TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE CD.VICTORIA

SECRETARÍA DE EDUCACIÓN PÚBLICA

SEP

TecnoINTELECTO

Órgano de Divulgación Científica

Una Publicación del Instituto Tecnológico de Cd. Victoria

Volumen 14 No. 1 Julio 2017 ISSN 1665-983X

INGENIERÍA Y TECNOLOGÍA

Estudio de herramientas de minería de datos para la tarea de clasificación. *H. M. Marin-Castro & P. E. Franco-Vázquez……..1*

Towards a definition and validation of a serious game evaluation process. *L. García-Mundo, J. Vargas-Enríquez, S. Martínez-Guerra, M. Genero & M. Piattini……10*

Análisis del coeficiente de transferencia de calor en las propiedades mecánicas y esfuerzos internos durante el temple en aceros de medio carbono utilizando FEM*. R. D. López-García, A. Maldonado-Reyes, M. A. Jiménez-García, C.E. López-García & J.A. Maldonado-Zúñiga……..…19*

Review of data integration architectures and their methods. *O. D. Fernández-Bonilla, M. González-García & R. Santaolaya-Salgado……..……..….27*

Modelos para la clasificación de frases clave en textos científicos. *G. Flores-Petlacalco, M. Tovar-Vidal, J. A. Reyes-Ortiz & A. P. Cervantes-Márquez……40*

Algoritmos para detectar la calidad de servicio en los dominios de restaurantes y laptops. *K. L. Vázquez-Flores, M. Tovar-Vidal, H. Castillo-Zacatelco & M. Rossainz-López…..……..51*

TecnoINTELECTO (ISSN 1665-983X y reserva: 04-2004-072626452400-102) es un órgano de divulgación científica de forma semestral del Tecnológico Nacional de México-Instituto Tecnológico de Cd. Victoria. Boulevard Emilio Portes Gil No. 1301, C. P. 87010, Cd. Victoria, Tamaulipas, México; Tels. (834) 153 20 00 Ext. 325. El contenido y la sintaxis de los artículos presentados son responsabilidad del autor (es). Editor Principal: División de Estudios de Posgrado e Investigación. Apoyo editorial-**informático: M.C. Aurora Y. Rocha Sánchez. Envío de documentos, consultas y sugerencias al correo electrónico: ludivinab@yahoo.com**, almagavetec@hotmail.com .Todos los derechos son reservados y propiedad del Tecnológico Nacional de Méxco-Instituto Tecnológico de Cd. Victoria-Sistema Nacional de Educación Superior Tecnológica. TecnoINTELECTO, 2017, Vol. 14 No. 1. Cd. Victoria, Tamaulipas, México.

Consúltanos en el Índice Latinoamericano www.latindex.org y en el Índice de Revistas Latinoamericanas en Ciencias PERIÓDICA www.dgb.unam.mx/periodica.html

ESTUDIO DE HERRAMIENTAS DE MINERÍA DE DATOS PARA LA TAREA DE CLASIFICACIÓN

H.M. Marin-Castro & P.E. Franco-Vázquez

Universidad Politécnica de Victoria, Av. Nuevas Tecnologías 5902, Parque Científico y Tecnológico de Tamaulipas, C.P. 87138, Cd Victoria, Tamaulipas, México.
hmarinc@upv.edu.mx , 12301478@upv.edu.mx

RESUMEN. En la actualidad existen diversas herramientas de minería de datos (HMDs) disponibles para utilizarse en diferentes aplicaciones. Sin embargo, algunas de estas herramientas tienen características limitadas, lo que ocasiona la construcción de modelos de regresión, clustering, clasificación o predicción con un desempeño variable. El objetivo de este trabajo de investigación es evaluar el desempeño de los modelos de clasificación construidos a partir del uso de algoritmos de clasificación provistos por HMDs. Las herramientas R y RapidMiner fueron seleccionadas por su popularidad y amplio uso en la academia e investigación. Para conocer el desempeño de los modelos de clasificación construidos se realizó una evaluación experimental de los algoritmos de clasificación Máquinas de Soporte Vectorial (SVM), Árboles de Decisión, RandomForest y K-Vecinos más cercanos (KNN), los cuales están disponibles en las dos herramientas. Estos algoritmos se evaluaron bajo las mismas condiciones en una aplicación web usando bases de datos de diabetes y cáncer de mama. Los resultados experimentales revelan que los algoritmos de clasificación Árboles de decisión y RandomForest usando la herramienta R ofrecen la mejor exactitud y la menor tasa de error utilizando las bases de datos de diabetes y cáncer de mama, respectivamente. Mientras que KNN y RandomForest tienen mejores resultados usando la herramienta RapidMiner.

PALABRAS CLAVE: HMDs, algoritmos de clasificación, diabetes, cáncer de mama, métricas de evaluación.

ABSTRACT. Currently, there are several datamining tools (HMDs) available for use in different applications. However, some of these tools have limited features, which leads to the construction of regression, clustering, classification or prediction models with variable performance. The objective of this research work is to evaluate the performance of the classification models constructed from the use of classification algorithms provided by HMDs. The R and RapidMiner tools were selected for their popularity and extensive use in academia and research. The performance of classification models build in this study was assessed by executing an experimental evaluation of classifiers Vector Support Machines (SVM), Decision Trees, RandomForest and K-Neighbors (KNN), which are included in the two datamining tools under study. The classification algorithms were evaluated under the same conditions in a web application using databases of diabetes and breast cancer. The experimental results reveal that the Decision Trees algorithm and RandomForest using the R tool offer the best accuracy and the lowest error rate using the databases of diabetes and breast cancer, respectively. While KNN and RandomForest perform better using the RapidMiner tool.

KEYWORDS: HMDs, classification algorithms, diabetes, breast cancer, evaluation metrics.

1. INTRODUCCIÓN

En la actualidad existe una gran cantidad de información almacenada en numerosas bases de datos. Sin embargo, trabajar con herramientas convencionales para el manejo de información contenida en dichas fuentes es limitado ya que la mayoría no incluye un proceso de análisis de la información y extracción de nuevo conocimiento. Hoy en día existen nuevas técnicas y disciplinas para extraer información útil a partir de distintas fuentes de información. Esto permite tomar decisiones importantes sobre un tema en particular. Estas técnicas han sido implementadas en herramientas construidas con el objetivo de aplicar la minería de datos o el proceso de descubrimiento de conocimiento (KDD). Las herramientas de minería de datos (HMDs) son también llamadas “máquinas de

aprendizaje", ya que permiten hacer análisis y evaluación de los datos (Han et al., 2011). La mayoría de las HDMs incluyen el manejo del proceso KDD. Este proceso consta de una serie de etapas a seguir para generar nuevo conocimiento: Selección de datos, Pre-procesamiento, Transformación, Minería e Interpretación/Evaluación. Las HMDs incluyen técnicas para cada una de las etapas del proceso KDD.

En este trabajo se realizó un estudio de las distintas HMDs existentes en el mercado para seleccionar de entre estas, aquella que tenga un alto desempeño con el uso de técnicas de clasificación. Las técnicas de clasificación consideradas en este trabajo fueron RandomForest, K-Vecinos más cercanos (KNN), Máquinas de soporte vectorial (SVMs) y Árboles de decisión. Estas técnicas son ampliamente conocidas y resulta fácil comprender su funcionamiento. Para poder comprobar el funcionamiento de las HMDs en la tarea de clasificación de datos se realizó una investigación sobre dos de las enfermedades predominantes dentro y fuera del país (diabetes y cáncer de mama). A partir de esta investigación se llevó a cabo un proceso de integración de las distintas fuentes de datos (bases de datos médicas) disponibles y relacionadas con las dos enfermedades. Una vez integradas las fuentes de datos se le aplicaron técnicas de pre-procesamiento de datos (transformación de datos) para posteriormente construir modelos de clasificación que sirvieran para predecir la clase (tiene o no la enfermedad) de nuevas instancias o ejemplos de pacientes.

Finalmente se desarrolló una aplicación web como herramienta de apoyo en la toma de decisiones. Esta herramienta permite cargar el modelo de clasificación construido y recibir nuevos datos o instancias para determinar a la clase a la que pertenecen. En nuestro caso, la herramienta permite predecir si el paciente, dadas sus características, puede o no tener diabetes o cáncer de mama según sea el caso. Este artículo está organizado de la siguiente manera: En la Sección 2 se describe el marco teórico de este trabajo, definiendo algunas de las técnicas de clasificación y métricas de evaluación utilizadas. La Sección 3 presenta la revisión realizada sobre las distintas HMDs. La Sección 4 muestra y discute la evaluación experimental de las HMDs seleccionadas en la tarea de clasificación. Finalmente la Sección 5 presenta las conclusiones de este trabajo.

2. MARCO TEÓRICO

Dentro de las técnicas de clasificación supervisadas que se utilizaron en este trabajo, están RandomForest, KNN, SVM y Árboles de Decisión, las cuales se describen a continuación.

2.1 RANDOMFOREST

RandomForest (Han et al., 2011) es un algoritmo predictivo que usa la técnica de Bagging para combinar diferentes árboles, donde cada árbol es construido con observaciones y variables aleatorias. En forma resumida, RandomForest sigue este proceso:

- Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes conjuntos de datos.
- Crea un árbol de decisión con cada conjunto de datos, obteniendo diferentes árboles, ya que cada conjunto contiene diferentes individuos y diferentes variables.
- Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (sin podar).
- Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los árboles predicen la observación como positiva y como "negativo" si la mayoría de los árboles predicen la observación como negativa.

2.2 KNN

KNN es un algoritmo de clasificación supervisado (como SVM). La idea principal de este algoritmo es que las nuevas instancias se clasifican en la clase más frecuente de sus K vecinos más próximos (Han et al., 2011). KNN pierde precisión si se utilizan datos ruidosos o con atributos irrelevantes. Si se utiliza KNN con datos categóricos, el algoritmo devuelve la categoría a la cual debería pertenecer la instancia desconocida. Si se utiliza KNN con datos continuos, el algoritmo devuelve la media de los valores de los vecinos. El uso de KNN en clasificación se fundamenta en el uso del voto (mayoría) para decidir el valor más adecuado. El voto puede ser con pesos o sin ellos. Si la clasificación es binaria, entonces es

preferiblemente elegir k impar para evitar empates.

2.3 SVM

Las Máquina de Soporte Vectorial (SVM) (Han et al., 2011) son estructuras de aprendizaje basadas en la teoría estadística del aprendizaje. Las SVM son aplicadas a resolver tareas tanto de clasificación como de regresión. Se basan en transformar el espacio de entrada en otro espacio de dimensión superior (infinita) en el que el problema puede ser resuelto mediante un hiperplano. Las SVM trabajan de la siguiente manera:

- Dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría se desconoce) pertenece a una categoría o a la otra.
- Como en la mayoría de los métodos de clasificación supervisada, los datos de entrada (los puntos) son vistos como un vector p-dimensional (una lista ordenada de *p* números).
- La SVM busca un hiperplano (ver Figura 1) que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior.

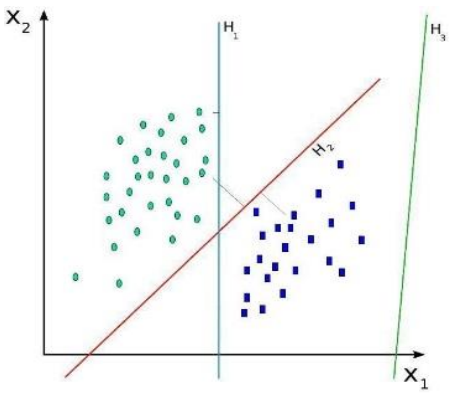

Figura 1. Separación de datos.

2.4 ÁRBOLES DE DECISIÓN

Es un modelo de predicción que representa en forma secuencial, condiciones y acciones (ver Figura 2). Los árboles de decisiones sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

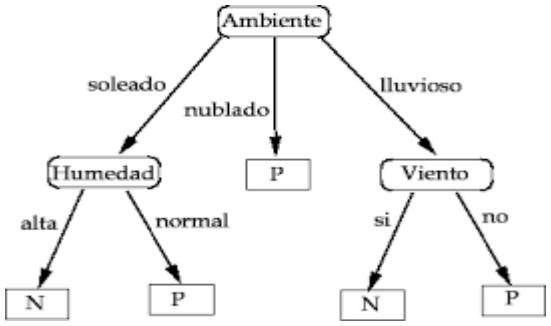

Figura 2. Árbol de decisión.

Los árboles proveen una visión gráfica de la toma de decisión necesaria, especifican las variables que son evaluadas, qué acciones deben ser tomadas y el orden en la cual la toma de decisión será efectuada. Cada vez que se ejecuta un árbol de decisión, solo un camino será seguido dependiendo del valor actual de la variable evaluada. Características continuas (reales) pueden ser clasificadas al permitir nodos que dividan una característica real en dos rangos. Pueden manejar ruido en datos de entrenamiento. Los árboles de decisión se utilizan en el diagnóstico médico, análisis de riesgo en crédito, clasificador de objetos para manipulador de robot, entre otras aplicaciones.

2.5 MÉTRICAS DE EVALUACIÓN DE TÉCNICAS DE CLASIFICACIÓN

La evaluación de un algoritmo de clasificación se puede realizar atendiendo a distintos aspectos:

- Precisión (porcentaje de casos clasificados correctamente).
- Eficiencia (tiempo necesario para construir/usar el clasificador).
- Robustez (frente a ruido y valores nulos).
- Escalabilidad (utilidad en grandes bases de datos).
- Interoperabilidad (el clasificador no es sólo una caja negra).
- Complejidad (del modelo de clasificación).

Para evaluar el rendimiento de las técnicas de clasificación se puede aplicar la matriz de

confusión. Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de la matriz de confusión es que facilita observar si el sistema está confundiendo dos clases. La tabla en la Figura 3 muestra la matriz de confusión para un clasificador de dos clases.

		Clase real	
		Clase referencia	***Clase no referencia***
Clase estimada	Clase referencia	TP	FP
	Clase no referencia	FN	TN

Figura 3. Matriz de confusión.

En la Figura 3:

- TP representa verdaderos positivos, esto es, casos correctamente clasificados y que pertenecen a la clase de referencia.
- TN representa verdaderos negativos, es decir, casos correctamente clasificados y que pertenecen a la clase no referencia.
- FP se refiere a falsos positivos, estos es, casos clasificados como pertenecientes a la clase referencia, pero su clase real es la clase no referencia.
- FN indica falsos negativos, es decir, casos clasificados como pertenecientes a la clase no referencia, pero su clase real es la clase referencia.

Una vez obtenida la matriz de confusión del modelo de clasificación se pueden aplicar métricas como la *Exactitud* y la *Tasa de Error* obtenidas a partir de la matriz para evaluar la calidad del clasificador. La *Exactitud* es una métrica para evaluar la efectividad del clasificador, expresada como el número de instancias clasificadas correctamente más el número de instancia que no debieron y no son clasificadas respecto al total de casos evaluados. La medida de *Exactitud*, no distingue entre el número de etiquetas correctas de diferentes clases (Ver Ecuación1).

$$Exactitud = \frac{tp+tn}{tp+tn+fp+fn} \quad (1)$$

También existen otras métricas para evaluar el rendimiento de los clasificadores como: la *Precisión* (precisión), el *Recuerdo* (recall) y la medida-F (f-measure) que se definen a partir de las ecuaciones 2, 3 y 4 respectivamente.

$$Precision = \frac{tp}{tp+fp} \quad (2)$$

$$Recuerdo = \frac{tp}{tp+fn} \quad (3)$$

$$Medida - F = 2 * \frac{Precision * Recuerdo}{Precision + Recuerdo} \quad (4)$$

3. REVISIÓN DE CARACTERÍSTICAS DE LAS HMDs

En la actualidad existen diversas HMDs que permiten hacer análisis sobre diferentes fuentes de información para generar nuevo conocimiento. En la Tabla 1 se muestra algunas de las HMDs más comúnmente conocidas y utilizadas en proyectos reales de Big Data, Data mining y Data Science. Así, también en la Tabla 1 se muestra el porcentaje de uso de las HMDs entre los usuarios en una encuesta realizada a 1880 votantes entre los años 2012 y 2013. Los resultados de esta encuesta fueron tomados del sitio web tecnológico KDnuggets (Piatetsky-Shapiro, 2013).

Tabla 1. Uso de HMDs gratuitas y disponibles entre la comunidad de usuarios.

Herramientas Open Source (Número de votantes)	Porcentaje de usuarios en 2013	Porcentaje de usuarios en 2012
RapidMiner/ RapidAnalytics free edition (737)	39.2	26.7
R (704)	37.4	30.7
Weka / Pentaho (269)	14.3	14.8
Python with any numpy/scipy/pandas/i Python packages (250)	13.3	14.9
KNIME free edition (110)	5.9	21.8
Rattle (84)	4.5	--
Orange (67)	3.6	5.3
Other free analytics/ data mining software (64)	3.4	4.9

GNU Octave (54)	2.9	--
Revolutions Analytics R free edition (46)	2.4	--
C4.5 /C5.0 (21)	1.1	1.6
F# (14)	0.7	0.6

De acuerdo a los resultados mostrados en la Tabla 1, se puede observar que dos de las herramientas que han tenido éxito en los últimos años, entre los usuarios son las herramientas RapidMiner y R, ya que ocupan los primeros lugares de uso. Por un lado, RapidMiner es una HMD disponible de forma gratuita y de código abierto, que cuenta con una interfaz de usuario muy útil y fácil de manejar. En esta herramienta los usuarios no tienen que escribir códigos ya que cuentan con muchas plantillas y otros elementos que permiten analizar los datos de forma sencilla. Esta herramienta permite realizar un pre-procesamiento de datos, hacer análisis predictivo utilizando varios clasificadores, así como construir modelos estadísticos. RapidMiner permite convertir los datos en acciones. Además, de ayudar a predecir los resultados futuros utilizando varios algoritmos de minería de datos y aprendizaje automático.

La versión original de RapidMiner fue desarrollada en java por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001 (Lee et. al., 2006-2016). En contra posición, R es una HMD y análisis estadístico que se utiliza por línea de comandos. El lenguaje R es muy utilizado por estudiantes y profesores para la tarea de investigación (Chambers et al., 2016). R-software es una herramienta popular de minería de datos, de código abierto. Cuenta con una serie de módulos y funciones predefinidas. Los usuarios tienen que escribir guiones para sus operaciones. R es una gran herramienta para gráficos estadísticos. También proporciona elementos necesarios para el modelado lineal y no lineal, pruebas estadísticas, clasificación y agrupación. En la Tabla 2 se muestran algunas de las características principales (desarrollador, lenguaje de programación, licencia, versión actual, línea de comandos, propósito principal, soporte y sitio web) de cinco de las HMDs más utilizadas de acuerdo a la encuesta de KDnuggets.

Tabla 2. Características principales de las Herramientas de Minería de Datos.

Carac terísticas	RapidMiner	R	Weka	Orange	KNIME
Desarrollador	RapidMiner, Alemania	Desarollador De la wordwide	Universidad De Waikato, Nueva Zelanda	Universidad de Ljubljana, Eslovenia	KNIME.com AG,Suiza
Lenguaje de Programación	Java	C, Fortran, R	Java	C++.Python, Qt framew	Java
Licencia	Open s. (v.5 or lower); dosed s., free Starter ed. (v.6)	Software libre, GNU GPL 2+	open source, GNU GPL 3	open source, GNU GPL 3	open source, GNU GPL 3
Versión Actual	6	3.02	3.6.10	2.7	2.9.1
GUI/Línea de comandos	GUI	Ambos	Ambos	Ambos	GUI
Propósito principal	Minería de datos en general	Computación y estadística	Minería de datos en general	Minería de datos en general	Minería de datos en general
Soporte	Para 200,000 usuarios	Muy grande para 2000,000 usuarios	Grande	Moderado	Moderado más de 15,000 usuarios
Sitio Web	https://rapidminer.com/	http://www.rdatamining.com/package	http://www.cs.waikato.ac.nz/ml/weka/	http://orange.biolab.si/	https://www.knime.org/

De acuerdo a la Tabla 2, se puede observar que cuatro de las cinco HMDs fueron desarrolladas utilizando el paradigma orientado a objetos, a excepción de R, el cual fue desarrollado en C, Fortran y el lenguaje R. Esta herramienta fue creada con el propósito de apoyar en cálculos estadísticos, lo cual permite obtener modelos más precisos. Además, el soporte de usuarios soportado por R es superior del resto de las HMDs.

