MINERÍA DE TEXTO COMO UNA HERRAMIENTA PARA LA BÚSQUEDA DE ARTÍCULOS CIENTÍFICOS PARA LA INVESTIGACIÓN

TEXT MINING AS A TOOL TO FIND SCIENTIFIC PAPERS FOR RESEARCH

Andrés Arias Consuegra*, Yancarlos Mattos Salazar*, Jeison Heredia García* & Diana Heredia Vizcaino**
{dianahv@unisimonbolivar.edu.co}
Universidad Simón Bolívar, Barranquilla-Colombia

Resumen Por medio de la extracción de datos tanto precisos como no relevantes sobre un texto o estructura de datos textual, se es posible generar nueva información y conocimiento, este análisis es realizado con la minería de texto. Si bien esta práctica puede conducir más apropiadamente hacia la generación y/o descubrimiento de nuevos datos, también es viable para la verificación y elaboración de datos a referenciar. Ya que su uso yace en el análisis de la información arrojando datos de manera indiscriminada, esto resulta útil a la hora de rectificar la compatibilidad de la información, entre dos o más medios de contenido textual. Se expondrán y explicarán las diversas maneras como a través de la minería de texto, ayuda en la elaboración de un artículo de investigación o un estado del arte, teniendo en cuenta el tópico a explicar en dicho texto, se utilizarán diferentes metodologías para implementar la minería de texto con el fin de analizar otros diferentes artículos potenciales referencias, teniendo así, un mejor enfoque sobre el material cimiento de un nuevo proyecto o artículo. Adicional a esto se mostraran pruebas realizadas en la herramienta de software Weka donde se evidencian los resultados que arroja el software de los artículos que se analizan por medio de este.

Palabras clave: *Minería de texto, Minería de datos, Estructura de datos*

Abstract By removing both accurate as irrelevant on a text or textual data structure, data is possible to generate new information and knowledge, this analysis is performed with text mining. While this practice can lead to more appropriately the generation and/or discovery of new data, it is also viable for verification and processing of reference data. Since its use lies in the analysis of information indiscriminately throwing data, this is useful when rectifying the compatibility of information between two or more modes of textual content. Will be presented and explained the various ways through text mining helps in the development of a research paper or a state of the art, given the topic to explain in this text, different methodologies are used to implement mining text in order to analyze potential references other than items, thus having a better approach to the material foundation of a new project or item. In addition to this testing in software Weka tool where translations of the software items are analyzed by this it is evident be shown.

Key-words: *Text mining, Data mining, Data structure*

Para referenciar este artículo (IEEE):

Artículo resultado de formación para la investigación

*Estudiante del programa de Ingeniería de Sistemas.

** *Tutora*, Profesora e investigadora del grupo INGEBIOCARIBE

[N] A. Arias, Y. Mattos, J. Heredia & D. Heredia, "Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación", *Investigación y Desarrollo en TIC,* vol. 7, no. 1, pp. xx-xx, 2016.

I. INTRODUCCIÓN

Las obtenciones de las informaciones contenidas en las diversas bases de datos de datos existente y todo tipo de información textual siempre será engorroso las búsquedas de aquellos datos que son de suma importancia para desarrollar de un tema en especial o algún tema de interés, esto se debe a la gran cantidad de texto contenido en este mismo, el cual se tarda mucho tiempo en la búsqueda y la extracción de algunos conceptos. Es por eso que sería útil importancia la implementación de ciertas herramientas que de manera eficaz y segura te permite llegar a ese mismo objetivo de esta manera y de la mano con la tecnología se puede lograr esto mismo. Estas herramienta son las minerías de texto que gracias a la implementación de un software te permite hacer todo esto que antes era engorroso sea algo muy sencillo y con todas las garantías [1].

Imagine un texto que contenga miles de conceptos que siendo todos importante pero solo se quiere uno o unos en especial, sería algo terrible verdad este es el caso de aquéllos usuarios que le gustan la literatura pero muchas veces quieren algún tema específico y no leer todo el contexto para llegar a él, suena ilógico que teniendo grandes herramientas tecnología para la extracción de dichos conceptos se tenga que recurrir a leer gran parte o todo el texto para llegar al objetivo. Es aquí donde la minería de texto hace su trabajo y hace esto algo muy ameno [2].

