UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

ESCUELA DE POST GRADO - UNIDAD DE POSTGRADO DE LA FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y SERVICIOS

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

Construcción automática y análisis de Modelos de Espacios de Palabras de n-gramas y su aplicación a tareas de procesamiento de lenguaje natural

Presentado por:
M.Sc. Víctor Manuel Cornejo Aparicio
Para optar el Grado Académico de
Doctor en Ciencias de la Computación

Docente Asesor:
Dr. Javier Tejada Cárcamo

Arequipa – Perú
2013

Dedicatoria

Quiero dedicarle esta tesis a mi familia, a mi esposa Sidanelia por su comprensión y apoyo, así como también a mis hijas Cosette y Alexa, mi hijo Nicolás, quienes con sus alegrías alimentan al motor de mis mayores esfuerzos, y de una forma muy particular a mi padre que con su ejemplo y dedicación supo formarme como una persona de bien, también quiero dedicarle este trabajo a mi madre cuya memoria siempre me acompaña

Agradecimientos

Quiero expresarle mi más sincero agradecimiento al Dr. Javier Tejada Cárcamo, quien con sus conocimientos supo guiar el desarrollo de mis acciones en pro de la conclusión de la presente tesis, quien en lo apretado de su agenda laboral supo hacerme un espacio para darme la pauta y así no perder el norte de lo que se debe hacer y cómo se debe hacer las tareas encomendadas.

RESUMEN

La presente tesis tiene por objetivo mejorar la calidad de vocablos relacionados semánticamente mediante la construcción automática y análisis de Modelos de Espacios de Palabras basados en n-gramas. Este método debe incluir vocablos que a su vez deben mejorar la precisión de tareas de procesamiento de lenguaje natural, específicamente la clasificación de textos, para ello se emplearon modelos ya existentes como base de conceptualización y se implementaron mejoras en el pre-procesamiento de los textos, tales como la extracción de verbos y sustantivos, posteriormente se trabajó la clasificación a tres niveles de n-gramas (monogramas, digramas y digramas ordenados horizontalmente), luego se efectuaron los experimentos con el corpus estandarizado “corpora Reuters 21578”, del cual se seleccionaron las ocho categorías más relevantes con las que se obtuvo un nivel de precisión del orden del 84.17%, con lo que se superó el 83% de precisión prevalente, lo cual permitió validar la propuesta.

ABSTRACT

This thesis aims to improve the quality of semantically related words using Automatic Construction and Model Analysis of Word Spaces based on n-grams. This method should include words which in turn should improve the accuracy of natural language processing tasks, specifically, text classification. For this purpose existing models were used as basis for conceptualization and implemented improvements in the pre-processing of texts such as verbs and nouns extraction. Later, classification of three levels of n-grams (monograms, digrams and digraphs arranged horizontally) was worked. Then conducted experiments with standardized corpus "Reuters corpora 21578" were developed. From these experiments, the eight most relevant categories were selected. We got 84.17% level of precision which exceeded the 83% current accuracy percentage, which allowed us to validate the proposal.

ÍNDICE DE CONTENIDOS

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La presente tesis comienza con la etapa formal la formulación de la realidad problemática en la clasificación de documentos, hasta la definición del problema central o núcleo del problema, el mismo que asume como objetivo básico el “Mejorar la calidad de vocablos relacionados semánticamente mediante la construcción automática y análisis de Modelos de Espacios de Palabras basados en n-gramas. Estos vocablos deben mejorar la precisión de tareas de procesamiento de lenguaje natural, tales como la clasificación de textos”, fundamentando el mismo de manera posterior.

Posteriormente se indaga respecto al estado del arte en la materia del asunto, lo cual se presenta en forma de conceptos formalizados por los respectivos autores referenciados, para luego presentar el resumen de las investigaciones más relevantes, donde se muestra que el resultado empleando el modelo de espacio de palabras es de 83% de precisión (Romero F. et al. 2008), cabe acotar que en las tareas de clasificación de textos, para que los resultados sean susceptibles de comparación, se debe emplear un corpus de textos estandarizados, los mismos que deben tener cierta representatividad. En esta tesis se empleó el corpus “corpora Reuters 21578”, donde los autores que trabajaron sobre él, deben hacer referencia a las categorías más relevantes, adema que deben presentar la pauta para la réplica de sus experiencias.

De una forma esquemática y procedimental se plantea la propuesta del método de clasificación basado en el modelo de espacio de palabras, para que paso a paso se pueda efectuar acciones consecutivas de pre-procesamiento y cálculo de proximidad de categorías predeterminadas de documentos, dicha mecánica emplea conceptos combinados de propuestas similares y aportes propios que pretenden mejorar el estatus de la precisión en la clasificación de documentos. En lo referente al uso de las propuestas existentes, se empleó el método vectorial de cálculo de proximidad por la ley de cosenos, y su mejora con la técnica de ganancia de información. En lo referente a aportes propios, se precisó el uso de verbos y sustantivos en la constitución de los patrones, se segmentaron los términos que constituyen los n-gramas a aquel grupo constituido por los términos del patrón y su comparación con los términos del documento a clasificar que se intersecten con los términos del patrón, y por último al

emplear más de un vocablo en la constitución de un término, se acoto con el ordenamiento horizontal de los vocablos que constituyeron el n-grama.

Seguidamente se presentan las matrices de confusión de los experimentos desarrollados en dos escenarios (corpus con documentos privados y corpus estandarizado “corpora Reuters 21578”), con los que se efectúan los experimentos a tres niveles de n-gramas (monogramas, digramas y digramas ordenados horizontalmente). En el primer escenario se muestran resultados finales del proceso de clasificación, no se ahondó más en el detalle de este escenario debido a que sus resultados no son susceptibles de comparación. En el segundo escenario las rondas de experimentos se dieron en cuatro grupos de experimentos, el primero en la aplicación del método prevalente de forma original. En el segundo grupo de experimentos se aplicó el método prevalente añadida la técnica de ganancia de información con un umbral local y total. En el tercer grupo de experimentos, se aplicó la propuesta de la tesis de forma original. En el cuarto grupo de experimentos se añadió a la propuesta de la tesis la técnica de ganancia de información con un umbral total

Finalmente se plantean las respectivas conclusiones donde se muestra el resultado central de 84.17% de nivel de precisión en la clasificación, la misma que se obtuvo en el tercer grupo experimental con el empleo de monogramas del segundo escenario. Y de forma consecutiva se proponen algunas recomendaciones emanadas del proceso de investigación, las cuales se espera que abran la posibilidad de nuevas investigaciones.

Capítulo I:
MARCO METODOLÓGICO

1.1. Descripción de la Realidad Problemática

En la actualidad, existe el tema muy discutido relativo al sentido de las palabras, temas de desambiguación o entendimiento de lo que trata de decir o expresar un texto determinado. Esto está sujeto a demasiados aspectos colaterales que van en el orden de posiciones lingüísticas y computacionales.

Los lingüistas sostienen que las palabras que forman un léxico o diccionario, se relacionan entre sí, lo que invitaría a creer que todas las palabras son matemáticamente combinables; estos lingüistas al combinar sus conocimientos definen el grado de relación que tienen los diversos términos, y que término se parece o asemeja más a otro (Tejada J, 2009).

Los computólogos emplean métodos automatizados que recolectan recursos textuales (corpus, diccionarios electrónicos, documentos, etc) en los que existen las relaciones de proximidad semántica. En la misma se puede apreciar que no todas las palabras se relacionan entre sí, ya que existen combinaciones de palabras que se repiten con alguna frecuencia, lo cual es computable y de acuerdo al nivel o grado de ocurrencia,

se determina la distancia de proximidad existente entre ellas. En el medio existen combinaciones (n-gramas) que se encuentran en léxicos, diccionarios digitales y en corpus textuales (Hernández M. 2007).

Para que una computadora pueda analizar y comprender el texto usa combinaciones semánticas de palabras. El entendimiento del texto es necesario para tareas de procesamiento de lenguaje natural, tales como traducción de textos, clasificación de textos, detección de plagios, etc.

Las instituciones, dentro de las actividades propias de su negocio, producen diferentes tipos de textos, los mismos que se generan de forma caótica. Para lograr optimizar las labores de quienes emplean estos documentos, es necesario ordenar dichos documentos, para lo que es indispensable clasificarlos, lo que les permitiría tener un ordenamiento estructurado más eficiente en su manejo y efectivo en sus resultados.

Los textos que en demasía se acumulan progresivamente contienen información específica, esta información para ser consultada, se procede a efectuar una lectura pormenorizada lo cual representa un esfuerzo mayor, si se trata de gran cantidad de documentos, lo adecuado sería contar con un resumen extraído de la fuente original del texto.

1.1.1. Problema Principal

Baja calidad de vocablos relacionados semánticamente obtenidos de Modelos de Espacios de Palabras tradicionales basados en n-gramas. Dicha información decrementa la eficacia y precisión de aplicaciones de procesamiento de lenguaje natural, tales como la clasificación de textos.

1.2. Objetivos

En este punto se denotan los objetivos de la presente tesis. Inicialmente se describe el objetivo principal, para de forma consecutiva determinar los respectivos objetivos específicos.

1.2.1. Objetivo Principal

Mejorar la calidad de vocablos relacionados semánticamente mediante la construcción automática y análisis de Modelos de Espacios de Palabras basados en n-gramas. Estos vocablos deben mejorar la precisión de tareas de procesamiento de lenguaje natural, tales como la clasificación de textos

1.2.2. Objetivos Específicos

a) Determinar las relaciones semánticas que influyen positiva o negativamente en la construcción de un modelo multidimensional de espacio de palabras con n-gramas
b) Aplicar las relaciones semánticas proporcionadas por modelos de espacio de palabras en la clasificación de textos
c) Evaluar y validar el modelo

1.3. Justificación de la Investigación

Los n-gramas son combinaciones de vocablos que se emplean en el modelado de espacio de palabras. Estos modelos determinan las relaciones de proximidad semántica entre grupos de palabras. Se usan para resolver la ambigüedad propia del lenguaje y mejorar la clasificación de textos, que permite obtener mejores resultados que los métodos lingüísticos. El uso de estos modelos de espacio de palabras en la actualidad emplean n-gramas de un vocablo, en esta tesis se plantea construir un modelo de espacio de palabras basado en n-gramas de una o dos palabras y aplicarlos a la clasificación de textos, para de esta manera establecer un nivel de aporte en el sentido de demostrar que tan bueno son los modelos de espacio de palabras con estas combinaciones.

1.4. Relevancia

La presente tesis estará direccionada hacia la construcción de un modelo de espacios multidimensionales de palabras basados en n-gramas con ventanas de uno o dos vocablos, con lo cual se constituyen los n-gramas para establecer las proximidades

semánticas, con ello se examinará que tan mejor puede ser el tratamiento del espacio de palabras, esto es muy prometedor por que los resultados pueden ser mucho mejores para el tratamiento de la clasificación de documentos o alguna otra aplicación, se ha elegido la clasificación de textos, debido a que estos encierran un conjunto determinado de palabras que son recurrentes, donde la naturaleza de las mismas puede tener una relación directa entre su tipo y las relaciones de proximidad, lo que a entender de los expertos debe ser de una forma automática y alejada de cualquier juicio con intervención humana.

1.5. Delimitación de la Investigación

La presente tesis se limita a la propuesta de un método automático, el mismo que permita clasificar textos sin intervención humana, empleando n-gramas construidos a partir de un modelo multidimensional de espacios de palabras basados en n-gramas de uno o dos vocablos,

1.6. Metodología

1.6.1. Tipo de investigación:

Es una investigación aplicada (F. Eliana 2009, Caballero A. 2005) de nivel exploratorio (Hernandez R. 2010), debido a que se hará uso de conocimiento a priori en la conceptualización básica, y se propondrá soluciones alternativas a problemas ya existentes.

1.6.2. Método de la investigación:

Se aplicara el método de investigación en acción (Bausela E. 2010, Kember D.2010), debido a que a medida que se va desarrollando el proceso de investigación, también se va construyendo los documentos que avalen y den sustento a la investigación, además en el proceso de abstracción de hará una descomposición del objeto materia de estudio identificando sus elementos componentes para que en una forma sucesiva y reiterada, se efectúen más

descomposiciones hasta llegar a un nivel adecuado, el mismo que permita lograr el entendimiento del objeto materia de estudio.

1.7. Marco Teórico de Referencia

En este punto plantearemos las concepciones básicas sobre las cuales se establece la presente tesis, las cuales detallamos a continuación:

- **Modelo de Espacio de Palabras (Word Space Model)**

  Es una representación espacial del significado de palabras, en la cual a cada vocablo se le asigna una localidad semántica tomando en cuenta sus propiedades de distribución en el lenguaje, de tal manera que la medición de la lejanía o cercanía entre dos vocablos determina su relación o similitud semántica (Tejada J. 2009).

- **Matriz de Co-Ocurrencia**

  Los datos que se recogen en una matriz de co-ocurrencia son los pesos, y los vectores de contexto que se definen como las filas o las columnas de la matriz. Esta matriz cuenta las veces que las palabras ocurren en algún tipo de conjunción, y normalmente se denota por F (de frecuencia). Puede ser palabra por palabra de la matriz “w x w” donde w son los tipos de palabras en los datos, o una palabra por los documentos de la matriz “W x D”, donde D son los documentos en los datos. La matriz de co-ocurrencia registra la frecuencia con la que aparece una palabra en algún texto o documento (Schutze H. 1997).

- **Proximidad**

  Las coordenadas de una palabra solamente nos dá su posición en el espacio n-dimensional, lo importante es la relación con otros términos, y es ahí donde se aplica la metáfora de la proximidad (Sahlgren M. 2006).

1.8. Secuencia de la Investigación

La secuencia estructural de la presente tesis se representa en cuatro capítulos donde el primero se expone las concepciones formales del planteamiento metodológico. En el segundo capítulo se presentan todos los aspectos conceptuales así como el estado del arte referidos a la lingüística computacional y clasificación de documentos. En el capítulo tercero se hace el planteamiento de la propuesta para lograr el objetivo planteado. En el capítulo cuarto se presentan los experimentos que permiten validar la propuesta de una forma cuantitativa empleando un corpus estandarizado. Y por último se plantean las respectivas conclusiones, recomendaciones y trabajos futuros.

1.9. Resultados Esperados

Mejorar la calidad de la clasificación de textos por medio de un método que emplea un modelo multidimensional de espacio de palabras basado en n-gramas.

Capítulo II:
MARCO TEÓRICO

2.1. Lingüística computacional

En este punto citaremos los aspectos teóricos propuestos por diversos autores respecto al significado e implicancia de la lingüística computacional, sobre la cual nos apoyaremos para la elaboración del presente trabajo.

2.1.1. La Lingüística General

La lingüística es la ciencia que estudia los lenguajes naturales (Fernandez M. 1999), para ser más precisos, abarca un amplio conjunto de diferentes ciencias relacionadas. La lingüística general estudia la estructura general de varios lenguajes naturales y descubre las leyes universales de su funcionamiento. La lingüística general es una ciencia fundamental, desarrollada por muchos investigadores durante los últimos dos siglos y está basada en gran parte en los métodos y resultados de los primeros gramáticos que desarrollaron este tema desde la antigüedad.

La lingüística computacional es un paradigma, en el cual, los ordenadores proporcionan una metáfora formal para modelar y probar las teorías que tratan de dar cuenta del funcionamiento del lenguaje, de cómo somos capaces de entender, qué procesos subyacen a la conducta lingüística (Villayandre M. 2010).

Estos modelos formales que describen el funcionamiento del lenguaje, no tienen por qué reproducir exactamente el funcionamiento de la mente humana. Basta que nos faciliten una plataforma para acercarnos a la comprensión de un fenómeno tan complejo y cotidiano como el lenguaje.

Dominar el lenguaje natural es una de las tareas más complejas que se le puede asignar a una computadora. El lenguaje humano es con frecuencia alusivo y ambiguo. Alusivo, porque las palabras pueden incorporar referencias a múltiples niveles. Hasta el 40% en ciertos tipos de textos, pueden resultar ambiguos para una computadora (Villayandre M. 2010).

A. Componentes generales de la lingüística

Los componentes generales de la lingüística son aquellos que se encuentran en todas las lenguas, son los aspectos que configuran a un medio de comunicación el carácter de lenguaje, estos cinco componentes se resumen en la figura siguiente:

$$Componentes\ Generales = \begin{cases} Fonología \\ Morfología \\ Sintaxis \\ Semántica \\ Pragmática \end{cases}$$

Figura 1: Componentes de la Lingüística General
Fuente: Elaboración Propia

Fonología.- estudia los sistemas fónicos de las lenguas, trata con los sonidos que componen el habla.

Morfología, estudia la estructura interna de las palabras individuales y las leyes concernientes a la formación de nuevas palabras.

Sintaxis, considera la estructura de las oraciones y las formas cómo las palabras individuales están conectadas entre sí.

Semántica.-, estas están estrechamente relacionadas. La semántica trata con el significado de las palabras individuales y textos enteros

Pragmática.- estudia las motivaciones de la gente para producir textos u oraciones específicas.

B. Componentes especializados de la lingüística

La lingüística dentro de su estudio, discrimina algunos aspectos en función a perspectivas que buscan el entendimiento de los diversos fenómenos de la lengua, los fenómenos propiamente dichos no serán abordados en esta tesis, solo se les mencionará a razón de darle un contexto y ubicación a los contenidos posteriores. Los componentes especializados de la lingüística se pueden ver de una forma más clara en la siguiente gráfica.

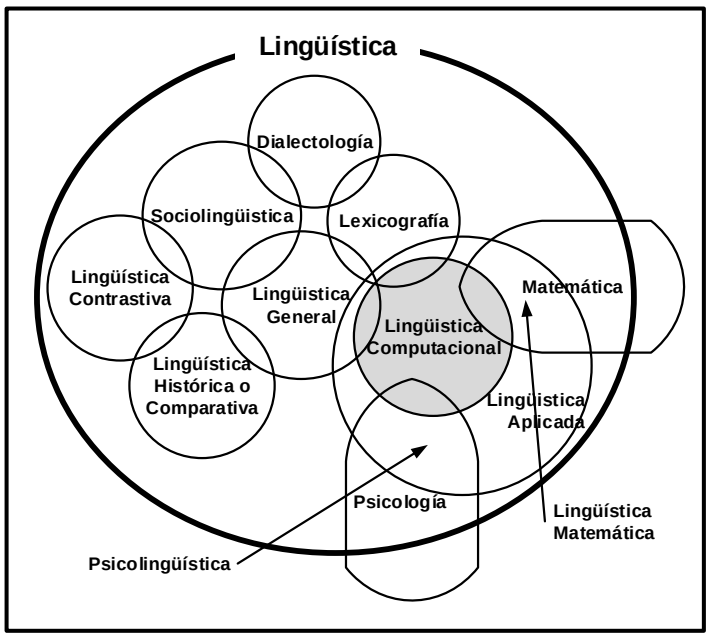


Figura 2: Estructura de la ciencia de la lingüística
Fuente: Hernández M. 2007

Lingüística histórica o comparativa.- estudia la historia de los lenguajes a través de la comparación entre ellos, por ejemplo, estudiando la historia de su similitud y diferencias. El segundo nombre es explicado por el hecho de que la comparación es el método principal en esta rama de la lingüística. La lingüística comparativa es incluso más antigua que la lingüística general, se originó en el siglo XVIII.

Lingüística contrastiva o lingüística tipológica.- clasifica una variedad de lenguajes de acuerdo a la similitud de sus características sin interesarse en el origen de los lenguajes.

La sociolingüística.- describe las variaciones de un lenguaje a través de la escala social. Es bien conocido que varios estratos sociales utilizan frecuentemente sublenguajes dentro de un lenguaje común, toda vez que una misma persona utiliza diferentes sublenguajes en diferentes situaciones.

La dialectología.- compara y describe los varios dialectos o sublenguajes de un lenguaje común, el cual es usado en diferentes áreas de un territorio donde algún lenguaje es usado oficialmente. Por ejemplo, en diferentes países de habla hispana, muchas palabras, combinaciones de palabras o incluso formas gramaticales son usadas diferentemente, sin mencionar las significativas diferencias en la pronunciación.

La lexicografía.- estudia el léxico o el conjunto de todas las palabras de un lenguaje específico, con sus significados, características gramaticales, pronunciación, etc, así como los métodos de compilación de varios diccionarios basados en dicho conocimiento.

La psicolingüística.- estudia el comportamiento del lenguaje de los seres humanos a través del significado de una serie de experimentos de tipo psicológico. Entre las áreas de especial interés, la psicolingüística estudia la enseñanza del lenguaje a los niños, enlaza la habilidad de lenguaje en

general y el arte del habla, así como otras características psicológicas conectadas con el lenguaje natural y lo expresado a través de él.

Lingüística matemática.- Hay dos diferentes vistas en la lingüística matemática. En la vista más estrecha, el término lingüística matemática es usado por la teoría de las gramáticas formales de un tipo especial llamadas gramáticas generativas. Esta es una de las primeras teorías puramente matemáticas dedicadas al lenguaje natural. Alternativamente, en la vista más amplia, la lingüística matemática es una intersección entre las matemáticas y la lingüística, por ejemplo, la parte matemática que toma el fenómeno lingüístico y las relaciones entre ellos como objetos de su posible aplicación e interpretación.

La lingüística aplicada.- desarrolla los métodos para la aplicación de las ideas y nociones de la lingüística general en la práctica humana. Hasta mediados del siglo XX, las aplicaciones de la lingüística estaban limitadas al desarrollo y mejoramiento de gramáticas y diccionarios impresos orientados al uso extensivo por no especialistas, así como los métodos racionales para la enseñanza de los lenguajes naturales, su ortografía y estilo. Este fue sólo un producto puramente práctico de la lingüística.

2.1.2. La Lingüística Computacional

La lingüística computacional es el estudio de los sistemas computacionales para comprender y generar el lenguaje natural. Aunque los objetivos de la investigación de la lingüística computacional son ampliamente variados, la motivación primaria ha sido siempre el desarrollo de sistemas específicos y prácticos que involucran al lenguaje natural. Existen cuatro clases de aplicaciones (Hernández M. 2007) que han sido centrales en el desarrollo de la lingüística computacional:

- Traducción de textos
- Extracción de información

- Búsquedas de Información
- Interfaces hombre máquina.

Según Moreno (Moreno A. 1998) la Lingüística Computacional es una disciplina que trata básicamente de dos cosas: lenguas naturales y computadoras. Muchas líneas de investigación comparten ambos objetivos aunque desde perspectivas diferentes. Como siempre hay que enfrentarse con el objeto de estudio y con la delimitación de las terminologías de las ciencias, hay que dejar claro que la lingüística computacional es equivalente al Procesamiento de Lenguaje Natural, y no es igual a la *lingüística informática* y la *ingeniería lingüística.* La Lingüística Computacional trata de la construcción de sistemas informáticos que procesen realmente estructuras lingüísticas y cuyo objetivo sea la simulación de la capacidad lingüística humana, independientemente de su carácter comercial o de investigación básica.

2.1.3. La lingüística computacional como parte de la lingüística

La lingüística computacional comparte con la lingüística general un interés por describir y descubrir cómo funciona el lenguaje y cómo podemos comunicarnos las personas, y difiere de la lingüística general en las herramientas que emplea para llevar a cabo sus investigaciones (Tejada C. 2009).

2.1.4. La lingüística computacional como rama de la informática

La lingüística computacional no consiste en solo estudiar el lenguaje y las lenguas, sino de hacerlo con el apoyo de computadoras, además de reproducir estos estudios en programas informáticos con algún fin práctico (Tejada C. 2009).

Desde la perspectiva de la Informática y de forma paralela a como sucede en el caso de la Lingüística, hay que destacar también que no es extraño hacer depender a la lingüística computacional, no directamente de la inteligencia artificial, sino de la parte de la inteligencia artificial que se ocupa

específicamente del lenguaje humano, el Procesamiento del Lenguaje Natural (en adelante PLN), subdisciplina que se caracteriza en general por presentar una orientación práctica y por estar centrada en el tratamiento de la lengua escrita (Tejada C. 2009).

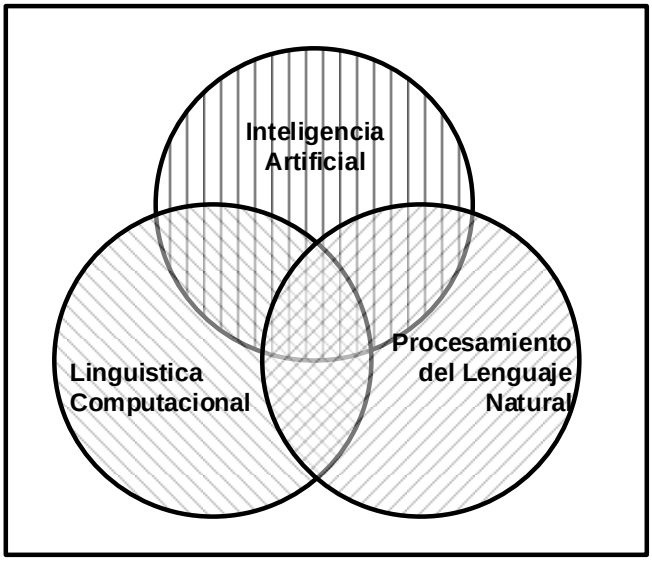


Figura 3: Relación de la LC con el PLN y la IA
Fuente: Elaboración propia

2.1.5. Objetivos de la lingüística computacional

Los objetivos teóricos, también llamados “científicos”, son independientes de cualquier aplicación y constituyen el ámbito de trabajo de la lingüística computacional teórica. Se concretan en:(Grisman R. 1991)

- Probar las gramáticas que propone la Lingüística Teórica.
- Investigar los procesos psicológicos que intervienen en la producción y comprensión del lenguaje dentro del marco general de la Ciencia Cognitiva.
- Estudiar la forma de representar el conocimiento general o del mundo.

Los objetivos aplicados, también llamados “tecnológicos” o “aplicaciones orientadas a la ingeniería”, tienen que ver con sistemas prácticos o programas informáticos específicos y constituyen el ámbito de trabajo de la lingüística computacional aplicada (Moreno A. 1998).

La lingüística computacional es una disciplina aplicada, en cuanto a qué el despegue tecnológico de los últimos años, asociado con el propio devenir metodológico en el campo de la Lingüística, ha provocado la atención al procesamiento artificial de las lenguas, al tratamiento informático de ingentes bases de datos lingüísticos, o a los medios automáticos de traducción, lo cual es ámbito de la lingüística computacional (Fernández M. 1999).

Actualmente, si la investigación en la lingüística computacional no ha avanzado con mayor rapidez no es a causa de limitaciones de tipo teórico, sino porque nuestros conocimientos sobre el lenguaje son todavía más pobres de lo que suponemos. Las aplicaciones computacionales dependen, hoy más que nunca, de una teoría lingüística que las avale y les proporcione el apoyo formal imprescindible para la gestión de sus datos (Moure T. 1996).

La lingüística computacional se caracteriza por su interdisciplinariedad, ya que es un dominio científico que utiliza los conocimientos de la lingüística pero también de otras disciplinas con las que tiene intersección (Tejada J. 2009).