4. EVALUACIÓN EXPERIMENTAL

En esta sección se presenta la evaluación experimental de los algoritmos de clasificación SVM, Árboles de Decisión, RandomForest y KNN provistos por dos de las HMDs más populares, R y RapidMiner. Las fuentes de datos que se utilizaron para realizar la evaluación del desempeño de los algoritmos de clasificación fueron bases de datos obtenidas del repositorio UCI Machine Learning (Lichman, 2013), que es un repositorio que cuenta con más de 300 bases de datos, donadas por diferentes instituciones que realizan estudios de investigación para diferentes áreas como la medicina.

En este trabajo se utilizaron las bases de datos médicas Pima Indians (Diabetes) y BreastCancerData (Cáncer de Mama), tomando en cuenta que tanto la diabetes como el cáncer de mama son dos de las principales enfermedades responsables de la muerte de miles de personas (INEGI, 2016). En México como en Estados Unidos de América (USA, por sus siglas en inglés) la diabetes es una enfermedad grave que ataca a jóvenes como adultos y en ambos países es una de las principales causas de defunciones (OMS, 2016).

También el cáncer de mama es una de las principales causas de muerte en ambos países y afecta en la mayoría de los casos a las mujeres (INEGI, 2015), (ENSANUT, 2014). Ambas enfermedades ocupan los primeros lugares de defunción en México y USA (OMS, 2016).

Tabla 3. Descripción de las Bases de Datos Pima Indians y BreastCancerData.

Características	Pima Indians	BreastCancer Data
# Instancias	768	699
# Atributos	9	10
Tipo de atributo	Numéricos	Numéricos
Clase	0,1	2=Benigno, 4=Maligno
Valores faltantes	Si	Si
Atributos	Número de embarazos, Concentración de glucosa, Presión en la sangre, Triceps, Insulina, Índice de masa corporal, Función pedigree, Edad, Clase (0 or 1)	Grueso del grupo, Uniformidad del tamaño de la célula, Uniformidad de la Forma Celular, Adhesión marginal, Tamaño de célula epitelial única, Núcleos, Cromatina suave, Nucleoli.

En la Tabla 3 se presentan las características de las bases de datos utilizadas, Pima Indians y BreastCancerData. En la evaluación se utilizó la matriz de confusión como herramienta para visualizar el desempeño de un clasificador. A partir de la matriz de confusión se obtuvieron las métricas de *Exactitud* y la *Tasa de Error*. Al realizar la evaluación de los clasificadores se puede deducir el desempeño de los modelos construidos sobre las bases de datos utilizadas.

Las Tablas 4-7 muestran el desempeño obtenido por los cuatro clasificadores RandomForest, KNN, Árboles de decisión y SVM disponibles en las herramientas RapidMiner y R.

Tabla 4. Desempeño de los clasificadores utilizando Pima Indians en RapidMiner.

	RapidMiner	
Algoritmo	**Exactitud**	**Tasa de Error**
RandomForest	70.34%	29.57%
KNN	100%	0%
Decision Tree C4.5	73.21%	26.79%
SVM	75%	25%

Tabla 5. Desempeño de los clasificadores utilizando Pima Indians en R.

	R	
Algoritmo	**Exactitud**	**Tasa de Error**
RandomForest	75.60%	24.40%
KNN	74.44%	25.56%
Decision Tree C4.5	**100%**	0%
SVM	79.45%	20.55%

Tabla 6. Desempeño de los clasificadores utilizando BreastCancer en RapidMiner.

	RapidMiner	
Algoritmo	**Exactitud**	**Tasa de Error**
RandomForest	94.76%	5.24%
KNN	94.56%	5.44%
Decision Tree C4.5	92.86%	7.14%
SVM	74.61%	25.39%

Tabla 7. Desempeño de los clasificadores utilizando BreastCancer en R.

	R	
Algoritmo	**Exactitud**	**Tasa de Error**
RandomForest	96.96%	3.04%
KNN	95.93%	4.07%
Decision Tree C4.5	70.20%	29.80%
SVM	96.35%	3.65%

Para la evaluación de los algoritmos utilizando la herramienta R se obtuvo una exactitud promedio a partir de la ejecución de 10 veces cada clasificador, ya que la selección de las instancias para los datos de entrenamiento y prueba se realiza mediante muestreo aleatorio.

De los resultados presentados en las Tablas 4 y 5, se muestra una variabilidad en la exactitud obtenida por los diferentes modelos construidos utilizando los mismos clasificadores y base de datos pero diferente HMD. En la Tabla 4 el modelo KNN resulto con el mejor desempeño. Sin embargo en la Tabla 5 el modelo de árbol de decisión resulta obtener mayor exactitud. Considerando que R en la mayoría de los casos obtiene mejores resultados en la construcción de modelos con un clasificador distinto comparado con RapidMiner. Por otro lado, en las Tablas 6 y 7 el mejor desempeño siempre se mantiene en el modelo construido con el algoritmo RandomForest. Esto posiblemente a la correcta discriminación existente entre las clases de la base de datos BreastCancer. De las pruebas realizadas se puede concluir que R es una herramienta que ofrece una mayor exactitud y confianza en la construcción de modelos de clasificación en comparación con otras HMDs. Por otro lado RadipMiner es una herramienta amigable. Sin embargo, los modelos construidos no siempre resultan tener el mejor desempeño en comparación con R.

4.1 ENTORNO DE DESARROLLO DE LA APLICACIÓN WEB

Para el proceso de evaluación de los algoritmos de clasificación se creó una aplicación web que sirvió de apoyo para observar los resultados por los clasificadores construidos en cada HMD. Además, la aplicación web permite a profesionales médicos ingresar datos de pacientes y predecir si los pacientes tienen o no diabetes o cáncer de mama de acuerdo a las características de dichos pacientes. Esto con el objetivo de ayudar en la toma de decisiones. Por ejemplo, para determinar el tratamiento que se utilizará para el paciente. La aplicación web está dividida en dos módulos: a) módulo de evaluación de clasificadores y módulo de predicción de enfermedad. Para desarrollar la aplicación se utilizaron las siguientes herramientas:

El entorno de desarrollo que se utilizó fue IDE NetBeans. Para el diseño de la aplicación web, se utilizó el framework MATERIALIZE CSS (Google, 2014-2016), el cual cuenta con hojas de estilo en cascada para cada uno de los componentes HTML. Este framework simplifica el proceso de creación de diseños web, ofrece una serie de recursos en CSS, fuentes, JS así como permite integrar JQuery. JQuery es una librería que permite dar dinamismo a las páginas web (Resign, 2016), por ejemplo, para utilizar un elemento “select”, se utiliza JQuery para aplicar el movimiento de selección.
Para realizar la conexión de la aplicación web con R se utilizó la tecnología JavaServer Pages (JSP), ya que permite incrustar código Java en

HTML. También se utilizó GlassFish Server 4.1 como servidor de aplicaciones integrado con NetBeans. Además, se utilizaron otras librerías como: REngine y RserveEngine para la conexión entre R y JSP, jfreeChart, jcommon para los gráficos y mysql-connector para la conexión de la BD entre JSP y phpMyAdmin. En la Figura 4 se muestra la interfaz gráfica principal de la aplicación web construida, a partir de la cual el usuario inicia sesión para posteriormente acceder a los módulos: a) evaluación y visualización del desempeño de los clasificadores, b) carga de datos de los nuevos pacientes y c) predicción de la clase de nuevos registros de pacientes.

Figura 4. Interfaz Principal.

El módulo de evaluación y visualización de clasificadores permite mostrar el rendimiento de los algoritmos. También en este módulo se genera un gráfico de barras en 3D del desempeño (exactitud) de los cuatro algoritmos utilizados por R, como se muestra en la Figura 5. A través de código R se obtienen los valores de las métricas de evaluación anteriormente mencionadas.

Figura 5. Gráfico de barras del desempeño de los clasificadores.

El módulo de carga de datos de pacientes permite ingresar datos de nuevos pacientes (número de embarazos, concentración de glucosa, presión, etc.) que no han sido clasificados con alguna de las enfermedades antes mencionadas. Este módulo (ver Figura 6) está diseñado para ser utilizado por profesionales médicos.

Figura 6. Formulario de carga de datos de nuevos pacientes.

Después de llenar el formulario es posible realizar la predicción utilizando alguno de los modelos de clasificación previamente construidos con las bases de datos Pima Indians y BreastCancerData (ver Figura 7).

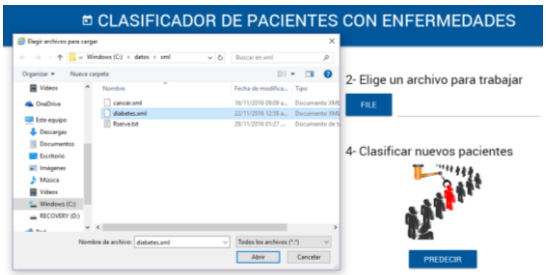

Figura 7. Selección de modelo de clasificación.

Por último, a partir de la aplicación web es posible visualizar los resultados y obtener un resultado por parte del clasificador con el mejor desempeño obtenido como el que se muestra en la Figura 8.

Figura 8. Resultado del proceso de predicción de clase del nuevo paciente.

5. CONCLUSIONES

Las técnicas de clasificación junto con la aplicación web desarrollada permiten tener un diagnóstico rápido y certero con base en el aprendizaje previo obtenido por los clasificadores. Éstos son evaluados mediante diferentes métricas de desempeño para poder elegir entre ellos, aquel que presente un mejor modelo de clasificación para hacer una buena predicción con nuevos registros. De acuerdo a los resultados obtenidos después de realizar varias ejecuciones utilizando las herramientas RapidMiner, Weka y R se concluye que R resulta ser la mejor opción para clasificar ya que obtiene resultados con mayor precisión. Además de considerar que los algoritmos de clasificación con el mejor desempeño son Árboles de decisión para predecir pacientes con diabetes y el algoritmo RandomForest para predecir pacientes con cáncer de mama.

6. REFERENCIAS

Chambers, J. et al. 2016. The R Project for Statistical Computing (https://www.r-project.org/). Bell Laboratories (formerly AT&T, now Lucent Technologies).

ENSANUT. 2014. Encuesta Nacional de Salud y Nutrición. Resultados Nacionales 2014 (http://ensanut.insp.mx/informes/ENSANUT 2014 ResultadosNacionales.pdf.)

Google. 2014-2016. Materialize (http://materializecss.com/). Carnegie Mellon University.

Han, J., Kamber, M., Pei, J. 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

INEGI. 2016. Instituto Nacional de Estadística y Geografía. Estadísticas de Mortalidad en México (http://www.inegi.org.mx/est/contenidos/proyectos/registros/vitales/mortalidad/tabulados/ConsultaMortalidad.asp).

INEGI. 2015. Instituto Nacional de Estadística y Geografía. Estadísticas de la lucha contra el cáncer (www.inegi.org.mx/saladeprensa/aproposito/2015/cancer2015_0.pdf).

Lee, P., Mierswa, I., et al. 2006 - 2016. RapidMiner: Data Science Platform | Machine Learning (https://rapidminer.com/). Rapid-I, University of Dortmund.

Lichman, M. 2013. UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). Irvine, CA: University of California, School of Information and Computer Science.

OMS (Organización Mundial de la Salud). 2016. Diabetes (http://www.who.int/topics/diabetes_mellitus/es/).

Piatetsky-Shapiro, P. 2013. KDnuggets. Data Science and AI Consulting (http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html).

Resig, J. 2016. jQuery 3.0.0 (https://jquery.com/), GPL and MIT.

TOWARDS A DEFINITION AND VALIDATION OF A SERIOUS GAME EVALUATION PROCESS

L. García-Mundo[1], J. Vargas-Enríquez[1], S. Martínez-Guerra[1], M. Genero[2] & M. Piattini[2]

[1]Tecnológico Nacional de México-Instituto Tecnológico de Ciudad Victoria, Boulevard Emilio Portes Gil, #1301, Pte. A.P. 175, C.P. 87010, Cd. Victoria, Tamaulipas, México.
lgarcm64@gmail.com, jvargd@gmail.com, sylvia.mtz.guerra@gmail.com,
[2]Universidad de Castilla-La Mancha, C/ Altagracia 50,13071, Ciudad Real, España.
Marcela.Genero@uclm.es, Mario.Piattini@uclm.es

ABSTRACT. A Serious Game is a game for purposes other than mere entertainment. Serious Games are currently in widespread use and their popularity has begun to increase steadily. The number of users of these systems is also growing day-by-day, signifying that their social impact is very high; it is precisely because of this reason that Serious Game quality evaluation is of the utmost importance. This motivated us to initiate a long-term research consisting of the definition and validation of a quality model that is specific to Serious Games. In a previous work a Product Quality model specifically for Serious Games, called QSGame-Model, was presented. Although, we have done examples of Serious Game quality evaluation using the QSGame-Model, we must stress that when these examples were conducted, we detected some problems related to the applicability of the QSGame-Model. We believe that the availability of an evaluation process that would indicate to us in detail how to perform these evaluations would mitigate some of the problems identified. The main goal of this paper is to present the preliminary version of the Serious Game evaluation process (QSGame-Evaluation Process), which has been defined by adapting the reference model for software product evaluation proposed in the ISO 25040 standard. This process can be applied to evaluate the quality of a Serious Game using the QSGame-Model and will be validated in the near future.

KEY WORDS: QSGame-Model, QSGame-Evaluation Process, ISO 25040.

1. INTRODUCTION

Serious Games (SGs) are games for purposes other than mere entertainment (Susi et al., 2007), which means that they have a serious purpose not only as regards education but also training, advertising or simulation. SGs are a fast-emerging area of opportunity, in addition to being a rapidly-growing market (Michael y Chen, 2005). With a global growth rate of almost 7% a year, it is forecasted that by 2018 worldwide revenue will reach 2.4 billion dollars (Tyson, 2014). SGs are a means to achieve relevant goals from both a personal and an institutional point of view. They may be used in fields as diverse as defense, education, scientific exploration, health care, emergency management, city planning, engineering, religion, and politics. What is more, the number of users of these systems is growing every day, pointing to their significant social impact. This is what makes the quality of SGs such a critical issue; they are not just another variety of software (in which it is already assumed that quality is important). They can have a major impact on many areas of society and on a huge amount of users; it is therefore our duty as researchers and computer professionals to ensure the quality of SGs. That led us to perform a systematic mapping study (SMS) in a previous piece of work (Vargas et al., 2014) the aim being to collect all the existing literature on SG quality. The results of this SMS indicate that although researchers are interested in the quality of these applications, we were unable to find an agreed on quality model that considers all the characteristics, sub-characteristics, attributes and measures that are applicable to any kind of SG. Based on these results we propose an SG product Quality Model called QSGame-Model (García-Mundo et al., 2015).

With the aim to evaluate the feasibility of the QSGame-Model, some examples of SGs quality evaluation were performed (García-Mundo et.al, 2016b; García-Mundo et al., 2016c; Valencia, 2015; Valencia et al., 2016). Although, we have done examples of SG quality evaluation using the QSGame-Model, we should stress that, while performing these examples, we detected some problems related

to the applicability of the QSGame-Model. We believed that the availability of an evaluation process, that would indicate to us in detail how to perform these evaluations, would mitigated some of the problems identified. Therefore, we considered appropriated to define an evaluation process which describes in detail the inputs, outputs, activities and tasks involved in the complete process. The main objective of this paper is to present a preliminary version of the SG evaluation process (QSGame-Evaluation Process), which has been defined by adapting the reference model for software product evaluation proposed in the ISO 25040 standard (ISO, 2011). The remainder of this document is organized as follows. As a background, Section 2, briefly introduces the ISO 25040 standard. Section 3 presents a description of the QSGame-Evaluation Process. Finally, our main conclusions and ideas for future work will be presented in Section 4.

2. BACKGROUND

The evaluation procedure proposed in this work, has been defined by taking the ISO 25040 standard (ISO, 2011) as a basis. In this section, a brief overview of the ISO 25040 standard, is presented.

The ISO 25040 proposes the reference model for software product evaluation, which contains requirements and recommendations for the evaluation of software product quality and clarifies the general concepts. It provides a process description for evaluating software product quality and states the requirements for the application of this process. The evaluation process can be used for different purposes and approaches. The process can be used for the evaluation of the quality of pre-developed software, commercial-off-the-shelf software or custom software and can be used during or after the development process (ISO, 2011). The process is composed of five activities (see Figure 1), and each of these activities is composed of several tasks. A brief description of each of these activities is showing bellow:

1. “Establish requirements of the evaluation”, the goal of this activity is establish the purpose of the evaluation and its subsequent mapping to characteristics, sub characteristics, and quality attributes to be evaluated, and metrics to calculate.
2. “Specify evaluation”, in this activity the evaluation modules and the decision criteria for quality measures are specified, in order to design the documents to be used during the evaluation
3. “Design the evaluation”, this activity consists of planning the evaluation activities taking into account the available human resources in order to know how the evaluation will be carried out.
4. “Execute the evaluation”, the execution of the evaluation tasks are carried out in this activity, by applying the decision criteria. During the execution of the evaluation, physical documents which contains the results of the evaluation are produced.
5. “Complete the evaluation”, this activity is related to the conclusion of the software product quality evaluation, the reviewing of the evaluation results and the creation of the evaluation report.

3. THE QSGAME-EVALUATION PROCESS

In this section, a description of the QSGame-Evaluation Process, to evaluate the quality of a SG using the QSGame-Model, is presented. This process has been defined by adapting the reference model for software product evaluation, proposed in the ISO 25040 standard (ISO, 2011).

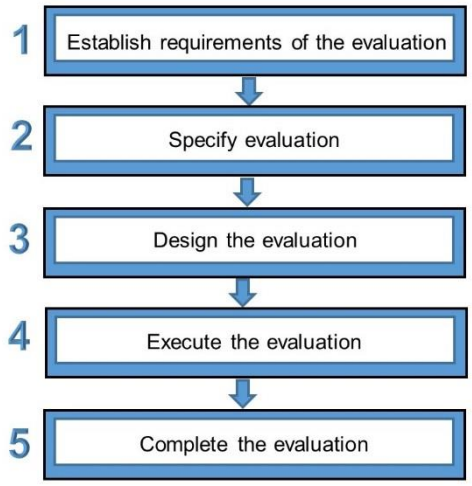

Figure 1. ISO 25040 process description for evaluating quality of software product.

3.1 Description of the QSGame-Evaluation Process

The QSGame-Evaluation Process consists of the following five activities (see Figure 2):

- A1 Specification of the evaluation requirements: The aim of this activity is to establish the requirements of the evaluation in order to delimit the scope of the evaluation, i.e. define the purpose of the evaluation and select the quality attributes to be evaluated.
- A2 Evaluation design: The objective of this activity is to design the checklist form, specifically considering what information will be collected during the evaluation of the SG, according to the specification requirements defined in A1.
- A3 Evaluator training: This activity aims the Evaluator designer to train the Evaluator how to perform the evaluation taking into account the QSGame-Model and the Serious Game Design Document (SGDD).
- A4 Execution of the evaluation: The aim of this activity is to execute the evaluation of the quality of an SG by filling in the checklist form generated in A2.
- A5 Calculation of the SG quality: The calculation of the SG quality can be done manually or automatically. The manual calculation, is performed by calculating the quality value of each measure, applying its measurement function defined in the QSGame-Model. In the other hand, the automatic calculation is performed by means of the QSGame-Tool (García-Mundo, 2016a) a Web tool developed with the aim of automating the applicability of the QSGame-Model.

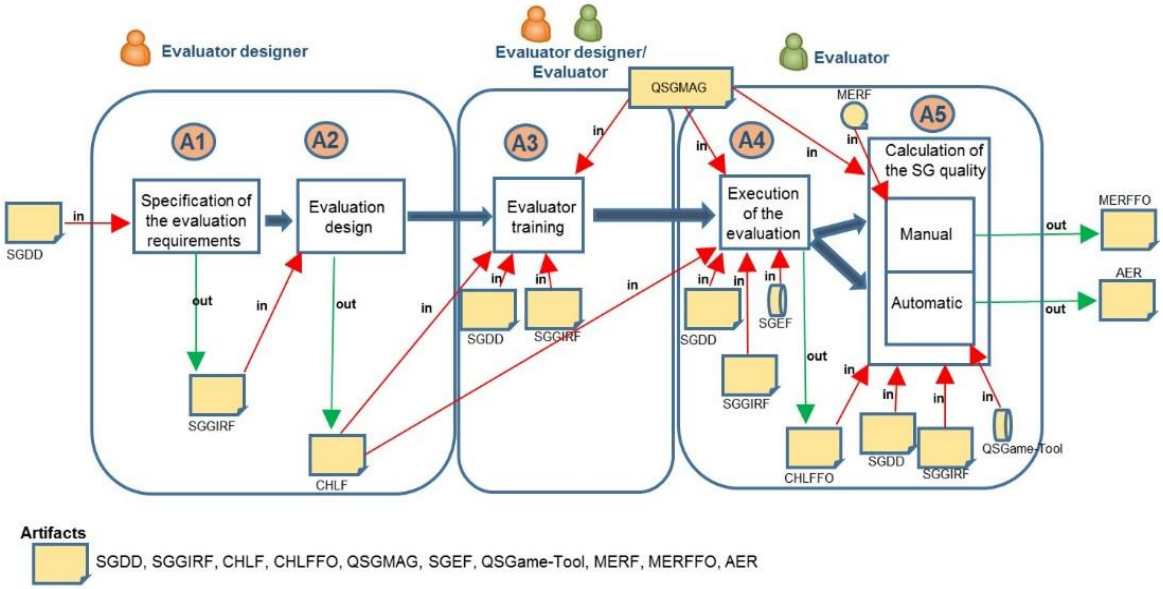

Figure 2. QSGame-Evauation process.