Este presente trabajo comprende el estudio de análisis de información a través de la minería de texto (*text mining*) también conocida como minería de datos de texto, la cual pueden ser utilizadas para la extracción de información de manera eficaz y precisa desde las diferentes bases de datos cuyos textos no se encuentran estructurados o semi-estructurados.

La importancia de la minería de texto es el descubrimiento de información desconocida, que por medio de la extracción de datos de dicho tema desde diversas fuentes existente.

A través de esta técnica es posible hacer que unas búsquedas sean eficientes con respecto a dicho tema y de esta manera la minería de texto es de suma importancia para el proceso de investigación u aquellas actividades donde se requiera el uso de información.

Esta investigación se realizó mediante la recopilación de datos desde diferentes bases de datos con muy buenas referencias, la cual da como resultados factibles y de mucha confianza en la redacción del contenido, además se contó el apoyo de personas idóneas en el tema de minería de texto.

II. PROCESO DE MINERÍA DE TEXTO

Como se ha estado exponiendo con anterioridad, Tal como la minería de datos puede describirse en términos generales como la búsqueda de patrones en los datos, minería de texto se trata buscando patrones en texto. Sin embargo, la similitud superficial entre las dos oculta verdaderas diferencias. La minería de datos puede ser completamente caracterizada como la extracción de información implícita, desconocida y potencialmente útil a partir de los datos. La información está implícita en los datos de entrada: se refiere a que se encuentra oculta, son datos e información desconocida, y se requieren ciertas técnicas, herramientas, y procesos automáticos de minería de datos para extraer información valiosa o significativa de manera exitosa. Con la minería de texto, sin embargo, la información que se extrae es clara y explícitamente indicado en el texto. No está oculta en todo-la mayoría de los autores ir a grandes esfuerzos para asegurarse de que se expresan de manera clara y sin ambigüedades y, desde un punto de vista humano, el único sentido en el que es "hasta ahora desconocido" es que las restricciones de recursos humanos hacen inviable para la gente a leer el texto sí mismos. El problema, por supuesto, es que la información no se expresa de una manera que es susceptible de procesamiento automático. La minería de texto se esfuerza para sacarlo del texto en una forma que es adecuada para el consumo por equipos directamente, sin necesidad de un intermediario humano [3].

III. SOFTWARE COMO MINERÍA DE TEXTO

Para el proceso de minería de texto, existen diversos software que de manera automática, implementan varias técnicas y metodologías aplicadas a la minería de texto para arrojar determinados resultados, los siguientes son algunos de los más utilizados basados en su desarrollador, funcionamiento y popularidad.

A. SAS Text miner

SAS Text Miner es un *plug-in* para el medio ambiente SAS Enterprise Miner. SAS Enterprise Miner ofrece un amplio

conjunto de herramientas de minería de datos que facilitan el aspecto predicción de minería de texto. La integración de SAS Text Miner en SAS Enterprise Miner combina datos de texto con variables de minería de datos tradicionales. Nodos de minería de texto se pueden incrustar en un diagrama de flujo del proceso SAS Enterprise Miner. SAS Text Miner soporta diversas fuentes de datos textuales: archivos de texto locales, el texto como las observaciones en los conjuntos de datos SAS o bases de datos externas y archivos en la Web [4].

SAS Text Miner 12.1 incluye los siguientes nodos que se pueden utilizar en su análisis de la minería de texto:

- Nodo para importar texto
- Nodo de análisis de texto
- Nodo filtro texto
- Nodo tema texto
- Nodo de clúster de texto
- Nodo texto constructor de regla

Juntos, los nodos Text Miner abarcan el análisis y los aspectos de exploración de la minería de texto y la preparación de los datos para la minería predictiva y una mayor exploración cuando utiliza otros nodos SAS Enterprise Miner. Se podría analizar la información de texto estructurado, y combinar la salida estructurado de los nodos Miner texto con otros datos estructurados si así se desea. Los nodos Text Miner son altamente personalizables y permiten elegir entre una variedad de opciones. Por ejemplo, el nodo de análisis de texto le permite analizar documentos para obtener información detallada acerca de los términos, frases y otras entidades de la colección. El nodo de clúster de texto le permite agrupar los documentos en grupos significativos y que informe los conceptos que usted descubrirá en los racimos. Ordenar, buscar, filtrar (subconjuntos), y la búsqueda de términos o documentos similares en todo mejorar el proceso de exploración [4].