2.1.6. Conocimiento lingüístico para la recuperación de información

Para alcanzar un conocimiento lingüístico, y poder recuperar información, se debe tomar en consideración los siguientes niveles que comprende un texto hablado o escrito, en el cual los seres humanos expresan ideas que se transmiten entre ellos (Contreras H. 2001).

$$Conocimiento\ Linguístico = \left\{ \begin{array}{l} Nivel\ Morfológico \\ Nivel\ Léxico \\ Nivel\ Sintáctico \\ Nivel\ Semántico \\ Nivel\ de\ Expansión\ de\ términos \\ Nivel\ de\ Producción\ de\ Vectores\ Semánticos \\ Nivel\ de\ Discurso \\ Nivel\ Pragmático \end{array} \right\}$$

Figura 4: Niveles de Conocimiento Lingüístico
Fuente: Elaboración propia

2.1.7. Formalismos para la representación del significado

Igual que ocurre con los demás niveles lingüísticos, la representación formal es un requisito indispensable para que el significado pueda ser abordado computacionalmente. Dicha representación debe ajustarse a una serie de parámetros (Juafsky D. et al. 2009).

- La verificabilidad, es decir, poder determinar la verdad o falsedad de un enunciado. Normalmente se hace comparando la representación del significado con una base de conocimientos sobre el mundo.
- La ausencia de ambigüedad, ya que esta puede dar lugar a múltiples interpretaciones, lo que supone un mayor coste de procesamiento computacional.
- Cierto grado de vaguedad, útil en ciertos casos.
- El empleo de una forma canónica, es decir, asignar la misma representación semántica a oraciones que expresan el mismo contenido proposicional, pese a que difieran en la forma.
- Capacidad para realizar inferencias, es decir, deducir información implícita (generalmente conocimiento del mundo) a partir de la representación del significado.
- Expresividad para que el sistema sea capaz de comprender una amplia gama de asuntos, no sólo los referidos a un dominio particular.

2.1.8. Procesamiento de Lenguaje Natural

Este concepto se desarrolló durante los inicios de la guerra fría y fue definido como el mecanismo que usaban los físicos Soviéticos para la traducción de documentos, uno de los primeros objetivos computacionales más investigados. Estos esfuerzos prematuros, por analizar y modelar el lenguaje humano, fueron caracterizados por una técnica sin conocimiento lingüístico y por el bajo rendimiento computacional de la época (Locke W. et al. 1955).

Es el uso de computadoras para entender lenguajes (naturales) humanos tales como inglés, francés, japonés, etc. Por entender no se quiere decir que el computador tenga pensamientos, sentimientos y conocimientos humanizados, sino que el computador pueda reconocer y usar información expresada en lenguaje humano (Covington M.).

Es aquel que encapsula un modelo del lenguaje natural en algoritmos apropiados y eficientes. En donde las técnicas de modelado están ampliamente relacionadas con eventos en muchos otros campos tales como: Ciencia de la computación, Lingüística, Matemática y Neurociencia (Manaris B. et al. 1996).

2.2. Modelo de Espacio de Palabras

En esta parte presentamos los aspectos más relevantes respecto al modelamiento de espacio de palabras, sobre lo cual se basa fundamentalmente la propuesta que se hace en esta tesis para mejorar el proceso de clasificación de documentos basado en la técnica que esta propone, cuyas definiciones describimos a continuación:

2.2.1. Definición

Es la representación espacial de los significados de las palabras, siendo las más cercanas las palabras relacionadas, en las diferentes dimensiones (Schütze H. 1993). Pero el modelo de espacio de palabras no sólo es la representación espacial de los significados; es también la manera en que el espacio es construido (Sahlgren M. 2006). Lo que hace el modelo de espacio de palabras único en comparación con otros modelos geométricos del significado es que el espacio es construido sin intervención humana, y sin conocimientos a priori o restricciones sobre la representación de las semejanzas. En el modelo de espacio de palabras, las semejanzas entre las palabras son extraídas automáticamente del propio lenguaje buscando el uso real que este les da (Grishman R. 1991).

2.2.2. Elementos semánticos de las palabras

Los elementos de los cuales están compuestos los términos o palabras está dado por los siguientes (Lakoff G. et al. 1999):

- Ubicación
- Dirección
- Proximidad

2.2.3. Metáfora geométrica del significado

El significado se representa como ubicaciones en un espacio semántico, y la semejanza semántica como proximidades entre las ubicaciones (Schütze H. 1993). La metáfora geométrica del significado no está basada en el razonamiento intelectual sobre el lenguaje (Moreno A. 1998). Esta representación requiere de un conjunto de palabras para determinar la proximidad semántica entre éstas.

2.2.4. Relaciones espaciales

La proximidad, es una relación espacial básica (Schütze H. 1993), en ella se denota que existe una aproximación semántica más fuerte en la medida que dos términos se encuentren más próximos, dicho de otra manera, cuando la distancia entre dos términos es menor, la relación es más fuerte.

2.2.5. Hipótesis distributiva del significado

El modelo de espacio de palabras usa datos estadísticos de las propiedades distributivas de las palabras, para poner las palabras en regiones similares con las propiedades distributivas similares, así estas proximidades reflejan la semejanza distributiva.

La idea fundamental detrás del uso de la información distributiva es la hipótesis distributiva:

Las palabras con las propiedades distributivas similares tienen significados similares.

2.2.6. Vector de contexto

Son aquellos que describen las ubicaciones en el espacio de contexto, en relación a la totalidad del contexto de las palabras (Schütze H. 1993).

2.2.7. Matriz de Coocurrencia

Los enfoques de Schütze, Qiu y Frei (Schütze H. 1993, Qiu Y. et al. 1993) fueron adoptados como estándar para los algoritmos de espacio de palabras, obteniendo los datos en una matriz de coocurrencia (realizando una cuenta de las coocurrencias) y los vectores de contexto están definidos como las filas o columnas de la matriz; esta matriz es llamada como matriz de coocurrencia y denotada como *F* (por frecuencia). Esta matriz se forma por la frecuencia de coocurrencia de una palabra en el contexto de otra palabra (Mandala R. 2000).

2.2.8. Antecedentes en espacio de palabras

Las palabras con significados similares se producen con palabras más próximas, si se dispone de suficiente material de texto (Schutze H. et al. 1997).

Uno de los primeros estudios, donde explícitamente formulan e investigan la hipótesis distributiva, obteniendo que palabras similares en el significado se producen en contextos similares (Rubenstein et al. 1965).

La metodología distributiva que es referencia para la hipótesis distributiva. En la metodología distributiva de Harris, se reducen un conjunto de hechos distributivos que establecen las entidades básicas de los fonemas de la lengua, los morfemas, unidades sintácticas y las relaciones distributivas entre ellas. Basado en que los miembros de las clases básicas de las entidades actúan

distribucionalmente de forma similar, y que por lo tanto pueden ser agrupados de acuerdo con su comportamiento distributivo (Harris Z. 1951).

2.2.9. Antecedentes en vectores de contexto

El enfoque de semántica diferencial de la representación del significado. En este enfoque las palabras son representadas por característica de los vectores donde los elementos son actitudes humanas, en pares de adjetivos (Osgood C. et al. 1957). El trabajo de Osgood influyó en la investigación conexionista que usa representación del significado (Smith et al. 1981):.

Se emplearon las llamadas micro-características para representar el significado de las palabras. Este conjunto de micro-características eran representadas como un vector, donde cada elemento corresponde a un nivel de activación para una micro-característica en particular. Este enfoque es similar al de Osgood, pero éste no recibió influencias del trabajo de Osgood (Waltz D. et al.1985).

Los vectores de contexto son definidos como filas y columnas de una matriz, en la cual los elementos se registran con el número de veces de coocurrencia (Manaris B. et al. 1996, Schütze H. 1992).

Un enfoque similar al de Schutze, la diferencia radica en el uso de la matriz, tomando un conjunto de palabras más amplio para obtener el número de coocurrencias (Qiu Y. et al. 1993).

Se observaron inconvenientes propios del enfoque de espacio de características. La idea se basa en un número limitado de características semánticas para describir el significado de las palabras. Introduce el término vector de contexto para describir la representación del espacio de características. Define al vector de contexto como un conjunto de características derivadas manualmente (Sahgren M. 2006).

2.3. Clasificación de Documentos

2.3.1. Definición de clasificación

Consiste en colocar un documento dentro de un grupo de clases previamente definidas (Coyotl R. 2007). La mayor parte del trabajo en esta área se ha enfocado en la clasificación de textos por su tema o tópico. Sin embargo, un documento también puede ser clasificado de acuerdo a su estilo (clasificación no-temática). En la clasificación no-temática se consideran tareas tales como la clasificación de opiniones, la detección de plagio, la atribución de autoría, la clasificación por género, etc.

La clasificación automática de textos, tiene por objetivo caracterizar los documentos en referencia a un número de categorías establecidas de acuerdo a su contenido, esto debido a que un documento cualquiera puede pertenecer a una, varias o todas, y hasta incluso ninguna de las categorías establecidas con anterioridad (Joachims T. 1998). Cuando se emplea el aprendizaje automático, la razón de aprender a partir de casos definidos de documentos, es que ello nos permita hacer asignaciones a las categorías de una forma automática.

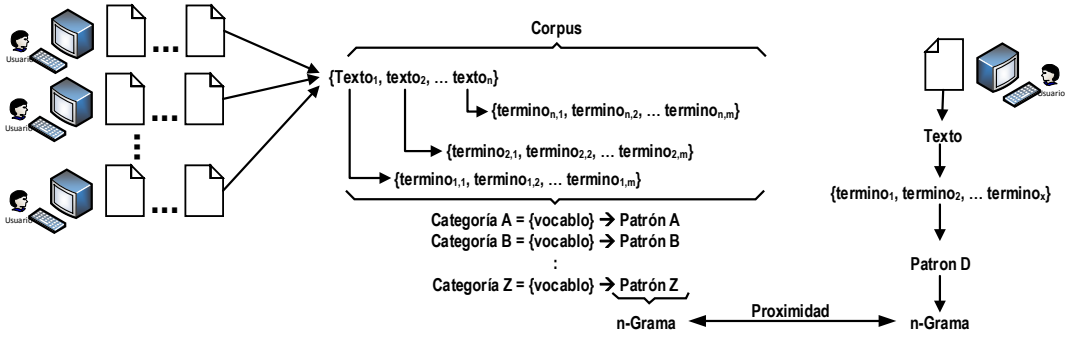


Figura 5: Esquema de la clasificación de documentos
Fuente: Elaboración propia

La clasificación de documentos es un proceso que consiste en coleccionar un conjunto de documentos elaborados por uno o más usuarios, los mismos que constituyen un corpus, cada documento es una colección de términos, los cuales al ser disgregados constituyen diferentes categorías, las mismas que son caracterizadas determinando un patrón que identifica a la categoría, dichas

categoría son determinadas por un patrón en forma de n-grama. Los usuarios al crear un documento forman un conjunto de términos que son disgregados para de forma similar determinar su patrón en forma de n-grama. De esta manera es posible determinar la proximidad entre ambos patrones. Por lo cual, al existir una mayor proximidad entre ambos patrones, es más probable que el documento corresponda a dicha categoría.

2.3.2. Algoritmos de clasificación de documentos

La clasificación de documentos no es un tema nuevo, es un aspecto ya tratado por diferentes autores de acuerdo a propuestas auténticas y/o adaptaciones efectuadas a métodos ya existentes, dentro de los diversos métodos de clasificación podemos mencionar los siguientes:

- Algoritmo probabilístico de Naive Bayes
- Algoritmo de Rocchio
- Algoritmo del vecino más próximo y variantes
- Algoritmos basados en redes neuronales
- Algoritmo basado en el modelo de espacio de palabras

A. Algoritmo probabilístico de Naive Bayes

El teorema de Bayes permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero, además de ser este algoritmo una de los más conocidos en el tema de clasificación en general, cuyo planteamiento data de hace mucho tiempo atrás (Marom M. 1961). Este algoritmo trata de estimar la probabilidad de que un documento pertenezca a una categoría. Dicha pertenencia depende de la posesión de una serie de características, de cada una de las cuales conocemos la probabilidad de que aparezcan en los documentos que pertenecen a la categoría en cuestión, dichas características son los términos que conforman los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada categoría, pueden obtenerse a partir de

los documentos de entrenamiento; para ello se utilizan las frecuencias de aparición en la colección de entrenamiento (Moens M. et al 1999, Lewis D. et al 1994).

B. Algoritmo de Rocchio

Este algoritmo es conocido y aplicado en la realimentación de consultas. Su planteamiento es simple: formulada y ejecutada una primera consulta, el usuario examina los documentos devueltos y determina cuáles le resultan relevantes y cuáles no. Con estos datos, el sistema genera automáticamente una nueva consulta, basándose en los documentos que el usuario señaló como relevantes o no relevantes. En este contexto, el algoritmo de Rocchio proporciona un sistema para construir el vector de la nueva consulta, recalculando los pesos de los términos de ésta y aplicando un coeficiente a los pesos de los la consulta inicial, otro a los de los documentos relevantes y otro distinto a los de los no relevantes (Rocchio J. 1971).

El mismo algoritmo de Rocchio proporciona un sistema para construir los patrones de cada una de las clases, tipos o categorías de documentos. Así, partiendo de una colección de entrenamiento, categorizada manualmente de antemano, y aplicando el modelo vectorial, podemos construir vectores patrón para cada una de las clases, considerando como ejemplos positivos los documentos de entrenamiento de esa categoría, y como ejemplos negativos los de las demás categorías.

Una vez que se tienen los patrones de cada una de las clases, el proceso de entrenamiento o aprendizaje está concluído. Para categorizar nuevos documentos, simplemente se estima la similitud entre el nuevo documento y cada uno de los patrones. El que arroja un índice mayor nos indica la categoría a la que se debe asignar ese documento.

C. Algoritmo del vecino más próximo - Nearest Neighbour

La idea sobre la que trabaja este algoritmo es calcular la similitud entre el documento a clasificar y cada uno de los documentos de entrenamiento, a cuál de éstos es mas parecido, esto nos estará indicando a qué clase o categoría debemos asignar el documeto que deseamos clasificar. Visto de una forma más práctica, el vecino más próximo puede aplicarse con cualquier programa de recuperación de tipo *best macth*. Lo más frecuente es utilizar alguno basado en el modelo vectorial, pero esto no es imprescindible. Lo necesario es que sea *best macth* y no de comparación exacta, como pueda ser el caso de los booleanos; el algoritmo se basa en localizar el documento más similar o parecido al que se desea clasificar. Para esto no hay más que utilizar ese documento como si fuera una consulta sobre la colección de entrenamiento, una vez localizado el documento de entrenamiento más similar, dado que éstos han sido previamente categorizados manualmente, sabemos a qué categoría pertenece y, por ende, a qué categoría debemos asignar el documento que estamos clasficando.

Una de las variantes más conocidas de este algoritmo es la del k-nearest neighbour o KNN que consiste en tomar los k documentos más parecidos, en lugar de sólo el primero. Como en esos k documentos se presume que existiran varias categorías, se suman los coeficientes de los de cada una de ellas. La que más puntos acumule, será la candidata elegida. El KNN une a su sencillez una eficacia notable. Obsérvese que el proceso de entrenamiento no es más que la indización o descripción automática de los documentos, y que tanto dicho entrenamiento como la propia categorización pueden llevarse a cabo con instrumentos bien conocidos y disponibles para cualquiera. KNN parece especialmente eficaz cuando el número de categorías posibles es alto, y cuando los documentos son heterogéneos y difusos (Gövert N. et al 1999).

D. Redes neuronales

De una manera genérica, una de las principales aplicaciones de las redes neuronales es el reconocimiento de patrones, por lo tanto es posible aplicarlo a problemas de categorización de documentos. Una red neuronal consta de varias capas de unidades de procesamiento o neuronas interconectadas; en el ámbito que nos ocupa la capa de entrada recibe términos, mientras que las unidades o neuronas de la capa de salida mapea clases o categorías. Las interconexiones tienen pesos, es decir, un coeficiente que expresa la mayor o menor fuerza de la conexión. Es posible entrenar una red para que, dada una entrada determinada (los términos de un documento), produzca la salida deseada (la clase que corresponde a ese documento). El proceso de entrenamiento consta de un ajuste de los pesos de las interconexiones, a fin de que la salida sea la deseada (Schutze H. et al 1995).

E. Algoritmo basado en el modelo de espacio de palabras

En el presente documento se aplicara este algoritmo de una forma más detallada, sin embargo esta forma de clasificar documentos consiste en extraer de un conjunto de documento categorizados, los términos que estos contienen, eliminar los términos que no aportan información respecto al motivo del texto, luego en base a los términos restantes, extraer conjugaciones de uno dos o mas palabras, con lo que se construye un n-grama (monogramas, digramas, etc) que representa el patrón de la categoría de la cual se extrajeron los textos que lo conforman. Posteriormente se procesa de forma similar el documento a clasificar, para de forma consecutiva calcular la proximidad entre ambos patrones, la categoría con la que resulte más próxima, se considerara como la categoría a la que pertenece el referido documento.

2.3.3. Proceso de clasificación de textos

El proceso de clasificación de textos está compuesto por dos etapas (entrenamiento y prueba), las cuales tienen una secuencia de pasos que se muestran en el siguiente esquema:

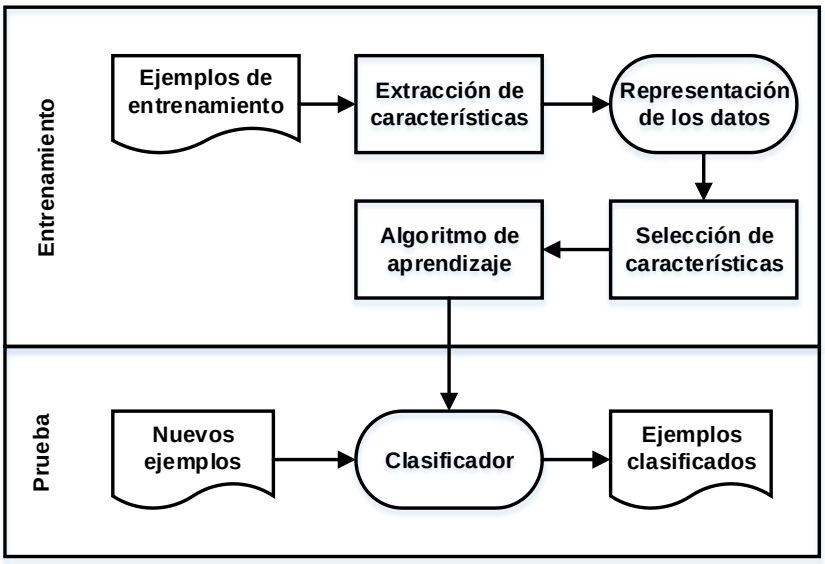


Figura 6: El paradigma de aprendizaje y la clasificación.
Fuente: Coyoltl R. 2007.

En la etapa de entrenamiento, se debe contar con un conjunto de documentos denominados de ejemplos de entrenamiento, por medio de los cuales se efectúa una extracción de características, estas se deben representar por medio de los datos que los componen, luego de ello se hace una selección de las características relevantes, para con ello ejecutar los algoritmos de aprendizaje que permitirá pasar a la etapa de prueba.

En la etapa de prueba, por medio de un conjunto de documentos nuevos (Ejemplos nuevos), se pone a evaluación la capacidad de discernir entre las características extraídas de estos documentos, y el que contienen los patrones de entrenamiento, para establecer de esta forma, el grado de asertividad que posee el modelo propuesto.

Cabe destacar que este modelo de clasificación, cumple con las etapas de la prueba experimental, donde se puede entender claramente que los ejemplos

de entrenamiento son un grupo control, y los nuevos ejemplos para efectuar las pruebas, son grupos objetivo, con la salvedad de que estos últimos ya han sido clasificados con anterioridad, así que de una forma certera se puede establecer una medida de efectividad o confianza, respecto a la propuesta en cuestión.

2.3.4. Representación de un documento

Cuando se desea hacer una clasificación de documentos de forma automática, el proceso de entrenamiento se lleva a cabo con un conjunto de documentos definido, al cual se denomina conjunto de entrenamiento, esto nos da una representación tipificada, esto es susceptible de que se clasifique por medio de algoritmos, una forma de hacer esto posible es por medio de un modelo vectorial, o modelado de espacio de palabras, lo cual es ampliamente usado en clasificación.

Dado un documento d_j, este es representado por un vector $\vec{d}_j = (w_{1j} \quad \cdots \quad w_{rj})$, donde w son los términos o palabras y r es el número total de palabras que se encuentran presentes en el documento, estas son parte de un diccionario particular, es usual que este número r, sea luego de haber filtrado o excluido a las palabras funcionales o vacías (Ass K. et al. 1999).

Existe otra manera de efectuar esta representación, y está dada por lo que se denomina como un *lema*, lo cual se efectúa con el propósito de que se contabilicen las palabras con el mismo sentido conceptual, o significado de raíz, para de manera consecuente se pueda asignarles su respectivo peso al término específico.

Existen varias formas de asignarles pesos a los términos, los mismos son:

- Ponderado booleano
- Ponderado por frecuencia de término
- Ponderado tf-idf

A. Ponderado booleano

Este tipo de ponderación acepta solamente uno de dos valores; uno (1) o cero (0), para los casos de si el término aparece o no aparece en el documento, lo cual se expresa de la forma siguiente:

$$w_{ij} = \begin{cases} 1\ si\ t_i\ aparece\ en\ d_j \\ 0\ \ En\ caso\ contrario \end{cases}$$

B. Ponderado por frecuencia de término

En este caso se contabiliza el número de veces que aparece un termino i en el documento d_j, lo cual se denota como f_{ij}, en este tipo de ponderación, cabe destacar que se interpreta la frecuencia, en el sentido del grado de importancia del término para el documento, esto quiere decir que a medida que el termino figura con más frecuencia, significa que el mismo es muy importante para el referido documento.

$$w_{ij} = f_{ij}$$

C. Ponderado tf-idf

Asigna el peso de la palabra i en el documento j, en proporción al número de ocurrencias de la palabra en el documento y en proporción inversa al número de documentos en la colección, para los cuales ocurre la palabra al menos una vez.

$$w_{ij} = f_{ij} * Log\left(\frac{N}{n_i}\right)$$

Donde N es el número de documentos en la colección y n_i es el número de documentos en los que el término aparece.

2.3.5. Reducción dimensional

Cuando se emplea los modelos de espacio de palabra, se configuran vectores de numerosas dimensiones cuando se emplean los conjuntos de entrenamiento, esto es complejo para un procesamiento eficiente. Es necesario reducir el número de elementos con los cuales trabajar (Coyotl R. 2007). Para reducir las dimensiones de los vectores generados por el modelo de espacio de palabras se efectúa una selección de un subconjunto de características, con la finalidad de encontrar los términos con mayor capacidad de discriminación.

La técnica de ganancia de información (Yang Y. et al. 1997) consiste en medir el número de bits de información obtenida, para predecir la categoría por medio de la presencia o ausencia de una palabra en el documento.

$$IG(t_i) = -\sum_{k=1}^{M} P(c_k)\log P(c_k) + P\,(t_i)\sum_{k=1}^{M} P(c_k|t_i)\log P(c_k|t_i) + P(\overline{t_i})\sum_{k=1}^{M} P(c_k|\overline{t_i})\log P(c_k|\overline{t_i})$$

Donde:

- IG: Ganancia de información (*information gain*)
- $c_1 \ldots c_k$: Conjunto de clases
- t_i: Término del cual se calculará la ganancia de información.
- M: Número de clases
- $P(c_k)$: Probabilidad de la clase c_k
- $P(t_i)$: Probabilidad de seleccionar un documento que contienen el término t_i
- $P(c_k|t_i)$: Probabilidad condicional de que un documento con el término t_i pertenezca a la categoría c_k
- $P(\overline{t_i})$: Probabilidad de seleccionar un documento que no contienen el término t_i
- $P(c_k|\overline{t_i})$: es la probabilidad condicional de que un documento con el término t_i no pertenezca a la categoría c_k

Calculando la ganancia de información de cada término, es posible identificar aquellos términos con mayor capacidad discriminante. Usualmente se seleccionan aquellos términos que sobrepasan un cierto umbral.

2.3.6. Corpus

Un corpus lingüístico es una colección de elementos lingüísticos seleccionados y ordenados de acuerdo con criterios lingüísticos explícitos con la finalidad de ser usado como muestra de la lengua, un corpus lingüístico consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos (Guerra P. 1998).

A efectos de trabajar en tareas de procesamiento de lenguaje natural, específicamente en las tareas de clasificación de documentos, se acuña el término de corpus estandarizado; el cual es un conjunto de documentos que están al alcance de la comunidad interesada en la clasificación de textos, los mismos que serán empleados por diversos investigadores a efectos de poder comparar resultados con sus pares en otras latitudes de una forma homogénea.

Existen varios corpus estandarizados, los mismos que son de libre acceso, dentro de los cuales podemos citar a:

- Reuters21578-Apte-90Cat
- Reuters21578-Apte-115Cat
- RCV1, RCV2
- Ohsumed
- Etc.

Los corpus son combinaciones semánticas de palabras, estas combinaciones han sido extraídas de una gran cantidad de recursos textuales, en los que el cómputo de las ocurrencias ha sido contabilizado e incluidas en forma de

frecuencia. Las combinaciones se establecen por la ventana de proximidad existente en los textos procesados, debe aclararse que en la determinación de la referida ventana, ha mediado un pre-procesamiento que excluye a las palabras vacías.

En el ámbito de procesamiento de lenguaje natural, un corpus lingüístico es una colección de textos en soporte electrónico, normalmente amplio, que contiene ejemplos reales de uso de una lengua tal y como es utilizada por los hablantes, con sus errores, peculiaridades y excepciones (Borja F. 2007).

Los corpus son de los tipos siguientes:

- Corpus de análisis
- Corpus de evaluación.

Un corpus de análisis o entrenamiento, se emplea para establecer hasta qué punto llega la capacidad discriminatoria de los n-gramas. Su diseño y los experimentos que se efectúan con éste han sido elaborados teniendo en cuenta los principales factores cuyo efecto se considera que puede afectar (reducir o anular) la capacidad de distinguir una marca identificativa de un vocablo o término específico.

El corpus de evaluación o control, permite llevar a cabo pruebas de evaluación de un proceso de clasificación, para ello se somete a comparación el patrón del documento con los patrones de las diversas categorías de documentos definidas.

Hay que entender e inscribir el empleo de corpus en Lingüística dentro de una perspectiva metodológica general que adopta el empirismo como forma de concebir el estudio de la lengua. En este sentido, el empleo de datos reales, de muestras de uso lingüístico, resulta el complemento ideal y la referencia ineludible en cualquier investigación que aspire a dar cuenta de algún aspecto relacionado con el lenguaje: los datos son los que apoyan o contradicen una

postura teórica, los que permiten inferir reglas y generalizaciones, los que proporcionan informaciones cuantitativas, etc. Y también constituyen el material necesario como punto de partida para el desarrollo de una aplicación práctica. En general, los principales parámetros para clasificar los corpus se centran en (Tejada J. 2009):

- La modalidad de la lengua
- El número de lenguas a que pertenecen los textos
- El tamaño o cantidad de textos que conforman el corpus
- Los límites del corpus
- La variedad lingüística o el grado de especialización de los textos
- El período temporal que abarcan los textos
- El tratamiento aplicado al corpus

A. Funciones de los corpus de clasificación

Los corpus utilizados en los sistemas de clasificación basados en aprendizaje automático cumplen tres funciones (Tomás D. 2009):

1. **Entrenamiento**. Las preguntas etiquetadas correctamente contenidas en el corpus de entrenamiento permiten al sistema aprender a clasificar nuevas instancias. El conjunto de entrenamiento debe ser representativo de las situaciones con las que se puede encontrar el sistema de clasificación durante su funcionamiento.