Each of the five activities listed above, are completed after performing several tasks (T1, T2, etc). The artifacts involved in these activities can be both input or output artifacts. The artifacts used in this process are shown in Table 1. More than one person may be responsible for each activity, and they can play the following roles:

- Evaluator designer: The Evaluator designer is responsible for specifying the evaluation requirements, designing the evaluation and training the Evaluator.
- Evaluator: The role of the Evaluator is to execute the evaluation and calculate the SG's quality either manually or automatically.

In the following sub-sections, the person responsible, the tasks and the artifacts involved in each of the activities of the QSGame-Evaluation Process, are described.

Table 1. Artifacts involved in the QSGame-Evaluation Process.

Artifact acronym	Artifact description
SGDD (Serious Game Design Document)	This document contains SG information required for SG evaluation and is related to the description of the game's features, which are among others, the requirements of the SG, its objectives, the profile of the SG (SG genre, history, narrative, etc.), the mechanics, the user interfaces, etc.
SGGIRF (Serious Game General Information Requirements Form)	This form contains a summary of the information needed to perform the quality evaluation of an SG using the QSGame-Model, either manually or automatically.
CHLF (CHeckLists Form)	CHLF is a form which contains each of the SG quality attributes that must be evaluated.
QSGMAG (QSGame-Model And Glossary)	This document contains a complete description of the QSGame-Model along with a glossary of relevant terms related to the SG context.
SGEF (SG Executable File)	This is an executable file of the SG to be evaluated.
CHLFFO (CHeckLists Form Filled Out)	CHLFFO is the CHLF filled out with the SG evaluation results.
QSGame-Tool	It is the tool developed for the automation of the QSGame-Model.
MERF (Manual Evaluation Results Form)	This form is an Excel file in which the quality values calculated manually for each of the measures of the QSGame-Model must be entered.
MERFFO (Manual Evaluation Results Form Filled Out)	MERFFO is the MERF form filled out with the quality values of the measures calculated manually.
AER (Automatic Evaluation Results)	Results of the automatic evaluation generated by the QSGame-Tool.

3.1.1 Specification of the evaluation requirements (A1)

Person responsible: Evaluation designer
Input: SGDD
Output: SGGIRF
The tasks involved in this activity are described in Table 2. The outcome of these tasks (SGGIRF) will be used as input by activity A2.

Table 2. Tasks involved in A1.

Task num.	Task action	Task description
T1	Establish the purpose of evaluation	In this task, the quality characteristics to be evaluated are determined, i.e., the Functional suitability characteristic, the Usability characteristic or both.

T2	Specify the SG functions	The SG functions that will be evaluated are specified, i.e., those functions that are directly related to the SG mechanics.
T3	Specify additional requirements	This task has the aim of specifying some additional requirements: whether the SG is oriented toward disabled users, whether the SG is a multiplayer game, whether or not the SG has levels of difficulty, whether the SG is developed in a realistic environment, whether the SG is a simulation game, along with whether the SG is an strategy game, in order to consider the randomization in performing tasks.
T4	Specify the SG screens and the SG interfaces data	If the need to evaluate the Usability quality characteristic has been established, it is necessary to register the following information: • The list of the user interfaces of the SG that will be evaluated. • The list of the screens of the SG that will be evaluated.
T5	Select the measures to be discarded	Depending on the SG requirement specification (T3 in this activity), select those measures to be discarded.
T6	Define information for evaluation	In this task, the numeric values required to perform the calculations of some of the quality measures of the SG are defined.

3.1.2 Evaluation design (A2)

Person responsible: Evaluation designer
Input: SGGIRF
Output: CHLF
Table 3 describes the tasks involved in this activity. The checklist form specifies the information that must be collected by the Evaluator when performing the quality evaluation of the SG and will be used as input in activities A3 and A4.

3.1.3 Evaluator training (A3)

Person responsible: Evaluation designer and evaluator
Input: QSGMAG, SGDD, SGGIRF and CHLF
Output: none
The tasks performed in activity A3 are described in Table 4.
The QSGMAG, SGDD, SGGIRF and CHLF will also be used as input in the next activity.

Table 3. Tasks involved in A2.

Task num.	Task action	Task description
T1	Design the checklist form	Define the structure of the checklist form depending on the specification of the evaluation requirements.

Table 4. Tasks involved in A3.

Task num.	Task action	Task description
T1	Introduce the QSGame-Model	An explanation overview of the QSGame-Model is presented to the Evaluator.

T2	Introduce the SG	The evaluation designer uses the SGDD to show the Evaluator the main functions and features of the SG.
T3	Explain CHLF	Explanations of how to perform the evaluation and how to fill out the CHLF, using the SGGIRF and the CHLF, are given to the Evaluator.
T4	Deliver evaluation material	The Evaluator is provided with: the QSGMAG, the SGDD, the SGGIRF and CHLF.

3.1.4 Execution of the evaluation (A4)

Person responsible: Evaluator
Input: QSGMAG, SGDD, SGGIRF and CHLF, and SGEF
Output: CHLFFO
The tasks performed in activity A4 are described in Table 5. The output of this activity (CHLFFO) contains the quality evaluation result of the SG and will be used as input in the next activity.

3.1.5 Calculation of the SG quality: manual or automatic (A5)

In the following sub-sections the tasks involved on the calculation of the quality measures both manually and automatically are described.

Table 5. Tasks involved in A4.

Task num.	Task action	Task description
T1	Download and install the SG to be evaluated	In the first task of this activity, the downloading and installation of the SGEF to be evaluated must be done.
T2	Be familiar with QSGMAG and SGDD documents	In this task, it is advisable to read the SGDD and the QSGMAG concepts related to SGs.
T3	Perform the SG functions	To be familiar with the SG, it is important execute it, browsing the whole SG and performing each of its functions.
T4	Evaluate general aspects of the SG	Answer the “GENERAL QUESTIONS” in the CHLF.
T5	Evaluate SG functions and interfaces	Answer the “QUESTIONS BY FUNCTION” and the “QUESTIONS OF THE INTERFACE IN WHICH THE FUNCTION IS EXECUTED” in the CHLF.
T6	Evaluate screens	Answer the “SCREEN QUESTIONS” in the CHLF.

3.1.5.1 Manual Calculation

The manual calculation of the SG quality value, is performed taking into account the SGGIRF.
Person responsible: Evaluator
Input: CHLFFO, QSGMAG, SGDD, SGGIRF, and MERF
Output: MERFFO

The tasks performed in activity A3 are described in Table 6. The MERFFO artifact contains the quality evaluation results for each quality measure calculated and is grouped by quality characteristic and sub-characteristic. The MERFFO evaluation results are the basis used to generate a complete report of the evaluation of an SG.

Table 6. Tasks involved in A5 manual calculation.

Task num.	Task action	Task description
T1	Calculate quality value measures of YES/NO questions	Count how many SG functions or how many SG interfaces comply with what it is asked. The resulting amount of this count represents the value of A in the measurement function of the quality measure being evaluated. The value of B is always the total SG functions being evaluated which are specified in the SGGIRF. As a hypothetical example, let us suppose we are evaluating a game which has 3 functions (value of B specified in the SGGIRF). In order to calculate the quality value for the attribute *coverage of progress*, the questionnaire requests the Evaluator to answer the following: *"To what extent do SG functions indicate how the player will progress during the game?*" If the Evaluator considers that the answer to this question is YES for just 2 (out of 3) of the functions of the game, then the total number of functions that meet it, are 2 (value of A). The measurement function used to calculate the quality value is X= A / B (García-Mundo et al., 2016c). Therefore, X = 2/3 = 0,667.
T2	Calculate quality value measures of numeric questions	In order to obtain the value of the variable A of the measurement function, the Evaluator must add each of numerical values collected by an SG function. The value of B is always the total number of SG functions being evaluated which are specified in the SGGIRF. As a hypothetical example, let us suppose that the game being evaluated has 3 functions and a total of 10 progress messages (value of B specified in SGGIR). In order to calculate the quality measure for the attribute "clarity of feedback messages" the questionnaire requests the Evaluator to answer the following for each of the game functions "How many of the progress messages in this function provide the player with a description of how progress has been achieved?" If the Evaluator considers that function 1 has 2, function 2 has 1 and function 3 has 2 clear progress messages, the total number of clear progress messages is 5 (value of A, which is the sum of 2+1+2). The measurement function used to calculate the quality value of the measure is X=A / B (García-Mundo et al., 2016c). Therefore, X= 5 /10=0,50.
T3	Enter quality values of the measures in an RF	Each quality measure value calculated must be entered in a MERF provided.

3.1.5.2 Automatic Calculation

The automatic calculation of the SG quality value, is performed taking into account the SGGIRF and the QSGame-Tool.
Person responsible: Evaluator
Input: CHLFFO, QSGMAG, SGDD, SGGIRF, and QSGame-Tool
Output: AER
The tasks performed in activity A5 automatic calculation are described in Table 7. The AER artifact contains the quality evaluation results generated automatically. These evaluation results are the quality value result of each of the measures, along with the result of the quality per characteristic and sub-characteristic. It also

includes graphics of these quality results which are the basis used to generate a complete report of the evaluation of an SG.

Table 7. Tasks involved in A5 automatic calculation.

Task num.	Task action	Task description
T1	Register information from the evaluation	Register the information needed to identify the evaluation. The company that owns the SG to be evaluated, a project associated with this company and an evaluation associated with that project must be registered.
T2	Input requirement specifications	Enter the SG evaluation requirements specified in the SGGIRF: the quality characteristics to be evaluated, the functions directly related to SG mechanics, measures to be applied, numeric values needed to perform calculations.
T3	Input data evaluation of YES/NO questions	For each YES/NO question, mark the selected choice.
T4	Input data evaluation of numeric questions	For each numeric question enter the numeric value answered.
T5	Generate the evaluation results	Select the QSGT option for the automatic generation of the quality values of the evaluation of each of the measures, per quality sub-characteristic and per quality characteristic. The QSGT also generates a graphical evaluation result per sub-characteristic and characteristic, along with a global quality result.

4. CONCLUSION AND FUTURE WORK

The current relevance of the quality of SGs, motivated us to define a Product Quality model that is specific to SGs called QSGame-Model (García-Mundo et al., 2015). The problems detected while some examples of SG quality evaluation were carried out (García-Mundo et.al, 2016b; García-Mundo et al., 2016c; Valencia, 2015; Valencia et al., 2016), revealed that the availability of an explicit SGs evaluation process, that would indicate to us in detail how to perform these evaluations, would mitigated some of the problems identified.

The main contribution of this paper is the description of a preliminary proposal of the SG evaluation process (QSGame-Evaluation Process), which is a process has been defined by adapting the reference model for software product evaluation proposed in the ISO 25040 standard (ISO, 2011). This process defines in detail the inputs, outputs, activities and tasks involved and can be applied to evaluate the quality of a SG using the QSGame-Model.

We believe the QSGame-Evaluation process proposed in this paper, along with the QSGame-Model proposed in a previous work (García-Mundo et al., 2015), will be useful as regards the facilitating the task of assessing the quality of the SG, as well as allowing SG developers to ensure, evaluate and improve the quality of the SGs they build.

In the near future, we plan to validate the QSGame-Evaluation Process in the following manner:

- Feasibility evaluation. By performing examples of the quality evaluation of SGs

using the QSGame-Model and by executing each activity defined in the QSGame-Evaluation Process during the conduct of these examples of evaluation.

- Empirically evaluation. By gathering empirical evidence about the perceived usefulness and the perceived ease of use of the QSGame-Evaluation process, based on the Technology Acceptance Model (TAM) model (Davis, 1989). This model is one of the most popular theoretical frameworks as regards understanding user acceptance or adoption of information systems (Lin et al., 2015).

5. ACKNOWLEDGMENT

We would like to thank the Tenológico Nacional de México-Instituto Tecnológico de Ciudad Victoria and PRODEP for granting the first two authors the scholarship that made it possible to complete the research work presented in this paper.

6. REFERENCES

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

García-Mundo, L. (2016a). QSGame-Model: A Serious Game Product Quality Model. (PhD in Computer Science), Universidad of Castilla-La Mancha, Ciudad Real, Spain.

García-Mundo, L., Genero, M., y Piattini, M. (2015). Towards a Construction and Validation of a Serious Game Product Quality Model. Paper presented at the 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games), Skövde, Sweden.

García-Mundo, L., Genero, M., y Piattini, M. (2016b). Applying a Serious Game Quality Model. Serious Games, Interaction, and Simulation (pp. 12-20): Springer.

García-Mundo, L., Vargas, J. A., Genero, M., Piattini , M., y Martínez-Guerra, S. I. (2016c). A Serious Game Product Quality Model: Construction and Application to an Educational Serious Game. TecnoIntelecto, 13(2), 48-66.

ISO. (2011). ISO/IEC 25040 Systems and Software Engineering – System and software Quality Requirements and Evaluation (SQuaRE) – Evaluation process. Ginebra, Suiza: International Organization and Standarization.

Lin, F.-T., Wu, H.-Y., y Tran, T. N. N. (2015). Internet banking adoption in a developing country: an empirical study in Vietnam. Information Systems and e-Business Management, 13(2), 267-287.

Michael, D., y Chen, S. (2005). Serious games: Games that educate, train, and inform. Boston, Ma: Muska & Lipman/Premier-Trade.

Susi, T., Johannesson, M., y Backlund, P. (2007). Serious games: An overview. Skövde, Sweden: University of Skövde.

Tyson, G. (2014). The 2013-2018 Worldwide Game-based Learning and Simulation-based Markets. Monroe, WA: Ambient Insight.

Valencia, D. (2015). Un Juego para Desarrollar Habilidades Convenientes en Desarrollo Global del Software. (Computer Science Degree), University of Castilla La-Mancha, Ciudad Real, Spain.

Valencia, D., Vizcaíno, A., Garcia-Mundo, L., Piattini, M., y Soto, J. P. (2016). GSDgame: A serious game for the acquisition of the competencies needed in GSD. Paper presented at the Global Software Engineering Workshops (ICGSEW), 2016 IEEE 11th International Conference on.

Vargas, J. A., García-Mundo, L., Genero, M., y Piattini, M. (2014). A systematic mapping study on serious game quality. Paper presented at the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE), London, UK.

ANALISIS DEL COEFICIENTE DE TRANSFERENCIA DE CALOR EN LAS PROPIEDADES MECÁNICAS Y ESFUERZOS INTERNOS DURANTE EL TEMPLE EN ACEROS DE MEDIO CARBONO UTILIZANDO FEM

R.D. López-García, A. Maldonado-Reyes, M. A. Jiménez-García, C.E. López-García & J.A. Maldonado-Zúñiga

Tecnológico Nacional de México-Instituto Tecnológico de Ciudad Victoria, Boulevard Emilio Portes Gil, #130, Pte. A.P. 175, C.P. 87010, Cd. Victoria, Tamaulipas, México. rdlgitcv@hotmail.com, arma_y2k@hotmail.com, m_jimenez81@yahoo.com.mx carloslopez7616@hotmail.com, angel-maldonado94@hotmail.com

RESUMEN. El temple es uno de los tratamientos térmicos más utilizados en la industria para incrementar la resistencia y dureza en los aceros mediante la obtención de la fase martensita. Sin embargo, es bien conocido que durante el proceso de enfriamiento se presenta una distorsión macroscópica en los componentes, originada por una rápida variación de los esfuerzos térmicos y transformaciones de fase. Si los esfuerzos internos son más grandes que el esfuerzo de cedencia del material se presenta una deformación plástica y dependiendo de su magnitud, la distorsión puede ser muy grande. Aquí la importancia de determinar las condiciones óptimas para el proceso de temple y de manera que se pueda obtener un componente con geometría y propiedades deseadas. Para ello, los modelos matemáticos para este tipo de procesos requieren considerar interacciones complejas tales como la composición química de la aleación, las transformaciones de fase, los mecanismos de transferencia de calor, el comportamiento mecánico, las condiciones de frontera y las propias condiciones del proceso. En la presente investigación se analizó el efecto del coeficiente de transferencia de calor (HTC) durante el temple en aceros y su efecto en las propiedades mecánicas y la generación de esfuerzos internos y la distorsión de componentes utilizados en la industria automotriz. Para ello se utilizó un modelo matemático basado en el método por elementos finitos (FEM) capaz de predecir con precisión las variables que intervienen en el proceso de temple. El modelo fue validado experimentalmente utilizando pruebas como dilatometría, tratamientos térmicos *in situ*, caracterización microestructural y medición de propiedades mecánicas. Los resultados de la simulación concluyeron que el modelo matemático utilizado es capaz de predecir las tendencias en la distribución de fases presentes y la microestructura final, la dureza del material, la distorsión presente y los esfuerzos residuales con una precisión aceptable.

PALABRAS CLAVE: Temple, FEM, HTC, Distorsión.

ABSTRACT. Quenching is one of the heat treatment most widely used in the Industry to increase strength and hardness in steel trough martensitic transformation. However, it is well known that during the cooling process a macroscopic distortion occurs in the components, originated by a very quick variation of the thermal stress and phase transformation. If this internal stress fields result greatest than material yielding stress a plastic deformation occurs, and depending of its magnitude, the distortion could be very evident. Hence the importance of determinate the optimal conditions for the quenching process and so that can obtain a component with desirable geometry and properties. To do this, mathematic model from these type of processes requires to consider complex interactions such as alloy chemical composition, phase transformations, mechanisms of heat transfer, mechanical behavior, border conditions and the own conditions of the process. At the present research the effect of the heat transfer coefficient (HTC) during tempering on steel and their effect on the mechanical properties and the generation of internal stresses and the distortion of components used in the automotive industry were analyzed. To achieve this, a mathematic model based on the method of finite elements (FEM) able to predict with accuracy the variables involved in quenching process was used. The model was validated experimentally using tests that include quench dilatometry, *in situ* heat treatments, microstructural characterization and measurement of mechanicals properties. The results of the simulation concluded that the mathematical model used

was able to predict the tendencies in distribution of the phases presents in the final microstructure, the hardness of the material, the distortion present and the residual strength with a remarkable accuracy.

KEYWORD: Quenching, FEM, HTC, Distortion.

1. INTRODUCCIÓN

Los tratamientos térmicos involucran diferentes etapas durante el proceso como son: calentamiento, tiempo de permanencia y enfriamiento. Es conocido que de la velocidad de enfriamiento dependerá la microestructura formada y las propiedades finales de la pieza. El tratamiento térmico de temple ha sido uno de los procesos más utilizados en la industria automotriz y aeroespacial para mejorar las propiedades de los materiales tales como: dureza, resistencia y rigidez. Esto se logra calentando el material hasta su temperatura de austenizado y posteriormente con altas velocidades de enfriamiento transformar a martensita. Por otro lado, en el templado están interrelacionados fenómenos físicos, mecánicos y metalúrgicos los cuales son difícil de controlar y predecir de manera experimental. Esto no solo conduce a no lograr las propiedades mecánicas deseadas sino que puede afectar la condición superficial de la pieza o componente distorsionando la misma (Totten et al., 1997).

El templado en los aceros promueve la generación de esfuerzos residuales de origen térmico y por la transformación austenita-martensita, cuya magnitud dependerá de la velocidad de enfriamiento. Velocidades lentas de enfriamiento favorecen la aparición de un bajo nivel de esfuerzos residuales. Sin embargo, para obtener altas propiedades mecánicas, son necesarias velocidades de enfriamiento que logren suprimir la formación de microconstituyentes como: ferrita, perlita y bainita, para esto son necesarias altas velocidades de enfriamiento que logren una completa transformación a martensita. Esta transformación es la principal causa para la formación de esfuerzos internos en el material que pueden dependiendo de su magnitud ser perjudiciales para la vida de la pieza o componente. Las transformaciones de fase en estado sólido dan lugar a una deformación volumétrica y una transformación de plasticidad. La primera se da por expansión de volumen de CCC de austenita a TCC de martensita, y la transformación de plasticidad ocurre por las deformaciones de las transformaciones de fase y los campos de esfuerzos existentes.