SAS Text Miner también le permite utilizar una macro de SAS que se llama% TMFILTER. Esta macro realiza un paso de pre procesamiento de texto y permite a los conjuntos de datos SAS que se crean a partir de los documentos que residen en el sistema de archivos o en las páginas Web. Pueden existir estos documentos en varios formatos propietarios [4].

B. IBM Analytics

El software IBM Business Analytics ofrece información completa, coherente y precisa que al momento de tomar desiciones, se confía para mejorar el rendimiento de un negocio.

IBM SPSS Modeler Text Analytics ofrece potentes capacidades analíticas de texto, que utilizan tecnologías lingüísticas avanzadas y Procesamiento del Lenguaje Natural (PLN) para procesar rápidamente una gran variedad de datos de texto no estructurados y, a partir de este texto, extraer y organizar los conceptos clave. Además, SPSS Modeler Text Analytics puede agrupar estos conceptos en categorías. Alrededor del 80% de los datos en poder dentro de una organización es en forma de documentos -por ejemplo, informes de texto, páginas web, correos electrónicos y notas de centros de llamadas. El texto es un factor clave para que una organización para obtener una mejor comprensión del comportamiento de sus clientes. Un sistema que incorpora NLP puede extraer de forma inteligente conceptos, incluyendo frases compuestas. Por otra parte, el conocimiento de la lengua subyacente permite la clasificación de los términos en grupos relacionados, tales como productos, organizaciones o personas, utilizando significado y contexto. Como resultado, se puede determinar rápidamente la pertinencia de la información a las necesidades. Estos conceptos y categorías extraídos se pueden combinar con los datos estructurados existentes, como la demografía, y se aplican a modelar en suite completa de IBM SPSS Modeler de herramientas de minería de datos para producir decisiones mejores y más enfocadas.

Sistemas lingüísticos son el conocimiento que tengan en cuanta más información contenidos en diccionarios, mayor será la calidad de los resultados. SPSS Modeler texto Analytics se entrega con un conjunto de recursos lingüísticos, como diccionarios de términos y sinónimos, bibliotecas y plantillas. Este producto permite además a desarrollar y refinar estos recursos lingüísticos a su contexto. Puesta a punto de los recursos lingüísticos a menudo es un proceso iterativo y es necesario para la recuperación concepto exacto y categorización. Las plantillas personalizadas, bibliotecas y diccionarios para dominios específicos, como CRM y genómica, también están incluidos [5].