2. **Validación**. Si fuera necesario, una porción del conjunto de entrenamiento puede emplearse para ajustar los parámetros del algoritmo de aprendizaje mediante la optimización de su funcionamiento en este subconjunto. Este proceso de optimización se puede también llevar a cabo sobre todo el conjunto de entrenamiento realizando una validación cruzada.

3. **Evaluación**. Una vez tenemos el sistema entrenado, el siguiente paso es evaluar su funcionamiento. Esta evaluación se lleva a cabo proporcionando al sistema un nuevo conjunto de ejemplos (preguntas) para los que el sistema determinará la clase a la que pertenecen. La clase que predice el sistema se contrasta con la clase real asignada previamente por un humano (o a través de algún tipo de medición) para obtener el rendimiento del clasificador. Para que la evaluación sea fiable, el conjunto de preguntas de evaluación debe ser diferente al empleado durante el entrenamiento. Al igual que ocurría en el punto anterior, podemos realizar una validación cruzada sobre el corpus de entrenamiento para evitar la necesidad de definir un corpus específico de evaluación. Esta técnica es especialmente útil cuando los corpus de entrenamiento son pequeños y no se desea dedicar una parte de ellos exclusivamente para la evaluación.

B. Requisitos de un corpus de referencia

Para que una colección de textos pueda ser considerada un corpus de referencia de una lengua, según el uso del término en la Ingeniería Lingüística actual, debe cumplir cuatro requisitos (McEnery T. et al. 2001):

1. **Debe ser representativo de la lengua.**

   Un corpus representativo es aquel formado por muestras suficientes que den cuenta de cómo es la lengua. Así, la representatividad del corpus depende del origen de las muestras que lo forman: qué procedencia tienen y en qué cantidad están representadas.

   Un corpus de propósito general es representativo de una lengua si consta de textos procedentes de fuentes diversas y cada uno con una cantidad de palabras compensada, de tal manera que no haya más textos de un dominio que otro.

Según la variedad lingüística que representen se establecen dos clases de corpus: corpus orales, que representan la variedad oral de las lenguas; y corpus representativos de la variedad escrita de las lenguas. Estos corpus escritos, además, suelen representar la variedad estándar.

2. **Debe tener un tamaño finito y compensado.**

Dado que todo corpus es finito, el tamaño y cantidad de muestras de cada variedad lingüística debe estar compensada, esto es, el corpus debe estar balanceado, de tal manera que las porciones de muestras textuales sean uniformes con relación a un criterio determinado.

3. **Debe estar en formato electrónico.**

Si los textos contenidos en los archivos contenedores, están en medio electrónico, independientemente del formato que este tenga, todos deben estar en archivos de computadora, puesto que esto garantizaría que siempre todo aquel que quiera probar sus proyectos con estos archivos, trabajaran bajo un único patrón estandarizado, con lo cual la comparación entre los resultados de un autor puedan ser comparados directamente contra los de otro.

4. **Debe ser una referencia estándar de la lengua que representa.**

Esta característica hace alusión a la utilidad de éste: un corpus se considera útil si es utilizado por diferentes investigadores con diferentes fines (no necesariamente previstos por los desarrolladores del corpus). Para que ello sea posible, el corpus debe ser referencia estándar en todos estos estudios o aplicaciones, de tal manera que las diferencias entre éstos no dependan de la construcción del corpus, sino de los métodos o procesos seguidos en su explotación.

2.3.7. N-gramas

Los n-gramas son conjunciones de vocablos que determinan un espacio multidimensional, estos se logran definir en la medida de la conjunción de vocablos existentes en un texto, un número constante por vez, el número de vocablos conjugados determinara el nivel o grado de n-grama empleado.

Los n-gramas son una técnica que se puede interpretar como un lenguaje estadístico de múltiples dimensiones (Masschussetts Institute 2003), donde se evalúan una gran diversidad de criterios o ejes dimensionales, estos al sobrepasar el espacio visual del ser humano, son materialmente imposibles de graficar en un espacio bidimensional, de tal forma que sean entendibles o su representación gráfica sea aceptable.

2.3.8. Lematización

La lematización es un proceso lingüístico que consiste en: dada una forma flexionada de un vocablo (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente es en determinar el vocablo que de un origen pueda generar diversas variaciones de género, cantidad, etc. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Es decir, el lema de una palabra es la palabra que nos encontraríamos como entrada en un diccionario tradicional: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos.

A. Stemming

El *Stemming* es un método para reducir una palabra a su raíz o a un *stem* (en inglés) o lema. Hay algunos algoritmos de *stemming* que ayudan en sistemas de recuperación de información.

B. Algoritmo de porter

Es un algoritmo que por medio de un análisis morfológico de las palabras, se reducen a una base común o raíz, sobre la cual se puede implementar programas que requieran de la lematización de textos. El algoritmo de Porter permite hacer stemming, esto es extraer los sufijos y prefijos comunes de palabras literalmente diferentes pero con una raíz común que pueden ser consideradas como un sólo término

2.3.9. Calculo de proximidad

El patrón de un tipo determinado de documento está dado por la conjunción de los vocablos relevantes que constituyen su corpus de documentos, esto ensamblado en la forma de algún n-grama, de la misma manera, el patrón de un documento por clasificar, debe estar definida por el mismo nivel de n-grama del patrón del tipo de documento con el cual se quiere determinar su proximidad.

En la medida que el patrón de documento a clasificar se aproxime más al patrón del tipo de documento, se puede afirmar que es muy probable que el documento pertenezca a dicho tipo clasificado.

Para estimar la proximidad de estos patrones, se puede aplicar la fórmula del coseno de dos vectores (Sahlgren M. 2006), la misma que presentamos a continuación:

$$\cos\_measure(\vec{w},\vec{r}) = \frac{\vec{w},\vec{r}}{|\vec{w}| \quad |\vec{r}|} = \frac{\sum_{j=1}^{n} w_j \times r_{i,j}}{\sqrt{\sum_{j=1}^{n}(w_j)^2} \times \sqrt{\sum_{j=1}^{n}(r_{i,j})^2}}$$

Donde $\vec{w}$ corresponde al vector patrón del documento a clasificar y $\vec{r}$ es el vector del patrón del tipo de documento i con el cual se quiere determinar su proximidad

2.3.10. Estado del arte del modelamiento de espacio de palabras en la clasificación de documentos

En lo que respecta a la clasificación de documentos, existen una gran variedad de artículos que trabajan la metódica de formas diversas y desde perspectivas particulares, esto quiere decir que describen sus técnicas basadas en problemas y necesidades propias, sin embargo, si se desea hacer una propuesta que contenga un aporte susceptible de comparación, se debe emplear un corpus de documentos estandarizados, que para fines del presente trabajo se empleará el corpus Reuters21578-Apte-90Cat (Moschitti A.).

Francisco Romero y otros plantean el tema “Filtrado de información mediante prototipos borrosos y perfiles basados en criterios de calidad de datos”, que es un modelo de filtrado basado en una estructura de categorías conceptuales. Donde la estructura se define partiendo de un conjunto de documentos no estructurados, es necesario seguir los siguientes pasos (Romero F. et al. 2008):

1. Calcular la calidad de datos de todos los documentos, y descartar aquellos cuyas mediciones no estén dentro de los rangos de aceptación establecidos en los requisitos de calidad de datos de los usuarios que no superen un umbral mínimo de calidad.

2. Preproceso lingüístico: Selección de las palabras que van a representar conceptualmente a las categorías de la estructura.

3. Representación de los documentos en base a los conceptos tratados en sus contenidos. Para ello se utiliza el modelo FIS-CRM (Olivas J. et al. 2003).

4. Agrupación de los documentos conceptualmente similares mediante clustering (Romero F. et al. 2006).

5. Extracción de conceptos claves.

6. Creación de una base de conocimiento utilizando Categorías Prototípicas Borrosas.

La estructura básica de filtrado se basa en las diferentes categorías extraídas mediante procesos de clustering. Estas categorías comprenden una serie de documentos conceptualmente similares. Las categorías estarán organizadas en una jerarquía que reflejará los diferentes niveles de especificidad tratados en los conceptos. Cada uno de esos documentos puede estar clasificado en diferentes categorías sobre las cuales poseerán un grado de pertenencia. A su vez, el resultado ofrecido empleando esta propuesta es del orden del 83% versus el 80% de las clasificaciones sin emplear criterios de calidad.

En el Departamento de Ciencias de la Computación e Ingeniería de la Universidad de California en los Estados Unidos, desarrollo una técnica de clasificación utilizando inferencia transductiva en clasificación de documentos con el corpus Reuters 21578, empleando la técnica de “Naibe Bayes” y “Soporte de Maquina Vectorial”, donde luego de presentar su propuesta presenta sus resultados, los mismos que resumimos en el siguiente cuadro (Branson K. 2001):

Técnica	Precisión promedio
Naive Bayes	55.2%
Transductive Naive Bayes	81.44%
Support Vector Machines	48.43%
Transductive Support Vector Machines	61.33%

Tabla 1: Resultados de la clasificación de Kristin Branson
Fuente: Branson K. 2001.

Capítulo III:
Método Propuesto

3.1. Premisas de la investigación

En el presente trabajo de tesis se inicia partiendo de un conjunto de premisas, las mismas que justificaran las acciones desarrolladas y cuyos mecanismos y resultados se presentan posteriormente.

Premisa 1: Los documentos tienen una naturaleza y estructura, los mismos que a su vez están constituidos por textos que son un conjunto de vocablos que son regularmente empleados en documentos de similar categoría.

Premisa 2: Los vocablos individualmente constituyen información, y estos a la vez que se asocian entre sí, incrementan el volumen de información, la misma que podría caracterizar en mejor manera a los documentos que los contengan.

Premisa 3: Cuando se emplean más de un vocablo, en un proceso de clasificación automática (n-gramas), puede darse el caso que una conjunción de vocablos (A, B), pueda presentarse como (B, A) en el mismo documento o uno similar del mismo tipo, lo cual en términos prácticos, constituiría una dispersión de las frecuencias asociadas

a la categoría definida, para lo cual, dado el caso se debería indexar horizontalmente los vocablos, y de esta forma evitar la dispersión de las frecuencias.

Premisa 4: Al constituirse los vocablos asociados de uno, dos o más, estos se deberán catalogar asociados al tipo de documento que les dio origen, una vez constituida la asociación y elaborado la concentración de frecuencias, estos vocablos se asumirán como únicos a efectos de desarrollar los cálculos requeridos para la determinación de las proximidades entre los vocablos y el tipo de documento asociado.

Premisa 5: Los corpus de entrenamiento no serán modificados o reevaluados a efectos de construir los n-gramas que configuren el patrón, esto debido a que al desarrollar una aplicación real, los corpus de entrenamiento ya pasaron filtros diversos que permiten superar esta fase, y no es dable evaluar constantemente cada vez que se requiera hacer una clasificación. Cabe aclarar que esta premisa contradice los postulados de la evaluación de un corpus, que muchas veces requiere de un ajuste del contenido a efectos de mejorar el proceso de entrenamiento de los algoritmos.

3.2. Esquema del Método Propuesto

El método propuesto está comprendido por dos etapas claramente definidas (entrenamiento y control). En la etapa de entrenamiento se debe contar con un conjunto de documentos $\{d_1, d_2, \dots d_n\}$, los cuales serán sometidos a un procesamiento que incluye las fases de pre-procesamiento, indexado y reducción de la dimensionalidad. Estas acciones nos permite lograr tres entregables: Léxico, Co-ocurrencias, Escalas de Clasificación. La etapa de control, emplea un conjunto de documentos $\{d'_1, d'_2, \dots, d'_m\}$, los cuales se someten a un proceso de clasificación empleando los entregables de la etapa de entrenamiento, en forma conjunta permite obtener un catálogo de documentos clasificados, el esquema del método propuesto se puede apreciar en la figura siguiente:

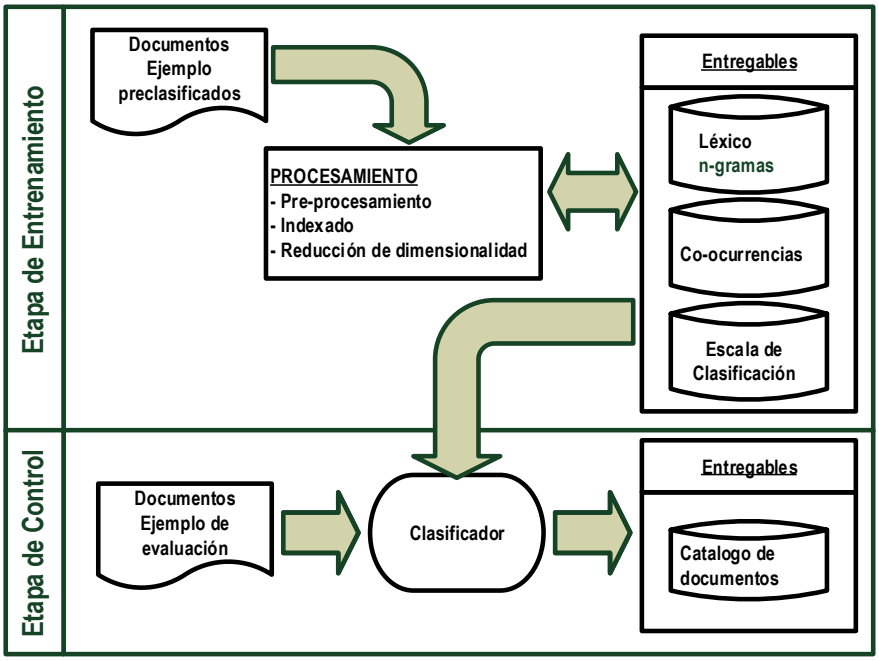


Figura 7: Esquema del Método
Fuente: Elaboración propia

Un léxico es un catálogo de términos estandarizado para un conjunto de vocablos, los mismos que deben estar relacionados de alguna manera con una raíz única, para que con ello al efectuar las tareas de clasificación, la reducción de los términos se vean entendidos en una raíz común, cuando ello sea pertinente.

A continuación se detallaran de una forma más específica y formalizada los pormenores de cada paso del método.

3.2.1. Etapa de Entrenamiento

En la etapa de entrenamiento, lo que se busca es por medio de los documentos ejemplo de entrenamiento, establecer los patrones de clasificación, para lo cual se obtienen los entregables descritos anteriormente, cuyos elementos describiremos a continuación:

A. Documentos Ejemplo Preclasificados

El conjunto de documentos de entrenamiento $\{d_1, d_2, \dots d_n\}$, son documentos preclasificados, estos deben ser cuantificados con anterioridad, para que en los procesos sucesivos, se puedan caracterizar

de una forma adecuada y esta caracterización sea susceptible de ser asociada a los tipos de documento que correspondan.

Por lo tanto; los documentos se deben presentar en un catálogo que impliquen la ubicación del documento y su tipo de documento preclasificado.

Una “clase de documento” es el nombre que identifique a los patrones de entrenamiento para el proceso de clasificación de textos, los mismos que están predefinidos y asociados al archivo que servirá para esta fase.

B. Procesamiento

El procesamiento está definido por tres fases, las cuales tienen un propósito propio que se detallará a continuación:

- Pre-procesamiento
- Indexado
- Reducción dimensional

1. Pre-procesamiento

Esta fase tiene el propósito de eliminar elementos textuales que no contienen información relevante. Esta información haría que se eleve los costos de procesamiento, así como la calidad de información, debido a que los patrones se harían más semejantes, por tanto dichos no serían muy efectivos al momento de ejecutar las rutinas de clasificación, en este caso se efectuara las siguientes acciones:

Etiquetas. Los documentos suelen contener elementos textuales en forma de cabeceras o pies de página, etiquetas en formatos html o xml, todos ellos serán retirados del texto para proseguir el procesamiento, dicho de otra forma, se debe borrar todos estos elementos y dejar el texto de forma continua.

Palabras Vacías. Las preposiciones, artículos, conjunciones y otros, son elementos textuales que abundan en un texto normal. Estas tampoco suelen aportar información, sus conjunciones son muy reiteradas, y de incluirlos en el proceso de construcción de los digramas y más, confundirían el propósito de estas combinaciones para una clasificación efectiva, por lo tanto estas palabras deben eliminarse.

Extracción de Verbos y Sustantivos. Las palabras cuando conforman un texto, en esencia comprenden su contenido en función a verbos y sustantivos. La particularidad de un texto está definido por estos, y su patrón de clasificación se perfecciona al incluir únicamente estos elementos. Por ello se debe reconstituir el texto con solamente estas palabras.

Lematización de Palabras. Las palabras empleadas en una redacción de texto normal, son usadas haciendo uso de sufijos para darles sentidos de acción, tiempo, etc, los mismos que en el contexto de la redacción original, hace un texto entendible fonéticamente, pero todos estos términos tienen una raíz, la misma que es general a todas y que encierra el significado base de los diversos términos. Por ello la lematización consiste en reducir las palabras a su lema o raíz, para de esta forma las conjunciones sean concentradas de una manera uniforme, por ejemplo hablar, hablará, hablando, hablo, etc., tiene su raíz en habl.

Los procesos a desarrollar podrían efectuarse tomando como base el conjunto de procesos que se ejemplifican en el siguiente módulo de procesos:

```
…
… …
… … …
sTexto = ExtraerTexto(“d:\Reuter21578\test\corn\0009622”)
sTexto = PreProcesamiento(sTexto)
… … …
```

```
… …

Funcion PreProcesamiento( texto)
       sTextoProcesado = ProcesarTexto(texto)
       sTextoProcesado = LematizarTexto(sTextoProcesado)
       Retornar sTextoProcesado
Fin Funcion

Funcion ProcesarTexto( texto)
       _texto = EliminaLinks(_texto)
       _texto = EliminarSignos(_texto)
       _texto = EliminarCaracteresEspeciales(_texto)
       _texto = EliminarApostrofe(_texto)
       _texto = EliminarVacias(_texto)
       _texto = EliminarSignosPuntuacion(_texto)
       _texto = EliminarNumeros(_texto)
       _texto = QuitarPalabrasConNumeros(_texto)
       _texto = ExtraerSustantivosVerbos(_texto)
       Retornar _texto
Fin Funcion
```

2. Indexado

Es un proceso de ordenamiento de las palabras existentes en todos los documentos, por ello esto puede ser representado por una matriz en la cual se incluyan todos los términos existentes en todos los documentos de la forma siguiente:

$$A = \begin{Bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{Bmatrix}$$

Figura 8: Representación de un n-grama
Fuente: Elaboración propia

Si A es un n-grama de grado n, significa que cada término está compuesto por n conjugaciones de vocablos, de acuerdo a la ubicación dentro de los documentos que constituyen el corpus, esto significaría que este n-grama está compuesto por m términos, donde cada línea de la matriz representaría un término.

Es simple concebir la idea de un ordenamiento vertical, lo que dicho de otro modo es el ordenamiento entre líneas diferentes de una matriz, esto de acuerdo a un criterio generalmente alfabético, pero lo que vale la

pena aclarar es el ordenamiento horizontal, esto implica que se debe ordenar dos o más vocablos conjugados de acuerdo al nivel de n-grama que se esté trabajando, esto significa que un termino t_{ij} contiene un conjunto de vocablos $\{v_1, \ v_2 \dots \ v_n\}$ donde v_x debe ser menor que v_{x+1}. Por ejemplo si tenemos el término t cuyo digrama que contiene los vocablos {"aaa", "bbb"}, en alguna parte del texto, lo cual tendría una frecuencia igual a uno, y en algún otro lugar tiene los vocablos {"bbb", "aaa"} con una frecuencia igual a uno, lo que se debería hacer es ordenar alfabéticamente estos vocablos quedándonos con {"aaa", "bbb"} y una frecuencia igual a dos, lo cual evitaría la dispersión de las frecuencias. Debido a que $a^2 + b^2$ es menor que $(a + b)^2$ al aplicar la ley de los cosenos en el cálculo de proximidad. Esto se puede apreciar de mejor manera en el gráfico siguiente:

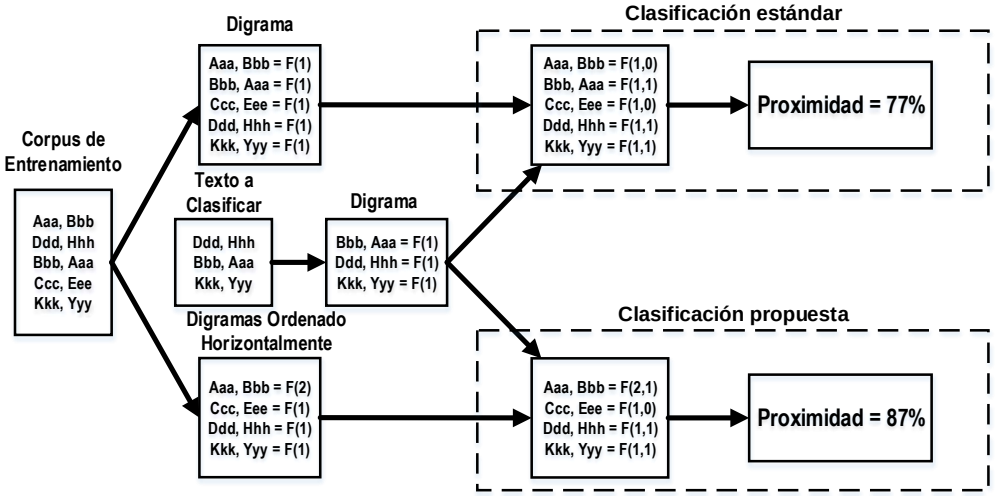


Figura 9: Ejemplo de clasificación
Fuente: Elaboración propia

Posterior a la indexación vertical y horizontal se debe crear tantas tablas como tipos de documentos puedan existir los mismos que deben contener la frecuencia de los *términos* que contenga el texto pre-procesado, el mismo que debe estar organizado de la forma siguiente:

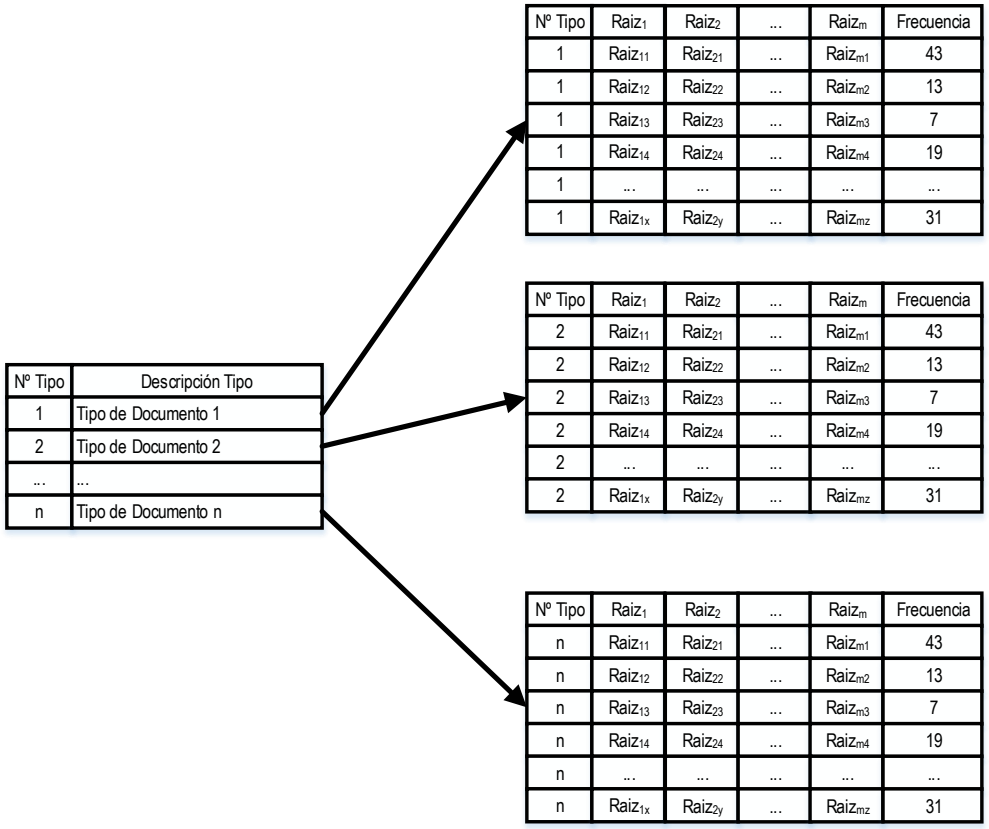


Figura 10: Esquema de la creación de tablas patrón por tipo de documento
Fuente: Elaboración propia

3. Reducción dimensional

En esta fase se deberá tener en cuenta la ganancia de información de los términos encontrados en el conjunto de documentos ejemplo de entrenamiento, para lo cual se deberá asumir los términos que concentren la mayor ganancia de información (IG_i), lo cual se trató en el punto 2.3.4., y trabajar con un umbral total o local según convenga.

Entiéndase como umbral total al nivel de ganancia de información representada por el total de las categorías existentes, esto significa que una vez calculada la ganancia de información de los términos en los n-gramas, se determinará un nivel o grado de participación segmentándose todo cálculo de proximidad a este grupo reducido de términos. Cabe aclarar que al aplicar esta técnica, alguna categoría podría quedar con pocos términos y hasta incluso con ninguno.

El umbral local es el nivel de ganancia de información empleada por cada categoría de las existentes, esto significa que dentro de cada grupo o categoría, se tomara un nivel de participación en función a dicha ganancia de información.

C. Entregables

En esta primera etapa se plantea tres entregables, cada uno con motivo o propósito propio los cuales son:

- Léxico
- Co-ocurrencias
- Escalas de clasificación

1. Léxico

El léxico es el conjunto de palabras contenidas en el volumen total de documentos, esta se trabajará en primer orden a un nivel de palabras individuales según las ocurrencias que se encuentren al explorar cada documento. En segundo orden se trabaja a nivel de raíz, dejando expresamente establecido el orden y conservando la relación entre la raíz y el conjunto de palabras asociadas de las cuales se determinó dicha raíz.

2. Co-ocurrencias

La matriz de co-ocurrencia es la correspondencia que existe entre las palabras en forma combinatoria, donde se puede apreciar que no todas las palabras se corresponden entre sí, esto se denota por la existencia de un valor cero en la intersección de estas palabras. En esta tesis acotaremos la denominación "*término*", para referirnos a las combinaciones en el momento de construir la matriz, esto debido a que indicar una palabra significaría una restricción implícita de comparar una a una estas palabras.