En el endurecimiento por temple, la microestructura del acero, las deformaciones y los esfuerzos residuales cambian constantemente en función de la temperatura. Si en cualquier punto del componente a un tiempo determinado se excede el límite de cedencia del material, se presentará flujo plástico no uniforme que dará como resultado la presencia de esfuerzos residuales en el componente, y dependiendo de su magnitud y si estos se encuentran a tensión o compresión pueden ser de beneficio o perjudiciales para la pieza. Si los esfuerzos residuales están a compresión en la superficie del material son de beneficio caso contrario si se encuentran a tensión producirán problemas en las propiedades de fatiga del componente (Nallathambi et al., 2010; Şimşir et a.l, 2008).

Debido a que el temple es un proceso multi-físico que involucra una interacción complicada de acoplamientos entre diferentes eventos físicos tales como: la transferencia de calor, las transformaciones de fase y la evolución de los esfuerzos internos, no existe una solución analítica a las ecuaciones gobernantes en particular considerando el caso tridimensional y en estado transitorio. En consecuencia el uso de simulación numérica con acoplamiento térmico-mecánico-metalúrgico es indispensable para lograr un mejor entendimiento de los efectos del tratamiento térmico de temple en la calidad de las piezas templadas, permitiendo evaluar propiedades que no pueden ser medidas experimentalmente (Kang et. al, 2007). Woodard et al, 1999, propusieron un elemento bidimensional mediante análisis por elemento finito (FEA) para el proceso de temple de cilindros de acero de tipo SAE 1080, mostrando una fuerte evidencia del calor latente con las transformaciones de fase. Homberg 1996, ha propuesto un modelo por

elementos finitos (FE) para una prueba Jominy para evaluar la historia térmica del enfriamiento incluyendo las transformaciones de fase. En el presente trabajo, se realizó una investigación experimental necesaria para implementar un modelo numérico mediante simulación por el método por elemento finito (FEM) utilizando el software DEFORM-3D, con la finalidad de evaluar el coeficiente de transferencia de calor (HTC) constante y en función de la temperatura, y su efecto en las propiedades mecánicas, la distorsión y la distribución de los esfuerzos residuales de una pieza de acero utilizada comúnmente en la industria automotriz en la fabricación de sistemas de suspensión.

2. DESARROLLO EXPERIMENTAL

El material estudiado es un acero del tipo SAE 5160, su composición química se presenta en la Tabla 1. Las dimensiones de la pieza utilizada fueron de 180 mm de longitud, 60 mm de ancho y 10 mm de espesor como se muestra en la Figura 1.

Tabla 1. Composición química del acero SAE 5160.

Elemento	C	Mn	Cr	Si	Cu	Ni	Fe
(% e.p.)	0.6	0.87	0.96	0.25	0.22	0.14	Balance

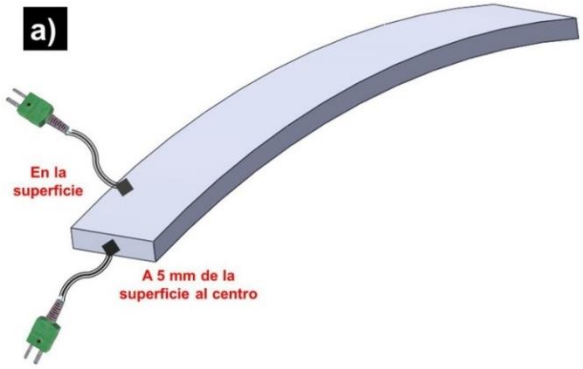

Figura 1. Geometría de la pieza templada y simulada en DEFORM-3D.

Tabla 2. Coeficiente de transferencia de calor en función de la temperatura.

Temperatura (°C)	Coeficiente de Transferencia de Calor HTC (W/m^2°C)
920	200
900	210
800	590
700	900
600	1100
500	1300
400	1280
300	800
250	750
200	1400
150	1100
100	250
50	220
25	200

Para el análisis de la distorsión, los esfuerzos residuales y las propiedades mecánicas, se utilizó un coeficientes de transferencia de calor constante 1700 W/m^2°C y uno en función de la temperatura, para este último la pieza fue instrumentada con termopares tipo K colocados en la superficie y en el centro de la pieza con la finalidad de medir la historia térmica, los datos obtenidos son mostrados en la Tabla 2.

Para el tratamiento térmico, las piezas de acero fueron calentadas en un horno tipo mufla marca Thermolyne a una velocidad de calentamiento de 3 °C/s hasta una temperatura de austenizado de 900 °C, el tiempo de permanencia fue de 520 s para lograr obtener un tamaño de grano ASTM de 8. Posteriormente las piezas fueron enfriadas en un baño de aceite a una temperatura de 50 °C, no se utilizó agitación y las piezas permanecieron en el baño de aceite un tiempo de 90 s. La historia térmica fue obtenida mediante una tarjeta de adquisición de datos marca OMEGA de 12 puertos. Una vez obtenida la historia térmica con base en aproximaciones y utilizando el software DEFORM-3D se obtuvo el HTC en función de la temperatura. La caracterización de fases se realizó por medio de técnicas tradicionales de metalografía y para revelar la microestructura se utilizó ácido nítrico. Mediante el software JMatPro se calculó la cinética de transformaciones de fase y la expansión lineal de la martensita. Los datos obtenidos fueron posteriormente refinados mediante pruebas de laboratorio a través de dilatometría por temple, utilizando un dilatómetro de temple de alta velocidad (L78 RITA Linseis Messgeräte, Germany). Se utilizó un sistema de

calentamiento por inducción y enfriamiento con helio, las muestras utilizadas en las pruebas fueron de 10 mm de longitud y 3 mm de diámetro, estas muestras fueron instrumentadas mediante termopares tipo-K para una obtención de 1000 datos por segundo.

2.1 Procedimiento de Simulación por FEM y Datos de Entrada

Para el desarrollo del modelo matemático por elementos finitos se utilizó el software DEFORM-3D (Scientific Forming Technology Corporation, Columbus, Ohio, USA) donde se emplea un modelo totalmente acoplado que involucra propiedades térmicas, mecánicas, físicas y metalúrgicas, en la determinación de las transformaciones de fase, esfuerzos residuales y desplazamiento del material templado. En la Tabla 3, se resumen los parámetros de simulación utilizados. Comenzaremos por el mallado del material es cual fue de 100000 elementos. Esto permitió obtener resultados detallados en partes específicas de la pieza. La ruta de inmersión de las piezas fue en la coordenada "Z" como se muestra en la Figura 1. La velocidad de inmersión fue de 40 mm/s. La temperatura inicial fue de 920°C en todos los nodos, y se estableció que todos los nodos estaban constituidos de austenita con un tamaño de grano austenítico ASTM 8.

Tabla 3. Parámetros de simulación.

Parámetros de simulación	Valores
Radio de Poisson	0.3
Método de interacción	Newton-Rapshon
Conductividad térmica	***k*** (para cada fase)
HTC constante (W/m-2°C)	1700
HTC *f*(T) (W/m-2°C)	Tabla 2
Pasos de simulación	900
Número de elementos	100000
Número de nodos	14156
Temperatura inicial (°C)	920
Temperatura ambiente (°C)	25
Temperatura del medio (°C)	50
Vel. de inmersión (mm/s)	40
Dirección de inmersión	"Z" vertical al baño
Tamaño de grano ASTM	8 (22 μm)

Las propiedades térmicas y mecánicas del material fueron calculadas como independientes de cada fase y obtenidas mediante JMatPro estas fueron: calor especifico (Cρ), conductividad térmica (*k*), entalpía (H), densidad (ρ), módulo de Young´s (E), relación de Poisson (υ), flujo de esfuerzos (S_r) y esfuerzo de cedencia (S_y).

3. RESULTADOS

Los resultados experimentales obtenidos de los tratamientos térmicos, la dilatometría por temple, y los resultados obtenidos de las simulaciones mediante JMatPro, se alimentó la base de datos de las propiedades térmicas, mecánicas y físicas en el software DEFORM-3D para realizar las simulaciones numéricas. La Figura 2, muestra el comportamiento térmico durante el enfriamiento de una pieza del acero estudiado templada a una velocidad de enfriamiento de 35°C/s, junto con la respuesta térmica obtenida mediante DEFORM-3D. Se puede apreciar que existe una relación bastante cercana de la curva de enfriamiento experimental con la simulada, el error no supera el 4%. Esto es un dato importante ya que da certeza y confiabilidad de los resultados arrojados por el software.

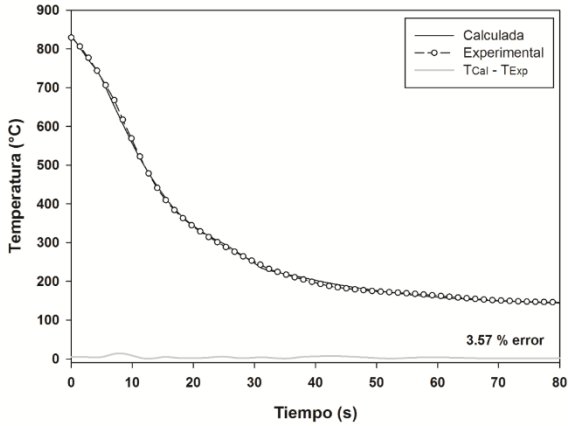

Figura 2. Curvas de enfriamiento experimental y calculada mediante DEFORM-3D del acero estudiado.

La Figura 3, muestra una comparación de las curvas de expansión lineal obtenida de manera simulada por JMatPro (línea continua) y validada mediante dilatometría por temple (línea punteada), durante el enfriamiento del acero SAE 5160 desde una temperatura de

920°C, con un tamaño de grano ASTM 8 y una velocidad de enfriamiento de 35 °C/s (velocidad máxima de enfriamiento alcanzada en el templado de estas piezas). Se puede observar que el comportamiento de ambas curvas de expansión son muy similares, se presenta un cambio importante a una temperatura aproximada de 260 °C que corresponde a la temperatura de inicio de la transformación martensítica (*Ms*), siendo a una temperatura de 140 °C donde finaliza la transformación de la martensita (*Mf*).

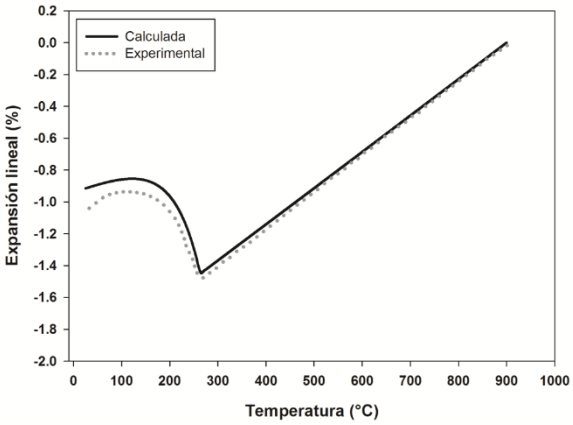

Figura 3. Curva de expansión lineal calculada y obtenida mediante dilatometría por temple del acero estudiado.

La Figura 4, muestra el diagrama Tiempo-Temperatura-Transformación (TTT) del inicio y fin de la transformación de perlita y bainita y el inicio de la transformación de la martensita, para el acero estudiado con un tamaño de grano austenítico antes del temple ASTM 8. Los resultados obtenidos para esta figura se calcularon utilizando el software JMatPro y posteriormente fueron validados a través de dilatometría por temple. Es importante señalar que además del incremento en el contenido de C en la aleación el incremento en el tamaño de grano austenítico de temple tiene una marcada importancia en los tiempos de enfriamiento. Al incrementar el tamaño de grano las curvas de estos diagramas son desplazados hacia la derecha del mismo lo que favorece al contar con mayor tiempo para la velocidad de enfriamiento y así evitar posibles grietas o fracturas producidas por alas velocidades de enfriamiento.

Los resultados presentes en los siguientes apartados corresponden a las simulaciones realizadas mediante DEFORM-3D con las variables en el HTC (constante =1700 W/m2°C y en función de la temperatura, Tabla 2) y velocidades de inmersión al baño de enfriamiento (0.02, 0.06 y 0.1 m/s).

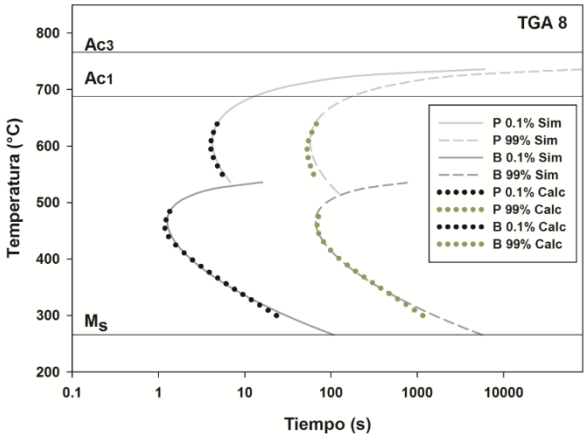

Figura 4. Diagrama TTT de la cinética de transformaciones de fase de perlita y bainita para un tamaño de grano austenítico ASTM 8.

3.1. Análisis térmico

La Figura 5, muestra la historia térmica durante el enfriamiento de las muestras de acero SAE 5160 templadas en aceite a 50 °C desde una temperatura de austenizado de 920 °C, para los HTC en *f(*T) y constante y 3 velocidades de inmersión. Se puede observar para el primer caso utilizando un HTC *f*(T) las curvas de la velocidad de enfriamiento tienen un comportamiento muy similar, las máximas velocidades de enfriamiento se encontraron en 32 °C/s para las velocidades de inmersión más altas utilizadas y de 28°C/s para la velocidad más baja. Sin embargo, cuando se utilizó un HTC constate las velocidades de enfriamiento fueron más altas, presentando la máxima velocidad de inmersión la mayor velocidad de enfriamiento, cabe señalar que aun con la velocidad de inmersión más baja 0.02 m/s se obtuvieron velocidades por encima de 40 °C/s superiores a las mostradas para el HTC *f*(T).

Es importante señalar, que con la obtención de mayores velocidades de enfriamiento se puede lograr obtener una transformación casi completa de austenita-martensita,

considerando la fracción de austenita retenida, permitiendo obtener mayores niveles de dureza en el componente.

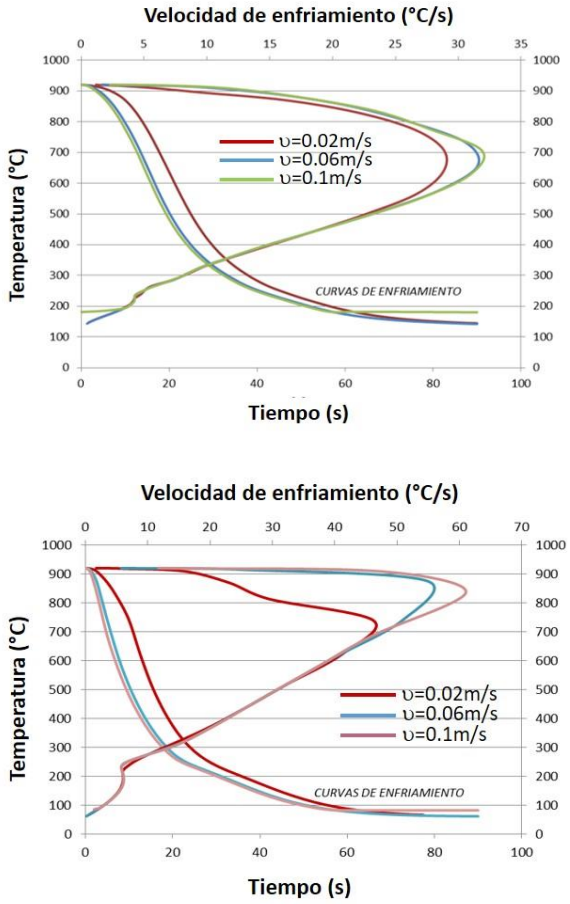

Figura 5. Historia térmica y curvas de velocidad de enfriamiento para 3 velocidades de inmersión y con un HTC *f* (T) y constante.

3.2. Transformaciones de fase y Propiedades mecánicas

La trasformación completa de la fase austenita-martensita, es la reacción que se busca en este tipo de tratamientos térmicos. La Figura 6, muestra el comportamiento de esta transformación para las velocidades de inmersión estudiadas y ambos HTC. Se puede observar, que tanto para la fracción total transformada y el tiempo de inicio de transformación, utilizando un HTC constante se obtienen más altas fracciones transformadas de martensita independientemente de la velocidad utilizada, los niveles alcanzados están por encima de 90%, las otras fases formadas que se observaron fueron austenita retenida y bainita en pequeñas concentraciones.

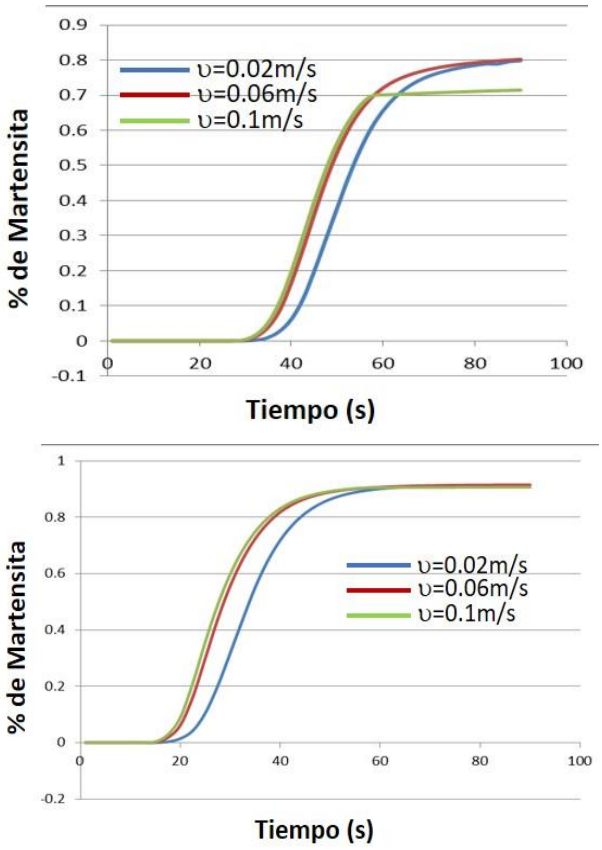

Figura 6. Fracción de volumen de martensita transformada para 3 velocidades de inmersión y para un HTC *f* (T) y constante.

Como era de esperarse por la fracción transformada de martensita obtenida, la dureza presente en estos aceros también se ve favorecida cuando se utiliza un HTC constante y más homogénea se aprecia a altas velocidades de inmersión, Figura 7.

3.3. Distorsión y Esfuerzos residuales

La distorsión generada en las piezas estudiadas se midió mediante el desplazamiento producido en la pieza con los datos obtenidos por DEFORM-3D, para las muestras templadas, con HTC f(T) y constante, y para las 3 velocidades de inmersión estudiadas. Cabe mencionar que estos desplazamientos para estos procesos son casi inevitables lo que se pretende es encontrar los parámetros óptimos para reducir este fenómeno.

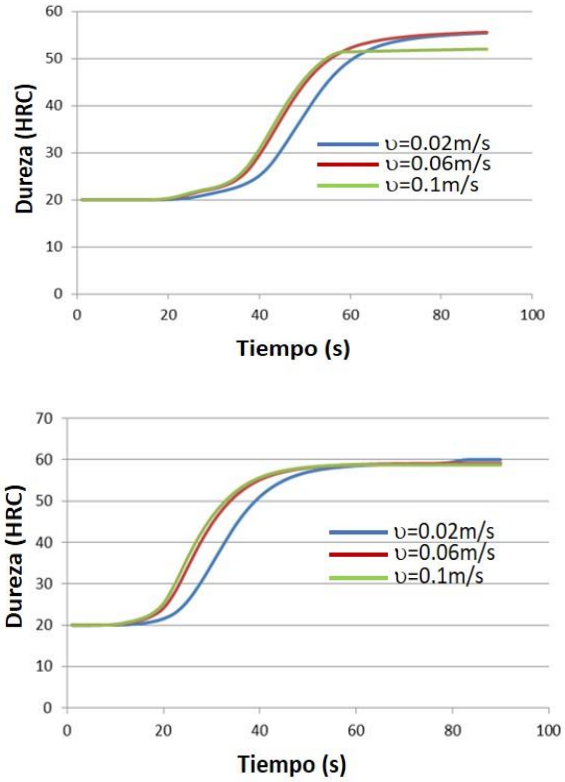

Figura 7. Dureza HRC en función de la velocidad de inmersión y para un HTC *f* (T) y constante.