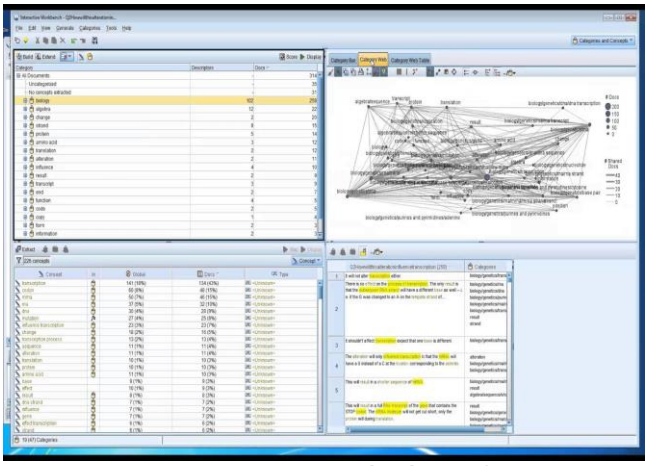

Figura 1. Primer screenshot de minería

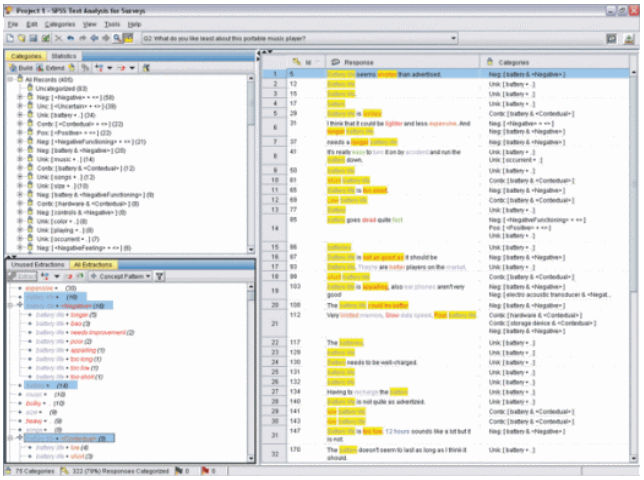

Figura 2. Segundo screenshot de minería

Análisis de texto, una forma de análisis cualitativo, es la extracción de información útil de texto para que las ideas claves o conceptos contenidos en este texto, puedan agruparse en un número apropiado de categorías. El análisis del texto se puede realizar en todos los tipos y longitudes de texto, aunque el enfoque para el análisis variará algo.

Registros o documentos más cortos son los más fáciles de clasificar, ya que no son tan complejas y por lo general contienen menos palabras y las respuestas ambiguas. Por ejemplo, con preguntas abiertas de la encuesta cortas, si se pregunta a la gente a nombrar sus tres actividades vacacionales favoritas, se podría esperar ver muchas respuestas cortas, como ir a la playa, visitar parques nacionales, o no hacer nada. Más aún, respuestas abiertas, por otro lado, puede ser muy complejo y muy largo, sobre todo si los encuestados son educados, motivados, y tienen el tiempo suficiente para completar un cuestionario. Si se le pide a la gente a hablar acerca de sus creencias políticas en una encuesta o tener un canal de blog de política, que se podría esperar algunos largos comentarios sobre todo tipo de temas y posiciones. La capacidad de extraer los conceptos clave y crear categorías interesantes de estas fuentes de texto más largos en un período muy corto de tiempo es una ventaja clave de utilizar IBM SPSS *Modeler Text Analytics.* Esta ventaja se obtiene a través de la combinación de técnicas lingüísticas y estadísticas automatizadas para producir los resultados más fiables para cada etapa del proceso de análisis de texto [5] [4].

La minería de texto se puede decir que es la agrupación, clasificación no supervisada de aquellos patrones (elementos de datos o vectores de características) en grupos. La minería de datos se puede decir que una de las técnicas más importantes y populares que permite el manejo de grandes cantidades de geoespacial de datos. Además cuanta con una gran variedad de tamaños de la información al momento de su procesamiento. La finalidad de la minería de texto es la identificación y seguimientos al procesamiento del lenguaje natural PNL el cual analiza y comprende e incluso genera texto [5] [4].

Lo primero que se debe tener en cuenta la función que cumple la minería de texto es la recuperación de información, es decir, con la extracción le dicha información se hace la construcción de una base de datos la cual permitirá el cotejo de estas mismas. Esto puede ser aplicado en diferentes campos en donde se encuentre gran cantidad de texto, desde mensajes de *Twitter* a una colección de artículos científicos, dependiendo del tema que sea de gran interés.

Luego de esta labor de recopilación, se reconocen entidades nombradas. Esto hace referencia a identificar las partes especificas del texto y todas aquellas pistas que me permitan generar información para llegar al objetivo y descartar aquellas palabras que tiene doble significad diferente a lo que se quiere [12].

La minería de texto es una de las mejores herramientas que permite la extracción de dicha información en un texto de interés de esta manera a través de una aplicación sistematizada se puede lograr obtener información en textos de forma rápida y de manera eficiente se pueden construir conceptos a partir de esta búsqueda minuciosa.