El propósito de esta tesis es representar dos o más palabras, en forma combinatoria de acuerdo a las ocurrencias de estas conjunciones gramaticales. Por ello cuando nos referimos a los "*términos*", según sea el caso podríamos referirnos a dos o más palabras asociadas y posteriormente conjugadas en sus relaciones espaciales. Por ejemplo: si se tratara del siguiente párrafo:

> Cuando me levanto por las mañanas y veo que un nuevo día inicia en mi vida, elevo una plegaria y ruego a Dios que me acompañe y guíe mis acciones al mejor de los términos posibles

Figura 11: Texto ejemplo
Fuente: Elaboración propia

Luego del pre-procesamiento este texto terminaría de la forma siguiente:

> levant mañan ver nuev inici vida elev plegaria rueg dios acompañ guiar accion termin

Figura 12: Texto ejemplo pre-procesado
Fuente: Elaboración propia

Un léxico normal o regular estaría formado por los términos presentes en los textos pre-procesados, esto asociado a su raíz, como se puede apreciar en la figura siguiente:

Raíz	Palabra
levant	levantar
	levanto
	levantaré
	levanta
	…
mañan	mañana
ver	ver
	verá
	veré
	…
nuev	nuevo
	nueva
	renueva
	…
inici	inicio
	iniciar
	Iniciare
	…

Raíz	Palabra
vida	vida
	…
elev	elevo
	…
plegaria	plegaria
	…
rueg	ruego
	…
dios	Dios
	…
acompañ	acompañe
	…
guiar	guíe
	…
accion	acciones
	…
termin	término
	…

Figura 13: Ejemplo de léxico normal
Fuente: Elaboración propia

En esta tesis, el léxico generado sería definido en una relación de asociación entre los diversos vocablos, la asociación seria determinada por dos o más vocablos, según sea la profundidad del análisis que se desea realizar, lo cual se puede apreciar de una forma más clara en la figura siguiente:

2 vocablos
levant -mañan
mañan -ver
ver -nuev
nuev -inici
inici -vida
vida -elev
elev -plegaria
plegaria -rueg
rueg -dios
dios -acompañ
acompañ -guiar
guiar -accion
accion -termin

3 vocablos
levant -mañan -ver
mañan -ver -nuev
ver -nuev -inici
nuev -inici -vida
inici -vida -elev
vida -elev -plegaria
elev -plegaria -rueg
plegaria -rueg -dios
rueg -dios -acompañ
dios -acompañ -guiar
acompañ -guiar -accion
guiar -accion -termin

Figura 14: Ejemplo de léxico de n vocablos
Fuente: Elaboración propia

Para tener un manejo más efectivo de la información contenida en este nuevo léxico, la indexación de este deberá ser efectuada con anterioridad y obtener una tabla que se muestra en la figura siguiente:

2 vocablos
accion -guiar
accion -termin
acompañ -dios
acompañ -guiar
dios -rueg
elev -plegaria
elev -vida
inici -nuev
inici -vida
levant -mañan
mañan -ver
nuev -ver
plegaria -rueg

3 vocablos
accion -acompañ -guiar
accion -guiar -termin
acompañ -dios -guiar
acompañ -dios -rueg
dios -plegaria -rueg
elev -inici -vida
elev -plegaria -rueg
elev -plegaria -vida
inici -nuev -ver
inici -nuev -vida
levant -mañan -ver
mañan -nuev -ver

Figura 15: Ejemplo de léxico de n vocablos indexado
Fuente: Elaboración propia

Posteriormente, se efectúa la construcción de la matriz de co-ocurrencias, esta matriz contiene la frecuencia de las ocurrencias en el conjunto de documentos de los términos (n-vocablos), como se presenta a continuación

	accion -guiar	accion - termin	acompañ -dios	acompañ -guiar	dios - rueg	elev - plegaria	elev - vida	inici - nuev	inici - vida	levant - mañan	mañan - ver	nuev - ver	plegaria -rueg
accion -guiar	0	1	9	15	5	9	4	7	15	7	13	15	5
accion -termin	7	0	11	9	7	5	15	9	16	10	8	3	2
acompañ -dios	6	11	0	11	17	2	17	17	12	3	6	10	15
acompañ -guiar	1	13	4	0	14	15	1	16	14	16	3	13	8
dios -rueg	5	12	11	6	0	12	14	3	15	12	2	8	9
elev -plegaria	11	17	6	12	5	0	3	3	0	5	0	8	14
elev -vida	7	17	15	3	9	10	0	5	1	7	15	7	12
inici -nuev	1	16	4	17	3	13	7	0	5	17	4	10	5
inici -vida	6	6	7	13	0	17	13	1	0	8	5	10	11
levant -mañan	17	10	6	3	0	15	9	11	16	0	5	12	14
mañan -ver	12	0	14	0	3	17	13	1	2	12	0	17	13
nuev -ver	17	17	7	5	0	17	16	0	1	6	12	0	4
plegaria -rueg	17	12	12	13	4	5	7	6	4	1	15	12	0

Figura 16: Matriz de co-ocurrencia de 2 vocablos
Fuente: Elaboración propia

	accion - acompañ -guiar	accion -guiar - termin	acompañ -dios - guiar	acompañ -dios - rueg	dios - plegaria -rueg	elev -inici -vida	elev - plegaria -rueg	elev - plegaria -vida	inici - nuev - ver	inici - nuev - vida	levant - mañan - ver	mañan - nuev - ver
accion -acompañ -guiar	0	3	15	5	5	3	11	14	15	5	3	16
accion -guiar –termin	15	0	8	6	16	5	10	14	0	10	15	16
acompañ -dios -guiar	3	16	0	0	13	7	3	5	4	6	12	15
acompañ -dios -rueg	3	0	12	0	5	3	8	3	10	6	16	2
dios -plegaria -rueg	9	6	15	16	0	0	11	3	5	0	14	1
elev -inici -vida	11	17	3	8	7	0	3	4	9	8	15	2
elev -plegaria -rueg	11	17	8	1	12	13	0	8	8	8	17	3
elev -plegaria -vida	6	7	4	12	10	6	13	0	16	16	2	6
inici -nuev -ver	8	0	3	5	1	3	0	5	0	17	14	7
inici -nuev -vida	3	8	3	5	3	5	7	12	14	0	5	6
levant -mañan -ver	13	14	2	13	7	2	13	15	6	11	0	7
mañan -nuev -ver	11	4	14	14	0	2	4	16	1	6	14	0

Figura 17: Matriz de co-ocurrencia de 3 vocablos
Fuente: Elaboración propia

3. Escalas de clasificación

La escala de clasificación es un catálogo referencial de tipos de documentos que pueden ser confeccionados de dos maneras. En primer orden puede establecerse una clasificación a priori efectuada unilateralmente por alguna persona entendida en la materia de los textos a clasificar. En segundo orden se puede hacer una clasificación automática, esto empleando alguna técnica de clasificación o agrupamiento de documentos en función a características proporcionadas por la conjunción de dos o más vocablos, la tabla que contenga este catálogo, debe tener definidas las columnas de tipo de documento y cantidad de documentos de entrenamiento.

Nº Tipo	Descripción Tipo	Nº Documentos
1	Tipo de Documento 1	
2	Tipo de Documento 2	
3	Tipo de Documento 3	
…		
n	Tipo de Documento n	

Tabla 2: Tabla de número de documentos por tipo de documento
Fuente: Elaboración propia

3.2.2. Etapa de Control

En esta etapa se debe poner a prueba de una forma controlada el método de clasificación propuesto, para ello se plantea una secuencia de elementos que se debe cumplir, los cuales se detallan a continuación:

A. Documentos Ejemplo de Evaluación

Los documentos ejemplo de evaluación $\{d'_1, d'_2, \dots, d'_m\}$, constituyen un grupo de textos que contienen un determinado tipo de documento, el cual será contrastado por la mecánica de clasificación que se haya establecido, para luego determinar el grado de precisión de ésta forma de clasificación.

B. Clasificación

En la etapa de clasificación, se procesa los documentos ejemplo en los términos de la clase de n-grama a emplear. Luego se conjugan con los n-gramas obtenidos en la etapa de entrenamiento. Posteriormente se establece la proximidad que pueda haber entre ambos grupos de conjunciones, por medio de la ley de los cosenos, para luego establecer a que tipo definido se aproxima más el documento en proceso de evaluación, esto permite determinar en un grado aceptable la clasificación. La fórmula que permite lograr esto se trató en el punto 2.3.7. la cual simplificamos a continuación:

$$p_{ij} = \frac{\sum_{k=1}^{n} E_k C_k}{\sqrt{\sum_{k=1}^{n} E_k^2} \quad \sqrt{\sum_{k=1}^{n} C_k^2}}$$

Donde

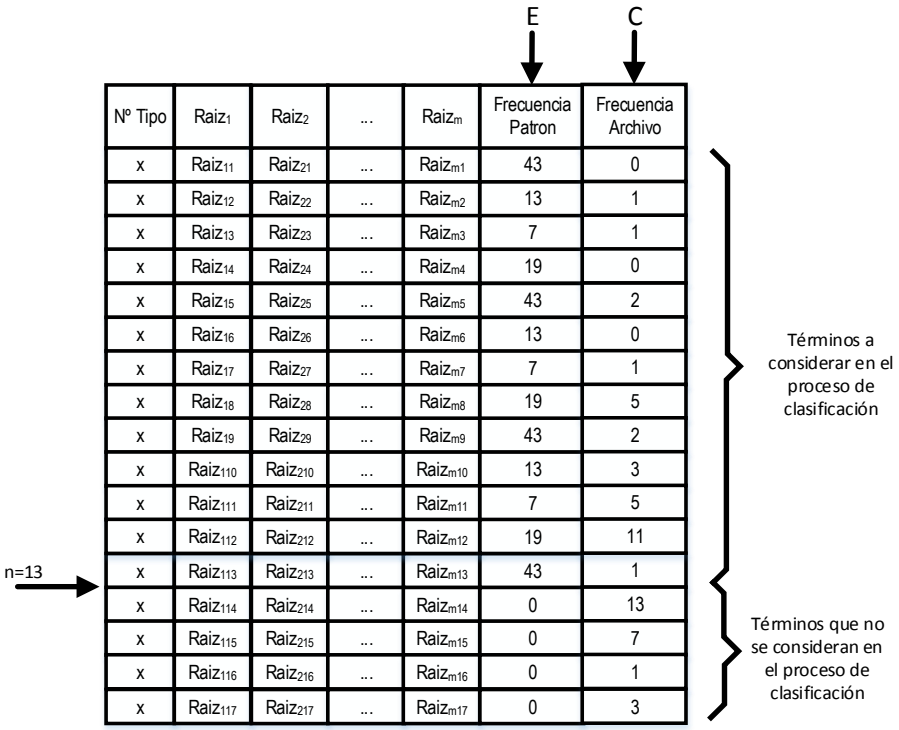
p_{ij} Grado de proximidad entre el documento i por clasificar y el patrón del documento j con el cual se está comparando.

E Frecuencias de los términos del patrón del tipo de documento j

C Frecuencias de los términos del documento i, que se intersectan con los términos del patrón del términos del documento j

n Número de términos que contiene el patrón de la categoría de documento tipo j

Es un hecho que el conjunto de términos que existan en el documento por clasificar, contenga términos que no se encuentren en el patrón de términos del n-grama del tipo de documento con el cual se esté comparando; en este caso las frecuencias de los términos del documento que no estén en el patrón del n-grama del tipo de documento con el cual se compara, no se deben considerar en el cálculo de proximidad, lo cual se ve en la figura siguiente.



Nº Tipo	$Raiz_1$	$Raiz_2$	...	$Raiz_m$	Frecuencia Patron	Frecuencia Archivo
x	$Raiz_{11}$	$Raiz_{21}$	...	$Raiz_{m1}$	43	0
x	$Raiz_{12}$	$Raiz_{22}$	...	$Raiz_{m2}$	13	1
x	$Raiz_{13}$	$Raiz_{23}$	...	$Raiz_{m3}$	7	1
x	$Raiz_{14}$	$Raiz_{24}$	...	$Raiz_{m4}$	19	0
x	$Raiz_{15}$	$Raiz_{25}$	...	$Raiz_{m5}$	43	2
x	$Raiz_{16}$	$Raiz_{26}$	...	$Raiz_{m6}$	13	0
x	$Raiz_{17}$	$Raiz_{27}$	...	$Raiz_{m7}$	7	1
x	$Raiz_{18}$	$Raiz_{28}$	...	$Raiz_{m8}$	19	5
x	$Raiz_{19}$	$Raiz_{29}$	...	$Raiz_{m9}$	43	2
x	$Raiz_{110}$	$Raiz_{210}$	...	$Raiz_{m10}$	13	3
x	$Raiz_{111}$	$Raiz_{211}$	...	$Raiz_{m11}$	7	5
x	$Raiz_{112}$	$Raiz_{212}$	...	$Raiz_{m12}$	19	11
x	$Raiz_{113}$	$Raiz_{213}$	...	$Raiz_{m13}$	43	1
x	$Raiz_{114}$	$Raiz_{214}$	...	$Raiz_{m14}$	0	13
x	$Raiz_{115}$	$Raiz_{215}$	...	$Raiz_{m15}$	0	7
x	$Raiz_{116}$	$Raiz_{216}$	...	$Raiz_{m16}$	0	1
x	$Raiz_{117}$	$Raiz_{217}$	...	$Raiz_{m17}$	0	3

Figura 18: Representación de la matriz de términos empleados en el proceso de clasificación
Fuente: Elaboración propia

De una forma esquemática en la figura siguiente se muestra un ejemplo de clasificación de un documento respecto a un corpus de un tipo definido, en este se puede apreciar que la proximidad es superior cuándo no se consideran los términos del documento a clasificar que no estén en el corpus del tipo de documento definido.

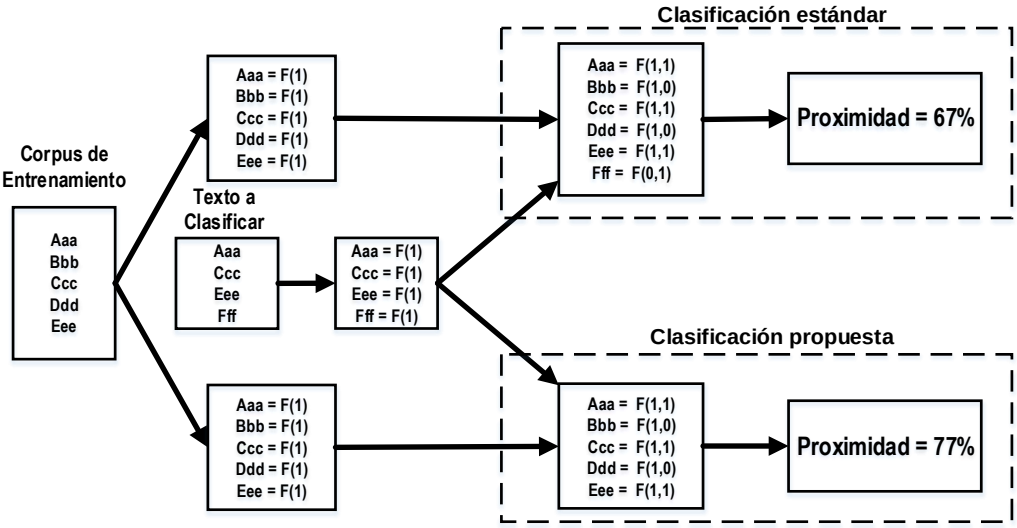


Figura 19: Ejemplo de clasificación
Fuente: Elaboración propia

C. Entregables

Al término de esta etapa, se tendrá un catálogo de documentos pre clasificados, comparados a la par con la clasificación automática que haya efectuado la mecánica de la propuesta, que en términos totales nos darán un grado de certeza sobre la precisión del método. De forma tangible se debe presentar una matriz de doble entrada (matriz de confusión) que además de presentar la cantidad de documentos clasificados, debe presentar la cantidad de documentos por tipo que fueron clasificados como los diferentes tipos definidos, esta matriz puede elaborarse como se muestra en la tabla siguiente:

	Total	Doc. Tipo 1	Doc. Tipo 2	Doc. Tipo 3	…	Doc. Tipo n
Doc. Tipo 1	A	a_1	a_2	a_3	…	a_n
Doc. Tipo 2	B	b_1	b_2	b_3	…	b_n
Doc. Tipo 3	C	c_1	c_2	c_3	…	c_n
…	…	…	…	…	…	…
Doc. Tipo n	N	n_1	n_2	n_3	…	n_n

Tabla 3: Matriz de confusión
Fuente: Elaboración propia

Posteriormente se debe sacar la proporción de la clasificación efectuada, para ello se debe calcular la proporción de la cantidad de documentos

clasificados por tipo, dividido entre el total de documentos de ese tipo que se emplearon para probar el método de clasificación

	Doc. Tipo 1	Doc. Tipo 2	Doc. Tipo 3	…	Doc. Tipo n
Doc. Tipo 1	**a_1/A**	a_2/A	a_3/A	…	a_n/A
Doc. Tipo 2	b_1/B	**b_2/B**	b_3/B	…	b_n/B
Doc. Tipo 3	c_1/C	c_2/C	**c_3/C**	…	c_n/C
…	…	…	…	…	…
Doc. Tipo n	n_1/N	n_2/N	n_3/N	…	**n_n/N**

Tabla 4: Matriz de confusión porcentual
Fuente: Elaboración propia

De la tabla anterior podemos extraer los resultados de la intersección de los tipos definidos de documentos que fueron clasificados correctamente, su promedio determinaría el grado de precisión del método, sin embargo, ello no impediría sacar conclusiones de los resultados individuales que presenten los demás casilleros.

En una aplicación real la clasificación se emplea para tener un acceso más rápido a la información requerida, permite hacer una búsqueda segmentada por las categorías definidas, en el caso de la clasificación de documentos, un documento puede pertenecer a más de una categoría, esto se determinará por un umbral de pertenencia que se determine de acuerdo a la exhaustividad con las que se desee hacer dicha búsqueda, en tal sentido se empleará el porcentaje de proximidad del tipo de documento en el cual se desee efectuar la búsqueda, en un sentido práctico se puede decir que se empleara la columna del tipo de documento para segmentar los documentos con los que se efectuara la búsqueda, como se muestra en la tabla siguiente:

	Doc. Tipo 1	Doc. Tipo 2	Doc. Tipo 3	…	Doc. Tipo n
Documento 1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	…	$P_{1,n}$
Documento 2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$	…	$P_{2,n}$
Documento 3	$P_{3,1}$	$P_{3,2}$	$P_{3,3}$	…	$P_{3,n}$
…	…	…	…	…	…
Documento m	$P_{m,1}$	$P_{m,2}$	$P_{m,3}$	…	$P_{m,n}$

Tabla 5: Catálogo de Documentos
Fuente: Elaboración propia

Capítulo IV:
Evaluación del Método Propuesto

4.1. Aspectos generales.

Las pruebas experimentales se elaboraron basados en dos escenarios, el primero conformado por un conjunto de documentos coleccionados y preclasificados provenientes de una oficina en una institución pública, el segundo se conformó por un conjunto de documentos que conforman un corpus estandarizado “Reuters21578-Apte-90Cat” (Moschitti A.), del cual se extrajeron ocho categorías, las mismas que contaban con un volumen adecuado para las pruebas a desarrollar.

4.2. Clasificación en el primer escenario – Documentos particulares

El escenario que comprende este corpus, está constituido por archivos propios de una oficina en una entidad pública, donde se tomó un conjunto de documentos para entrenamiento y para prueba, el mismo que se detalla a continuación (Cornejo V. 2012):

Motivo	Tipo de documento	Número de Archivos	
		Entrenamiento	Prueba
1	Cartas de Presentación	86	60
2	Asensos	128	60
3	Contratos	153	60
4	Procedimientos Administrativos de Grado	74	60
5	Eventos Académicos	83	56
	Total =	524	296

Tabla 6: Composición del corpus privado
Fuente: Elaboración propia

Efectuado el proceso de entrenamiento, y luego de construir el corpus de monogramas y digramas correspondientes a los modelos de documentos con los motivos establecidos, se procedió a efectuar el proceso de clasificación empleando para ello los monogramas y digramas, en un primer momento empleando estos n-gramas de forma original y en un segundo tiempo aplicando una reducción dimensional.

Los resultados obtenidos en el caso de los monogramas sin la aplicación de reducción dimensional, se muestran en el siguiente cuadro, en esa tabla se puede observar que el proceso de clasificación es de un grado precisión del 82.76%, lo cual nos indica que no es del todo aplicable para casos que requiera un nivel de confiabilidad superior.

Motivo	1		2		3		4		5		% precisión
	F	%	F	%	F	%	F	%	F	%	
1	48	80.00%	7	11.67%	0	0.00%	1	1.67%	4	6.67%	80.00%
2	5	8.33%	45	75.00%	0	0.00%	7	11.67%	3	5.00%	75.00%
3	0	0.00%	1	1.67%	55	91.67%	3	5.00%	1	1.67%	91.67%
4	3	5.00%	0	0.00%	0	0.00%	51	85.00%	6	10.00%	85.00%
5	4	7.14%	1	1.79%	0	0.00%	5	8.93%	46	82.14%	82.14%
										Promedio	82.76%

Tabla 7: Tabla de clasificación de documentos con monogramas sin aplicar reducción dimensional
Fuente: Elaboración propia

Los resultados obtenidos en el caso de los digramas sin la aplicación de reducción dimensional, se muestran en el cuadro siguiente, en este se presentan resultados nada alentadores para el empleo de digramas como técnica de clasificación, pues solo alcanza un nivel de precisión promedio del orden del 77.50%, lo cual podría juzgarse erróneamente de forma apresurada como una técnica no confiable.

Motivo	1		2		3		4		5		% precisión
	F	%	F	%	F	%	F	%	F	%	
1	45	75.00%	8	13.33%	0	0.00%	1	1.67%	6	10.00%	75.00%
2	8	13.33%	40	66.67%	0	0.00%	3	5.00%	9	15.00%	66.67%
3	0	0.00%	6	10.00%	52	86.67%	2	3.33%	0	0.00%	86.67%
4	5	8.33%	2	3.33%	0	0.00%	43	71.67%	10	16.67%	71.67%
5	3	5.36%	1	1.79%	0	0.00%	3	5.36%	49	87.50%	87.50%
										Promedio	77.50%

Tabla 8: Tabla de clasificación de documentos con digramas sin aplicar reducción dimensional
Fuente: Elaboración propia

Los resultados obtenidos en el caso de los monogramas con la aplicación de reducción dimensional, reduce notablemente la precisión de los monogramas, esto con un nivel de precisión promedio del orden del 70.60%, lo cual se puede apreciar en el cuadro que a continuación se presenta:

Motivo	1		2		3		4		5		% precisión
	F	%	F	%	F	%	F	%	F	%	
1	43	71.67%	9	15.00%	0	0.00%	3	5.00%	5	8.33%	71.67%
2	12	20.00%	37	61.67%	0	0.00%	6	10.00%	5	8.33%	61.67%
3	0	0.00%	1	1.67%	53	88.33%	6	10.00%	0	0.00%	88.33%
4	14	23.33%	2	3.33%	0	0.00%	37	61.67%	7	11.67%	61.67%
5	6	10.71%	4	7.14%	0	0.00%	7	12.50%	39	69.64%	69.64%
										Promedio	70.60%

Tabla 9: Tabla de clasificación de documentos con monogramas aplicando reducción dimensional
Fuente: Elaboración propia

Los resultados obtenidos en el caso de los digramas con la aplicación de reducción dimensional, mejoran notablemente la precisión del proceso de clasificación, alcanzando un promedio del orden del 94.95%, dichos resultados se evidencian en el cuadro que se presenta a continuación.

Motivo	1		2		3		4		5		% precisión
	F	%	F	%	F	%	F	%	F	%	
1	53	88.33%	4	6.67%	0	0.00%	1	1.67%	2	3.33%	88.33%
2	1	1.67%	57	95.00%	0	0.00%	1	1.67%	1	1.67%	95.00%
3	0	0.00%	0	0.00%	59	98.33%	1	1.67%	0	0.00%	98.33%
4	1	1.67%	0	0.00%	0	0.00%	58	96.67%	1	1.67%	96.67%
5	1	1.79%	0	0.00%	0	0.00%	1	1.79%	54	96.43%	96.43%
										Promedio	94.95%

Tabla 10: Tabla de clasificación de documentos con digramas aplicando reducción dimensional
Fuente: Elaboración propia

4.3. Clasificación en el segundo escenario – Corpus estandarizado.

En este segundo escenario se empleó el corpus “corpora Reuters21578-Apte-90Cat” (Moschitti A.). El núcleo de cualquier experimentación de categorización de texto es la precisión final y la posibilidad de compararlo con trabajos anteriores. El corpus Reuters ofrece esta posibilidad, ya que se ha usado en gran medida en el trabajo de la categorización de textos. Las categorías se expresan en diferentes directorios. En cada directorio se almacenan el conjunto de archivos asociados a la categoría de destino (tipo de documento). La división de entrenamiento / prueba es proporcionada por medio de dos directorios principales diferentes (*training* y *test*).

Según David Lewis Reuters es actualmente una de las colecciones de prueba más ampliamente utilizadas para la investigación en categorización de textos. Los datos fueron recogidos inicialmente y etiquetados por Carnegie Group, Inc. y de Reuters, Ltd. en el curso del desarrollo del sistema de categorización e interpretación de textos (Lewis D.).

En este corpus y de acuerdo a las recomendaciones que presenta la fuente, se ha tomado como unidad experimental ocho categorías de documentos los mismos que presentamos a continuación.

Documento Tipo	Cantidad de Archivos	
	Entrenamiento	Prueba
• Acq	1650	719
• Crude	389	189
• Earn	2877	1087
• Grain	433	149
• Interest	347	131
• money-fx	538	179
• trade	369	117
• unknown	1830	280
Total =	8433	2851
	11284	

Tabla 11: Cantidad de documentos por tipo para entrenamiento y clasificación
Fuente: Elaboración propia

Con el corpus obtenido para el proceso de clasificación, se plantearon los siguientes experimentos que de una forma concreta nos permitirá evaluar y sacar las respectivas conclusiones al respecto del proceso de clasificación.

- Clasificación estándar con el método vectorial tradicional.
- Clasificación estándar con el método vectorial tradicional incluyendo el proceso de ganancia de información total y local.
- Clasificación con el método propuesto.
- Clasificación con el método propuesto incluyendo el proceso de ganancia de información total.

En cada uno de los casos presentados anteriormente se practicará con la evaluación de los siguientes n-gramas:

- Monogramas
- Digramas
- Digramas Ordenados Horizontalmente

4.3.1. Clasificación estándar con el método vectorial tradicional

A. Monogramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar sin la conjunción de vocablos, esto quiere decir, empleando monogramas.