La Figura 8, muestra el desplazamiento obtenido por las pruebas de simulación paras las condiciones estudiadas. Se puede observar, que cuando se utiliza un HTC *f*(T) la mayor distorsión o desplazamiento se presenta para velocidades de inmersión más altas alcanzando valores mayores a 2 mm de desplazamiento. Para velocidades más lentas la distorsión disminuye. Para el caso cuando se utiliza un HTC constante el desplazamiento al final del proceso es muy similar pero menor que cuando se utilizó un HTC *f*(T) con excepción de la más baja velocidad de inmersión.

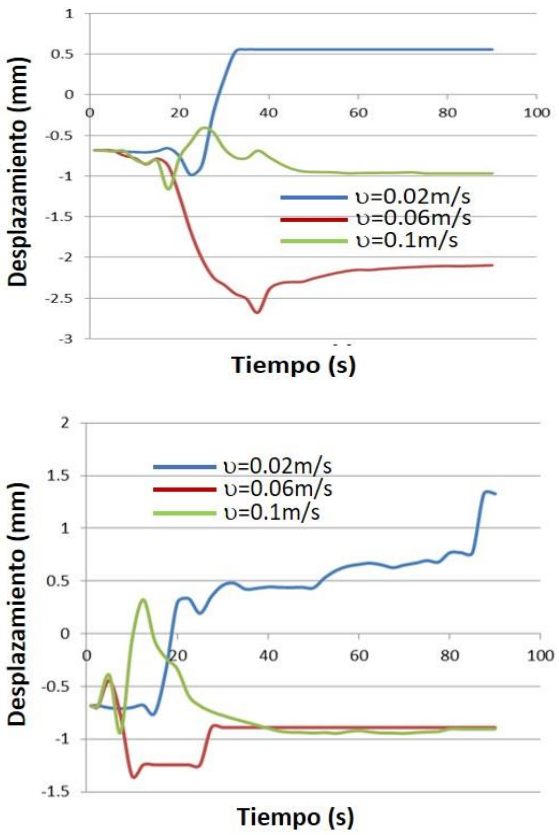

Figura 8. Desplazamiento nodal para 3 velocidades estudiadas y para un HTC *f*(T) y constante.

En la Figura 9, se presentan los esfuerzos internos generados bajo las condiciones estudiadas. Se puede observar que no existe una relación pronunciada entre el HTC utilizado y el desplazamiento presente. Si bien los menores valores de esfuerzos que son los deseados, se obtuvieron para un HTC constante solo fue para una condición de inmersión. Lo que significa que existe un amplio campo de oportunidades para seguir investigando las condiciones que nos puedan generar esfuerzos residuales lo más bajo posibles.

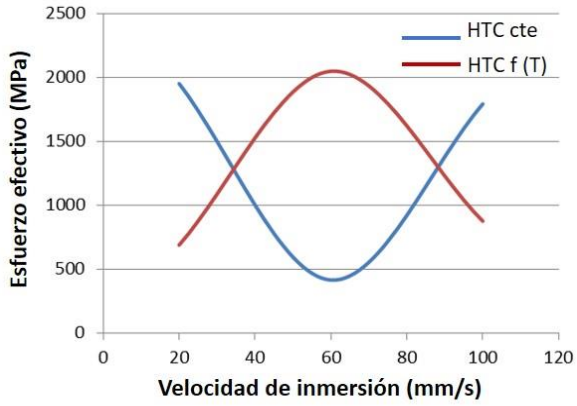

Figura 9. Esfuerzo efectivo en función de la velocidad de inmersión, para un HTC f(T) y constante.

4. CONCLUSIONES

De los resultados obtenidos a través de la simulación en DEFORM-3D, donde se estudió el efecto del HTC para diferentes velocidades de inmersión para un acero SAE 5160 se puede concluir lo siguiente:

1. La velocidad de inmersión presenta un efecto directo sobre la velocidad de enfriamiento de las piezas independientemente del HTC utilizado, para altas velocidades de inmersión mayor velocidad de enfriamiento en las piezas templadas.
2. La mayor cantidad de fracción transformada de martensita se presenta a menores velocidades de inmersión es decir 0.02 m/s, para ambos HTC utilizados, y la fracción de Bainita por lo tanto esta en menor concentración. Esto es lo ideal para estos aceros.
3. La dureza para estos aceros presenta mejores condiciones para un HTC constante y velocidades bajas de inmersión, lo cual está de acuerdo con la cantidad de martensita obtenida en estas condiciones.
4. La mayor distorsión en estos componentes se presenta bajo condiciones de bajas velocidades de inmersión, donde en los primeros 30 s de enfriamiento se presenta el mayor desplazamiento, indicando su influencia por la reacción térmica durante el enfriamiento.
5. Los esfuerzos internos más altos se presentaron para un HTC *f*(T), para un HTC constante los esfuerzos disminuyeron.
6. Es conveniente para este tipo de procesos, lograr obtener un HTC constante es decir tratar de mantener el baño de inmersión a una temperatura lo suficientemente baja para disipar el calor de la pieza, para disminuir los esfuerzos residuales y prevenir una distorsión macroscópica.

5. AGRADECIMIENTOS

Los autores agradecen al CINVESTAV Unidad Saltillo, por la prestación de la licencia del software DEFORM-3D, para el desarrollo de este proyecto.

6. REFERENCIAS

Hömberg D. A. 1996. Numerical simulation of the Jominy end-quench test. Acta Materialia.44, 4375-4385 pp.

Kang S. H. and Im Y. T. 2007, Three-dimensional Thermo-Elastic-Plastic Finite Element Modeling of Quenching Process of Plain-Carbon Steel in Couple with Phase Transformation, International Journal of Mechanical Sciences, 49, 423-439 pp.

Nallathambi A. K, Kaymak Y, Specht E and Bertram A. 2010. Sensitivity of Material Properties on distortion and Residual Stresses During Metal quenching Processes, *Journal of Materials Processing Technology*, 210, 204-211 pp.

Şimşir C. and Gür C. H. 2008. A FEM Based Framework for Simulation of Thermal Treatments: Application to Steel Quenching, Computational Materials Science, 44, 588-600 pp.

Totten G. E and Howes M. A. H. 1997. Distortion of Heat Treated Components, Chapter 5, Steel Heat Treatment Handbook, Marcel Dekker, 292 p.

Woodard P. R, Chandrasekar S and Yang H.T. 1999. Analysis of Temperature and Microstructure in the Quenching of Steel Cylinders. Metallurgical and Materials Transactions. 30B, 815-822 pp.

REVIEW OF DATA INTEGRATION ARCHITECTURES AND THEIR METHODS

O. D. Fernández-Bonilla, M. González-García & R. Santaolaya Salgado

Tecnológico Nacional de México-Centro Nacional de Investigación y Desarrollo Tecnológico CENIDET, Interior Internado Palmira S/N, Col. Palmira, 62490, Cuernavaca, Morelos, México.
odfb81@gmail.com, moises@cenidet.edu.mx, rene@cenidet.edu.mx

ABSTRACT. The development of data integration systems has been an emerging research subject in recent years, roused by the growing number of data repositories of different areas. For developing these systems, there are two types of architectures, materialized and virtualized. In a materialized approach, the data is extracted from assorted sources, pre-processed (according to user's needs) and then stored for further use. In a virtualized approach, an intermediary software create a virtual database, translate a user global query into local source query among the available sources, unify the query results and then presents it as a single one. However, despite the used approach for developing a data integration system, three integration tasks must be done in order to obtain homogeneous data: syntactic integration, schema integration and semantic integration. To solve these integration tasks, virtual and materialized approaches have different methods to solve them. After choosing an integration method and getting results, some techniques to assess the results are needed, in order to evaluate if the obtained data is useful for the intended purpose. This article objective is to show the different methods for data integration and how they work. As a result we show a comparison between previous works and their results using the different methods for data integration.

KEYWORDS: data integration architectures, ontologies, data validation.

1. INTRODUCTION

Obtaining information from data repositories has been a recent subject of study. This was triggered by the availability of numerous data sets. When developing knowledge based system using heterogeneous data repositories, selecting architecture for data integration is required in order to obtain data from heterogeneous repositories. There are two types of architectures for integrating data: materialized and virtualized. The results of the former approach are data repositories with all the data available, updated until the time of pre-process and ready for further analyses. The latter approach uses intermediary software that locally queries each available data source and then combine the results it into a single answer. The results of this approach are data updated until the moment of execution of the intermediary software.

If a user's needs to pre-process data before use it, a materialized approach is more suited since it allows additional processing or analysis over the resulting data; including execution of inference rules without compromising queries performance. However, since a materialized approach pre-process data first, they took a lot of time for delivering a result, therefore is not suitable for users who need current data in a certain period of time. On the other hand, a virtual approach has the advantage that queries are executed against the most current version of the data, which is significant if the data is frequently updated and users need current data at any time [45]. However, a virtual implementation may suffer of query performance penalties if entailment inference rules were applied to the repositories to infer new information or if the data quantities were too big.

Despite the used approach, both have to deal with schematic, syntactic and semantic heterogeneity at some point. Schema integration deals with solving mapping between data sources schemas and the global schema. Syntactic integration deals with changing file formats and data model. Semantic integration is tasked with combining and completing data, solving semantic issues. Since developing a data integration architecture and choosing its respective method is crucial for any knowledge discovery system, a knowledge manager would need all the information of each architecture and their methods in order to choose one that satisfy his/her needs. Therefore, we need to evaluate not only the architectures and methods but also the available resources at

hand, what types of data results are needed and how to measure them. The rest of the paper is organized as follows. Section 2 defines data integration, the heterogeneity problems and requirements to take into account when developing a data integration system. In Section 3, the materialized and virtualized approaches and their respective methods are described. In Section 4 methods to validate obtained results in data integration systems are shown. In section 5 previous works are presented and, finally, section 6 presents the conclusions.

2. DATA INTEGRATION OVERVIEW

A combination of multiple information systems that aims to combine a set of data sources, so that they form a unified new whole and give users the impression of interacting with one single information source. To that end we use data integration which purpose is to manipulate data across multiple data sources. Nevertheless, the literature offer many different definitions of how data integration problems are managed.

In [44], the author gathered several definitions from different papers. One definition describes data integration as the merging of data from different sources to provide the user with a unified view of data. The users should be provided with uniform interface to data sources but these data sources need to remain independent. Other definition describes data integration as the situation where several data sources, each one with an associated local schema, are integrated to form a single virtual database with an associated global schema transforming several source data models to a global schema data model. Regardless the different approaches to describe data integration, all these definitions agree with that data integration provides the ability to manipulate data transparently across multiple data sources. Considering the previous assertions, data integration systems are relevant to many fields including medical facts supervision, software engineering, geographical information systems and business applications.

However, integrating heterogeneous data sources is a tedious task that requires a deep analysis of the sources of data and their internal schemas in order to integrate them. This is known as the data integration process; this process deals with different heterogeneity conflicts. These problems are [33]:

- Different naming conventions, duplicated and/or inconsistent data and different units for values. These problems are complex because semantics may be embodied in data models, conceptual schemas, application programs an data itself. These difficulties concerning semantics are the reasons for many still open research challenges in the area of data integration [8].
- Structuring same information in different ways and different data types.
- Differences between data models, language representations and file formats.

2.1 Characteristics of a data integration system

While the goal of any data integration system is to deliver a homogeneous and unified view on data from heterogeneous sources, how they do it, what resources they need and how they give results is another matter. So before choosing a data integration approach and its respective methods, these characteristics must be considered [36, 44]:

- The architectural view of an information system. How data will be viewed or consulted by end users.
- The content and functionality of the component systems.
- The type of information that is contained by component systems, simple or compound. Simple types of information are alphanumeric and numeric data. Compound types of information are structured and semi-structured.
- Intended use of the results.
- Performance requirements. How fast and updated the data needs to be delivered.
- Metadata use and management

3. MATERIALIZED AND VIRTUALIZED APPROACHES

Before describing the methods for each approach, the elements to evaluate them are show. This evaluation was based on the evaluation elements given by [20, 42].

- Level of integration. The method does syntactic, schema and semantic integration.
- Update of source data. How often does the method update the data from their sources.
- Complexity. How hard is the developing, configuration and maintaining of a method.
- Time of execution. How much time the system requires for delivering a result.
- Results. Type of results the system give.

3.1 Materialized approaches and their methods

The main objective of a materialized approach is to extract the information from the heterogeneous sources, pre-process and send it to a central repository. The main advantage of this approach, is the processing of the data before providing it to the user, so further analysis can be done to the data. The methods used in materialized approaches are [3, 42]:

- Data warehouse: Data is extracted, transformed and loaded into a designated data warehouse. Then, further analysis can be done, such as online analytical processing, data mining, use of data cubes, and data reports. Data warehouses can divide data using data marts.
- Operational data stores: A warehouse with fresh data is built by immediately propagating updates in local data sources to the data store. Therefore, current data is available for decision support. Contrasting with data warehouses, data is neither cleansed nor aggregated and metadata is not supported.
- Portals: A portal is a form of uniform data access; portals are personalized doorways to the Internet or Intranet, where each user is provided with information according to his or her information needs. Data is not pre-processed and users need to have knowledge of SQL.
- Manual integration: Users directly interact with all relevant information systems and manually integrate selected data. That is, users have to deal with different user interfaces and query languages. Additionally, users need to have detailed knowledge on location, logical data representation, and data semantics.

Table 1 show an evaluation of each method, using the factors previously described.

3.2 Virtualized approaches and their methods

A virtualized approach main objective is to query heterogeneous data sources using mediator software, combine the obtained results and then send it to the user as a single outcome. Their main advantage is that they do not make copies of source data and show the latest version of the available data sources, consequently giving faster and current results. The methods for virtualized approaches are [3, 43]:

- Mediated query systems: They present a uniform data access solution by providing a single point for read-only querying access to various data sources [41]. The mediator software contains a global query processor employed to send subqueries to the local data sources.
- Federated dataset systems: Are fully-fledged database manager systems; that is, they implement their own data model, support global queries as well as global transactions and global access control. They achieve a uniform data access solution by logically unifying data from underlying local database manager systems.
- Peer to peer integration or P2P: It is a decentralized approach for integration of distributed, autonomous databases. Data is mutually shared and integrated through mappings between local schemas.

Table 2 show an evaluation of each method, using the factors previously described.

3.3 Use of ontologies in data integration approaches

Ontologies are used in data integration systems because they provide a common schema, domain vocabulary for heterogeneous data, metadata can be embedded with the model and an explicit machine-understandable conceptualization of a domain using a formal language [32, 33].

The main advantage of using ontologies in data integration methods are that despite the

architecture approach for data integration (materialized or virtualized) an ontology can be added to the solution without changing how the solution works. There are three approaches for using ontologies as described in [37]: a) global ontology approach, b) multiple ontology approach and c) hybrid approach.

- **Global ontology approach**. All source schemas are directly related to a shared global ontology that provides a uniform interface to the user. However, this approach requires that all sources have nearly the same view on a domain, with the same level of granularity.
- **Multiple ontology approach**. Each data source is defined by its individual local ontology separately. Instead of using a global ontology, local ontologies are mapped to each other. For this purpose, an additional representation formalism is necessary for defining the inter-ontology mappings. The inter-ontology mapping identifies semantically corresponding terms of different source ontologies.
- **Hybrid ontology approach**. First, a local ontology is built for each schema, which is not mapped to other local ontologies, but to a global shared ontology. Then new sources can be easily added with no need for modifying existing mappings. It also supports the acquisition and evolution of ontologies. The use of a shared vocabulary makes the source ontologies comparable.

Each approach for ontology based data integration has its advantages and disadvantages. Based on [34], Table 3 describes them. Ontologies need to be developed in order to be used in any of the mentioned data integration architectures. There are many well-known methodologies for building ontologies, [38, 39, 40]. Among them, five activities in common, for building ontologies were identified.

- Specification: Identification of the purpose and the scope of the ontology.
- Conceptualization: A conceptual model of the ontology is constructed. It consists of the different concepts, relations and properties that can occur in the domain.
- Formalization: The conceptual model is translated into a formal model.
- Implementation: The formal model is implemented in a knowledge representation language.
- Maintenance: The implemented ontology has to be constantly evaluated, updated and corrected.

Table 1. Methods for materialized approaches.

Method	Level of integration	Update of source data	Complexity	Execution time	Results
Data warehouse	Syntactic, semantic and schema	Rarely	Medium	Medium	A repository ready for further analysis
Operation data store	Syntactic and schema	Very often	Low	Very fast	A repository of data for reports or graphics
Portals	Schema and syntactic	Very often	High	Fast	A web page with the results
Manual integration	Syntactic, semantic and schema	Very rarely	Very high	Very slow	A tailored repository according to users' needs

Table 2. Methods for virtualized approaches.

Method	Level of integration	Update of source data	Complexity	Execution time	Results
Mediated query systems	Syntactic and schema	Often	High	High	A webpage showing the results as a local query.
Federated databases	Syntactic	Often	High	High	Results for each local database.
P2P	Syntactic and schema	Often	High	Medium	Data is for read only

Having in mind how ontologies work in data integration and how to build them, Table 4 shows how the different architectures for data integration use them. The first column includes the integration method; the second column includes the type of ontology they use and the third column includes how the method uses ontologies.

Table 3. Advantages and disadvantages of the different ontologies architectures.

Criteria	Ontology-based data integration architectures		
	Global	Multiple	Hybrid
Evaluation of semantic heterogeneity	Useful for systems, which have the same view on a domain	Useful for systems which have the same view on a domain	Useful for systems, which have different views on a domain
Appending new data Sources	Some modification is necessary in the global ontology	Supports appending of new data sources with some adaptions in other ontologies	New data sources can easily be added without the need of modification
Elimination of data sources	Some modification is necessary in the global ontology	Supports an opportunity to remove the data source with some adaptions in other ontologies	Data sources can easily be removed without the need of modification
Comparison of multiple ontologies	Impossible	Difficult, because the lack of a common vocabulary	Simple, because ontologies use a global shared vocabulary
Implementation effort	Straight-forward	Costly	Reasonable

Table 4. Integration methods and type of ontologies used by each one.

Method	Type of ontology	Used for
Data warehouse	Single or hybrid	Building central data integration systems Defining vocabulary Managing data related to integration process
Operation data store	Single	Define vocabulary

Portals	Single	Building central data repository
Manual integration	Single	Manually solve semantic issues
Mediated query systems	Multiple	Defining local data sources
Federated databases	Multiple	Defining local data sources
Peer-to-peer integration	Multiple	Defining source peers Defining vocabulary of each peer

3.4 Metadata in data integration systems

During the integration process of a materialized or virtualized approach, metadata is produced. Metadata can be about, from where the data was obtained, how query the data or how to process it, etc. Therefore, in any integration process, it is necessary a way to manage, process and keep metadata. Ontologies can be used for metadata management. For example an ontology can keep track of what type of methods where used for pre-processing data. Ontologies can also give a detailed description of data sources or describe the destiny repository [2]. According to [21, 22, 23, 24, 25] metadata is an important part of any integration process. Nonetheless, there is not a standard for storing the metadata obtained during the life cycle of a data integration process, as described in [25, 27, 28].

In addition, during analysis of previous works (see [30, 31, 32]) it was found that despite the obtained results from their researches, they do not describe or mention what they did with the metadata generated during the integration process. During the planning of a data integration process, metadata is essential for these processes, so it is reasonable to say that not considering it is a mistake. In a data integration process, without metadata, the user would have to manually manage and solve semantic heterogeneities and data transformations issues, despite the applied data integration method. Also without metadata, the user would have to define formal vocabulary of the domain or the business rules. In addition, the user would have to manually do the mapping between attributes and the respective data transformations and manually search the meaning of each word, therefore increasing the possibility of human error. The way the diverse methods use metadata is shown in Table 5.

Table 5. Metadata use.

Method	Use of metadata
Data warehouse	Description of source data, data processing and business rules
Operation data store	Description of source data
Portals	For making queries
Manual integration	Description of integration methods and for making queries
Mediated query systems	Description of source data and making local queries
Federated databases	Description of source data and making local queries
Peer-to-peer integration	Description of a peer and source data

4. DATA VALIDATION IN DATA INTEGRATION SYSTEMS

Once data is integrated, it needs to be verified to see if it is useful for the planned purpose. Integrated data is not necessarily validated data, therefore if the data is not validated; it could cause unintended results in the final application. So a way to measure quality of data is needed. Data quality is defined as a multidimensional concept, where each dimension represents the views, criteria, or measurement attributes for data quality problems that can be assessed, interpreted, and possibly improved individually. [26]. Given this concept, measurement attributes are needed in order to validate the results in a data integration system. According to [29], these attributes are completeness, accuracy and consistency. Completeness measures the amount of data, in terms of both the number of

attribute values and the number of attributes. However, not all attributes and their respective values are equally important. Therefore, when completing missing data, the method must specify if the focus is data density or data coverage.