Existe software (aplicaciones) que permiten la trasformación de textos encuesta estructurada en datos cuantitativos y de esa manera se puede obtener una visión utilizando el análisis de sentimiento. Esta técnica clasifica respuestas e integra los resultados obtenidos con otros datos recopilados en la encuesta para una mejor comprensión y análisis de datos estadísticos que le dan mayor asertividad a dichas encuestas [6] [8].

C. Minersoft

Es una aplicación (software) que tiene como objetivo el análisis de la infraestructura de la nube y Grid de forma que se de fácil acceso para los usuario. Este software se apoya en la búsqueda basada en palabras claves, para el desarrollo de esta aplicación de instalaron software en los nodos a gran escala, cuadricula federada y cloud de infraestructuras informáticas.

Este es un gran desafío que surge de la naturaleza no estructurada de un software y carencia de los metadatos relacionados con el software. Esta aplicación se puede decir que es una cosechadora de visita en dicha infraestructura de la red y la nube en la cual recopila una gran información para luego ser cotejada [9].Cada búsqueda tiene sus pro y sus contras pero todo de pende de lo que desee el usuario.

1. Buscando en repositorios de software

El sistema de GURU. GURU adoptó la similitud del coseno métrica para que coincida con las consultas realizadas en todo el contexto de archivos del software, que través de un modelo de probabilidad (cantidad de información).

GURU una de sus funciones es la recuperación o búsqueda de dicho texto través de la vinculación de diferentes páginas de ayuda API cuya información es de mayor calidad que la de otros componentes [9].

2. Extracción de funciones de un cargo usando minería de texto en correos electrónicos

Este software fue creado con una metodología para la extracción de datos, de una gran cantidad de correos electrónicos cuya información son pertenecientes a empresas legales mentes constituidas. Esta tarea es efectuada por dicho software que a través de una herramienta de procesamiento de un lenguaje natural y minería de texto. La cual arrojo como resultado la documentación del 65% de las funciones que son analizada en la fase de evaluación. Este software permite la obtención y recuperación de dicha información de todos aquellos empleados que abandonas su trabaja de manera inesperada, todo esto a través de la minería de texto [10] [11].

3. Metodología CRISP-DM

Este modelo CRISPO-DM fue creado en el año de 1996 en el campo de la minería de Datos. Esta metodología tiene como finalidad la extracción de un subconjunto C de una colección de documentos C, donde todos aquellos documentos tienen la representación de unos correos electrónico.

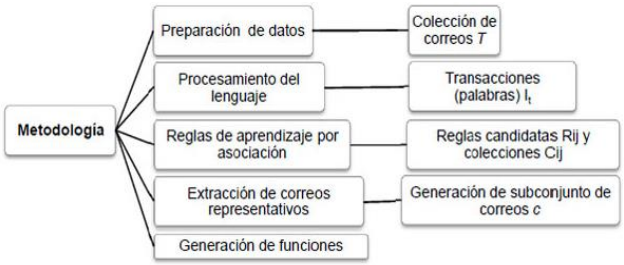

Figura 3. Metodología propuesta para extracción de funciones [11]

- **Preparación de datos**

Tras las obtenciones de los archivos que contengan informaciones de correos electrónicos de todos los periodos que desea obtener dicha información, en donde la base de datos se alimentan del aria de la organización a la cual pertenece cada correo electrónico, de igual forma muestra si dicho correo es de entrada o salida. Esta información se analiza de forma estadísticamente que se aloja en una información temporal de dicho correo que contienen el Mes, Día, Hora, de esta forma realiza un cotejo de estos documentos [11-15].

4. Implementación de la minería de texto con la herramienta Weka

A continuación, se presentarán a manera de pantallazos la forma de utilizar esta herramienta y los resultados que arroja con la minería de algunos artículos.