Tipo de Documento	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	719	646	13	23	1	7	4	6	19
crude	189	2	175	5	1	1	1	4	0
earn	1087	41	3	1015	0	2	3	10	13
grain	149	0	3	3	125	4	3	11	0
interest	131	2	1	0	1	97	19	5	6
money-fx	179	0	2	1	0	39	124	9	4
trade	117	1	0	4	4	2	3	103	0
unknown	280	24	1	8	1	13	8	11	214

Tabla 12: Matriz de confusión en frecuencias de clasificación estándar aplicando monogramas
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	Crude	earn	grain	interest	money-fx	trade	unknown
acq	89.85%	1.81%	3.20%	0.14%	0.97%	0.56%	0.83%	2.64%
crude	1.06%	92.59%	2.65%	0.53%	0.53%	0.53%	2.12%	0.00%
earn	3.77%	0.28%	93.38%	0.00%	0.18%	0.28%	0.92%	1.20%
grain	0.00%	2.01%	2.01%	83.89%	2.68%	2.01%	7.38%	0.00%
interest	1.53%	0.76%	0.00%	0.76%	74.05%	14.50%	3.82%	4.58%
money-fx	0.00%	1.12%	0.56%	0.00%	21.79%	69.27%	5.03%	2.23%
trade	0.85%	0.00%	3.42%	3.42%	1.71%	2.56%	88.03%	0.00%
unknown	8.57%	0.36%	2.86%	0.36%	4.64%	2.86%	3.93%	76.43%
Promedio	83.44%							

Tabla 13: Matriz de confusión en porcentajes de clasificación estándar aplicando monogramas
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los monogramas empleados en el proceso de clasificación estándar, se puede apreciar que las categorías “Acq”, “Crude”, “Earn”, “Grain” y “Trade”, superaron el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 69.27%.

B. Digramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar conjugando dos vocablos, esto quiere decir, empleando digramas.

Tipo de Documento	Total	acq	crude	earn	grain	interest	money-fx	trade	Unknown
Acq	719	641	15	4	2	6	13	19	19
Crude	189	18	135	0	6	1	2	24	3
Earn	1087	83	3	967	0	2	15	5	12
Grain	149	13	1	0	120	2	0	13	0
Interest	131	3	1	0	0	96	24	3	4
money-fx	179	15	1	0	1	41	106	12	3
Trade	117	6	0	0	3	1	6	96	5
Unknown	280	34	2	0	3	12	15	27	187

Tabla 14: Matriz de confusión en frecuencias de clasificación estándar aplicando digramas
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	89.15%	2.09%	0.56%	0.28%	0.83%	1.81%	2.64%	2.64%
crude	9.52%	71.43%	0.00%	3.17%	0.53%	1.06%	12.70%	1.59%
earn	7.64%	0.28%	88.96%	0.00%	0.18%	1.38%	0.46%	1.10%
grain	8.72%	0.67%	0.00%	80.54%	1.34%	0.00%	8.72%	0.00%
interest	2.29%	0.76%	0.00%	0.00%	73.28%	18.32%	2.29%	3.05%
money-fx	8.38%	0.56%	0.00%	0.56%	22.91%	59.22%	6.70%	1.68%
trade	5.13%	0.00%	0.00%	2.56%	0.85%	5.13%	82.05%	4.27%
unknown	12.14%	0.71%	0.00%	1.07%	4.29%	5.36%	9.64%	66.79%
Promedio	76.43%							

Tabla 15: Matriz de confusión en porcentajes de clasificación estándar aplicando digramas
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas empleados en el proceso de clasificación estándar, se puede apreciar que las categorías “Acq” y “Earn”, superaron el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 59.22%.

C. Digramas Ordenados Horizontalmente

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar conjugando dos vocablos que posteriormente se ordenan alfabéticamente entre sí, esto quiere decir, empleando digramas ordenados horizontalmente.

Tipo de Documento	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	719	642	13	4	2	6	13	19	20
crude	189	18	137	0	6	1	2	24	1
earn	1087	87	3	961	0	3	17	5	11
grain	149	13	1	0	120	1	0	14	0
interest	131	3	1	0	0	96	23	6	2
money-fx	179	15	1	0	2	38	110	11	2
trade	117	6	0	0	3	1	6	96	5
unknown	280	33	1	0	3	12	16	26	189

Tabla 16: Matriz de confusión en frecuencias de clasificación estándar aplicando digramas ordenados horizontalmente

Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	89.29%	1.81%	0.56%	0.28%	0.83%	1.81%	2.64%	2.78%
crude	9.52%	72.49%	0.00%	3.17%	0.53%	1.06%	12.70%	0.53%
earn	8.00%	0.28%	88.41%	0.00%	0.28%	1.56%	0.46%	1.01%
grain	8.72%	0.67%	0.00%	80.54%	0.67%	0.00%	9.40%	0.00%
interest	2.29%	0.76%	0.00%	0.00%	73.28%	17.56%	4.58%	1.53%
money-fx	8.38%	0.56%	0.00%	1.12%	21.23%	61.45%	6.15%	1.12%
trade	5.13%	0.00%	0.00%	2.56%	0.85%	5.13%	82.05%	4.27%
unknown	11.79%	0.36%	0.00%	1.07%	4.29%	5.71%	9.29%	67.50%
Promedio	76.88%							

Tabla 17: Matriz de confusión en porcentajes de clasificación estándar aplicando digramas ordenados horizontalmente

Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas empleados en el proceso de clasificación estándar, se puede apreciar que las categorías “Acq” y “Earn”, superaron el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 61.45%.

4.3.2. Clasificación estándar con el método vectorial tradicional incluyendo el proceso de ganancia de información local

A. Monogramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar sin la conjugación de vocablos, esto quiere decir, empleando monogramas, donde se le complemento con la técnica de ganancia de información aplicado de forma local en cada categoría con un umbral promedio por tipo de documento.

Tipo de Documento	Total	acq	crude	earn	Grain	interest	money-fx	trade	unknown
Acq	719	327	46	20	27	130	20	106	43
Crude	189	4	147	1	5	6	8	17	1
Earn	1087	31	8	974	5	36	10	13	10
Grain	149	0	0	1	129	0	0	19	0
interest	131	27	2	0	1	70	24	1	6
money-fx	179	28	5	3	1	70	66	5	1
Trade	117	6	8	2	15	8	20	53	5
unknown	280	28	16	0	9	66	28	18	115

Tabla 18: Matriz de confusión en frecuencias de clasificación estándar aplicando monogramas con ganancia de información local
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	45.48%	6.40%	2.78%	3.76%	18.08%	2.78%	14.74%	5.98%
crude	2.12%	77.78%	0.53%	2.65%	3.17%	4.23%	8.99%	0.53%
earn	2.85%	0.74%	89.60%	0.46%	3.31%	0.92%	1.20%	0.92%
grain	0.00%	0.00%	0.67%	86.58%	0.00%	0.00%	12.75%	0.00%
interest	20.61%	1.53%	0.00%	0.76%	53.44%	18.32%	0.76%	4.58%
money-fx	15.64%	2.79%	1.68%	0.56%	39.11%	36.87%	2.79%	0.56%
trade	5.13%	6.84%	1.71%	12.82%	6.84%	17.09%	45.30%	4.27%
unknown	10.00%	5.71%	0.00%	3.21%	23.57%	10.00%	6.43%	41.07%
Promedio	59.51%							

Tabla 19: Matriz de confusión en porcentajes de clasificación estándar aplicando monogramas con ganancia de información local
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los monogramas empleados en el proceso de clasificación estándar con ganancia de información local, se puede apreciar que las categorías “Earn” y “Grain” superaron el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 36.87%.

B. Digramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar conjugando dos vocablos, esto quiere decir, empleando digramas, donde se le complementó con la técnica de ganancia de información aplicado de forma local en cada categoría con un umbral promedio por tipo de documento.

Tipo de Documento	Total	acq	crude	earn	Grain	interest	money-fx	trade	unknown
Acq	719	599	24	17	5	18	14	13	29
Crude	189	4	179	0	1	0	1	3	1
Earn	1087	26	19	979	5	13	13	11	21
Grain	149	0	1	0	133	0	2	13	0
Interest	131	1	2	1	1	99	15	8	4
money-fx	179	7	3	0	2	55	93	13	6
Trade	117	0	2	0	4	3	6	99	3
Unknown	280	22	4	1	9	31	18	21	174

Tabla 20: Matriz de confusión en frecuencias de clasificación estándar aplicando digramas con ganancia de información local
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	83.31%	3.34%	2.36%	0.70%	2.50%	1.95%	1.81%	4.03%
crude	2.12%	94.71%	0.00%	0.53%	0.00%	0.53%	1.59%	0.53%
earn	2.39%	1.75%	90.06%	0.46%	1.20%	1.20%	1.01%	1.93%
grain	0.00%	0.67%	0.00%	89.26%	0.00%	1.34%	8.72%	0.00%
interest	0.76%	1.53%	0.76%	0.76%	75.57%	11.45%	6.11%	3.05%
money-fx	3.91%	1.68%	0.00%	1.12%	30.73%	51.96%	7.26%	3.35%
trade	0.00%	1.71%	0.00%	3.42%	2.56%	5.13%	84.62%	2.56%
unknown	7.86%	1.43%	0.36%	3.21%	11.07%	6.43%	7.50%	62.14%
Promedio	78.95%							

Tabla 21: Matriz de confusión en porcentajes de clasificación estándar aplicando digramas con ganancia de información local
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas empleados en el proceso de clasificación estándar con ganancia de información local, se puede apreciar que las categorías “Acq”, “Crude”, “Earn”, “Grain” y “Trade” superaron el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 51.96%.

C. Digramas Ordenados Horizontalmente

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar conjugando dos vocablos ordenados alfabéticamente, esto quiere decir, empleando digramas ordenados horizontalmente, donde se le complementó con la técnica de ganancia de información aplicada de forma local en cada categoría con un umbral promedio por tipo de documento.

Tipo de Documento	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
Acq	719	610	23	14	4	12	8	17	31
Crude	189	7	178	0	1	1	1	1	0
Earn	1087	51	12	960	2	16	15	11	20
Grain	149	1	0	0	135	0	0	13	0
Interest	131	1	1	1	1	97	17	7	6
money-fx	179	6	3	0	2	50	101	14	3
Trade	117	0	2	0	6	3	6	99	1
unknown	280	23	11	1	7	35	10	19	174

Tabla 22: Matriz de confusión en frecuencias de clasificación estándar aplicando digramas ordenados horizontalmente con ganancia de información local
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	84.84%	3.20%	1.95%	0.56%	1.67%	1.11%	2.36%	4.31%
crude	3.70%	94.18%	0.00%	0.53%	0.53%	0.53%	0.53%	0.00%
earn	4.69%	1.10%	88.32%	0.18%	1.47%	1.38%	1.01%	1.84%
grain	0.67%	0.00%	0.00%	90.60%	0.00%	0.00%	8.72%	0.00%
interest	0.76%	0.76%	0.76%	0.76%	74.05%	12.98%	5.34%	4.58%
money-fx	3.35%	1.68%	0.00%	1.12%	27.93%	56.42%	7.82%	1.68%
trade	0.00%	1.71%	0.00%	5.13%	2.56%	5.13%	84.62%	0.85%
unknown	8.21%	3.93%	0.36%	2.50%	12.50%	3.57%	6.79%	62.14%
Promedio	79.40%							

Tabla 23: Matriz de confusión en porcentajes de clasificación estándar aplicando digramas ordenados horizontalmente con ganancia de información local
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas ordenados horizontalmente empleados en el proceso de clasificación estándar con ganancia de información local, se puede apreciar que las categorías “Acq”, “Crude”, “Earn”, “Grain” y “Trade” superaron el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 56.42%.

4.3.3. Clasificación estándar con el método vectorial tradicional incluyendo el proceso de ganancia de información total

A. Monogramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar sin la conjugación de vocablos, esto quiere decir, empleando monogramas, donde se le complementó con la técnica de ganancia de información aplicada de forma total en todas las categorías con un umbral promedio general.

Tipo de Documento	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
Acq	719	442	45	27	17	11	31	37	109
crude	189	17	153	1	4	1	8	3	2
Earn	1087	72	13	951	6	4	5	4	32
Grain	149	35	3	0	95	0	2	10	4
interest	131	70	8	1	2	27	7	10	6
money-fx	179	102	8	0	1	17	27	18	6
trade	117	25	9	0	14	2	14	43	10
unknown	280	74	21	4	19	17	12	13	120

Tabla 24: Matriz de confusión en frecuencias de clasificación estándar aplicando monogramas con ganancia de información total
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	61.47%	6.26%	3.76%	2.36%	1.53%	4.31%	5.15%	15.16%
crude	8.99%	80.95%	0.53%	2.12%	0.53%	4.23%	1.59%	1.06%
earn	6.62%	1.20%	87.49%	0.55%	0.37%	0.46%	0.37%	2.94%
grain	23.49%	2.01%	0.00%	63.76%	0.00%	1.34%	6.71%	2.68%
interest	53.44%	6.11%	0.76%	1.53%	20.61%	5.34%	7.63%	4.58%
money-fx	56.98%	4.47%	0.00%	0.56%	9.50%	15.08%	10.06%	3.35%
Trade	21.37%	7.69%	0.00%	11.97%	1.71%	11.97%	36.75%	8.55%
unknown	26.43%	7.50%	1.43%	6.79%	6.07%	4.29%	4.64%	42.86%
Promedio	51.12%							

Tabla 25: Matriz de confusión en porcentajes de clasificación estándar aplicando monogramas con ganancia de información total
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los monogramas empleados en el proceso de clasificación estándar con ganancia de información total, se puede apreciar que la categoría “Earn” superó el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 15.08%.

B. Digramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar con la conjugación de dos vocablos, esto quiere decir, empleando digramas, donde se

le complementó con la técnica de ganancia de información aplicada de forma total en todas las categorías con un umbral promedio general.

Tipo de Documento	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
Acq	719	652	12	10	7	4	4	8	22
crude	189	4	175	0	3	1	1	5	0
Earn	1087	53	4	1002	1	2	4	7	14
Grain	149	1	1	0	140	1	0	6	0
interest	131	2	2	1	2	90	23	9	2
money-fx	179	7	2	0	3	48	99	14	6
trade	117	1	2	0	11	2	6	92	3
unknown	280	36	4	0	11	24	8	18	179

Tabla 26: Matriz de confusión en frecuencias de clasificación estándar aplicando digramas con ganancia de información total
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	90.68%	1.67%	1.39%	0.97%	0.56%	0.56%	1.11%	3.06%
crude	2.12%	92.59%	0.00%	1.59%	0.53%	0.53%	2.65%	0.00%
earn	4.88%	0.37%	92.18%	0.09%	0.18%	0.37%	0.64%	1.29%
grain	0.67%	0.67%	0.00%	93.96%	0.67%	0.00%	4.03%	0.00%
interest	1.53%	1.53%	0.76%	1.53%	68.70%	17.56%	6.87%	1.53%
money-fx	3.91%	1.12%	0.00%	1.68%	26.82%	55.31%	7.82%	3.35%
trade	0.85%	1.71%	0.00%	9.40%	1.71%	5.13%	78.63%	2.56%
unknown	12.86%	1.43%	0.00%	3.93%	8.57%	2.86%	6.43%	63.93%
Promedio	79.50%							

Tabla 27: Matriz de confusión en porcentajes de clasificación estándar aplicando digramas con ganancia de información total
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas empleados en el proceso de clasificación estándar con ganancia de información total, se puede apreciar que las categorías “Acq”, “Crude”, “Earn” y “Grain” superó el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 55.31%.

C. Digramas Ordenados Horizontalmente

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación estándar con la conjugación de dos vocablos que se ordenaron alfabéticamente entre sí, esto quiere decir, empleando digramas ordenados horizontalmente, donde se le complementó con la técnica de ganancia de información aplicada de forma total en todas las categoría con un umbral promedio general.

Tipo de Documento	Total	acq	crude	Earn	grain	interest	money-fx	trade	Unknown
Acq	719	650	11	11	6	5	5	11	20
Crude	189	7	171	0	3	1	1	6	0
Earn	1087	60	8	987	1	7	4	6	14
Grain	149	0	2	0	139	0	2	6	0
interest	131	2	1	1	2	92	21	7	5
money-fx	179	7	1	0	3	49	103	14	2
Trade	117	0	3	0	9	1	5	95	4
unknown	280	28	4	1	11	25	8	20	183

Tabla 28: Matriz de confusión en frecuencias de clasificación estándar aplicando digramas ordenados horizontalmente con ganancia de información total
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	90.40%	1.53%	1.53%	0.83%	0.70%	0.70%	1.53%	2.78%
crude	3.70%	90.48%	0.00%	1.59%	0.53%	0.53%	3.17%	0.00%
earn	5.52%	0.74%	90.80%	0.09%	0.64%	0.37%	0.55%	1.29%
grain	0.00%	1.34%	0.00%	93.29%	0.00%	1.34%	4.03%	0.00%
interest	1.53%	0.76%	0.76%	1.53%	70.23%	16.03%	5.34%	3.82%
money-fx	3.91%	0.56%	0.00%	1.68%	27.37%	57.54%	7.82%	1.12%
trade	0.00%	2.56%	0.00%	7.69%	0.85%	4.27%	81.20%	3.42%
unknown	10.00%	1.43%	0.36%	3.93%	8.93%	2.86%	7.14%	65.36%
Promedio	79.91%							

Tabla 29: Matriz de confusión en porcentajes de clasificación estándar aplicando digramas ordenados horizontalmente con ganancia de información total
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas ordenados horizontalmente empleados en el proceso de clasificación estándar con ganancia de información total, se puede apreciar que las categorías “Acq”, “Crude”, “Earn” y “Grain” superó el 83.44% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 57.54%.

4.3.4. Resumen de la clasificación estándar

A continuación se presenta una tabla que resume las frecuencias de los archivos de prueba clasificados empleando el método estándar de clasificación, a lo que se le agregó la técnica de ganancia de información en un ámbito local por categorías con un umbral promedio por tipo de documento y total por todas las categorías con un umbral promedio general.

Tipo de Documento	Total de Archivos de Prueba	Clasificación Estándar			Clasificación con Ganancia de Información					
					Local			Total		
		Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
Acq	719	646	641	642	327	599	610	442	652	650
Crude	189	175	135	137	147	179	178	153	175	171
Earn	1087	1015	967	961	974	979	960	951	1002	987
Grain	149	125	120	120	129	133	135	95	140	139
interest	131	97	96	96	70	99	97	27	90	92
money-fx	179	124	106	110	66	93	101	27	99	103
Trade	117	103	96	96	53	99	99	43	92	95
unknown	280	214	187	189	115	174	174	120	179	183

Tabla 30: Resumen en frecuencias de clasificación estándar
Fuente: Elaboración propia

De la tabla presentada anteriormente, se puede calcular los porcentajes de categorización por tipos definidos, los mismos que se muestran en la tabla siguiente:

Tipo de Documento	Clasificación Estándar			Clasificación con Ganancia de Información					
				Local			Total		
	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
Acq	89.85%	89.15%	89.29%	45.48%	83.31%	84.84%	61.47%	90.68%	90.40%
Crude	92.59%	71.43%	72.49%	77.78%	94.71%	94.18%	80.95%	92.59%	90.48%
Earn	93.38%	88.96%	88.41%	89.60%	90.06%	88.32%	87.49%	92.18%	90.80%
Grain	83.89%	80.54%	80.54%	86.58%	89.26%	90.60%	63.76%	93.96%	93.29%
Interest	74.05%	73.28%	73.28%	53.44%	75.57%	74.05%	20.61%	68.70%	70.23%
money-fx	69.27%	59.22%	61.45%	36.87%	51.96%	56.42%	15.08%	55.31%	57.54%
Trade	88.03%	82.05%	82.05%	45.30%	84.62%	84.62%	36.75%	78.63%	81.20%
Unknown	76.43%	66.79%	67.50%	41.07%	62.14%	62.14%	42.86%	63.93%	65.36%
Promedio	83.44%	76.43%	76.88%	59.51%	78.95%	79.40%	51.12%	79.50%	79.91%

Tabla 31: Resumen en porcentajes de clasificación estándar
Fuente: Elaboración propia

De la tabla anterior se puede apreciar que la clasificación que ofrece en promedio la mayor precisión, es la que se efectúa de forma estándar sin la conjunción de vocablos, dicho de otra forma, empleando monogramas, la misma que obtuvo un nivel de precisión del orden del 83.44%.

4.3.5. Clasificación con el método propuesto

A. Monogramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación propuesta sin la conjugación de vocablos, esto quiere decir, empleando monogramas.

Documento Tipo	Total	acq	crude	earn	grain	interest	money-fx	trade	Unknown
acq	719	661	15	4	1	6	6	8	18
crude	189	5	175	2	1	1	1	3	1
earn	1087	45	4	995	0	3	4	17	19
grain	149	0	4	0	127	3	3	11	1
interest	131	0	1	0	0	101	19	5	5
money-fx	179	3	1	0	0	40	125	7	3
trade	117	1	2	0	4	2	3	105	0
unknown	280	26	1	2	1	14	13	12	211

Tabla 32: Matriz de confusión en frecuencias de clasificación propuesta aplicando monogramas
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	91.93%	2.09%	0.56%	0.14%	0.83%	0.83%	1.11%	2.50%
crude	2.65%	92.59%	1.06%	0.53%	0.53%	0.53%	1.59%	0.53%
earn	4.14%	0.37%	91.54%	0.00%	0.28%	0.37%	1.56%	1.75%
grain	0.00%	2.68%	0.00%	85.23%	2.01%	2.01%	7.38%	0.67%
interest	0.00%	0.76%	0.00%	0.00%	77.10%	14.50%	3.82%	3.82%
money-fx	1.68%	0.56%	0.00%	0.00%	22.35%	69.83%	3.91%	1.68%
trade	0.85%	1.71%	0.00%	3.42%	1.71%	2.56%	89.74%	0.00%
unknown	9.29%	0.36%	0.71%	0.36%	5.00%	4.64%	4.29%	75.36%
Promedio	84.17%							

Tabla 33: Matriz de confusión en porcentajes de clasificación propuesta aplicando monogramas
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los monogramas empleados en el proceso de clasificación propuesto, se puede apreciar que las categorías “Acq”, “Crude”, “Earn”, “Grain” y “Trade” superaron el 84.17% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 69.83%.

B. Digramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación propuesta con la conjugación de vocablos, esto quiere decir, empleando digramas.

Documento Tipo	Total	acq	crude	Earn	grain	interest	money-fx	trade	unknown
Acq	719	652	15	12	1	6	5	11	17
crude	189	14	145	1	4	1	2	22	0
earn	1087	67	10	957	0	7	1	33	12
grain	149	13	1	0	119	3	1	12	0
interest	131	6	0	0	0	91	30	3	1
money-fx	179	23	0	0	1	38	108	8	1
trade	117	2	0	0	2	1	8	104	0
unknown	280	26	7	1	3	20	15	15	193

Tabla 34: Matriz de confusión en frecuencias de clasificación propuesta aplicando digramas
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	90.68%	2.09%	1.67%	0.14%	0.83%	0.70%	1.53%	2.36%
crude	7.41%	76.72%	0.53%	2.12%	0.53%	1.06%	11.64%	0.00%
earn	6.16%	0.92%	88.04%	0.00%	0.64%	0.09%	3.04%	1.10%
grain	8.72%	0.67%	0.00%	79.87%	2.01%	0.67%	8.05%	0.00%
interest	4.58%	0.00%	0.00%	0.00%	69.47%	22.90%	2.29%	0.76%
money-fx	12.85%	0.00%	0.00%	0.56%	21.23%	60.34%	4.47%	0.56%
trade	1.71%	0.00%	0.00%	1.71%	0.85%	6.84%	88.89%	0.00%
unknown	9.29%	2.50%	0.36%	1.07%	7.14%	5.36%	5.36%	68.93%
Promedio	77.87%							

Tabla 35: Matriz de confusión en porcentajes de clasificación propuesta aplicando digramas
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas empleados en el proceso de clasificación propuesto, se puede apreciar que las categorías “Acq”, “Earn”, y “Trade” superaron el 84.17% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 60.34%.

C. Digramas Ordenados Horizontalmente

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación propuesta con la conjugación de vocablos ordenados alfabéticamente entre si, esto quiere decir, empleando digramas ordenados horizontalmente.

Documento Tipo	Total	acq	crude	earn	grain	interest	money-fx	trade	Unknown
acq	719	659	11	13	1	6	5	10	14
crude	189	14	144	1	5	1	2	22	0
earn	1087	69	10	961	0	5	3	29	10
grain	149	10	1	0	125	1	1	11	0
interest	131	6	0	0	0	94	27	4	0
money-fx	179	21	0	0	1	41	109	7	0
trade	117	2	0	0	3	1	8	103	0
unknown	280	29	7	0	3	20	14	18	189

Tabla 36: Matriz de confusión en frecuencias de clasificación propuesta aplicando digramas ordenados horizontalmente
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	91.66%	1.53%	1.81%	0.14%	0.83%	0.70%	1.39%	1.95%
crude	7.41%	76.19%	0.53%	2.65%	0.53%	1.06%	11.64%	0.00%
earn	6.35%	0.92%	88.41%	0.00%	0.46%	0.28%	2.67%	0.92%
grain	6.71%	0.67%	0.00%	83.89%	0.67%	0.67%	7.38%	0.00%
interest	4.58%	0.00%	0.00%	0.00%	71.76%	20.61%	3.05%	0.00%
money-fx	11.73%	0.00%	0.00%	0.56%	22.91%	60.89%	3.91%	0.00%
trade	1.71%	0.00%	0.00%	2.56%	0.85%	6.84%	88.03%	0.00%
unknown	10.36%	2.50%	0.00%	1.07%	7.14%	5.00%	6.43%	67.50%
Promedio	78.54%							

Tabla 37: Matriz de confusión en porcentajes de clasificación propuesta aplicando digramas ordenados horizontalmente
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas ordenados horizontalmente empleados en el proceso de clasificación propuesto, se puede apreciar que las categorías “Acq”, “Earn”, y “Trade” superaron el 84.17% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 60.89%.

4.3.6. Clasificación con el método propuesto incluyendo el proceso de ganancia de información total

A. Monogramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación propuesta sin la conjugación de vocablos, esto quiere decir, empleando monogramas, la misma que se complementó con la técnica de ganancia de información aplicada a todas las categorías de forma total con un umbral promedio general.

Documento Tipo	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	719	487	39	30	29	30	33	25	46
crude	189	15	152	2	3	3	8	4	2
earn	1087	339	14	577	101	16	8	9	23
grain	149	9	1	1	119	1	6	12	0
interest	131	75	4	1	6	22	9	11	3
money-fx	179	75	5	0	14	50	21	12	2
trade	117	22	10	2	17	7	17	35	7
unknown	280	102	10	1	11	25	19	10	102

Tabla 38: Matriz de confusión en frecuencias de clasificación propuesta aplicando monogramas empleando ganancia de información total
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	67.73%	5.42%	4.17%	4.03%	4.17%	4.59%	3.48%	6.40%
crude	7.94%	80.42%	1.06%	1.59%	1.59%	4.23%	2.12%	1.06%
earn	31.19%	1.29%	53.08%	9.29%	1.47%	0.74%	0.83%	2.12%
grain	6.04%	0.67%	0.67%	79.87%	0.67%	4.03%	8.05%	0.00%
interest	57.25%	3.05%	0.76%	4.58%	16.79%	6.87%	8.40%	2.29%
money-fx	41.90%	2.79%	0.00%	7.82%	27.93%	11.73%	6.70%	1.12%
trade	18.80%	8.55%	1.71%	14.53%	5.98%	14.53%	29.91%	5.98%
unknown	36.43%	3.57%	0.36%	3.93%	8.93%	6.79%	3.57%	36.43%
Promedio	47.00%							

Tabla 39: Matriz de confusión en porcentajes de clasificación propuesta aplicando monogramas empleando ganancia de información total
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los monogramas empleados en el proceso de clasificación propuesto con ganancia de información total, se puede apreciar que ninguna categoría superó el 84.17% de precisión, siendo la categoría “Money-fx” la que obtuvo la menor precisión con un 11.73%.