- Coverage of the data is focused only on attribute values inside the database.
- Density of the data is defined as the number of attributes represented in the database compared to the required real-world object attributes.

Completeness measure is assessed using formula (1). Given a set of attribute values AtrVal(A1,...,An), formed by all the attribute values in a repository from a set Ai. We can define the set N_{Ai} as the set with all non-null attribute values in AtrVal.

$$N_{Ai} = \{ data \varepsilon Atrval \mid NotNull (data: A_i) \} \quad (1)$$

$$C(A_i) = \frac{|N_{Ai}|}{Atrval} \quad (2)$$

Accuracy measures correctness of data, that is, whether the data conform to the real world value. In addition, it specifies the distance to the actual real world value or the ideal degree of detail of an attribute value. To evaluate accuracy for a particular value n, the real world value n or a reference value is required.

Consistency is defined as the degree to which data satisfies data constraints. Therefore, consistency measures the uniqueness of object representations in terms of both the number of unique objects and the number of unique attributes of the objects, according to a set of rules or constraints In order to evaluate data consistency, the set of tuples from a repository R must satisfy a set of business rules BR. This is expressed in (3):

$$CS_{BR} = \{data \varepsilon R \mid Satisfies (data, BR)\} \quad (3)$$

Where CS_{BR} is the set of all attributes that satisfy the set of business rules BR. Data consistency is achieved when all data in the repository satisfies the business rules. If data violate any of these rules, data is not consistent. Therefore, data consistency can only be fully achieved or not achieved.

5. PREVIOUS WORKS

Several research works reported the development of a data integration system in order to obtain knowledge from data repositories in a domain of interest. All of these works used materialized or virtualized integration architectures. Therefore, in this section a brief description of previous works, with focus in what they did, how they did it and how they validated their results is presented. Next, a comparison of the previous works and their most significant characteristics is shown in Table 6. Finally, we describe how these works are related. In [1] the authors proposed a knowledge discovery method by integrating schemas and instances from a biomedical data repository. They use a mediator software architecture. By using virtual schemas, they mapped the attributes into a single ontology and then after that, they gave suggestions to a user about how to process the obtained data using a data pre-process ontology. After pre-processing the data according to the user specifications, they stored the information into a database. To validate their results they compared their obtained patterns with their proposed method against the patterns obtained without processing data and patterns obtained with schema integration. After the patterns are obtained, they used precision formula to compare them.

In [3] the authors developed the OntoCloud System. They used on open source software and open standards of a dynamic ontology based database integration system with inference capabilities. Ontocloud provides dynamic access to a consolidated database global view of several database sources using ontologies to consolidate heterogeneous data. Ontocloud use four ontologies. The global ontology lists the classes and properties in which the global database will be represented, as well as annotations. The federation ontology specifies the source databases and which classes and properties of the global ontology they implement. The mapping ontology relates tables and columns from a source database to basic concepts on the global ontology. The inference ontology maps derive concepts to basic concepts through an ontology alignment file. To validate these results they compared their system against three different systems. This proved that their systems are better by showing what their work did that the others did not.

In [4] the authors developed MOMIS-STATIS,

an approach for data integration based on ontologies. The objective was to create a comprehensive application suite, which allows enterprises to simplify the mapping process between data schemas based on semantics. To implement their method they first obtained a neutral schema representation, then they annotated the local sources using ontologies and finally they make the semantic mapping using an ontology. They showed the results of how their application maps but they do not validated their results.

In [5] the authors present Searchy, an agent based platform which uses heterogeneous semantic wrappers, integrates information from arbitrary sources and translates them into semantic terms. Searchy is an agent-based solution in which agents contain several wrappers that fetch information in local formats, translate them into a semantic standard format and integrates the information spread across several information systems. This agent is a wrapper platform that eases wrapper development providing an execution environment and several features. The application was proved in a website with success but they do not showed how they validated the obtained results.

In [6] the authors proposed InteGRail, a system for integrating data from the European railway system. They used a hybrid ontology approach and data conversion for integrating the information of the diverse railway systems. They proved their application by checking the time of three different queries with success but they did not verify the quality of the obtained data.

In [7] the authors propose an architecture for assisting part of the integration process from geographic sources. Their integration process is based on two main sets of tasks – non-logic and logic. The former is meant at finding similarities based on structural and syntactic analyzes of geographic data; and the latter is used to calculate inferences from semantics of data by using ontologies. They proved their application with two different geographic information systems with success but they did not verify the quality of the obtained data.

In [8] the authors described the current approaches for integrating data in the biomedical field using semantic web technologies. They described problems and divided them into semantic, syntactic and structural heterogeneity problems. They describe how ontologies can be used in data integration in biomedical data repositories. They show the three ontology integration approaches and how to implement them.

In [9] the authors integrated data from prostate cancer databases. Their aim was to transform diverse attributes from databases into a single one and then send them to a unified database that contains all the information. To do it, they used the information obtained from diverse prostate cancer databases and then used a mapping subsystem and finally a data query subsystem. They proved that their application was successful by making a series of queries and showed that without their systems the user would have to manually query and integrate the results.

In [10] the authors presented their experience in the search of knowledge in software engineering repositories. They represented several available software repositories and the information their attributes contain. At the end, they showed a comparison between repositories and described current issues when searching for data in them.

In [11] the authors developed a website for searching data using data links for navigating. Users needed to know SQL in order to extract information from the repositories. This research provided schema and metadata browsing capabilities, but the user must manually obtain it and the user must know what he or she is looking for.

In [12] the author described an experiment for searching knowledge in two repositories, the ISBSG and CSBSG. By using supervised learning and un-supervised learning knowledge methods, he proved that supervised is better to obtain knowledge. The author measured the results using the accuracy and F-Measure formulas. The author mentioned that before using knowledge discovery methods, he cleaned and normalized the data repositories. At the end, he established that with more data, it is possible to obtain better results but unless it is done, it is not certain.

In [13] the researchers obtained information about several software development projects

from diverse enterprises and then used the obtained information to estimate new projects. To obtain the information, a questionnaire was sent to diverse software enterprise. After obtaining the information, it was manually analyzed. However, the authors stablished that their data was too limited for meaningful statistical studies.

In [14] the authors developed an Asset Description Metadata Schema for Software (ADMS.SW). ADMS.SW is an ontology developed for describing software packages, releases and projects. It can be applied to describe packages in free distributions using semantic web methods. They proved their method by linking data of different projects in one package.

In [15] the authors proposed the use of domain ontologies in a software development project for reducing miscommunication problems and allowing project leaders to have a better understanding of the project and allowing project members to communicate more efficiently. They used a web interface for information retrieval and modifications.

In [16] the authors evaluated three approaches for extracting fault data from an open source system repository. Their main objective was to prove that it is very difficult to extract fault-fixing data from repositories, especially using automatic tools. After the experimentation, they proved that an expert is always necessary to validate information obtained from data repositories.

In [35] the authors presented a data integration system of autonomous data sources. Their system main advantage was the use of temporal metadata properties as the input for the integration algorithm presented in this work. Once data is integrated, they used their algorithm, to obtain more precise data in the integrated repository. They proved their results against three previous works and measured the results using three aggregation methods.

Table 6, describes the reviewed related works. The first column is the reference to the article, the second column is the used data integration architecture in the analyzed article, the third column shows if the article used ontologies and what type of ontology they used, the fourth column shows the data validation methods used to verify the obtained results and the last column describes the type of obtained results.

From an analysis of Table 6, these similarities were found:

- In works [1, 3, 9] the authors proved their results by comparing them with previous projects, but data quality was not assured.
- In [4, 5, 6, 7, 14, 16] the authors proved that their application gave faster or better results than a manual approach, but they did not verified data quality.
- In [8, 10, 11, 12, 13] the authors showed the importance of data integration; how ontologies can be used to help in the integration process and that the data integration problem at the semantic level is still and open issue.
- In [12, 16] they were the only authors that mentioned they cleaned the data before integrating it.
- In [35] the authors described the importance of measuring precision in a data integration process.
- In [17, 18, 19] authors described the importance of cleaning the data.

As it can be seen from the focus of previous works, when the objective was to give the data as fast and current as possible, a virtual approach was implemented. On the other hand, when further data processing was needed on the results or the results were to be used for additional data analysis methods, a materialized approach was implemented. Nevertheless, we can also conclude that none of the works measures their results with formal methods to evaluate quality of data.

6. CONCLUSIONS

In this paper, we have shown the data integration concept, available architectures for development and their characteristics. Therefore, in order to select the most appropriate choice for a data integration system, we considered several characteristics. For example, required results, available resources, preferred communication for system components and functionality, metadata management and methods to verify quality of data.

Table 6. Related works results and used integration architecture.

Author	Used architecture	Use of ontologies	Data validation	Obtained results
Anguita et al.	Virtual approach	Hybrid ontologies	Compared their results with other methods	Read only data
Diogo et al.	Virtual approach	Multiple ontologies	Compared their results with other methods	Read only data
Beneventano et al.	Virtual approach	Multiple ontologies	Faster data mapping than previous manual mapping	Read only data
Barrero et al.	Virtual approach	Global ontology	Used within another system with success	Read only data
Verstichel et al.	Materialized approach	Hybrid ontologies	Faster query results than previous systems	Data processed according to user needs
Buccella et al.	Materialized approach	Multiple ontologies	Used within another system with success	Data processed according to user needs
Min et al.	Virtualized approach	Multiple ontologies	Faster query results than previous systems	Data processed according to user needs
Howison et al.	Materialized approach	No ontologies	No validation	Data processed according to user needs
Zhang et al.	Materialized approach	No ontologies	Using accuracy and F-Measure	Data processed according to user needs
Morisio et al	Materialized approach	No ontologies	Manual analysis according to their needs	Data processed according to user needs
Berger et al.	Materialized approach	Global ontology	No validation	Data processed according to user needs
Wongthongtham et al.	Materialized approach	No ontologies	No validation	Data processed according to user needs
Hall et al.	Materialized approach	No ontologies	Manual analysis according to their needs	Data processed according to user needs
Salguero et al.	Virtualized approach	No ontologies	Compared their results with other methods	Read only data

A characteristic that most works studied and that was not contemplated is how metadata will be managed during data integration process. We showed that all data integration methods use metadata at some point, so a method to manage metadata is essential. Only in [2], is acknowledged that metadata is an essential part of an integration process. Another characteristic that was not considered in the researched works is data quality verification. After seeing the diverse works, despite what type of architecture and method the authors used, none of them applied data quality measurements. Many of them focused on describing how they executed better than previous systems or how fast they got their

results. However, these aspects are not related to data quality. Therefore, it is a mistake not to consider how and when to measure these aspects during a data integration process. With respect to the use of ontologies, of the 14 studied works 8 used ontologies. This proofs that the use of ontologies as a domain knowledge representation and as a tool is increasing. However, of the 8 research that used ontologies, only 2 used hybrid methods, showing that despite that hybrid ontologies avoid the problems of global and multiple ontologies, their implementation is still an open issue.

Finally 8 research works gave pre-processed data, meaning that the data needed to be cleaned before giving it to the user. As we said it before, one of the problems of current data integration systems is how users and groups formatted the data they donate. Even if the data make sense for them it would not make sense for other users unless they gave a semantic structure or a personalized dictionary that describes the meaning of the concepts, their relationships, data types and allowed instances. Ontologies are a current effort to give a shared meaning to data, though to implement them is still hard because of the need of an ontology developer expert and a knowledge domain expert.

As for future works, we will develop a data integration system oriented to software project development using free data. This project will use hybrid ontology and methods for measuring the obtained results to assure the quality of the integrated data.

7. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to all those somehow involved in this study and to the anonymous reviewers of this paper for their valuable comments. We would also like to thank CONACyT for the financial support to this research.

8. REFERENCES

1. Anguita A., Barreiro J. M., Crespo J., De la Calle G., García-Remesal M., Maojo V., Martín L., Pazos J., Pérez-Rey D., Rodríguez-Patón A., Silva A.: “Integración y Preprocesamiento Basado en Ontologías de Repositorios de Datos Biomédicos Distribuidos”. INFORSALUD 2007 – X Congreso Nacional de Informática de la Salud, Madrid, Marzo 2007.
2. Ismael Navas-Delgado, José F. Aldana-Montes.: “Extending SD-Core for Ontology-based Data Integration”. Journal of Universal Computer Science, vol. 15, no. 17 2009 pp 3201-3230.
3. Diogo F.C. Patrao, Helena Brentani, Marcelo Finger, Renata Wassermann.: "Ontocloud Clinical Information Ontology Based Data Integration System”. ONTOBRAS 2013 Belo Horizonte, Brazil September.
4. Domenico Beneventano, Mirko Orsini, Laura Po, Serena Sorrentino.: “The MOMIS STASIS approach for Ontology-Based Data Integration”. Workshop Interoperability through Semantic Data and Service Integration Camogli Genova, Italy. June 25, 2009.
5. David F. Barrero, Maria D. R-Moreno.: “Information Integration in Searchy: An Ontology and Web Services Based Approach”. International Journal of Computer Science and Applications. 7, 2, 2010 pp. 14 – 29.
6. Stijn Verstichel, Femke Ongenae a, Leanneke Loeve, Frederik Vermeulen, Pieter Dings, Bart Dhoedt, Tom Dhaene, Filip De Turck.: “Efficient data integration in the railway domain through an ontology-based methodology”. Transportation Research Part C 19 2011 617–643.
7. Agustina Buccella, Alejandra Cechich.:”Towards Integration of Geographic Information Systems”. Electronic Notes in Theoretical Computer Science. Number 168 2007 pp 45–59.
8. Roland Kienast, Christian Baumgartner.: “Semantic Data Integration on Biomedical Data Using Semantic Web Technologies, Bioinformatics - Trends and Methodologies”. Dr. Mahmood A. Mahdavi (Ed.), ISBN: 978-953-307-282-1, InTech, DOI: 10.5772/21086. Available from: http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/semantic-data-integration-on-biomedical-data-using-semantic-web-technologies

9. Hua Min, Frank J. Manion, Elizabeth Goralczyk, Yu-Ning Wong, Eric Ross, J. Robert Beck.: "Integration of prostate cancer clinical data using an ontology". Journal of Biomedical Informatics 42 2009 pp. 1035–1045.
10. D. Rodriguez, I. Herraiz, and R. Harrison, "On software engineering repositories and their open problems," in The International Workshop on Realizing AI Synergies in Software Engineering (RAISE'12), Zurich, Switzerland, 2012. June 2012 pp 52-56.
11. James Howison, Megan Conklin, Kevin Crowston.: "FLOSSmole: A collaborative repository for FLOSS research data and analyses". International Journal of Information Technology and Web Engineering. Volume 1 Issue
12. Wen Zhang, Ye Yang, and Qing Wang.: "A Study on Software Effort Prediction Using Machine Learning Techniques". ENASE, CCIS 275, pp. 1–15, 2013. Berlin Heidelberg 2013.
13. Maurizio Morisio, Michel Ezran, Colin Tully.: "Success and Failure Factors in Software Reuse". IEEE Transactions on Software Engineering. 28, 4, May 2002 pp 340-357.
14. Olivier Berger, Christian Bac.: "Linked Data descriptions of Debian source packages using ADMS.SW". International Conference on Open Source Systems July 2013 pp 168-181.
15. Wongthongtham, P.; Chang, E.; Dillon, T. "Ontology Modelling Notations for Software Engineering Knowledge Representation", Digital EcoSystems and Technologies Conference, DEST '07. Inaugural IEEE-IES, pp 339 – 345 2007.
16. Tracy Hall, David Bowes, Gernot Liebchen, Paul Wernick.: "Evaluating Three Approaches to Extracting Fault Data from Software Change Repositories". 11th International Conference, PROFES 2010 Limerick, Ireland, pp. 107–115 2010.
17. Nguyen Hung Son.: "Data cleaning and Data preprocessing". Available at http://www.mimuw.edu.pl/~son/datamining/datamining.htm
18. Richard D. De Veaux, David J. Hand.: "How to Lie with Bad Data. Statistical Science". 20,. 3, 2005 pp. 231–238.
19. Deepshikha Aggarwal, V. B. Aggarwal.: "An Optimum Model for the Retrieval of Missing Values for Data Cleansing using Regression Analysis". International Journal of Computer Applications. 117, 2, May 2015, pp 35-39.
20. A. Halevy and C. Li, "Information integration research: Summary of nsf idm workshop breakout session," NSF IDM Workshop, 2003. Seattle, Washington published in September 2004.
21. Osmar R. Zaïane.: "Principles of Knowledge Discovery in Data Chapter 3: Data Pre-processing". Available at http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/materials.shtml#clean 2004.
22. Rahm Erhard Do, Hong-Hai.: "Data Cleaning: Problems and Current Approaches". IEEE Techn.Bulletin on Data Engineering, 23, 4, Dec. 2000 pp 3-13.
23. Jens Bleiholder, Felix Naumann.: "Data Fusion". ACM Comput. Surv. 41, 1, Article 1, December 2008, pp 1-41
24. Hong Hai Do, Erhard Rahm.: "On Metadata Interoperability in Data Warehouses". Techn. Report 1-2000, Department of Computer Science, University of Leipzig. http://dol.uni-leipzig.de/pub/2000-13.
25. Murtadha M. Hamad, Alaa Abdulqahar Jihad.: "The Role of Metadata for Effective Data Warehouse". Journal. of University of Anbar for pure science. 6, 2, 2012, pp 95-100.
26. Ling Liu, M. Tamer Ozsu. "Encyclopedia of Database Systems". Springer Science+Business Media, LLC 2009.
27. Liane Carneiro, Angelo Brayner.: "X-META: A Methodology for Data Warehouse Design with Metadata Management". In: Proceedings of the 4th International Workshop on Design and Management of Data Warehouses. Volume 2, 2002, pp 13-22.
28. David Marco.: "Building and Managing the Metadata Repository: A Full Lifecycle Guide".Wiley Computer Publishing 2000.

29. Xin (Luna) Dong, Felix Naumann.: "Data Fusion – Resolving Data Conflicts for Integration". 35th International Conference on Very Large Data Bases VLDB, Lyon, France. August 2009.

30. Menzies 2003]Tim Menzies, Justin S. Di Stefano. "More Success and Failure Factors in Software Reuse". IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 29, NO. 5, MAY 2003.

31. Matthew Van Antwerp, Greg Madey.: "Advances in the SourceForge Research Data Archive". Workshop on Public Data about Software Development (WoPDaSD) at the 4th International Conference on Open Source Systems, Milan, Italy, September 2008.

32. Isabel F. Cruz Huiyong Xiao.: "The Role of Ontologies in Data Integration". Journal of Engineering Intelligent Systems. 13, 4, 2005 pp 245-252.

33. Michel Gagnon.: "Ontology-Based Integration of Data Sources". 10th International Conference on Information Fusion, Quebec, Quebec. 2007 pp 1-8. July.

34. Virginija Uzdanaviciute, Rimantas Butleris.: "Ontology-based Foundations for Data". The First International Conference on Business Intelligence and Technology BUSTECH 2011. Rome, Italy. September 2011 pp 34-40.

35. Alberto Salguero, Francisco Araque.: "Integration of Similar Evolving Data Sources for Supporting Decision Making Tasks". Journal of Universal Computer Science. 16, 1, 2010, pp 22-36.

36. Patrick Ziegler, Klaus R. Dittrich.: "Three Decades of Data Integration— All Problems Solved?" 18th IFIP World Computer Congress (WCC 2004), Building the Information Society/Springer. Toulouse, France. pp 3-12 August 2004.

37. Helena Sofia Pinto, João P.Martins .:"Knowledge and Information Systems". New York, NY. 6, 4, July 2004, pp 441 – 464.

38. Denny Vrandecic, Sofia Pinto, Christoph Tempich and York Sure.: "The DILIGENT knowledge processes". Journal of Knowledge Management VOL. 9 NO. 5, pp. 85-96. 2005.

39. Mariano Fernández, Asunción Gómez-Perez, Natalia Juristo.:"METHONTOLOGY: From Ontological Art Towards Ontological Engineering". Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering, Providence, Rhode Island. July 1997, pp. 33–40.

40. Uschold, M., King, M.:"Towards a methodology for building ontologies". Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal, Quebec, Canada, pp 20–25 August 1995.

41. Gio Wiederhold.: "Mediators in the Architecture of Future Information Systems". IEEE Computer. 25, 3, 1992, pp. 38-49

42. Hristo Hristov.:"Choosing Approach for Data Integration" Information System and Grid Technologies, Sofia Bulgaria. pp 98-113, June 2012.