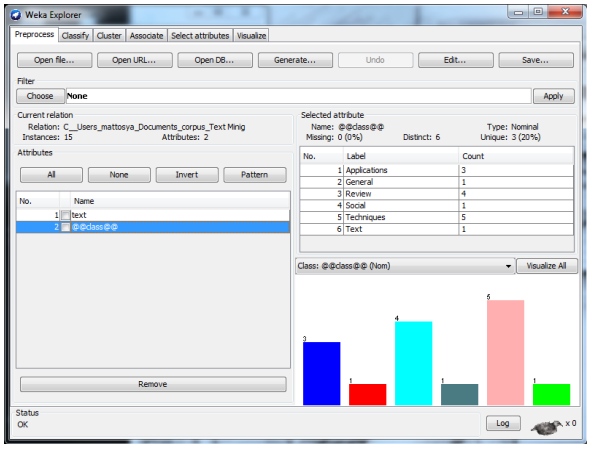

Figura 4. Primer screenshot de Weka Explorer

En este pantallazo muestra el análisis preliminar y a manera de categorización los directorios en formas de tabulación y otros aspectos a tener en cuenta.

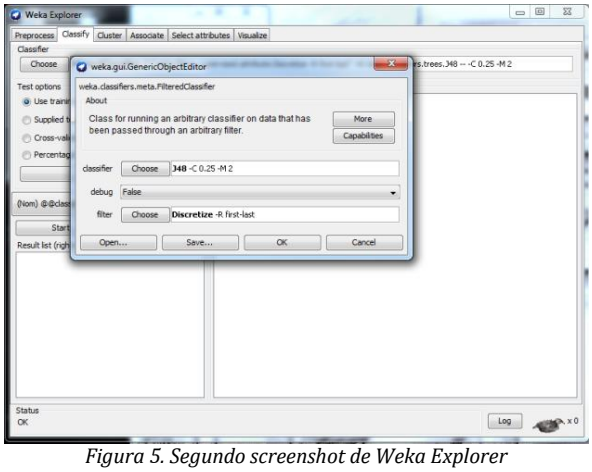

Figura 5. Segundo screenshot de Weka Explorer

Luego se procede a escoger el diagrama que es este caso se eligió el J48 tal como aparece en el pantallazo, y en el cual el filtro será el que expresará las ramificaciones del árbol.

Para el filtro se escogió la opción *StringToWordVector*, que permite la conversión de todos los caracteres.

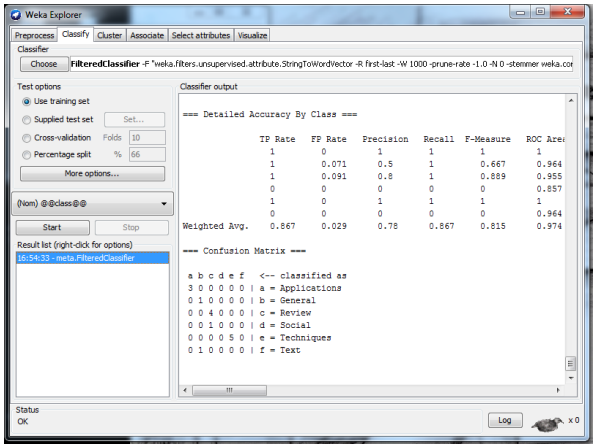

Figura 6. Tercer screenshot de Weka Explorer

Ya realizando los pasos anteriores se procede a realizar el siguiente proceso, que tiene como fin utilizar el FilteredClasiffier, lo que hace este atributo es contar las cantidades de palabras y de esta manera asignarles un valor a las que más se repiten y a su vez permite clasificar los archivos de la matriz y cuales no lograron [16-19].

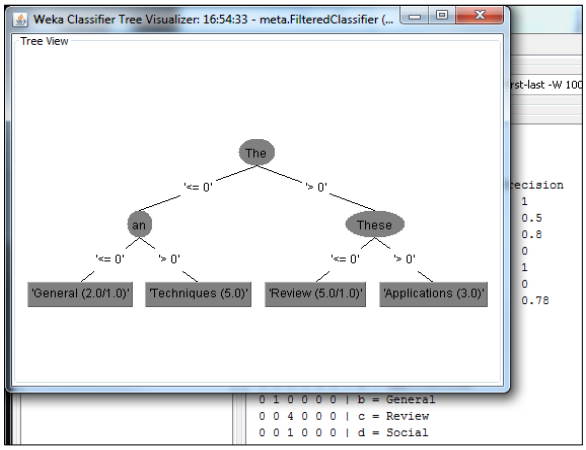

Figura 7. Cuarto screenshot de Weka Explorer

En este pantallazo, el árbol muestras una ramificación con sus respectivos conectores ya a manera de minería de texto.