B. Digramas

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación propuesta con la

conjugación de dos vocablos, esto quiere decir, empleando digramas, la misma que se complementó con la técnica de ganancia de información aplicada a todas las categorías de forma total con un umbral promedio general.

Documento Tipo	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
Acq	719	652	13	8	6	3	6	11	20
Crude	189	3	180	0	3	0	2	0	1
Earn	1087	71	11	965	1	5	9	4	21
Grain	149	0	1	0	135	1	1	10	1
Interest	131	25	3	3	1	57	24	7	11
money-fx	179	28	3	2	1	27	94	15	9
Trade	117	2	6	0	7	4	8	89	1
Unknown	280	32	4	1	3	11	17	15	197

Tabla 40: Matriz de confusión en frecuencias de clasificación propuesta aplicando digramas empleando ganancia de información total
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	90.68%	1.81%	1.11%	0.83%	0.42%	0.83%	1.53%	2.78%
crude	1.59%	95.24%	0.00%	1.59%	0.00%	1.06%	0.00%	0.53%
earn	6.53%	1.01%	88.78%	0.09%	0.46%	0.83%	0.37%	1.93%
grain	0.00%	0.67%	0.00%	90.60%	0.67%	0.67%	6.71%	0.67%
interest	19.08%	2.29%	2.29%	0.76%	43.51%	18.32%	5.34%	8.40%
money-fx	15.64%	1.68%	1.12%	0.56%	15.08%	52.51%	8.38%	5.03%
trade	1.71%	5.13%	0.00%	5.98%	3.42%	6.84%	76.07%	0.85%
unknown	11.43%	1.43%	0.36%	1.07%	3.93%	6.07%	5.36%	70.36%
Promedio	75.97%							

Tabla 41: Matriz de confusión en porcentajes de clasificación propuesta aplicando digramas empleando ganancia de información total
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas empleados en el proceso de clasificación propuesto con ganancia de información total, se puede apreciar que las categorías “Acq”, “Crude”, Earn” y “Grain” superaron el 84.17% de precisión, siendo la categoría “Interest” la que obtuvo la menor precisión con un 43.51%.

C. Digramas Ordenados Horizontalmente

A continuación se presenta una tabla que expresa los resultados de la clasificación de documentos con las ocho categorías del corpus estandarizado, expresados en número de archivos clasificados por cada tipo, donde dicha clasificación se efectuó empleando la técnica de clasificación propuesta con la conjugación de dos vocablos que posteriormente se ordenaron alfabéticamente entre sí, esto quiere decir, empleando digramas ordenados horizontalmente, la misma que se complementó con la técnica de ganancia de información aplicada a todas las categorías de forma total con un umbral promedio general.

Documento Tipo	Total	acq	crude	earn	grain	interest	money-fx	trade	unknown
Acq	719	656	11	4	2	2	7	11	26
Crude	189	6	178	0	1	1	1	1	1
Earn	1087	72	9	966	1	5	9	5	20
Grain	149	2	1	0	135	0	1	10	0
Interest	131	25	4	3	2	56	23	8	10
money-fx	179	29	5	2	2	36	86	9	10
Trade	117	3	6	0	8	5	7	84	4
unknown	280	33	8	1	3	13	15	6	201

Tabla 42: Matriz de confusión en frecuencias de clasificación propuesta aplicando digramas ordenados horizontalmente empleando ganancia de información total
Fuente: Elaboración propia

De la tabla presentada anteriormente, se deriva la matriz de confusión expresada en cantidades porcentuales de la comparación de los archivos de prueba con las ocho categorías, la misma que se muestra a continuación.

Tipo de Documento	acq	crude	earn	grain	interest	money-fx	trade	unknown
acq	91.24%	1.53%	0.56%	0.28%	0.28%	0.97%	1.53%	3.62%
crude	3.17%	94.18%	0.00%	0.53%	0.53%	0.53%	0.53%	0.53%
earn	6.62%	0.83%	88.87%	0.09%	0.46%	0.83%	0.46%	1.84%
grain	1.34%	0.67%	0.00%	90.60%	0.00%	0.67%	6.71%	0.00%
interest	19.08%	3.05%	2.29%	1.53%	42.75%	17.56%	6.11%	7.63%
money-fx	16.20%	2.79%	1.12%	1.12%	20.11%	48.04%	5.03%	5.59%
trade	2.56%	5.13%	0.00%	6.84%	4.27%	5.98%	71.79%	3.42%
unknown	11.79%	2.86%	0.36%	1.07%	4.64%	5.36%	2.14%	71.79%
Promedio	74.91%							

Tabla 43: Matriz de confusión en porcentajes de clasificación propuesta aplicando digramas ordenados horizontalmente empleando ganancia de información total
Fuente: Elaboración propia

De la anterior tabla que contiene la matriz de confusión porcentual de los digramas ordenados horizontalmente empleados en el proceso de clasificación propuesto con ganancia de información total, se puede apreciar que las categorías “Acq”, “Crude”, Earn” y “Grain” superaron el 84.17% de precisión, siendo la categoría “Interest” la que obtuvo la menor precisión con un 42.75%.

4.3.7. Resumen de la clasificación con el método propuesto

A continuación se presenta una tabla que resume las frecuencias de los archivos de prueba clasificados empleando el método propuesto de clasificación, a lo que se le agregó la técnica de ganancia de información en un ámbito total por todas las categorías con un umbral promedio.

Tipo de documento	Archivos procesados	Clasificación con la propuesta			Clasificación con la propuesta y ganancia de Información		
		Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
Acq	719	661	652	659	487	652	656
crude	189	175	145	144	152	180	178
Earn	1087	995	957	961	577	965	966
Grain	149	127	119	125	119	135	135
interest	131	101	91	94	22	57	56
money-fx	179	125	108	109	21	94	86
trade	117	105	104	103	35	89	84
unknown	280	211	193	189	102	197	201

Tabla 44: Resumen en frecuencias de clasificación propuesta
Fuente: Elaboración propia

De la tabla presentada anteriormente, se puede calcular los porcentajes de categorización por tipos definidos, los mismos que se muestran en la tabla siguiente:

Tipo de documento	Clasificación con la propuesta			Clasificación con la propuesta y ganancia de Información		
	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
acq	91.93%	90.68%	91.66%	67.73%	90.68%	91.24%
crude	92.59%	76.72%	76.19%	80.42%	95.24%	94.18%
earn	91.54%	88.04%	88.41%	53.08%	88.78%	88.87%
grain	85.23%	79.87%	83.89%	79.87%	90.60%	90.60%
interest	77.10%	69.47%	71.76%	16.79%	43.51%	42.75%
money-fx	69.83%	60.34%	60.89%	11.73%	52.51%	48.04%
trade	89.74%	88.89%	88.03%	29.91%	76.07%	71.79%
unknown	75.36%	68.93%	67.50%	36.43%	70.36%	71.79%
Promedio	84.17%	77.87%	78.54%	47.00%	75.97%	74.91%

Tabla 45: Resumen en porcentajes de clasificación propuesta
Fuente: Elaboración propia

De la tabla anterior se puede apreciar que la clasificación empleando la propuesta que ofrece en promedio la mayor precisión, es la que se efectúa sin la conjunción de vocablos, dicho de otra forma, empleando monogramas, la misma que obtuvo un nivel de precisión del orden del 84.17%.

4.4. Evaluación de recursos textuales empleados

Como se pudo apreciar en el punto 4.3. El corpus empleado cuenta con un total de 11284 archivos correspondientes a un total de 15437 archivos, lo quiere decir que se está empleando un total de 73.10% del cuerpo total de archivos propuestos por Reuters. De una forma más analítica se puede ver en el cuadro siguiente:

Categorías	Numero de Archivos Entrenamiento		Numero de Archivos Prueba	
	Cantidad	%	Cantidad	%
Empleadas	8433	73.89%	2851	70.85%
No Empleadas	2980	26.11%	1173	29.15%
Total	11413		4024	

Tabla 46: Constitución de archivos del Corpus Reuters 21578-90Cat por segmentos de entrenamiento y prueba
Fuente: Elaboración propia

Del cuadro precedente podemos ver que de los 11413 archivos de entrenamiento existentes en las 91 categorías, se está empleando 8433, lo que constituye una representatividad del orden del 73.89% de los archivos de entrenamiento. Del mismo modo se está empleando 2851 archivos de prueba de un total de 4024 archivos, lo que constituye un 70.85% de los archivos de prueba.

Nº	Tipo de Documento	Cantidad de Archivos		Nº	Tipo de Documento	Cantidad de Archivos	
		Entrenamiento	Prueba			Entrenamiento	Prueba
1	acq	1650	719	47	money-fx	538	179
2	alum	35	23	48	money-supply	138	34
3	barley	37	14	49	naphtha	2	4
4	bop	75	30	50	nat-gas	75	30
5	carcass	50	18	51	nickel	8	1
6	castor-oil	1	1	52	nkr	1	2
7	cocoa	55	18	53	nzdlr	2	2
8	coconut	4	2	54	oat	8	6
9	coconut-oil	4	3	55	oilseed	124	47
10	coffee	111	28	56	orange	16	11
11	copper	47	18	57	palladium	2	1
12	copra-cake	2	1	58	palm-oil	30	10
13	corn	181	56	59	palmkernel	2	1
14	cotton	39	20	60	pet-chem	20	12
15	cotton-oil	1	2	61	platinum	5	7
16	cpi	69	28	62	potato	3	3
17	cpu	3	1	63	propane	3	3
18	crude	389	189	64	rand	2	1
19	dfl	2	1	65	rape-oil	5	3
20	dlr	131	44	66	rapeseed	18	9
21	dmk	10	4	67	reserves	55	18
22	earn	2877	1087	68	retail	23	2
23	fuel	13	10	69	rice	35	24
24	gas	37	17	70	rubber	37	12
25	gnp	101	35	71	rye	1	1
26	gold	94	30	72	ship	197	89
27	grain	433	149	73	silver	21	8
28	groundnut	5	4	74	sorghum	24	10
29	groundnut-oil	1	1	75	soy-meal	13	13
30	heat	14	5	76	soy-oil	14	11
31	hog	16	6	77	soybean	78	33
32	housing	16	4	78	strategic-metal	16	11
33	income	9	7	79	sugar	126	36
34	instal-debt	5	1	80	sun-meal	1	1
35	interest	347	131	81	sun-oil	5	2
36	ipi	41	12	82	sunseed	11	5
37	iron-steel	40	14	83	tea	9	4
38	jet	4	1	84	tin	18	12
39	jobs	46	21	85	trade	369	117
40	l-cattle	6	2	86	unknown	1830	280
41	lead	15	14	87	veg-oil	87	37
42	lei	12	3	88	wheat	212	71
43	lin-oil	1	1	89	wpi	19	10
44	livestock	75	24	90	yen	45	14
45	lumber	10	6	91	zinc	21	13
46	meal-feed	30	19	Total =		11413	4024

Tabla 47: Constitución de archivos del Corpus Reuters 21578-90Cat en general
Fuente: Elaboración propia

Posteriormente a la selección de las ocho categorías con las que se evaluara la propuesta, se procede a efectuar los experimentos que emplearon el siguiente volumen de recurso textual.

Documento Tipo	Nº Archivos	Cantidad de Vocablos			% Reducción de vocablos
		Nº Palabras	Depurado	Verbos y Sustantivos	
Acq	2369	307193	157354	128698	81.79%
Crude	578	125529	63135	54091	85.68%
Earn	3964	386684	174747	129847	74.31%
Grain	582	119627	52618	46360	88.11%
Interest	478	82608	40652	36387	89.51%
money-fx	717	139053	68274	60324	88.36%
Trade	486	125110	61491	54174	88.10%
unknown	2110	316280	158747	136247	85.83%
Total =	11284	1602084	777018	646128	83.15%

Tabla 48: Constitución de archivos del Corpus Reuters 21578-90Cat por categorías seleccionadas y cantidad de vocablos
Fuente: Elaboración propia

Del cuadro precedente podemos observar que de un total de 1 602 084 vocablos se emplean de forma regular 777 018 vocablos con el método estándar. Cuando se emplea la propuesta, luego de reducir el cuerpo de los documentos a verbos y sustantivos; se emplean 646 128 vocablos lo que constituye una proporción de 83.15%.

Documento Tipo	Sin Propuesta			Con Propuesta		
	Monogramas	Digramas	Digramas O.H.	Monogramas	Digramas	Digramas O.H.
acq	8285	64764	60911	4357	50462	46278
crude	4220	27761	26395	2960	23702	22253
earn	7824	50154	47051	3843	38182	34956
grain	3635	25616	24059	2710	22312	20660
interest	2566	15865	15040	1958	13855	12976
money-fx	3589	27201	25680	2652	23639	22038
trade	3776	27837	26399	2838	24258	22728
unknown	7602	68809	64198	4508	57307	52329
Total =	41497	308007	289733	25826	253717	234218

Tabla 49: Constitución de archivos del Corpus Reuters 21578-90Cat por vocablos procesados con y sin propuesta
Fuente: Elaboración propia

En el cuadro precedente se puede observar la cantidad de términos empleados en la constitución de los monogramas, digramas y digramas ordenados horizontalmente, luego de haber efectuado las depuraciones de los documentos, así como también el caso de la propuesta con la extracción de los verbos y sustantivos para constituir los n-grama citados.

Documento Tipo	Con Propuesta		
	Monogramas	Digramas	Digramas O.H.
acq	52.59%	77.92%	75.98%
crude	70.14%	85.38%	84.31%
earn	49.12%	76.13%	74.29%
grain	74.55%	87.10%	85.87%
interest	76.31%	87.33%	86.28%
money-fx	73.89%	86.90%	85.82%
trade	75.16%	87.14%	86.09%
unknown	59.30%	83.28%	81.51%
Total =	62.24%	82.37%	80.84%

Tabla 50: Constitución de archivos del Corpus Reuters 21578-90Cat en proporción a la reducción de términos empleados por la propuesta
Fuente: Elaboración propia

Del cuadro anterior podemos observar que con el empleo de la propuesta se reducen al 62.24 % de la constitución de los términos en la elaboración de los monogramas, un 82.37% de términos en la elaboración de los digramas y finalmente 80.84 % de términos en la elaboración de los digramas ordenados horizontalmente.

En forma general podemos observar que con la extracción de verbos y sustantivos, no solo se reducen la cantidad de términos en la constitución de los n-gramas, sino también se caracterizan mejor los patrones de clasificación de los tipos de documento.

Conclusiones

Primero. La separación de los vocablos constituidos por verbos y sustantivos, logró mejorar la calidad de los términos empleados en el proceso de clasificación de documentos. El uso de estos términos permitió por medio de monogramas, mejorar la precisión de las tareas de clasificación de documentos en el marco del procesamiento de lenguaje natural.

Segundo. Se determinó que los verbos y sustantivos caracterizan efectivamente el patrón de un texto para procesos de clasificación de documentos, lo cual permite tener mejores resultados en el proceso de clasificación de éstos.

Tercero. Al establecer una mecánica para el proceso de clasificación de documentos, y acotando elementos propios de la propuesta, se logró superar la efectividad del proceso de clasificación en el orden del 84.17% de precisión promedio.

Cuarto. La exclusión de términos existentes en los n-gramas de los documentos a clasificar que no estén en los n-gramas del patrón de los tipos de documentos mejoró los resultados de la clasificación.

Recomendaciones

Primero. Se puede apreciar en el proceso de clasificación con un corpus estandarizado; que la ganancia de información puede disminuir el grado de precisión lo que contradictoriamente puede verse en el caso del primer escenario, lo que nos permite recomendar evaluar los diferentes métodos y sus agregados en los grupos de entrenamiento, para con ello en el caso de una aplicación real, efectuar una clasificación con un método combinado; Esto puede resultar en suma una clasificación más efectiva.

Segundo. Se puede observar que el ordenamiento horizontal incrementa la efectividad de los digramas, aunque en el caso del escenario estandarizado, no se pueda superar la efectividad de los monogramas. Se recomienda profundizar el estudio incrementando los experimentos a trigramas o superiores niveles de n-gramas para ver si estos alcanzan mayor efectividad que los monogramas.

Tercero. Se recomienda profundizar los experimentos a efectos de determinar por qué los patrones de clasificación pierden consistencia con la aplicación de ganancia de información en el escenario estandarizado, si más por el contrario teóricamente el patrón debería perfeccionarse al este estar conformado por los términos más valorados probabilísticamente.

Cuarto. Al extraer los verbos y sustantivos se recomienda eliminar los sustantivos propios debido a que si el tamaño del corpus de entrenamiento es demasiado extenso y los archivos están compuestos por estos términos, esto ocasionaría que el patrón se haga menos representativo del tipo o categoría de documento.

Trabajos Futuros

En los trabajos que se proyectan a futuro está la aplicación de la propuesta presentada en la presente tesis en un sistema para detectar duplicidad de proyecto de investigación en la Unidad de Investigación de la Facultad de Ingeniería de Producción y Servicios de la Universidad Nacional de San Agustin de Arequipa.

Desarrollar una investigación aplicada sobre el corpus RCV2, con el cual ya se cuenta en la versión en español, a la que se efectuará una diferenciación como Alessandro Moschitti, y sentar un precedente en la categorización de documentos con el modelamiento de espacio de palabras en español.

Hacer un estudio más profundo de la ganancia de información y la extracción de verbos y sustantivos, proponiendo una nueva técnica de clasificación por relevancia de términos luego de la construcción de los n-gramas, dentro del marco del modelo de espacio de palabras.

Desarrollar utilitarios para facilitar el procesamiento de lenguaje natural en español, tales como el morfotagger u otros similares, diccionario semántico en español con una confluencia de sinónimos y antónimos a una raíz común.

GLOSARIO DE TÉRMINOS

Contexto: El marco entorno a una palabra y las palabras circundantes, usado para ayudar explicar el significado de la palabra.

Coocurrencia: Se da cuando las palabras están juntas entre sí, entonces se dice que hay coocurrencia entre esas palabras.

CIIS 2013: Congreso Internacional de Ingeniería de Software, Octubre del 2013, Universidad La Salle Arequipa

COMTEL 2012: IV Congreso Internacional de Computación y Telecomunicaciones", Octubre 2012, Universidad Inca Garcilazo de la Vega, Lima

IA: Inteligencia artificial

Lema: Es definido por el termino raíz de un conjunto de términos que en esencia expresan lo mismo.

Palabra vacía: (también llamadas palabras funcionales) son las preposiciones, conjunciones, artículos, etc.

Palabra: Es cada uno de los segmentos limitados por pausas o espacios en la cadena hablada o escrita

Termino: Representa una palabra, raíz o conjunto de palabras o raíces asociadas.

Textos forenses: Texto extraído de una fuente específica, la misma que debe tener contenido textual y autoría.

Vector: Es un elemento de un espacio de vectores y es definido por n componentes o coordenadas, que describen la ubicación en el espacio n-dimensional.

Vocablo: Palabra, como expresión de una idea

Lingüística informática: es una disciplina que abarca el uso de computadores con relación al lenguaje y a las lenguas. Incluye todo tipo de herramientas que ayuden al estudio de las

lenguas y de la lingüística. La lingüística Computacional es una parte de la lingüística informática.

Ingeniería lingüística: se refieren a las aplicaciones potencialmente comerciales que implican el uso de nuevas tecnologías. Incluye la edición electrónica (diccionarios, libros), los productos multimedia, etc.

PLN: Procesamiento de Lenguaje Natural

LC: Lingüística Computacional

Referencias Bibliográficas

Ass K. y Eikvil L. (1999), "*Text categorization: a survey*", Technical Report 941, Norwegian Computing Center, Noruega.

Bausela Herreras; Esperanza (2010), "*La docencia a través de la investigación–acción*", Universidad Nueva Esparta, Disponible en: http://www.une.edu.ve/uneweb2005/servicio_comunitario/investigacion-accion.pdf

Branson; Kristin M. (2001), "*A Naibe Bayes Classifer Using Transductive Inference for Text Classificactions*", Dept. of Computer Science and Engineering, University of California, San Diego.

Caballero Romero; Alejandro (2005), "*Guías metodológicas para los planes y tesis de maestría y doctorado*", Lima, UGRAPH

Contreras Z.; Hilda Yelitza (2001), "*Procesamiento del Lenguaje Natural basado en una gramática de estilos para el idioma español*", Propuesta Tesis Doctoral, Universidad de los Andes, Facultad de Ingenierías, Postgrado en Computación.

Cornejo Aparicio; Víctor Manuel, Copara Zea; Jenny (2012), "*Análisis comparativo de la aplicación de monogramas y digramas en la clasificación de documentos*", IV Congreso Internacional de Computación y Telecomunicaciones - COMTEL 2012, Disponible en: http://www.comtel.pe/comtel2012/callforpaper2012/P36C.pdf

Covington; Michael, "*Natural Language Processing for Prolog Programmers*". Artificial Intelligence Programs The University of Georgia Athens, Georgia. Prentice Hall, Englewood Cliffs. NewJersey 07632.

Coyotl Morales; Rosa Maria (2007), "*Clasificación Automática de Textos considerando el Estilo de Redacción*", Tesis de Maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica – INAOE, Tonantzintla, Pue.

F. de Martínez; Elena (2009), Guía: "*Tipos de investigaciones*", Universidad Metropolitana, Caracas Venezuela

Fernández Pérez; Milagros (1999), "*Introducción a la Lingüística: Dimensiones del lenguaje y vías de estudio*", Ed. Ariel, Barcelona – España.

Gövert; Norbert, Lalmas; Mounia, Fuhr; Norbert (1999), *"A probabilistic description-oriented approach for categorising Web documents"*. In Susan Gauch and Il-Yeol Soong, editors, Proceedings of the Eighth International Conference on Information and Knowledge Management, pages 475–482, New York, ACM

Grishman; Ralph (1991), "*Introducción a la lingüística computacional*", Ed. Visor, Madrid.

Hernández Cruz; Macario (2007), "*Generador de los grafos conceptuales a partir del texto en español*", Tesis Magistral, Instituto Politécnico Nacional - Centro de Investigación en Computación, México.

Instituto Tecnológico de Massachussetts - Departamento de Ingeniería Eléctrica e Informática (2003), Guía de práctica "*Modelado del Lenguaje*", USA.

Joachims T. (1998), "*Text Categorization with Support Vector Machines: Learning with many relevant features*", 10th European Conference on Machine Learning, Edición 1298, pp 137-142, Dorint-Parkhotel, Chemnitz, Germany.

Jurafsky; Daniel, Martin; James H. (2009), "*Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*", Ed. Prentice Hall

Kember; David, Gow; Lyn (2010), "*Action research as a form of staff developnet in higher education, Kluwer Academic Press Publisher, Netherlands*", traducido por Pedro D. Lafourcade del Instituto de perfeccionamiento y Estudios Superiores, Montevideo Uruguay, Disponible en: http://ipes.anep.edu.uy/documentos/libre_asis/materiales/Investigacion%20accion.pdf

Lakoff; G., Johnson; M. (1999), "*Philosophy in the flesh: The embodied mind and its challenge to western thought*", In Basic Books, New York.

Lewis; David D., Test Collections - Reuters-21578, Disponible en: http://www.daviddlewis.com/resources/testcollections/reuters21578/

Lewis; Davis D., Ringuette; Marc (1994), *"A comparison of two learning algorithms for text categorization"*. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 81–93, Las Vegas, US.

Locke; W.N., Booth; A.D. (1995), "*Machine Translation of Languages*", Technology Press of MIT ans Wiley, Cambridge, Mass.

Manaris; Bill, Slator; Brian (1996), "*Interactive Natural Language Processing: Building on Success*", Computer, IEEE.

Mandala; Rila, Tokunaga; Takenobu, Tanaka; Hozumi (2000), "*Query expansion using heterogeneous thesauri*", Information Processing and Management, Vol 36,

Maron; M. (1961), *"Automatic indexing: an experimental inquirí"*. Journal of the ACM, 8:404–417

McEnery; Tony, Wilson; Andrew (2001), "*Corpus Linguistics. An Introduction*". Second Edition. Edinburgh University Press, Edinburgh.

Moens; Marie Francine, Dumortier; Jos (1999), *"Automatic categorization of magazine articles"*. In P. de Bra and L. Hardman, editors, Conferentie Informatiewetenschap 1999, Amsterdam.

Moreno Sandoval; Antonio (1998), "*Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos*". Madrid. Editorial Sintesis.

Moschitti; Alessandro, "*Text Categorization Corpora*", Disponible en: http://disi.unitn.it/moschitti/corpora.htm,

Moure; Teresa, Llisterri; Joaquim (1996), "*Lenguaje y nuevas tecnologías: el campo de la lingüística computacional*", en Servicio de Publicaciones e Intercambio Científico, Universidade de Santiago de Compostela, España.

Navarro Colorado; Francisco de Borja (2007), "*Metodología, construcción y explotación de corpus anotados semántica y anafóricamente*", Ph.D. Tesis, Universidad de Alicante, España.

Olivas; J.A., Garcés; P.J., Romero; F.P. (2003 pp. 201-219) "*An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents*". Int. Jour. of Approx. Reasoning (Soft Computing in Recognition and Search) Vol. 34

Osgood; Charles E., George; J. Suci, Tannenbaum; Percy H. (1957), "*The Measurement of Meaning*", Illinois, Estados Unidos, University of Illinois Press

Pérez Guerra; J. (1998), "*Introducción a la lingüística de corpus. Un ejercicio con herramientas informáticas aplicadas al análisis textual*", Santiago de Compostela: Tercera Edición.

Qiu; Yonggang, Frei; Hans Peter (1993, pp. 160-169), "*Concept-based query expansion, In Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval*", Pittsburgh, USA.

Rocchio; J. J. (1971), *"Relevance feedback in information retrieval. In Gerard Salton"*, editor, The SMART Retrieval System. Experiments in Automatic Document Processing, pages 313–323. Prentice Hall, Englewoods Cliffs, N. J.

Romero; F.P., Olivas; J.A., Garcés; P.J. (2006 pp. 1040 - 1045) "*A soft Approach to Hybrid Models for Document Clustering*". Proceedings of the Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'06, Paris Les Cordeliers, France. Vol 1.

Romero; Francisco P., Caballero; Ismael, Olivas; Jose A., Verbo; Eugenio (2008, pp 643-649), "*Filtrado de información mediante prototipos borrosos y perfiles basados en criterios de calidad de datos*", XIV Congreso Español sobre Tecnologías y Lógica Fuzzy, Cuencas Mineras (Mieres – Langreo), España.

Rubenstein; Herbert, Goodenough, John B. (1965 pp 627–633). "*Contextual correlates of synonymy. Communications*", the Association of Computing Machinery, 8(10)

Sahlgren; Magnus (2006), "*The Word-Space Model - Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*", Tesis doctoral, Stockholm University Department of Linguistics Computational Linguistics Stockholm, Sweden - National Graduate School of Language

Technology Gothenburg University, Gothenburg, Sweden - Swedish Institute of Computer Science Userware Laboratory Kista, Sweden.

Sahlgren; Magnus (2006), "*Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*", Ph.D. dissertation, Stockholm University, Sweden.