43. Maurizio Lenzerini.: "Principles of Peer to Peer Data integration". Proceedings of the Third International Workshop on Data Integration over the Web, Riga, Latvia. Volume 4 June 2004.

44. Patrick Ziegler, Klaus R. Dittrich.: "Data Integration — Problems, Approaches, and Perspectives". Conceptual Modelling in Information Systems Engineering Springer pp 39-58. 2007.

45. Ladjel Bellatreche, Guy Pierra, Nguyen Xuan Dung, Dehainsala Hondjack.: "An Automated Information Integration Technique using an Ontology-based Database Approach". Proceedings in the International Conference on Concurrent Engineering (ISPE CE 2003), Madeira, Portugal. July 2003.

MODELOS PARA LA CLASIFICACIÓN DE FRASES CLAVE EN TEXTOS CIENTÍFICOS

G. Flores-Petlacalco[1], M. Tovar-Vidal[1], J. A. Reyes-Ortiz[2] & A. P. Cervantes-Márquez[1]

[1]Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación Puebla, Puebla, México, gerardo.florespe@alumno.buap.mx, mtovar}@cs.buap.mx, patty}@cs.buap.mx
[2]Universidad Autónoma Metropolitana, Azcapotzalco, 02200, México. jaro@correo.azc.uam.mx

RESUMEN. Una frase clave engloba la idea general de un texto, además contiene implícitamente los recursos que el autor usó a lo largo del desarrollo de su investigación para lograr su objetivo, de ahí la importancia de crear modelos de clasificación que permitan agrupar las frases clave de acuerdo con su contenido. En este trabajo, se hace una comparación entre dos enfoques propuestos para la clasificación de frases clave en textos científicos con el objetivo de reconocer los recursos, materiales o procedimientos que el autor empleo en su trabajo de investigación, el primer enfoque se basa en el contexto de la palabra clave para hacer la clasificación, mientras que el segundo enfoque recopila más características que abarcan propiedades de la frase clave, su contexto y medidas de pesado de términos. En la clasificación se usan modelos basados en Naïve Bayes, Máquinas de Soporte Vectorial, Zero R y Árboles de Decisión, para hacer una comparación de rendimiento entre ellos y los resultados de otros autores que realizaron la misma tarea.

PALABRAS CLAVE: Aprendizaje automático, clasificación, frases clave, Naïve Bayes.

ABSTRACT. A key phrase has a general idea of a text, in addition, have implicitly resources used by the author in the development of his research to achieve his objective, hence the importance of creating classification models for key phrases using content. In this paper, a comparison between two proposed approaches for the classification of key phrases on scientific texts in order to know the resources, materials or procedures used by the author in its research work. The first approach is only based on the context of the key phrase for making a classification, the second approach makes a compilation of more features that extends more properties of the key phrase, the context and heavy of term measures in the classification. We use models based on Naïve Bayes, Support Vector Machines, Zero R and Decision trees for making a comparison between them and with the results of others authors in the same task.

KEY WORDS: Machine Learning, Classification, Key phrases, Naïve Bayes.

1. INTRODUCCIÓN

En un artículo científico existen ciertas sentencias que son de suma importancia dentro del texto pues contienen la idea central del trabajo realizado, tales sentencias se conocen como frases clave. Estas frases clave también capturan implícitamente los recursos que el autor usó para llegar al objetivo planteado. El reconocer los recursos empleados dentro de un texto es una tarea de interés para el área del Procesamiento de Lenguaje Natural que en años recientes aumentó de importancia gracias a su complejidad y la utilidad que tiene al ayudar a la mejor comprensión del conocimiento. Desde el año 2012 SemEval, un congreso de análisis semántico propone una lista de tareas que involucran el Procesamiento de Lenguaje Natural y sus áreas de investigación. En la edición 2017 fueron propuestas un total de doce. Entre ellas la tarea número 10, “Extracting Keyphrases and Relations from Scientific Publications” (Extracción de frases clave y relaciones desde publicaciones científicas) (Augenstein et. al., 2017) que contiene tres subtareas:

A. Identificación de frases clave.
B. Clasificación de frases clave.
C. Extracción de relaciones semánticas entre dos frases clave identificadas.

En este trabajo de investigación, se propone un enfoque para resolver la subtarea B “Clasificación de frases clave”, que tiene como objetivo hacer una clasificación de frases clave para identificar recursos dentro de un texto científico; la clasificación se realiza sobre tres tópicos:

1. Materiales. Son frases claves que indican recursos materiales usados en la investigación.
2. Procesos. Son frases claves relacionadas con algún modelo científico, algoritmo o proceso.
3. Tarea. Son frases claves que denotan aplicaciones, objetivos o tareas a realizar en la investigación.

Para realizar la tarea de clasificación se proponen y comparan dos enfoques, el primero sólo considera el contexto de la frase clave, mientras que el segundo enfoque además de considerar el contexto toma en cuenta características de la frase clave y medidas de pesado de términos. Las características se obtienen de frases clave previamente clasificadas en un conjunto de entrenamiento, que posteriormente se introducen a un programa que obtiene un modelo de clasificación y se evalúa el modelo clasificando frases clave identificadas en un conjunto de prueba. Finalmente, evaluamos el rendimiento de los modelos propuestos y comparamos los resultados con otros trabajos del estado del arte que realizan la misma tarea.

El presente trabajo se organiza de la siguiente manera, en la sección 2 se presentan algunos trabajos relacionados, en la sección 3 se presentan las bases teóricas de clasificación, en la sección 4 se explican los enfoques propuestos, en la sección 5 se muestran los resultados experimentales obtenidos de los enfoques y la comparación con otros sistemas del estado del arte y finalmente se presentan las conclusiones del trabajo.

2. TRABAJOS RELACIONADOS

Investigadores han puesto interés en trabajos relacionados con la tarea de clasificación de frases clave, a continuación, se presentan algunos: (Witten et. al., 1999) explican un sistema de extracción de frases clave usando Naïve Bayes y máquinas de aprendizaje automático. Las frases clave son extraídas de textos y posteriormente se clasifican por medio de un modelo creado usando la medida de pesado de términos *TF-IDF* como características principales.

(Wang y Li, 2017) en su investigación proponen un sistema que extrae características lingüísticas y de contexto para realizar una clasificación de frases clave en tópicos. Propone un sistema de clasificación por tres vías usando Máquinas de Soporte Vectorial (SVM) y la librería *scikit-learn*.

(Kim, 2014) describe una serie de experimentos con Redes Neuronales Convulsiónales (CNN) construidas con ayuda de la librería *word2vec*. En la investigación se descubre que una red CNN con solo una capa de convulsión funciona adecuadamente en la tarea de clasificación. Además, resalta el uso de vectores de palabras pre-entrenados como una evidencia de buen funcionamiento dentro del área de aprendizaje profundo en el área de Procesamiento de Lenguaje Natural.

(Eger et. al., 2017) presentan un sistema de clasificación de frases clave que emplea tres diferentes enfoques de clasificación: Redes Neuronales Convulsiónales (CNN), un clasificador basado en *MLP meta-clasifier* y otro basado en *Bi-LSTM*. Las características se extraen usando ingeniería de características de forma que sean dependientes del dominio para hacer más sencilla la tarea de clasificación.

(Liu et. al., 2017) presenta MayoNLP´s un sistema para la clasificación y extracción de relaciones semánticas entre frases clave de textos científicos. En su investigación exploran la similitud semántica y patrones de las frases clave usando modelos pre-entrenados de inserción de palabras. Los patrones conseguidos se combinan con otras características para hacer un Reconocimiento de Entidades Nombradas por medio de Máquinas de Soporte Vectorial, obteniendo buenos resultados.

(Segura-Bedmar et. al., 2017) presentan un sistema para la clasificación de frases clave en textos científicos. Los autores usan un sistema de reconocimiento de entidades basándose en campos aleatorios condicionales, utilizan la herramienta BANNER y características léxicas. Para la clasificación exploran el uso de redes semánticas UML y proponen una relación entre los tipos de frases clave y los grupos semánticos UML, basado en esta relación semántica, crean un diccionario para cada tipo de frase clave y luego con base a las características obtenidas se determina a que grupo semántico pertenece al realizar la clasificación.

3. CLASIFICADORES AUTOMÁTICOS

Los clasificadores tienen la función de agrupar objetos o instancias de acuerdo con sus características en común. Estas características deben ser propias de los objetos a clasificar pues serán usadas para clasificar nuevos objetos de acuerdo con la similitud que se tenga entre las características de la clase y el objeto a clasificar. (Garrido Satué, 2013) menciona que clasificar un objeto consiste en asignarlo a una de las clases existentes de acuerdo con la correspondencia de las características usadas durante el entrenamiento del clasificador. Para crear las clases, se necesitan definir sus fronteras y estas fronteras se calculan mediante un proceso de entrenamiento que usan las características de instancias con su clase bien definida.

Un proceso de clasificación consta generalmente de cuatro pasos:

- Reunir muestras de objetos con las clases definidas. De estas muestras se obtienen vectores de características.
- Se entrena el clasificador. Los vectores de características obtenidos se usan para calcular las fronteras entre las clases existentes.
- Se extraen las mismas características de objetos que se quieran clasificar.
- El clasificador usa las fronteras obtenidas para decidir a qué clase pertenece los objetos que deseamos clasificar.

Existen diferentes tipos de clasificadores, (Ramirez García et. al., 2015) explica tres tipos:

- Naïve Bayes. Clasificador probabilístico que usa el teorema de Bayes para estimar la probabilidad posterior $P(y|x)$ de la clase y dada la variable x. Naïve Bayes se centra en las probabilidades que se refiere a la verosimilitud de x dando el valor y. Por esto, se considera un clasificador generativo.
- Máquinas de Soporte Vectorial. Es un clasificador binario discriminante, dirigido a encontrar el hiperplano óptimo que separa los dos posibles valores de una variable etiquetada de acuerdo al espacio de características presentadas. Se refiere a hiperplano óptimo al margen entre las instancias positivas y negativas de un conjunto de datos de entrenamiento que contiene *N* observaciones.
- Árboles de decisión. Estos árboles describen el conjunto de reglas de forma jerárquica implementando una estructura de decisión. Se compone de hojas y nodos, donde cada hoja registra una respuesta (clase) y cada nodo las condiciones de la clase que le corresponde el valor único de la hoja. Se construyen de forma recursiva con datos de un conjunto de entrenamiento.

Además, en nuestra investigación consideramos un cuarto clasificador conocido como ZeroR, este clasificador simplemente predice la categoría mayoritaria dejando a un lado las predicciones y es útil para determinar un punto de referencia para calcular el desempeño entre otros métodos de clasificación, (Devasena, 2014). Este clasificador está incluido en la herramienta de minería WEKA[1], (Frank et. al., 2016).

4. ENFOQUES PROPUESTOS

Para la tarea de clasificación de frases clave se emplean dos enfoques. El primer enfoque se basa únicamente en el contexto de una palabra clave, mientras que el segundo está basado en características que explotan la estructura del texto y medidas de pesado de términos tal como *IDF*. El *IDF* es una medida que indica la frecuencia de un término *i* en el resto de la colección y se calcula con la Ecuación (1) donde N indica el número de documentos de la colección n_i el número de documentos donde aparece el término *i*, (Vilares, 2008).

$$IDF_i = log(N/n_i) \qquad (1)$$

4.1. Primer Enfoque

Este enfoque únicamente usa el contexto de la palabra clave como característica para hacer su predicción. En este caso se siguen los siguientes pasos que se ilustran en la Tabla 5:

1. Lectura de archivos. Se leen los extractos de texto del artículo junto con su respectivo archivo de anotaciones que contiene las frases clave clasificadas. (Ver Tabla 5, renglón 1).

[1] https://weka.wikispaces.com/ZeroR

2. Pre-procesamiento. El extracto de texto se parte en oraciones usando la función *sent_tokenize*[2] de la librería NLTK de Python. (Ver Tabla 5, renglón 2).
3. Mapeo de frases clave. Se obtienen las posiciones de las frases clave dentro del texto y las posiciones de inicio y fin de cada sentencia. (Ver Tabla 5, renglón 3).

Con base a las posiciones se determina si una frase clave pertenece a una sentencia y si es así se colocan en un conjunto con formato:

[Frase clase 1, Frase clave 2,... [Sentencia]]

4. Extracción del contexto. De cada conjunto que se obtiene en el paso anterior se sigue el siguiente procedimiento para extraer el contexto de la frase clave (Ver Tabla 5, renglón 4).
   a. Si la frase clave se encuentra al inicio de la sentencia, se extrae la frase clave más dos palabras a la izquierda de la misma.
   b. Si la frase clave se encuentra al final de la sentencia, se extrae la frase clave más dos palabras a la derecha de la misma.
   c. Si la frase clave se encuentra en una posición media de la sentencia, se extrae la frase clave más una palabra a la izquierda y otra a la derecha.
5. Creación del modelo. El contexto extraído como característica se complementa con la categoría de la frase clave y se forman tuplas con el siguiente formato:

…
Palabra-1 Frase_Clave Palabra-2, Categoría
Frase_Clave Palabra-1 Palabra-2, Categoría
Palabra-1 Palabra-2 Frase_Clave, Categoría
…

Información que utilizan los clasificadores de *WEKA* para producir el modelo de clasificación (Ver Tabla 5, renglón 5).

6. Evaluación del modelo. El modelo de clasificación fue evaluado sobre extractos y frases clave sin clasificar, es decir, desde un conjunto de datos de prueba para predecir alguna de las clases. A los datos de prueba se les aplica el mismo procedimiento de los pasos 1 hasta el 4 para obtener tuplas del tipo:

…
Palabra-1 Frase_Clave Palabra-2
Frase_Clave Palabra-1 Palabra-2
Palabra-1 Palabra-2 Frase_Clave

Donde no existe un campo de categoría ya que está la predice el clasificador de acuerdo al modelo. Este conjunto de tuplas es la entrada a la función de clasificación creada en el paso 5, para obtener la predicción de la clase o categoría a la que pertenece cada frase clave. Finalmente, las predicciones se comparan con un *gold*-estándar proporcionado por los organizadores de la Tarea 10 de SemEval-2017.

4.2. Segundo Enfoque

A diferencia del primer enfoque, en el segundo se consideran más características aparte del contexto donde se encuentra la frase clave tales como su etiquetado de Partes de la Oración (*PoS* en inglés) y medidas de pesado de términos. En este enfoque se realizan los siguientes pasos y se ilustra en la Tabla 6:

1. Lectura de archivos. Se leen los extractos de texto junto con el archivo de anotaciones que contiene las frases clave previamente clasificadas del conjunto de entrenamiento (Ver Tabla 6, renglón 1).
2. Pre-procesamiento. A cada extracto de texto y a sus frases clave correspondientes se eliminan las palabras vacías (*stopwords*[3]) y los símbolos de puntuación, para dejar únicamente símbolos alfanuméricos y griegos, puesto que se detectó que ciertas frases clave dentro del conjunto de entrenamiento las contienen y se consideraron de importancia (Ver Tabla 6, renglón 2).
3. Extracción de características. Usando el conjunto de frases clave junto con su correspondiente extracto de texto se extraen las siguientes características que se listan a continuación y se muestra una lista en la Tabla 1 (Ver Tabla 6, renglón 3):
   a. Frase clave pre-procesada. A la frase clave se le aplica el paso 2.

[2] http://www.nltk.org/_modules/nltk/tokenize.htm

[3] https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

b. Contiene letras capitales. Si la frase clave consta de letras capitales el valor de la característica es *True*, de lo contrario es *False*.

c. Contiene símbolos griegos. Mediante expresiones regulares se determina la existencia de símbolos griegos en la frase clave, si es así es *True*, en caso contrario es *False*.

d. Número de palabras. Valor numérico que determina el número de palabras de la frase clave.

e. Posición de inicio y fin. Se localiza la frase clave dentro del extracto del texto y se obtienen las posiciones relativas a su inicio y fin.

f. Contexto de la frase clave. Se localiza la posición de la frase clave en el extracto y se extrae una palabra a la izquierda y una palabra a la derecha, en caso de que no haya palabra a la izquierda o a la derecha se coloca un valor nulo.

g. Etiquetado. La frase clave identificada y su contexto son etiquetadas usando el paquete *pos_tag*[4] de NLTK, los valores retornados por la función son colocados como características.

h. Palabras a la izquierda y a la derecha. Las palabras extraídas del contexto se separan de la palabra clave y se añaden como características con su respectivo valor PoS.

i. Suma de *IDF*. Se extrae el *IDF* de todos los términos del conjunto de datos y el valor se calcula con la Ecuación (2).

$$\sum_{i=1}^{n} IDF(w_i) \qquad (2)$$

Donde w_i son las palabras que conforma la frase clave.

4. Creación del modelo. Las características se recopilan de los datos de entrenamiento y se forma la matriz de características con el formato ARFF que utiliza el programa *WEKA* para obtener los modelos de clasificación. En este enfoque se decidió crear más modelos aparte del que produce el clasificador Naïve Bayes, entre los cuales están los Árboles de Decisión, Máquinas de Soporte Vectorial y Zero (Ver Tabla 6, renglón 4).

5. Evaluación de la clasificación. Con los modelos creados en el paso 4 aplicados a un conjunto de datos de prueba, se procede a evaluar las predicciones utilizando las métricas de *Precisión, Exhaustividad y Medida-F_1*. Estas predicciones son comparadas con el *gold*-estándar.

5. RESULTADOS Y DISCUSION

Los enfoques presentados se evaluaron clasificando un conjunto de frases clave y se calificó el rendimiento del sistema usando las métricas de *Precisión* (3), *Exhaustividad* (4) y *Medida-F_1* (5) (Tolosa & Bordignon, 2008) determinado por los aciertos obtenidos con respecto al *gold*- estándar.

Tabla 1. Lista de características consideradas para el segundo enfoque.

	Características
Palabra clave	Palabra clave limpia Contiene letras capitales Contiene símbolos griegos Número de palabras Posición de inicio y fin Etiquetado PoS
Contexto	Contexto Etiquetado PoS del contexto Palabra a la derecha y a la izquierda de la frase clave Etiquetado PoS de la palabra a la izquierda o a la derecha de la frase clave
Medidas	Suma de *IDF* de cada elemento de la frase clave

$$Precisión(S) = \frac{Cantidad\ de\ términos\ relevantes\ recuperados}{Cantidad\ de\ términos\ recuperados} \qquad (3)$$

$$Exhaustividad(S) = \frac{Cantidad\ de\ términos\ relevantes\ recuperados}{Cantidad\ de\ términos\ relevantes} \qquad (4)$$

[4] http://www.nltk.org/book/ch05.html

$$Medida\text{-}F_1(S) = \frac{2}{\frac{1}{P(S)} + \frac{1}{R(S)}} \quad (5)$$

5.1. Conjunto de Datos

El corpus fue proporcionado por los organizadores de la tarea 10 de SemEval 2017 y contiene 500 artículos en el idioma inglés del acervo libre de la página web *ScienceDirect*[5]. Incluye temas de Ciencias de la Computación, Ciencias de Materiales y Física. El corpus se divide en tres partes: 50 artículos son para desarrollar los sistemas, 350 para las pruebas y los últimos 100 como conjunto de evaluación. Los artículos del conjunto de desarrollo y de pruebas contienen 3 tipos de archivos: un archivo XML con todo el contenido del artículo de la página *ScienceDirect*, un archivo TXT con un extracto del mismo que funciona de espacio de trabajo y un archivo ANN con anotaciones acerca de las frases claves identificadas y su correspondiente categoría. Para el conjunto de evaluación se tienen los mismos archivos, sin embargo, los ANN únicamente contienen las palabras clave sin categoría. Las categorías son: material, tarea y proceso.

5.2. Resultados

En ambos enfoques planteados, se usaron cuatro clasificadores Naïve Bayes, Árboles de Decisión, Zero R y Máquinas de Soporte Vectorial. Los modelos fueron creados con la herramienta WEKA**.** Para el segundo enfoque, se usó el complemento *StringToWordVector*[6] para mayor compatibilidad con los modelos. Los resultados de ambos enfoques son presentados en la Tabla 2 para el primer enfoque y en la Tabla 3 para el segundo.

El tiempo de entrenamiento para los clasificadores en ambos enfoques fue similar para los Árboles de Decisión, Zero R y Naïve Bayes, mientras que para el clasificador basado en Máquinas de Soporte Vectorial el aumento de características del segundo enfoque comparado con el primer enfoque afectó de manera directa el tiempo de entrenamiento de este clasificador, sin embargo, este aumento de tiempo también dio un aumento en su capacidad predictiva.