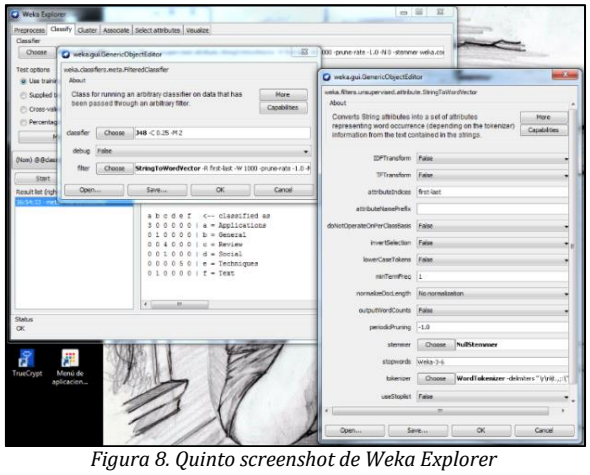

Figura 8. Quinto screenshot de Weka Explorer

Luego se procede a configurar el clasificador, esto se puede hacer dando click en el filtro el cual te permite establecer el tipo y la forma de los conectores de la ramificación del árbol de decisiones.

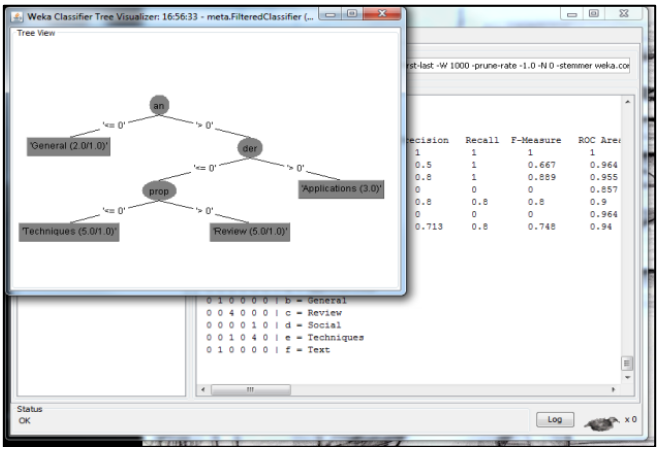

Figura 9. Sexto screenshot de Weka Explorer

De esta forma quedaría el árbol y sus ramificaciones después de la previa configuración.

IV. CONCLUSIONES

En vista del avance tecnológico, el cual poco a poco acoge en su manto, diferentes campos de estudio, es lógico afirmar que con la ayuda del desarrollo tecnológico, las invenciones y descubrimientos se componen y producen progresivamente de manera exponencial. Cuando se desease relatar, expresar, explicar o mostrar una propuesta, esta debe estar soportada en archivos y trabajos previos verídicos, siendo así entonces, una forma de colaborar con el progreso del estudio, la implementación de la minería de texto a un punto analítico, donde con su metodología y técnica, se puedan tomar decisiones más acertadas sobre qué tipo de material será cimiento o soporte de un trabajo o propuesta venidera.

Aplicando la minería de texto, se analizarán diferentes estructuras de datos que comúnmente se usan como sustento argumentativo en los trabajos o redacciones, de tal forma que se pueda conocer que tan acertado y útil es el soporte en análisis, siendo así, una referencia de carácter significativo, donde su contenido puede ser utilizado para generar un descubrimiento o nuevo conocimiento con más precisión.

V. REFERENCIAS

[1] S. Neha K, Introduction of Text mining and an Analysis of Text mining Techniques. PARIPEX. Vol. 2. No. 2. pp. 56-57. 2012.