Sarnpieri; Roberto Hernández, Fernández Collado; Carlos, Bapista Lucio; Pilar (2010), "*Metodología de la investigación*", 4ta Ed. , Mc Grow Hill, Mexico.

Schütze; H. (1992, pp. 787-796), "*Dimensions of meaning. In Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*", IEEE Computer Society Press.

Schütze; H. (1993, pp. 895-902), "*Word space In Conference on Advances in Neural Information Processing Systems*", San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Schutze; H., Hull; David A., Pedersen; Jan Q. (1995), *"A comparasion of classifers and document representations for the routing problema*", SIGIR 95

Schutze; H., Pedersen; J. (1997), "*A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing and Management*", Foundations of Statistical Natural Language Processing, Stanford University and Xerox Palo Alto Research Center, USA.

Smith; E. E., Medin; D. L. (1981), "*Categories and Concepts*". Harvard University Press, Cambridge, MA

Tejada Cárcamo; Javier (2009), "*Construcción Automática De Un Modelo De Espacio De Palabras Mediante Relaciones Sintagmáticas y Paradigmáticas*", Tesis Doctoral, Instituto Politécnico Nacional - Centro de Investigación en Computación, México, D.F.

Tomás Díaz; David (2009), "*Sistemas de clasificación de preguntas basados en corpus para la búsqueda de respuestas*", Dr. Tesis, Universidad de Alicante, Departamentos de Lenguajes y Sistemas Informáticos, España.

Villayandre Llamazares; Milka (2010), "*Aproximación a la Lingüística Computacional*", Tesis doctoral, Universidad de León, León, España.

Waltz; D. L., Pollack; J. B. (1985 pp. 51–74), “*Massivelyparallel parsing: A strongly interactive model of naturallanguage interpretation*”. Cognitive Science, 9.

Yang; Y., Pedersen; J. (1997), “*A comparative study on feature selection in text categorization*”, 14th International Conference on Machine Learning. Morgan Kaufmann Publishers, ICML-97, San Francisco, USA.

Zellig S; Harris (1951), “*Methods in Structwal Unguistics*”, Chicago, University of Chicago Press

Anexo Nº 1: Paper COMMTEL 2012 y Articulo revista UAP

"Análisis comparativo de la aplicación de monogramas y digramas en la clasificación de documentos"

M.Sc. Víctor Manuel Cornejo Aparicio[1,2] Jenny Copara Zea[2]

vcornejo5@hotmail.com, jen_copara@hotmail.com

[1]Universidad Nacional San Agustín de Arequipa
Av. Independencia S/N - Cercado
[2]Universidad Alas Peruanas Filial Arequipa
Urb. Daniel Alcides Carrión G-14 Dist. José Luis Bustamante y Rivero

Arequipa – Perú

1. **Resumen**: *El presente trabajo presenta el resumen del trabajo de investigación en el área de procesamiento de lenguaje natural (Natural Language Processing), la aplicación de modelos de espacios de palabras (Word Space Model) en la clasificación automática supervisada de documentos, empleando monogramas y digramas, donde el propósito fundamental es la comparación de la efectividad de la clasificación de estos ngramas.*

2. **Abstract**: *This paper presents a summary of the research in the area of Natural Language Processing, the application of Word Space Model in the supervised automatic classification of documents, using monograms and bigrams, where the primary purpose is to compare the effectiveness of the classification of these ngramas*

3. **Palabras claves**: Procesamiento de Lenguaje Natural, Modelo de Espacio de Palabras, Clasificación de Documentos, nGramas

Introducción

En las diversas instituciones elaboran documentos de diversa índole, estos son regularmente redactados en forma regular por personas bien definidas, por su cargo o responsabilidad, además de tener formatos de redacción estandarizados, dicho de otra forma, las personas ocupan cargos, y en ellos redactan oficios, cartas, memorándums, informes, etc. Todo esto da origen a un conjunto de características que de alguna manera configuran un estereotipo, sumado al factor de que cada persona tiene un vocabulario limitado, y es recurrente en el uso de diversas palabras al redactar sus documentos.

Todos los aspectos anteriormente descritos, constituyen un conjunto de patrones que son susceptibles de emplear para el reconocimiento de los tipos de documentos que generan. En el procesamiento de lenguaje natural, existe la técnica del modelamiento del espacio de palabras, el mismo que trata de la asociación de los diversos vocablos a los documentos que los contiene, lo cual constituye en suma un patrón de clasificación, en este contexto surge la duda de que si un vocablo esta directamente asociado en algún grado de importancia con un tipo de documento, y si la asociación de dos vocablos que constituyen mayor cantidad de datos, lo que daría a entender una mayor cantidad de información, y que por consiguiente, podría aportar mayor precisión en el tratamiento de la clasificación de documentos, esta es la problemática que trata de abordar el presente artículo, que básicamente tratará de demostrar que tan bueno es tratar de efectuar trabajos de clasificación empleando monogramas y digramas de palabras lematizadas.

Trabajos Previos

Un equipo constituido en la Universidad Europea de Madrid – CEES, que trabaja el área de procesamiento de lenguaje natural, ha creado una herramienta denominada CADOC: "herramienta de clasificación automática de documentos" [Gómez, 2003], su proceso esta enfocado en la extracción del texto de los documentos, el indexado de los mismos y su posterior tratamiento con el weka, para posteriormente efectuar la clasificación, su herramienta permite hacer una clasificación definida por el usuario.

Arturo Montejo Ráez, en el Departamento de Informática de la Universidad de Jaén de España, trabajo su tesis doctoral titulada "Clasificación Automática de Textos en el Dominio de la Física de Altas Energías" [Montejo 2005], quien desarrolló principalmente sus investigaciones en el Laboratorio Europeo para la Investigación Nuclear (CERN), su planteamiento sobre clasificación de documentos lo trabajo en base a tres disciplinas: Recuperación de Información (RI), Procesamiento del Lenguaje Natural (PLN) y algoritmos de Aprendizaje Automático (Machine Learning - ML), empleo la

introducción de información bibliográfica como factor importante en los resultados del proceso de clasificación

Peláez J.I. y Sánchez P. del Dpto. Lenguajes y Ciencias de la Computación, E.T.S.I. Informática. Campus de Teatinos. Universidad de Málaga España, conjuntamente que con La Red D. del Dpto. de Informática de la Universidad Nacional del Nordeste - Corrientes. Argentina, han desarrollado el trabajo denominado "Un Clasificador de Texto Por Aprendizaje" [Peláez 2002], quienes trabajan en el área de telemedicina, donde buscan elaborar un clasificador estomatológico de pacientes para según esta, sean derivados a áreas especializadas de acuerdo a la categorización definida. Su aprendizaje esta basándose en los pre diagnósticos establecidos por un profesional médico no especializado, y un diccionario de términos estomatológicos, es capaz de clasificar nuevos pre diagnósticos en las especialidades

Clasificación automática de documentos

3.1. Premisas de la investigación

En el presente trabajo se inicia partiendo de un conjunto de premisas, las mismas que justificaran las acciones desarrolladas y cuyos mecanismos y resultados se presentan en este artículo.

Premisa 1: Los documentos tienen una naturaleza y estructura, los mismos que a su vez están constituidos por textos que son un conjunto de vocablos que son regularmente empleados en documentos de similar categoría.

Premisa 2: Los vocablos individualmente constituyen información, y estos a la vez que se asocian entre si, incrementan el volumen de información, la misma que podría caracterizar en mejor manera a los documentos que los contengan.

Premisa 3: Cuando se emplean más de un vocablo, en un proceso de clasificación automática (n-gramas), puede darse el caso que una conjunción de vocablos (A, B), pueda presentarse como (B, A) en el mismo documento o uno similar del mismo tipo, lo cual en términos prácticos, constituiría una dispersión de las frecuencias asociadas a la categoría definida, para lo cual, dado el caso se debería indexar horizontalmente los vocablos, y de esta forma evitar la dispersión de las frecuencias.

Premisa 4: Al constituirse los vocablos asociados de uno, dos o más, estos se deberán catalogar asociados a la tipo de documento que les dio origen, una vez constituida la asociación y elaborado la concentración de frecuencias, estos vocablos se asumirán como únicos a efectos de desarrollar los cálculos requeridos para la determinación de las proximidades entre los vocablos y el tipo de documento asociado.

3.2. Proceso de clasificación

El proceso de clasificación expresado en una forma muy breve, se presenta en la figura 1, la misma que consta de dos etapas plenamente diferenciadas, la etapa de entrenamiento y la etapa de clasificación.

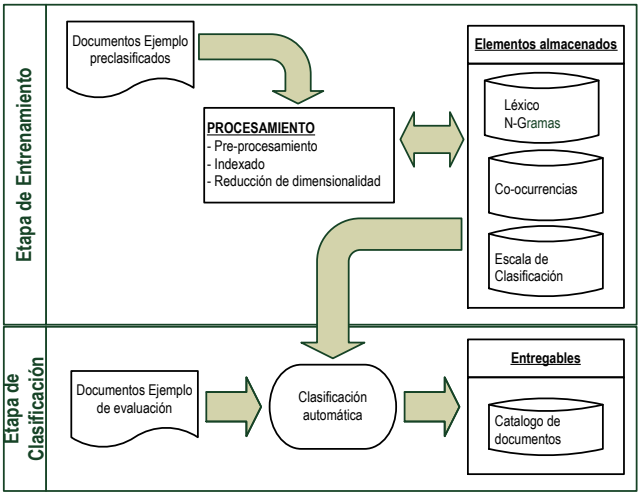


Figura 1: Esquema del proceso de clasificación

Debido a que el propósito del presente articulo es presentar la comparación de usar monogramas o digramas en el proceso de clasificación automática, detallaremos únicamente aquellos aspectos relevantes a dicho proceso.

Para iniciar el proceso de entrenamiento, es necesario contar con un número determinado de documentos preclasificados, esto para que en el entrenamiento, se puedan crear los nGramas de forma asociada al tipo definido.

El procesamiento consta de tres sub etapas, que son: Pre-procesamiento, Indexado y Reducción dimensional, en este contexto, durante el pre-procesamiento, es necesario limpiar el texto de forma tal que se pueda emplear un texto limpio de enlaces, caracteres especiales, números, símbolos u otros elementos que no aporten mayor información, una vez logrado este aspecto, se procede a efectuar un lematizado del texto, que en suma nos entrega un texto listo para ser empleado en las tareas siguientes. Esto se puede apreciar en la figura 2 que presentamos a continuación:

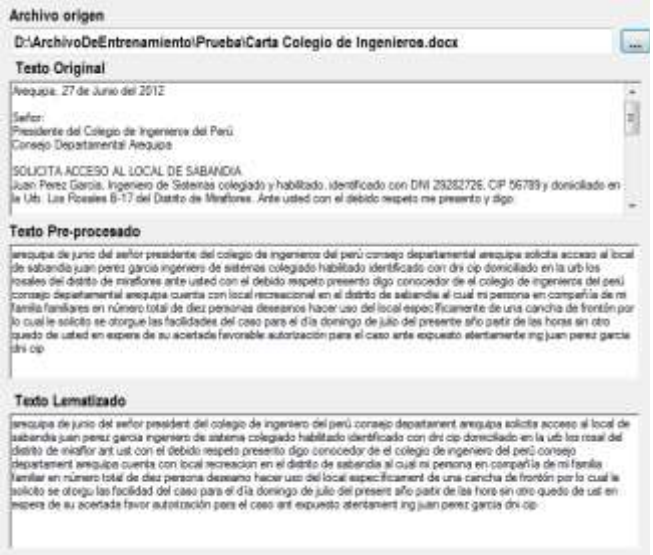


Figura 2: Texto pre-procesado

El indexado de la información se da en dos términos, el primero se da cuando se trata de asociaciones de vocablos de más de un término (digrama o superior) en sentido horizontal, esto debido a que las asociaciones (A, B) y (B, A), es la misma conjugación, mantenerlas separadas, constituiría una dispersión de la información., y en segundo plano es un indexado vertical donde se ordena en forma creciente los términos con una jerarquía definida de la forma siguiente: Tipo de documento, raíz 1, raíz 2, …raíz n, cabe recordar que el presente trabajo, solo presenta el caso de los monogramas y digramas, los mismos que se muestran en los cuadros 1 y 2.

Doc	Raiz	Frec.
Carta	arequip	3
Carta	juni	1
Carta	señor	1
Carta	president	1
Carta	colegi	2
Carta	ingenier	3
Carta	per	2
…	…	…
Carta	autorizac	1
Carta	expuest	1
Carta	atent	1
Carta	ing	1

Doc	Raiz1	Raiz2	Frec.
Carta	arequip	juni	1
Carta	juni	señor	1
Carta	president	señor	1
Carta	colegi	president	1
Carta	colegi	ingenier	2
Carta	ingenier	per	2
Carta	consej	per	2
…	…	…	…
Carta	atent	expuest	1
Carta	atent	ing	1
Carta	ing	juan	1
Carta	dni	garci	1

Cuadro 1: Tabla de ejemplo de monogramas y digramas indexado horizontalmente

Doc	Raiz	Frec.
Carta	acces	1
Carta	acertad	1
Carta	año	1
Carta	arequip	3
Carta	atent	1
Carta	autorizac	1
Carta	canch	1
Carta	cas	2
…	…	…
Carta	sistem	1
Carta	solicit	2
Carta	total	1
Carta	urb	1
Carta	uso	1

Doc	Raiz1	Raiz2	Frec.
Carta	acces	local	1
Carta	acces	solicit	1
Carta	acertad	esper	1
Carta	acertad	favorab	1
Carta	año	part	1
Carta	año	present	1
Carta	arequip	cuent	1
Carta	arequip	departament	2
Carta	arequip	juni	1
Carta	arequip	solicit	1
…	…	…	…
Carta	present	respet	1
Carta	president	señor	1
Carta	ros	urb	1

Cuadro 2: Tabla de ejemplo de monogramas y digramas indexado verticalmente

Posteriormente se efectúa la reducción dimensional, la misma que reducirá notablemente el número de términos con los cuales se trabajara la matriz de coocurrencia, esto es efectuado básicamente para no sobrecargar los algoritmos al aplicar la clasificación y reducir su tiempo de ejecución.

El proceso de entrenamiento en su núcleo central se trabaja con un algoritmo que contenga tres parámetros básicos: El texto lematizado a procesar, el nGrama que se desea construir, y el identificador del tipo de documento al que pertenece el texto, todo ello se elabora en los pasos siguientes:

```
Procedimiento EntrenarNGrama
Parametros: Texto    'Texto pre-procesado y lematizado
   nGrama  'Tipo de nGrama a Procesar [1] Monograma, [2] Digrama
   IdDocTipo  'Tipo de documento al que corresponde el entrenamiento
   Raiz = ExtraerPalabra(Texto)
   Mientras Raiz <> ""
      IdRaiz =MatricularRaiz(Raiz)
      InsertarRaiz(Idraiz, vRaiz)
      vRaizOrdenado = OrdenarVector( vRaiz)
      MatricularNGrama(IdDocTipo, vRaizOrdenado)
      Raiz = ExtraerPalabra(Texto)
   Fin Mientras
Fin Procedimiento
```

Experimentos y Resultados

Uno de los experimentos trabajados en la investigación respecto al tema planteado se efectuó con un conjunto de documentos definidos en el siguiente cuadro:

Tipo de Documento	Cantidad	Muestra
Oficio	180	40
Carta	342	45
Solicitud	188	41
Memorandum	179	40
Contrato	177	40
Informe	187	41
Recibo	919	49

Cuadro 3: Tabla de cantidad de tipos de documentos empleados

Para el proceso de clasificación se empleara una muestra aleatoria, la misma que se determino en base a la siguiente formula estadística, y cuyos resultados se muestran en el cuadro 3:

$$n = \frac{N}{1 + \frac{e^2(N-1)}{z^2 pq}}$$

Donde:

n: Tamaño de la muestra que deseamos obtener
N: Tamaño conocido de la población
e: 0.05

z: 1.65
p: 5%
q: 95%

De la muestra seleccionada, según los tipos de documentos, se seleccionaron cinco motivos de clasificación, se efectuó esta acción, pues en algún momento se trato de establecer la concordancia con la estructura de los documentos, los motivos que se seleccionaron fueron

1. Cartas de presentación
2. Ascensos
3. Contratos de prestación de servicios
4. Procedimientos administrativos de grado
5. Eventos Académicos

Efectuado el proceso de entrenamiento, y luego de construir el corpus de monogramas y digramas correspondientes a los modelos de documentos con los motivos establecidos, se procedió a efectuar el proceso de clasificación empleando para ello los monogramas y digramas, en un primer momento empleando estos ngramas de forma original y en un segundo tiempo aplicando una reducción dimensional.

Los resultados obtenidos en el caso de los monogramas sin la aplicación de reducción dimensional, se muestran en el cuadro número 4, en esa tabla se puede observar que el proceso de clasificación es aceptable pero existe un grado precisión del 83%, lo cual nos indica que no es del todo aplicable para casos que requiera un nivel de confiabilidad superior.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	48	80%	7	12%	0	0%	1	2%	4	7%	60
2	5	8%	45	75%	0	0%	7	12%	3	5%	60
3	0	0%	1	2%	55	92%	3	5%	1	2%	60
4	3	5%	0	0%	0	0%	51	85%	6	10%	60
5	4	7%	1	2%	0	0%	5	9%	46	82%	56
										Total	296

Cuadro 4: Tabla de clasificación de documentos con monogramas sin aplicar reducción dimencional

Los resultados obtenidos en el caso de los digramas sin la aplicación de reducción dimensional, se muestran en el cuadro número 5, en este se presentan resultados nada alentadores para el empleo de digramas como técnica de clasificación, pues solo alcanza un nivel de precisión promedio del orden del 78%, lo cual podría juzgarse erróneamente de forma apresurada como una técnica no confiable.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	45	75%	8	13%	0	0%	1	2%	6	10%	60
2	8	13%	40	67%	0	0%	3	5%	9	15%	60
3	0	0%	6	10%	52	87%	2	3%	0	0%	60
4	5	8%	2	3%	0	0%	43	72%	10	17%	60
5	3	5%	1	2%	0	0%	3	5%	49	88%	56
										Total	296

Cuadro 5: Tabla de clasificación de documentos con digramas sin aplicar reducción dimencional

Los resultados obtenidos en el caso de los monogramas con la aplicación de reducción dimensional, reduce notablemente la precisión de los monogramas, esto con un nivel de precisión promedio del orden del 71%, lo cual se puede apreciar en el cuadro número 6 que se presenta a continuación.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	43	72%	9	15%	0	0%	3	5%	5	8%	60
2	12	20%	37	62%	0	0%	6	10%	5	8%	60
3	0	0%	1	2%	53	88%	6	10%	0	0%	60
4	14	23%	2	3%	0	0%	37	62%	7	12%	60
5	6	11%	4	7%	0	0%	7	13%	39	70%	56
										Total	296

Cuadro 6: Tabla de clasificación de documentos con monogramas aplicando reducción dimencional

Los resultados obtenidos en el caso de los digramas con la aplicación de reducción dimensional, mejoran notablemente la precisión del proceso de clasificación, alcanzando un promedio del orden del 95%, dichos resultados se evidencian en el cuadro número 7, el mismo que se presenta a continuación.

Motivo	1		2		3		4		5		Total
	F	%	F	%	F	%	F	%	F	%	
1	53	88%	4	7%	0	0%	1	2%	2	3%	60
2	1	2%	57	95%	0	0%	1	2%	1	2%	60
3	0	0%	0	0%	59	98%	1	2%	0	0%	60
4	1	2%	0	0%	0	0%	58	97%	1	2%	60
5	1	2%	0	0%	0	0%	1	2%	54	96%	56
										Total	296

Cuadro 7: Tabla de clasificación de documentos con digramas aplicando reducción dimencional

Conclusiones

Luego de efectuadas las pruebas experimentales donde se entrenaron y luego clasificaron los documentos en un ambiente controlado y supervisado se puede decir que a medida que se incrementa la diversidad de tipos de documento con estructuras diversas o no muy bien definidas, los monogramas arrojan mejores resultados cuando no se aplica la reducción dimensional, esto también se puede observar tomando como medida de comparación el tamaño del corpus generado, y puede decirse que a medida que el corpus de documentos crece los monogramas son mas efectivos. Pero con

documentos bien estructurados, y con una reducción dimensional, los digramas mejoran su rendimiento y precisión.

Podría parecer que la estructura de los documentos es irrelevante, puesto que al pre-procesar los textos contenidos, esta estructura se pierde, lo cual no es del todo correcto, en formato la estructura de los párrafos puede perderse, pero la secuencia de los vocablos relevantes permanece, lo cual se pudo evidenciar en tipos con estructura muy rígida, como es el caso de los contratos, donde en todos los casos su clasificación fue efectiva, en los oficios esto sucedió en el noventa por ciento, y así disminuye en cuanto la estructura se hace mas diversa como es el caso de las cartas

Se puede sugerir trabajos futuros en el orden de determinar el impacto de la estructura de los documentos en el proceso de clasificación, así como la generación personalizada de corpus por autor, para ver la autenticidad de los documentos, también seria pertinente establecer umbrales de reducción dimensional por ganancia de información por cada motivo.

Para artículos futuros se esta experimentando con corpus Reuters21578-Apte-90Cat y Reuters21578-Apte-115Cat, para repetir el análisis exento de estructura, y con ello concretar un juicio de mayor precisión.

Referencias

[Montejo, 2010] Montejo A., Perea J.M., Martín M. y Ureña A., "Uso de la detección de bigramas para categorización de texto en un dominio científico", Revista Procesamiento de Lenguaje Natural, No 44 (2010).

[Cavnar 1994] Cavnar W. and Trenkle J., "N-Gram-Based Text Categorization", 3rd Annual Symposium on Document Analysis and Information Retrieval

[Gómez, 2003] Isidro Gómez Mompó, Jaime Lozano Muñoz, Diego Martínez Salazar, Luis Muñoz Góngora, Diego Ramírez Adrados, "CADOC: herramienta de clasificación automática de documentos", Universidad Europea de Madrid – CEES, Disponible en: http://www.esp.uem.es/jmgomez/plenum/plenum3/03.pdf , Junio 2003

[Montejo 2005] Arturo Montejo Ráez, Tesis Doctoral "Automatic Text Categorization of Documents in the High Energy Physic Domain", Departamento de Informática de la Universidad de Jaén de España, Diciembre del 2005.

[Peláez 2002] Peláez J.I. y Sánchez P. "Un Clasificador de Texto Por Aprendizaje", Revista Inteligencia Artificial, Vol 6, No 15 (2002), disponible en: http://aepia.lcc.uma.es/index.php/ia/article/view/756

[Magnus 2006] Magnus Sahlgren, Tesis Doctoral "The Word-Space Model - Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces", Stockholm University, 2006

[Tejada 2009] Javier Tejada Cárcamo, Tesis doctoral "Construcción automática de un modelo de espacio de palabras mediante relaciones sintagmáticas y paradigmáticas", Instituto Politécnico Nacional, Centro De Investigación En Computación, México 2009

Anexo Nº 2: Paper CIIS 2013 y Articulo Revista Universidad La Salle Arequipa

Construcción automática y análisis de Modelos de Espacios de Palabras de n-gramas y su aplicación a tareas de procesamiento de lenguaje natural

Dr(c). Víctor Manuel Cornejo Aparicio
vcornejo5@hotmail.com
Dr. Javier Tejada Cárcamo
jtejada@itgrupo.net
Universidad Nacional San Agustín de Arequipa
Arequipa - Perú

RESUMEN

El presente trabajo de investigación tiene por objetivo presentar como mejorar la calidad de vocablos relacionados semánticamente mediante la construcción automática y análisis de Modelos de Espacios de Palabras basados en n-gramas. Este método debe incluir vocablos que a su vez deben mejorar la precisión de tareas de procesamiento de lenguaje natural, específicamente la clasificación de textos, para ello se emplearon modelos ya existentes como base de conceptualización y se implementaron mejoras en el pre-procesamiento de los textos, tales como la extracción de verbos y sustantivos, posteriormente se trabajó la clasificación a tres niveles de n-gramas (monogramas, digramas y digramas ordenados horizontalmente), luego se efectuaron los experimentos con el corpus estandarizados "corpora Reuters 21578", del cual se seleccionaron las ocho categorías más relevantes con las que se obtuvo un nivel de precisión del orden del 84.17%, con lo que se superó el porcentaje prevalente y lo cual permitió validar la propuesta.

PALABRAS CLAVES: Modelos de espacios de palabras (MEP), Clasificación de documentos, Reuters 21578

ABSTRACT

The present research work aims to present how to improve the quality of semantically related words using the automatic construction and analysis of models based word spaces n-grams. This method should include words which in turn should improve the accuracy of tasks natural language processing, text classification specifically for this purpose existing models were used as a basis for conceptualization and implemented improvements in the pre-processing of texts such as verbs and nouns extraction, classification subsequently worked at three levels of n-grams (monograms, digrams and digrams arranged horizontally), then conducted experiments with standardized corpus "Reuters corpora 21578", which is selected the eight most relevant categories for which there was a level of precision of the order of 84.17%, which exceeded the rate prevalent and which allowed us to validate the proposal.

KEY WORD: Word Space Model (WSM), Text Clasification, Reuters 21578

1. INTRODUCCION

El presente trabajo resume la propuesta del trabajo de tesis doctoral en computación del autor del artículo y su asesor de tesis, en el marco del procesamiento de lenguaje natural, específicamente en la tarea de clasificación de documentos. Este tipo de investigación está enmarcada dentro de la inteligencia artificial, aquí no se trata de hacer una comparación del modelo de espacio de palabras con otras técnicas que pueden aplicase a este problema, lo que se trata de hacer es proponer una mejora en la calidad de los vocablos que constituyen los términos en la construcción de los patrones de clasificación de los diferentes tipos de documentos, las etapas de un proceso de clasificación no difieren de la que se aplican en todas las técnicas prevalentes, como son la etapa de entrenamiento y control, para ello se emplea el corpus "corpora Reuter 21578".

En el trabajo se diferencia primordialmente de propuestas anteriores; en el uso de verbos y sustantivos en la construcción de los patrones de clasificación con lo que se mejoró la calidad de los vocablos que constituyen los mencionados patrones, adicionalmente se aporta la idea de un ordenamiento horizontal en el caso de que se emplee términos compuestos por más de un vocablo como es el caso de los digramas o superiores; donde dicho ordenamiento mejora aún más la precisión de la clasificación.

2. ESTADO DEL ARTE

En lo que respecta a la clasificación de documentos, existen una gran variedad de artículos que trabajan la metódica de forma unilateral, esto quiere decir que describen sus técnica basados en problemas y necesidades propias, sin embargo, si se desea hacer una propuesta que contenga un aporte susceptible de comparación, se debe emplear un corpus de documentos estandarizados, que para fines del presente trabajo se empleara el corpus Reuters21578-Apte-90Cat.

