.5222 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.75      0          1         0.75      0.857
0.934    MM1M1
                 1         0.2        0.714     1         0.833
0.9      MM1M2
                 0.85      0          1         0.85      0.919
0.973    MM1M3
Weighted Avg.    0.867     0.067      0.905     0.867     0.87
0.936

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  5  0 |  a = MM1M1
  0 20  0 |  b = MM1M2
  0  3 17 |  c = MM1M3
```

- **Anexo N°13: Resultados de clasificaciones: Bayesiano ingenuo con MM5 y todos los atributos**

```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          29               48.3333 %
Incorrectly Classified Instances        31               51.6667 %
Kappa statistic                          0.225
Mean absolute error                      0.3445
Root mean squared error                  0.5869
Relative absolute error                 77.5043 %
Root relative squared error            124.499  %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.35      0.225      0.438     0.35      0.389
0.582    MM5M1
                 0.85      0.4        0.515     0.85      0.642
0.719    MM5M2
                 0.25      0.15       0.455     0.25      0.323
0.634    MM5M3
Weighted Avg.    0.483     0.258      0.469     0.483     0.451
0.645

=== Confusion Matrix ===

  a  b  c   <-- classified as
  7  9  4 |  a = MM5M1
  1 17  2 |  b = MM5M2
  8  7  5 |  c = MM5M3
```

- **Anexo N°14: Resultados de clasificaciones: Máquina de soporte vectorial con MM5 y todos los atributos**

```
=== Run information ===

Number of kernel evaluations: 820 (97.098% cached)


Time taken to build model: 0.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          24               40      %
Incorrectly Classified Instances        36               60      %
Kappa statistic                          0.1
```

```
Mean absolute error                      0.4259
Root mean squared error                  0.527
Relative absolute error                 95.8333 %
Root relative squared error            111.8034 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.1       0          1         0.1       0.182
0.551    MM5M1
                 0.25      0.15       0.455     0.25      0.323
0.531    MM5M2
                 0.85      0.75       0.362     0.85      0.507
0.55     MM5M3
Weighted Avg.    0.4       0.3        0.605     0.4       0.337
0.544

=== Confusion Matrix ===

  a  b  c   <-- classified as
  2  3 15 |  a = MM5M1
  0  5 15 |  b = MM5M2
  0  3 17 |  c = MM5M3
```

- **Anexo N°15: Resultados de clasificaciones: Bayesiano ingenuo con MM5 y los mejores atributos**

```
   === Stratified cross-validation ===
=== Summary ===
Test mode:10-fold cross-validation


Correctly Classified Instances          46               76.6667 %
Incorrectly Classified Instances        14               23.3333 %
Kappa statistic                          0.65
Mean absolute error                      0.1624
Root mean squared error                  0.3523
Relative absolute error                 36.5398 %
Root relative squared error             74.7434 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.35      0          1         0.35      0.519
0.849    MM5M1
                 1         0.025      0.952     1         0.976
0.987    MM5M2
                 0.95      0.325      0.594     0.95      0.731
0.825    MM5M3
```

```
Weighted Avg.    0.767     0.117      0.849     0.767     0.742
0.887

=== Confusion Matrix ===

  a  b  c   <-- classified as
  7  0 13 |  a = MM5M1
  0 20  0 |  b = MM5M2
  0  1 19 |  c = MM5M3
```

- **Anexo N°16: Resultados de clasificaciones: Máquina de soporte vectorial con MM5 y los mejores atributos**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          47               78.3333 %
Incorrectly Classified Instances        13               21.6667 %
Kappa statistic                          0.675
Mean absolute error                      0.2778
Root mean squared error                  0.36
Relative absolute error                 62.5    %
Root relative squared error             76.3763 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC
Area  Class
                 0.45      0          1         0.45      0.621
0.849    MM5M1
                 0.9       0          1         0.9       0.947
0.95     MM5M2
                 1         0.325      0.606     1         0.755
0.838    MM5M3
Weighted Avg.    0.783     0.108      0.869     0.783     0.774
0.879

=== Confusion Matrix ===

  a  b  c   <-- classified as
  9  0 11 |  a = MM5M1
  0 18  2 |  b = MM5M2
  0  0 20 |  c = MM5M3
```