[2] R. Lau, S. Liao, R. Kwok, K. Xu, Y. Xia and Y. Li, 'Text mining and probabilistic language modeling for online review spam detection', ACM Trans. Manage. Inf. Syst., Vol. 2. No. 4. pp. 1-30, 2011.

[3] Support.sas.com, 'SAS Text Miner', 2015. [Online]. Available: http://support.sas.com/software/products/txtminer/#s1=1. [Accessed: 17- Nov- 2015].

[4] S. Surveys, 'IBM - SPSS Text Analytics for Surveys', Www-03.ibm.com, 2015. [Online]. Available: http://www-03.ibm.com/software/products/en/spss-text-analytics-surveys. [Accessed: 17- Nov- 2015].

[5] M. Dikaiakos, A. Katsifodimos and G. Pallis, 'Minersoft', TOIT, vol. 12, no. 1, pp. 1-34, 2012.

[6] S. VijayGaikwad, A. Chaugule and P. Patil, 'Text Mining Methods and Techniques', International Journal of Computer Applications, vol. 85, no. 17, pp. 42-45, 2014.

[7] R. Lau, S. Liao, R. Kwok, K. Xu, Y. Xia and Y. Li, 'Text mining and probabilistic language modeling for online review spam detection', ACM Trans. Manage. Inf. Syst., vol. 2, no. 4, pp. 1-30, 2011.

[8] R. Chau and C. Yeh, 'A multilingual text mining approach to web cross-lingual text retrieval', Knowledge-Based Systems, vol. 17, no. 5-6, pp. 219-227, 2004.

[9] X. Xiao, H. Xu and S. Xu, 'Using IBM SPSS modeler to improve undergraduate mathematical modelling competence', Comput Appl Eng Educ, vol. 23, no. 4, pp. 603-609, 2015.

[10] V. Reitano, 'Book Review: Review of SAS Enterprise Miner Textbooks', Social Science Computer Review, vol. 33, no. 3, pp. 415-417, 2014.

[11] C. Aggarwal, 'Mining text and social streams', SIGKDD Explor. Newsl., vol. 15, no. 2, pp. 9-19, 2014.

[12] M. Wauer, J. Meinecke, D. Schuster, A. Konzag, M. Aleksy and T. Riedel, 'Semantic Federation of Product Information from Structured and Unstructured Sources', International Journal of Business Data Communications and Networking, vol. 7, no. 2, pp. 69-97, 2011.

[13] C. P. Sumathi and N. Priya, 'Analysis of an Automatic Text Content Extraction Approach in Noisy Video Images', International Journal of Computer Applications, vol. 69, no. 4, pp. 6-13, 2013.

[14] D. Fagan, 'Emily Dickinson's Unutterable Word', The Emily Dickinson Journal, vol. 14, no. 2, pp. 70-75, 2005.

[15] A. Hassan, A. Abu-Jbara, W. Lu and D. Radev, 'A Random Walk "Based Model for Identifying Semantic Orientation', Computational Linguistics. Vol. 40. No. 3. pp. 539-562. 2014.

[16] A. Neet and H. Singh, 'Web Data Mining: Survey', IJETT, vol. 10, no. 3, pp. 144-147, 2014.

[17] K. Felizardo, G. Andery, F. Paulovich, R. Minghim and J. Maldonado, 'A visual analysis approach to validate the selection review of primary studies in systematic reviews', Information and Software Technology. Vol. 54. No. 10. pp. 1079-1091, 2012.

[18] B. Londoño González and P. Sánchez, "Algoritmo Novedoso Para la Detección de Tareas Repetitivas en el Teclado", Investigacion e Innovación en Ingenierias, vol. 3, no. 2, 2015. DOI: 10.17081/invinno.3.2.2031

[19] M. Jimeno, Y. De la Hoz and J. Wilches, "Wireless ECG and PCG Portable Telemedicine Kit for Rural Areas of Colombia", Investigación e Innovación en Ingenierías, vol. 2, no. 2, 2014.DOI: 10.17081/invinno.2.2.2044