Francisco P. Romero, Ismael Caballero, Jose A. Olivas, Eugenio Verbo, plantean el tema "Filtrado de información mediante prototipos borrosos y perfiles basados en criterios de calidad de datos", plantea un modelo de filtrado basado en una estructura de categorías conceptuales. Donde la estructura se define partiendo de un conjunto de documentos no estructurados, es necesario seguir los siguientes pasos: 1. Calcular la calidad de datos de todos los documentos, y descartar aquellos cuyas mediciones no estén dentro de los rangos de aceptación establecidos en los requisitos de calidad de datos de los usuarios que no superen un umbral mínimo de calidad, 2. Preproceso lingüístico: Selección de las palabras que van a representar conceptualmente a las categorías de la estructura, 3. Representación de los documentos en base a los conceptos tratados en sus contenidos. Para ello se utiliza el modelo FIS-CRM , 4. Agrupación de los documentos conceptualmente similares mediante clustering , 5. Extracción de conceptos claves, 6. Creación de una base de conocimiento utilizando Categorías Prototípicas Borrosas. La estructura básica de filtrado se basa en las diferentes categorías extraídas mediante procesos de clustering. Estas categorías comprenden una serie de documentos conceptualmente similares. Las categorías estarán organizadas en una jerarquía que reflejará los diferentes niveles de especificidad tratados en los conceptos. Cada uno de esos documentos puede estar clasificado en diferentes categorías sobre las cuales poseerán un grado de pertenencia. A su vez, el resultado ofrecido empleando esta propuesta es del orden del 83% versus el 80% de las clasificaciones sim emplear criterios de calidad.

3. CUERPO DEL CONOCIMIENTO

3.1.Premisas de la investigación

En el presente trabajo se inicia partiendo de un conjunto de premisas, las mismas que justificaran las acciones desarrolladas y cuyos mecanismos y resultados se presentan en este artículo.

Premisa 1: Los documentos tienen una naturaleza y estructura, los mismos que a su vez están constituidos por textos que son un conjunto de vocablos que son regularmente empleados en documentos de similar categoría.

Premisa 2: Los vocablos individualmente constituyen información, y estos a la vez que se asocian entre sí, incrementan el volumen de información, la misma que podría caracterizar en mejor manera a los documentos que los contengan.

Premisa 3: Cuando se emplean más de un vocablo, en un proceso de clasificación automática (n-gramas), puede darse el caso que una conjunción de vocablos (A, B), pueda presentarse como (B, A) en el mismo documento o uno similar del mismo tipo, lo cual en términos prácticos, constituiría una dispersión de las frecuencias asociadas a la categoría definida, para lo cual, dado el caso se debería indexar horizontalmente los vocablos, y de esta forma evitar la dispersión de las frecuencias.

Premisa 4: Al constituirse los vocablos asociados de uno, dos o más, estos se deberán catalogar asociados al tipo de documento que les dio origen, una vez constituida la asociación y elaborado la concentración de frecuencias, estos vocablos se asumirán como únicos a efectos de desarrollar los cálculos requeridos para la determinación de las proximidades entre los vocablos y el tipo de documento asociado.

Premisa 5: Los corpus de entrenamiento no serán modificados o reevaluados a efectos de construir los n-gramas que configuren el patrón, esto debido a que al desarrollar una aplicación real, los corpus de entrenamiento ya pasaron filtros diversos que permiten superar esta fase, y no es dable evaluar constantemente cada vez que se requiera hacer una clasificación. Cabe aclarar que esta premisa contradice los postulados de la evaluación de un corpus, que muchas veces requiere de un ajuste del contenido a efectos de mejorar el proceso de entrenamiento de los algoritmos.

3.2.Esquema del Método Propuesto

El método propuesto está comprendido por dos etapas claramente definidas. En la etapa de entrenamiento se debe contar con un conjunto de documentos $\{d_1, d_2, \dots\ d_n\}$, los cuales serán sometidos a un procesamiento que incluye las fases de pre-procesamiento, indexado y reducción de la dimensionalidad. Estas acciones nos permite lograr tres entregables: Léxico, Co-ocurrencias, Escalas de Clasificación. La etapa de control, emplea un conjunto de documentos $\{d'_1, d'_2, \dots, d'_m\}$, los cuales se someten a un proceso de clasificación empleando los entregables de la etapa de entrenamiento, en forma conjunta permite obtener un catálogo de documentos clasificados, el esquema del método propuesto se puede apreciar en el esquema siguiente:

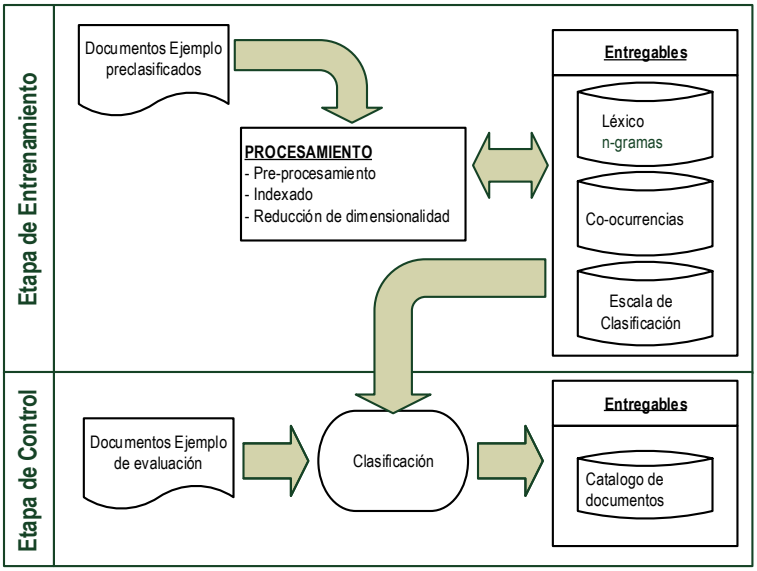


Figura Nro 1: Esquema del Método

En la etapa de entrenamiento, lo que se busca es por medio de los documentos ejemplo de entrenamiento, establecer los patrones de clasificación, para lo cual se obtienen los entregables descritos anteriormente.

El conjunto de documentos de entrenamiento $\{d_1, d_2, \dots d_n\}$, son documentos preclasificados, estos deben ser cuantificados con anterioridad, para que en los procesos sucesivos, se puedan caracterizar de una forma adecuada y esta caracterización sea susceptible de ser asociada a los tipos de documento que correspondan.

Una "clase de documento" es el nombre que identifique a los patrones de entrenamiento para el proceso de clasificación de textos, los mismos que están predefinidos y asociados al archivo que servirá para esta fase.

El procesamiento está definido por tres fases, las cuales tienen un propósito propio que se detallará a continuación:

El pre-procesamiento tiene el propósito de eliminar elementos textuales que no contienen información relevante. Esta información haría que se eleve los costos de procesamiento, así como la calidad de información, debido a que los patrones se harían más semejantes, por tanto dichos patrones no serían muy efectivos al momento de ejecutar las rutinas de clasificación, estas tareas comprenden la eliminación de etiquetas y similares, palabras vacias, pero lo más relevante es la extracción de verbos y sustantivos para la generación de los patrones de clasificación, previo lematizado de los mismos. por ejemplo:

```
… …
… … …
sTexto = ExtraerTexto("d:\Reuter21578\test\corn\0009622")
sTexto = PreProcesamiento(sTexto)
… … …
… …

Funcion PreProcesamiento( texto)
       sTextoProcesado = ProcesarTexto(texto)
       sTextoProcesado = LematizarTexto(sTextoProcesado)
       Retornar sTextoProcesado
Fin Funcion
```

```
Funcion ProcesarTexto( texto)
       _texto = EliminaLinks(_texto)
       _texto = EliminarSignos(_texto)
       _texto = EliminarCaracteresEspeciales(_texto)
       _texto = EliminarApostrofe(_texto)
       _texto = EliminarVacias(_texto)
       _texto = EliminarSignosPuntuacion(_texto)
       _texto = EliminarNumeros(_texto)
       _texto = QuitarPalabrasConNumeros(_texto)
       _texto = ExtraerSustantivosVerbos(_texto)
       Retornar _texto
Fin Funcion
```

Es simple concebir la idea de un ordenamiento vertical, lo que dicho de otro modo es el ordenamiento entre líneas diferentes de una matriz, esto de acuerdo a un criterio generalmente alfabético, pero lo que vale la pena aclarar es el ordenamiento horizontal, lo cual implica que se debe ordenas dos o más vocablos conjugados de acuerdo al nivel de n-grama que se esté trabajando, esto significa que un termino t_{ij} contiene un conjunto de vocablos $\{v_1, \quad v_2 \dots \quad v_n\}$ donde v_x es menor que v_{x+1}

Posterior a la indexación vertical y horizontal se debe crear tantas tablas como tipos de documentos puedan existir los mismos que deben contener la frecuencia de los términos que contenga el texto pre-procesado, el mismos que debe estar organizado de la forma siguiente:

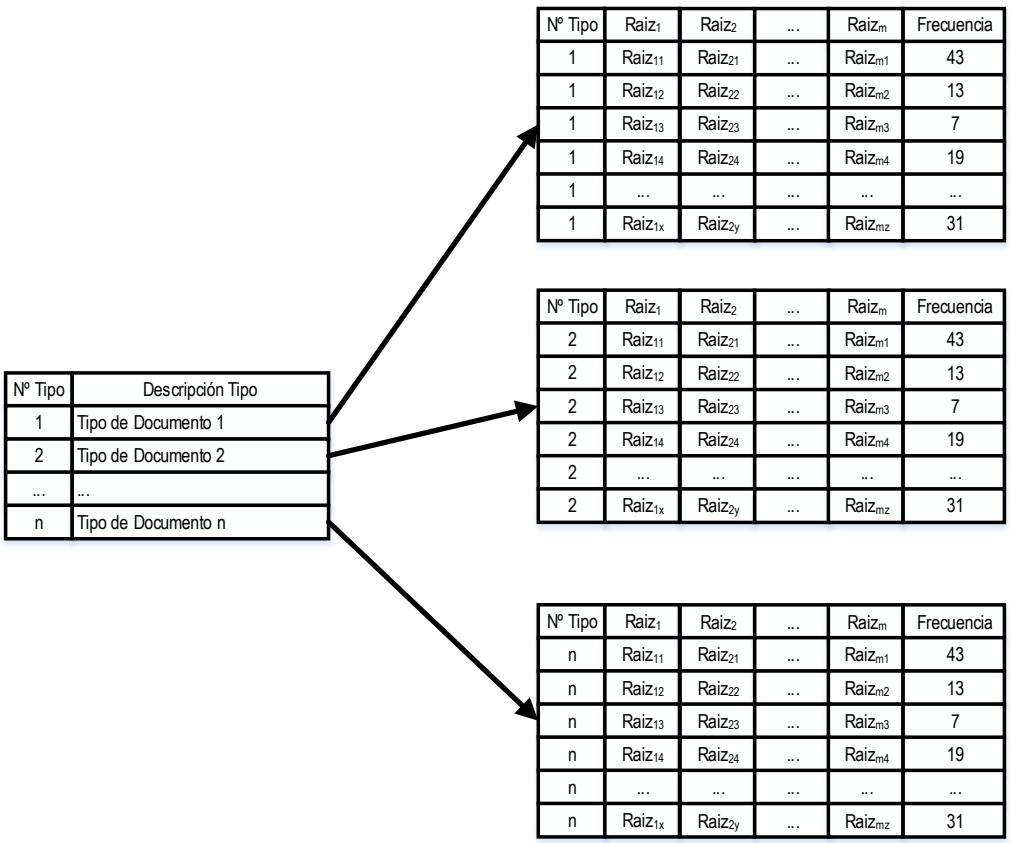


Figura Nro 2: Esquema de la creación de tablas patrón por tipo de documento

3.3.Reducción dimensional

En esta fase se deberá tener en cuenta la ganancia de información de los términos encontrados en el conjunto de documentos ejemplo de entrenamiento, para lo cual se deberá asumir los términos que concentren la mayor ganancia de información (IG_i)

$$IG(t_i) = -\sum_{k=1}^{M} P(c_k)\log P(c_k) + P(t_i)\sum_{k=1}^{M} P(c_k|t_i)\log P(c_k|t_i) + P(\overline{t_i})\sum_{k=1}^{M} P(c_k|\overline{t_i})\log P(c_k|\overline{t_i})$$

Donde:

IG:	Ganancia de información (information gain)
$c_1 \dots c_k$:	Conjunto de clases
t_i:	Término del cual se calculara la ganancia de información.
M:	Número de clases
$P(c_k)$:	Probabilidad de la clase c_k
$P(t_i)$:	Probabilidad de seleccionar un documento que contienen el término t_i
$P(c_k\|t_i)$:	Probabilidad condicional de que un documento con el término t_i pertenezca a la categoría c_k
$P(\overline{t_i})$:	Probabilidad de seleccionar un documento que no contienen el término t_i
$P(c_k\|\overline{t_i})$:	es la probabilidad condicional de que un documento con el término t_i no pertenezca a la categoría c_k

3.4.Clasificación

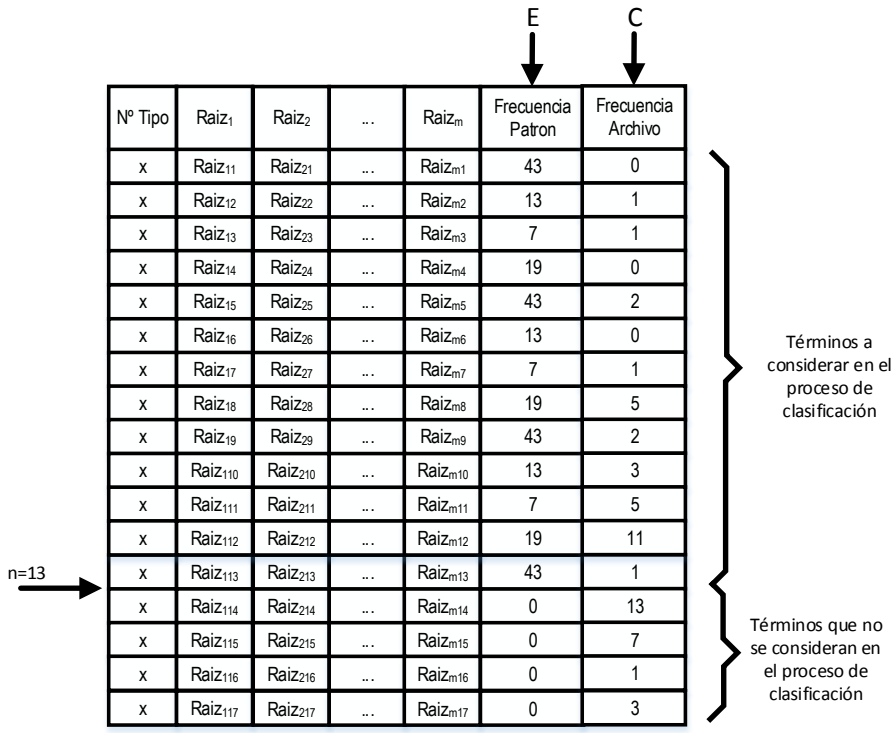
En la etapa de clasificación, se procesa los documentos ejemplo en los términos de la clase de n-grama a emplear, luego se conjugan con los n-gramas obtenidos en la etapa de entrenamiento, posteriormente se establece la proximidad que pueda haber entre ambos grupos de conjunciones, por medio de la ley de los cosenos, para luego establecer a que tipo definido se aproxima más el documento en proceso de evaluación, esto permite determinar en un grado aceptable la clasificación.

$$p_{ij} = \frac{\sum_{k=1}^{n} E_k C_k}{\sqrt{\sum_{k=1}^{n} E_k^2}\ \sqrt{\sum_{k=1}^{n} C_k^2}}$$

Donde

p_{ij}	Grado de proximidad entre el documento i por clasificar y el patrón del documento j con el cual se está comparando.
E	Frecuencias de los términos del patrón del tipo de documento j
C	Frecuencias de los términos del documento i, que se intersectan con los términos del patrón del términos del documento j
n	Número de términos que contiene el patrón de términos del documento tipo j

Es un hecho que el conjunto de términos que contenga el documento por clasificar, incluya términos que no se encuentren en el patrón de términos del n-grama del tipo de documento con el cual se este comparando; en este caso, las frecuencias de los términos del documento que no estén en el patrón del n-grama del tipo de documento con el cual se comprar, no se deben considerar en el cálculo de proximidad, lo cual se ve en el gráfico siguiente.



Nº Tipo	$Raiz_1$	$Raiz_2$	…	$Raiz_m$	Frecuencia Patron	Frecuencia Archivo
x	$Raiz_{11}$	$Raiz_{21}$	…	$Raiz_{m1}$	43	0
x	$Raiz_{12}$	$Raiz_{22}$	…	$Raiz_{m2}$	13	1
x	$Raiz_{13}$	$Raiz_{23}$	…	$Raiz_{m3}$	7	1
x	$Raiz_{14}$	$Raiz_{24}$	…	$Raiz_{m4}$	19	0
x	$Raiz_{15}$	$Raiz_{25}$	…	$Raiz_{m5}$	43	2
x	$Raiz_{16}$	$Raiz_{26}$	…	$Raiz_{m6}$	13	0
x	$Raiz_{17}$	$Raiz_{27}$	…	$Raiz_{m7}$	7	1
x	$Raiz_{18}$	$Raiz_{28}$	…	$Raiz_{m8}$	19	5
x	$Raiz_{19}$	$Raiz_{29}$	…	$Raiz_{m9}$	43	2
x	$Raiz_{110}$	$Raiz_{210}$	…	$Raiz_{m10}$	13	3
x	$Raiz_{111}$	$Raiz_{211}$	…	$Raiz_{m11}$	7	5
x	$Raiz_{112}$	$Raiz_{212}$	…	$Raiz_{m12}$	19	11
x	$Raiz_{113}$	$Raiz_{213}$	…	$Raiz_{m13}$	43	1
x	$Raiz_{114}$	$Raiz_{214}$	…	$Raiz_{m14}$	0	13
x	$Raiz_{115}$	$Raiz_{215}$	…	$Raiz_{m15}$	0	7
x	$Raiz_{116}$	$Raiz_{216}$	…	$Raiz_{m16}$	0	1
x	$Raiz_{117}$	$Raiz_{217}$	…	$Raiz_{m17}$	0	3

Figura Nro 3: Representación de la matriz de términos empleados en el proceso de clasificación

3.5.Presentación de resultados

A continuación se presenta una tabla que resume las frecuencias de los archivos de prueba clasificados empleando el método estándar de clasificación, a lo que se le agrego la técnica de ganancia de información en un ámbito local por categorías con un umbral promedio y total por todas las categorías con un umbral promedio

Tipo de Documento	Total de Archivos de Prueba	Clasificación Estándar			Clasificación con Ganancia de Información					
					Local			Total		
		Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
Acq	719	646	641	642	327	599	610	442	652	650
Crude	189	175	135	137	147	179	178	153	175	171
Earn	1087	1015	967	961	974	979	960	951	1002	987
Grain	149	125	120	120	129	133	135	95	140	139
interest	131	97	96	96	70	99	97	27	90	92
money-fx	179	124	106	110	66	93	101	27	99	103
Trade	117	103	96	96	53	99	99	43	92	95
unknown	280	214	187	189	115	174	174	120	179	183

Tabla Nro 1: Resumen en frecuencias de clasificación estándar

De la tabla presentada anteriormente, se puede calcular los porcentajes de categorización por tipos definidos, los mismos que se muestran en la tabla siguiente:

Tipo de Documento	Clasificación Estándar			Clasificación con Ganancia de Información					
				Local			Total		
	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
Acq	89.85%	89.15%	89.29%	45.48%	83.31%	84.84%	61.47%	90.68%	90.40%
Crude	92.59%	71.43%	72.49%	77.78%	94.71%	94.18%	80.95%	92.59%	90.48%
Earn	93.38%	88.96%	88.41%	89.60%	90.06%	88.32%	87.49%	92.18%	90.80%
Grain	83.89%	80.54%	80.54%	86.58%	89.26%	90.60%	63.76%	93.96%	93.29%
Interest	74.05%	73.28%	73.28%	53.44%	75.57%	74.05%	20.61%	68.70%	70.23%
money-fx	69.27%	59.22%	61.45%	36.87%	51.96%	56.42%	15.08%	55.31%	57.54%
Trade	88.03%	82.05%	82.05%	45.30%	84.62%	84.62%	36.75%	78.63%	81.20%
Unknown	76.43%	66.79%	67.50%	41.07%	62.14%	62.14%	42.86%	63.93%	65.36%
Promedio	83.44%	76.43%	76.88%	59.51%	78.95%	79.40%	51.12%	79.50%	79.91%

Tabla Nro 2: Resumen en porcentajes de clasificación estándar

De la tabla anterior se puede apreciar que la clasificación que ofrece en promedio la mayor precisión, es la que se efectúa de forma estándar sin la conjunción de vocablos, dicho de otra forma, empleando monogramas, la misma que obtuvo un nivel de precisión del orden del 83.44%

A continuación se presenta una tabla que resume las frecuencias de los archivos de prueba clasificados empleando el método propuesto de clasificación, a lo que se le agrego la técnica de ganancia de información en un ámbito total por todas las categorías con un umbral promedio

Tipo de documento	Archivos procesados	Clasificación con la propuesta			Clasificación con la propuesta y ganancia de Información		
		Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
Acq	719	661	652	659	487	652	656
crude	189	175	145	144	152	180	178
Earn	1087	995	957	961	577	965	966
grain	149	127	119	125	119	135	135
interest	131	101	91	94	22	57	56
money-fx	179	125	108	109	21	94	86
trade	117	105	104	103	35	89	84
unknown	280	211	193	189	102	197	201

Tabla Nro 3: Resumen en frecuencias de clasificación propuesta

De la tabla presentada anteriormente, se puede calcular los porcentajes de categorización por tipos definidos, los mismos que se muestran en la tabla siguiente:

Tipo de documento	Clasificación con la propuesta			Clasificación con la propuesta y ganancia de Información		
	Monograma	Digrama	Digrama Ordenado	Monograma	Digrama	Digrama Ordenado
acq	91.93%	90.68%	91.66%	67.73%	90.68%	91.24%
crude	92.59%	76.72%	76.19%	80.42%	95.24%	94.18%
earn	91.54%	88.04%	88.41%	53.08%	88.78%	88.87%
grain	85.23%	79.87%	83.89%	79.87%	90.60%	90.60%
interest	77.10%	69.47%	71.76%	16.79%	43.51%	42.75%
money-fx	69.83%	60.34%	60.89%	11.73%	52.51%	48.04%
trade	89.74%	88.89%	88.03%	29.91%	76.07%	71.79%
unknown	75.36%	68.93%	67.50%	36.43%	70.36%	71.79%
Promedio	84.17%	77.87%	78.54%	47.00%	75.97%	74.91%

Tabla Nro 4: Resumen en porcentajes de clasificación propuesta

De la tabla anterior se puede apreciar que la clasificación empleando la propuesta que ofrece en promedio la mayor precisión, es la que se efectúa sin la conjunción de vocablos, dicho de otra forma, empleando monogramas, la misma que obtuvo un nivel de precisión del orden del 84.17%

4. CONCLUSIONES

La separación de los vocablos constituidos por verbos y sustantivos, logró mejorar la calidad de los términos empleados en el proceso de clasificación de documentos. El uso de estos términos permitió por medio de monogramas, mejorar la precisión de las tareas de clasificación de documentos en el marco del procesamiento de lenguaje natural.

Se determinó que los verbos y sustantivos caracterizan efectivamente el patrón de un texto para procesos de clasificación de documentos, lo cual permite tener mejores resultados en el proceso de clasificación de éstos.

En el presente trabajo se evidencia que los monogramas superan la presición de los digramas y digramas ordenados horizontalmente, sin embargo hay que precisar dos cosas muy importantes; este trabajo se aplica sobre un corpus estandarizado (Reuters 21578), donde en la fase experimental los monogramas superaron la precisión en la clasificación. En trabajos paralelos que se efectuaron sobre casos reales, los digramas ordenados horizontalmente con ganancia de información total superaron la precisión de los monogramas y digramas. Esto nos conlleva a evaluar la mejor opción en un caso real, según se presenten y/o configuren los patrones de clasificación.

5. TRABAJOS FUTUROS

Los trabajos en los que se están ampliando el presente estudio es a otros corpus estandarizados como el RCV1, y RCV2 entre otros, además de aplicar los conceptos básicos a otras tareas de procesamiento de lenguaje natural como la generación de resúmenes y búsquedas semánticas.

En forma paralela se está trabajando en la construcción de un sistema de gestión de proyectos de investigación en la oficina de investigación de la Facultad de Ingeniería de Producción y Servicios de la Universidad Nacional de San Agustín de Arequipa, donde el tema pertinente a este articulo radica en encontrar la similitud de los proyectos que presentan los docentes en los semestres como parte de su actividad no lectiva.

REFERENCIAS

[1] Alessandro Moschitti moschitti, “TEXT CATEGORIZATION Corpora”, [Online] available: http://disi.unitn.it/moschitti/corpora.htm,

[2] Ass K. y Eikvil L. “Text categorization: A survey”, Technical Report 941, Norwegian Computing Center, Noruega, Junio 1999.

[3] David Kember y Lyn Gow, “Action research as a form of staff developnet in higher education, Kluwer Academic Press Publisher, Netherlands”, traducido por Pedro D. Lafourcade del Instituto de perfeccionamiento y Estudios Superiores, Montevideo Uruguay, Mayo 2010, [Online] available: http://ipes.anep.edu.uy/documentos/libre_asis/materiales/Investigacion%20accion.pdf

[4] David D. Lewis, Test Collections - Reuters-21578, [Online] available: http://www.daviddlewis.com/resources/testcollections/reuters21578/

[5] Esperanza Bausela Herreras, “La docencia a través de la investigación–acción”, Universidad Nueva Esparta, Mayo 2010, [Online] available: http://www.une.edu.ve/uneweb2005/servicio_comunitario/investigacion-accion.pdf
[6] F.P. Romero; J.A. Olivas; P.J. Garcés: “A soft Approach to Hybrid Models for Document Clustering”. Proceedings of the Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'06, Paris Les Cordeliers, France. Vol 1, pp. 1040 - 1045, 2006.
[7] Francisco de Borja Navarro Colorado, “Metodología, construcción y explotación de corpus anotados semántica y anafóricamente”, Ph.D. Tesis, Universidad de Alicante, España 2007
[8] Hilda Yelitza Contreras Z., “Procesamiento del Lenguaje Natural basado en una gramática de estilos para el idioma español”, Propuesta Tesis Doctoral, Universidad de los Andes, Facultad de Ingenierías, Postgrado en Computación, 2001
[9] Javier Tejada Cárcamo, “Construcción Automática De Un Modelo De Espacio De Palabras Mediante Relaciones Sintagmáticas Y Paradigmáticas”, Tesis Doctoral, Instituto Politécnico Nacional - Centro de Investigación en Computación, México, D.F., Junio 2009
[10] Joachims T. “Text Categorization with Support Vector Machines: Learning with many relevant features”, 10th European Conference on Machine Learning, Edición 1298, pp 137-142, Dorint-Parkhotel, Chemnitz, Germany 1998
[11] Magnus Sahlgren, “The Word-Space Model - Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces”, Tesis doctoral, Stockholm University Department of Linguistics Computational Linguistics Stockholm, Sweden - National Graduate School of Language Technology Gothenburg University, Gothenburg, Sweden - Swedish Institute of Computer Science Userware Laboratory Kista, Sweden, 2006