En la Tabla 2 del primer enfoque se muestra que los clasificadores basados en Árboles de Decisión, Zero R y Máquinas de Soporte Vectorial obtuvieron los mismos puntajes en las tres categorías, dando el mismo rendimiento, mientras que Naïve Bayes logró detectar las tres categorías con resultados competitivos por estar encima del resultado aleatorio (*Random).*

En la Tabla 3, que corresponde al segundo enfoque, se aprecia que el clasificador Naïve Bayes obtuvo mejores resultados en la evaluación, por el contrario, Zero R obtuvo un pobre rendimiento ya que clasificó todo en una sola categoría dejando resultados nulos en las otras dos. De manera individual, para la categoría de material el clasificador de máquina de soporte vectorial obtuvo 0.65 de *Medida-F_1*, en el caso de la categoría de tarea (*task*) el clasificador Naïve Bayes logró mejores resultados.

Las tablas muestran que Naïve Bayes obtiene mejores resultados en ambos enfoques. En el caso del primer enfoque, los restantes clasificadores obtienen resultados bajos, incluso algunos sólo llegan a clasificar en una sola categoría, pero al clasificador Naïve Bayes solo le basta el contexto para clasificar de forma adecuada y con resultados altos.

En el segundo enfoque, Naïve Bayes subió en rendimiento, esto corresponde al aumento de características usadas aparte del contexto, sin embargo, este aumento de características influye de manera importante al resto de los clasificadores que mejoraron notablemente, llegando incluso al doble de clasificaciones correctas, la excepción fue Zero R que mejoro muy poco, pero sigue clasificando sólo en una categoría.

[5] http://www.sciencedirect.com

[6] http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/StringToWordVector.html

Tabla 2. Resultados del primer enfoque por tipo de clasificador.

	Naïve Bayes			Árboles de Decisión			Zero R			Máquinas de Soporte Vectorial		
	P	***E***	***F_1***	***P***	***E***	***F_1***	***P***	***E***	***F_1***	***P***	***E***	***F_1***
Material	0.688	0.400	0.506	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Tarea	0.379	0.340	0.359	0.444	1.000	0.615	0.444	1.000	0.615	0.444	1.000	0.615
Proceso	0.533	0.755	0.624	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Promedio	0.587	0.559	0.547	0.197	0.444	0.273	0.197	0.444	0.273	0.197	0.444	0.273

Tabla 3. Resultados del segundo enfoque por tipo de clasificador.

	Naïve Bayes			Árboles de Decisión			Zero R			Máquinas de Soporte Vectorial		
	P	***E***	***F_1***	***P***	***E***	***F_1***	***P***	***E***	***F_1***	***P***	***E***	***F_1***
Material	0.699	0.625	0.660	0.673	0.549	0.605	0.477	1.000	0.646	0.581	0.738	0.650
Tarea	0.199	0.368	0.258	0.550	0.148	0.233	0.000	0.000	0.000	0.657	0.056	0.103
Proceso	0.589	0.559	0.574	0.536	0.721	0.615	0.000	0.000	0.000	0.507	0.530	0.518
Promedio	0.608	0.574	0.588	0.603	0.590	0.577	0.228	0.477	0.308	0.561	0.549	0.513

Por último, en la Tabla 4 se hace una comparativa de rendimiento entre los enfoques propuestos y los equipos que participaron en la tarea de acuerdo a lo reportado por (Augenstein et al., 2017) en el documento de descripción de SemEval 2017 Tarea 10, "Extracting Keyphrases and Relations from Scientific Publications". De acuerdo a los resultados obtenidos con la *Medida-F_1* se observa que el segundo enfoque utilizando el clasificador Naïve Bayes logró un resultado muy competitivo cercano a los dos primeros equipos que ganaron la competencia, 0.58 de *Medida-F_1*, superando por lo menos al resultado aleatorio (*Random*) y a otros equipos además del primer enfoque (con 0.54 de *Medida-F_1*).

Tabla 4. Comparativa de los enfoques propuestos con otros sistemas.

Equipo	***F_1***
MayoNLP	0.67
UKP/EELECTION	0.66
Segundo Enfoque (Naïve Bayes)	**0.58**
Primer Enfoque (Naïve Bayes)	**0.54**
LABDA	0.51
BUAP	0.45
upper bound	0.85
Random	0.23

Los sistemas ganadores de la competencia hacen uso de características léxicas y de contexto de la frase clave con la combinación de clasificadores como Máquinas de Soporte Vectorial para el caso del sistema MayoNLP, (Liu et. al., 2017) y Redes Neuronales Convulsionales para el sistema UKP/EELECTION (Eger et. al., 2017).

En el caso de los enfoques propuestos, los mejores resultados se logran con el clasificador Naïve Bayes. Considerando sólo los enfoques propuestos, se observa que el incremento en la cantidad de características permite al segundo enfoque identificar más clases; como es el caso de los resultados de SVM, que indica que el contexto y las características de peso como atributos de las frases clave incrementan la precisión de los resultados.

El rendimiento reportado en las tablas anteriores muestran que ambos enfoques tienen resultados satisfactorios en la clasificación de frases clave para la detección de recursos usados en textos científicos. Además, muestran la posibilidad de ser aplicados a otras tareas como pueden ser la clasificación de frases clave en tópicos, trasladarlos a la clasificación de documentos o la identificación de frases clave por medio de la clasificación, tal como se muestra en (Augenstein et al., 2017) en la subtarea A

5. CONCLUSIONES

La extracción de frases claves en textos no estructurados es una de las tareas del Procesamiento del Lenguaje Natural. Las frases claves permiten identificar la idea central de un texto. Por lo tanto, en este trabajo de investigación se presentan dos enfoques para la clasificación de frases clave en textos científicos usando diferentes clasificadores automáticos, el objetivo de la clasificación es reconocer recursos, materiales o procedimientos que el autor realizó y usó durante el desarrollo de su trabajo de investigación y que están de forma implícita en las frases clave.

Se propusieron dos enfoques, el primer enfoque únicamente usa el contexto de la frase clave como característica para crear el modelo de clasificación logrando el 0.54 de *Medida-F_1*. En el caso del segundo enfoque con el clasificador Naïve Bayes se obtiene una leve mejora de 0.58 de *Medida-F_1* con respecto al primer enfoque y esto corresponde al aumento de características consideradas en la clasificación. Ambos enfoques utilizan los clasificadores: Máquinas de Soporte Vectorial, Árboles de decisión y Zero R. El último obtiene un bajo rendimiento, mientras que el clasificador Naïve Bayes obtiene mejores resultados en la clasificación de las tres clases o categorías.

Los resultados conseguidos muestran que los enfoques presentados, clasifican de forma satisfactoria las frases clave permitiendo diferenciar los recursos usados en la investigación y que están implícitamente mencionados en ellas. Asimismo, el buen rendimiento abre la posibilidad de aplicar el mismo modelo para otras tareas aparte de la clasificación de frases clave entre las que pueden mencionar están la detección o clasificación de frases clave en tópicos.
Por último, se realiza una comparación con los resultados obtenidos por otros equipos que participaron en la Tarea 10 de SemEval-2017, que realizan la misma subtarea. En este caso se alcanza el tercer lugar con los resultados del segundo enfoque, superando al resultado aleatorio, lo que confirma que el enfoque es competitivo.

Como trabajo a futuro se plantea añadir más características en las que se considere el contexto a nivel semántico y el uso de otros clasificadores como redes neuronales, además de trasladar el enfoque de clasificación a otras tareas aparte de la identificación de recursos en textos científicos.

6. AGRADECIMIENTOS

Esta investigación es apoyada por el Fondo Sectorial de Investigación para la Educación, proyecto CONACyT CB-257357. Por el proyecto ID 00478 VIEP-BUAP y por el proyecto PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854.

7. REFERENCIAS

Eger, S., Do Dinh, E.-L., Kuznetsov, I., Kiaeeha, M., & Gurevych, I. (2017). EELECTION at SemEval-2017 Task 10: Ensemble of nEural Learners for kEyphrase ClassificaTION. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 942-946.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 546–555.

Devasena C, L. (2014). Effectiveness Analysis of ZeroR, RIDOR and PART Classifiers for Credit Risk Appraisal. *International Journal of Advances in Computer Science and Technology (IJACST), Vol.3 , No.11*, 06-11.

Frank, E., Hall, M., & Witten, I. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques".* Morgan Kaufmann, Fourth Edition.

Garrido Satué, M. (2013). Teoría de clasificadores. En M. Garrido Satué, *Reconocimiento de señales de tràfico para un sistema de ayuda a la conducción* (págs. 13-24). Sevilla: Departamento de Ingeniería de Sistemas y Automática.

H. Witten, I., W. Paynter, G., Frank, E., Gutwin, C., & G. Nevill-Manning, C. (1999). KEA: Practical automatic keyphrase extraction. *Proceedings of the fourth ACM conference on Digital libraries*, 254-255.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Liu, S., Shen, F., Chaudhary, V., & Liu, H. (2017). MayoNLP at SemEval 2017 Task 10: Word Embedding Distance Pattern for Keyphrase Classification in Scientific Publications. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 956–960.

Ramirez García, M., Carrillo Ruiz , M., & Sánchez López, A. (2015). Combinación de clasificadores para el análisis de sentimientos. *Research in Computing Science 94 (2015)*, 193–206.

Segura-Bedmar, I., Colón-Ruiz, C., & Martínez, P. (2017). LABDA at SemEval-2017 Task 10: Extracting Keyphrases from Scientific Publications by combining the BANNER tool and the UMLS Semantic Network. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 947–950.

Tolosa, G. H., & Bordignon, F. R. (2008). *Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos.* Buenos Aires: Universidad Nacional de Luján.

Vilares, J. (2008). *Introducción a la recuperación de información.* La Coruña: Facultad de Informática de La Coruña.

Wang, L., & Li, S. (2017). PKU_ICL at SemEval-2017 Task 10: Keyphrase Extraction with Model Ensemble and External Knowledge. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 934-937.

Tabla 1. Pasos del primer enfoque.

1) Lectura de archivos

Archivo del texto	Archivo de anotaciones
Three-dimensional digital subtraction angiographic (3D-DSA) images from diagnostic cerebral angiography were obtained at least one day prior to embolization in all patients. The raw data of 3D-DSA in a DICOM file were used for creating a 3D model of the target vessel segment. …	… - Three-dimensional digital subtraction angiographic - Process - 3D-DSA - Process - cerebral angiography - Process …

2) Pre-Procesamiento

Extracto:
Three-dimensional digital subtraction angiographic (3D-DSA) images from diagnostic cerebral angiography were obtained at least one day prior to embolization in all patients. The raw data of 3D-DSA in a DICOM file were used for creating a 3D model of the target vessel segment …

Sentencia:
1) Three-dimensional digital subtraction angiographic (3D-DSA) images from diagnostic cerebral angiography were obtained at least one day prior to embolization in all patients.
2) The raw data of 3D-DSA in a DICOM file were used for creating a 3D model of the target vessel segment.

3) Mapeo de frases clave

Posiciones de las sentencias:

Sentencia	Inicio	Fin
Three-dimensional digital subtraction angiographic (3D-DSA) images from diagnostic cerebral angiography were obtained at least one day prior to embolization in all patients.	1	174
The raw data of 3D-DSA in a DICOM file were used for creating a 3D model of the target vessel segment.	175	277

Posiciones de las frases clave:

Frases clave	Inicio	Fin
Three-dimensional digital subtraction angiographic	0	50
3D-DSA	52	58
cerebral angiography	83	103

Conjunto:
[Three-dimensional digital subtraction angiographic, 3D-DSA, cerebral angiography, [Three-dimensional digital subtraction angiographic (3D-DSA) images from diagnostic cerebral angiography were obtained at least one day prior to embolization in all patients]]

4) Extracción del contexto

Frase clave	Texto	Extracción
Three-dimensional digital subtraction angiographic	Three-dimensional digital subtraction angiographic (3D-DSA) images from …	Three-dimensional digital subtraction angiographic (3D-DSA) images
3D-DSA	… subtraction angiographic (3D-DSA) images from diagnostic…	angiographic (3D-DSA) images

5) Creación del modelo

Modelo:
…
Three-dimensional digital subtraction angiographic (3D-DSA) images, Process angiographic (3D-DSA) images, Process
…
Frase clave, Palabra-1, Palabra-2

Tabla 2. Pasos segundo enfoque.

1) Lectura de archivos

Archivo del texto	Archivo de anotaciones
Complex Langevin (CL) dynamics [1,2] provides an approach to circumvent the sign problem in numerical simulations of lattice field theories with a complex Boltzmann weight, since it does not rely on importance sampling…	… -Complex Langevin -CL -Sign problem in the thermodynamic limit - complexified configuration space …

2) Pre-Procesamiento

Texto:
Complex Langevin (CL) dynamics [1,2] provides an approach to circumvent the sign problem in numerical simulations of lattice field theories with a complex Boltzmann weight, since it does not rely on importance sampling…

Texto procesado:
Complex Langevin CL dynamics 12 provides approach circumvent sign problem numerical simulations lattice field theories complex Boltzmann weight since rely importance sampling…

Frases clave

Original	Procesada
Complex Langevin	Complex Langevin
CL	CL
Sign problem in the thermodynamic limit	Sign problem thermodynamic limit

3) Extracción de características

Texto:
Complex Langevin CL dynamics 12 provides approach circumvent sign problem numerical simulations lattice field theories complex Boltzmann weight since rely importance sampling…

Frase clave:
Complex Langevin

Caracteristica	Valor
Palabra clave limpia	Complex Langevin
Contiene letras Capitales	False
Contiene símbolos griegos	False
Numero de palabras	2
Posición de Inicio y fin	0 , 16
Etiquetado *PoS*	NNP NNP
Contexto	NULL Complex Langevin CL
Etiquetado *PoS* del contexto	NULL NNP NNP NNP
Palabra a la izquierda y derecha	NULL, CL
Etiquetado *PoS*	NULL, NNP
Medida *IDF*	3.15

4) Creación del modelo

Modelo:
…
Complex Langevin, False, False, 2, 0,16, NNP NNP, NULL Complex Langevin CL, NULL NNP NNP NNP NULL, CL, NULL, NNP, 3.15, Process
…

ALGORITMOS PARA DETECTAR LA CALIDAD DE SERVICIO EN LOS DOMINIOS DE RESTAURANTES Y LAPTOPS

K. L. Vázquez-Flores, M. Tovar-Vidal, H. Castillo-Zacatelco & M. Rossainz-López

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Ciudad Universitaria, 72570, Puebla, Mexico. krnlet@gmail.com, mtovar@cs.buap.mx, hilda@cs.buap.mx, rossainz@cs.buap.mx

RESUMEN. En este trabajo, se lleva a cabo un estudio de minería de opiniones o análisis de sentimientos, que es un área del Procesamiento de Lenguaje Natural y sub-disciplina entre la recuperación de información y la lingüística computacional. Su objetivo es detectar los sentimientos expresados en un texto u opinión. En este trabajo, se realiza la detección de polaridad y categoría en un conjunto de opiniones de usuarios hacia el dominio de restaurantes en idioma Español e idioma Inglés, así como opiniones de laptops en idioma Inglés. Dos algoritmos son presentados como propuestas de solución, en el primero, es utilizado el modelo vectorial y en el segundo se utiliza un algoritmo de clasificación automática. La investigación se realizó con el principal propósito de resolver una tarea propuesta en SemEval 2016 y se utilizaron los datos propuestos por ellos.

PALABRAS CLAVE: Análisis de sentimientos, clasificación automática, modelo vectorial.

ABSTRACT. In this work, a study is carried out on the opinion mining or sentiment analysis, an area of the Natural Language Processing and sub-discipline between information retrieval and computational linguistics that detects the opinions and sentiments expressed in a text. In this work, we perform the detection of polarity and category in a set of opinions of users towards the domains of restaurants in the Spanish and English language, as well as opinions on laptops in the English language. Two algorithms are presented as solution proposals, in the first the vectorial model is used and in the second an automatic classification algorithm is used. The research is done with the aim of solving a task proposed in SemEval 2016, the data proposed by them were used.

KEY WORDS: Sentiment Analysis, Automatic Classification, Vectorial Model.

1. INTRODUCCIÓN

El análisis de sentimientos, extracción de opiniones o minería de opiniones es un área del Procesamiento del Lenguaje Natural que se enfoca en detectar la información subjetiva de un texto y en clasificarla (Jiménez. et al., 2014). El análisis de sentimientos se define como el estudio computacional de opiniones, sentimientos y emociones expresadas en textos. Su propósito es determinar la actitud del escritor ante determinados productos o situaciones; identificar el aspecto que genera la opinión, el tipo de sentimiento y su orientación semántica (positivo, negativo, neutro o conflicto). En este trabajo se presenta el análisis de sentimientos dirigido hacia la determinación de la calidad de los servicios recibidos, es decir, sobre opiniones o críticas que los usuarios aportan por algún servicio recibido, dichas opiniones expresadas a través de internet, en este caso hacia restaurantes y laptops. Las opiniones analizadas están escritas en el idioma Español e Inglés para el dominio de restaurantes y en el idioma Inglés para el dominio de laptops. El propósito principal es el etiquetado de las opiniones de acuerdo a una categoría y polaridad. Como una solución, se propone el uso de aprendizaje supervisado con diferentes algoritmos de aprendizaje automático.

Como parte de SemEval 2016[7] , en este artículo, se propone una solución para la tarea número 5 (Análisis de Sentimientos Basada en Aspectos), subtarea 2: Nivel-Texto ABSA[8] . El objetivo es que dado un conjunto de reseñas de clientes sobre una entidad objetivo (por ejemplo: una computadora portátil o un restaurante), el objetivo es identificar un

[7] Http://alt.qcri.org/semeval2016/
[8] http://alt.qcri.org/semeval2016/task5/

conjunto de tuplas {aspecto, polaridad} que resumen las opiniones expresadas en cada reseña. La polaridad puede ser positiva, negativa, neutral o conflicto. Dado el conjunto de opiniones por entidad objetivo, nuestro propósito es determinar el grado de satisfacción que tienen los clientes que emitieron su opinión ante el servicio recibido y clasificar automáticamente las nuevas opiniones recibidas por otros clientes con un grado de confianza alto, que le permita a la entidad (por ejemplo, el restaurante) determinar el grado de satisfacción de sus clientes que emiten la reseña u opinión.

El resto del documento está estructurado como sigue. En la sección 2 se presentan los trabajos relacionados, en la sección 3 se muestra la propuesta de solución. En la sección 4 se presentan los resultados obtenidos y en la subsección 4.1 se describen los conjuntos de datos que fueron utilizados para los experimentos de los algoritmos propuestos y finalmente en la sección 5 se presentan las conclusiones.

2. TRABAJOS RELACIONADOS

En esta sección se describen brevemente algunos de los documentos revisados para la minería de opiniones, en dominios de datos diferentes y métodos. La investigación presentada en (Rothfels y J. Tibshirani, 2010), proponen el uso de la clasificación no supervisada para opiniones en películas de idioma inglés. Utilizan una métrica llamada orientación semántica de una frase[9]. Los términos PMI (Pointwise Mutual Información) de la ecuación fueron calculados consultando una base de datos, donde las frases consideradas fueron bigramas consistiendo de un adjetivo y un sustantivo, lo que los hace que probablemente tengan un contenido semántico. Un documento es clasificado como positivo si la suma de las orientaciones semánticas de sus términos es mayor que cero, y negativo en otro caso. También eligieron mantener el algoritmo de clasificación iterativa de Zagibalov y Carrol, (Zagibalov y J. Carrol, 2008). Alcanzaron el 65.5% de exactitud con su propio sistema.

En el caso de (Sangeetha, R.F., 2016), utilizan la técnica de aprendizaje automático con datos de twitter, donde determinaron la opinión sobre un producto.

El pre-procesamiento de datos es la parte más importante en el proceso e incluye el reemplazado de emoticones,[10] URL y hash-tags. En la extracción de datos, la técnica usada fue basada a nivel de oración y consiste en la tokenización, etiquetado POS (Part Of Spech) y Stopwords, en la selección de características usaron SentiWordNet para identificar el peso entre -1 y +1, incluyendo características de negación, ese puntaje indica si el tweet tiene polaridad positiva o negativa. En la clasificación utilizaron Máquina de Soporte Vectorial, el sistema calcula el porcentaje de positivo y negativo de cada tweet de un producto en particular.

En (Go, A., et al., 2009.), se introduce un enfoque para la clasificación de sentimientos positivos y negativos, el resultado de un algoritmo de aprendizaje automático mediante supervisión remota, donde el conjunto de datos consiste en mensajes de twitter con emoticones. En el enfoque que proponen prueban con diferentes clasificadores: basado en palabras clave, Naïve Bayes, Entropía Máxima y Máquina de Soporte Vectorial (SVM) y como características utilizaron: unigramas, bigramas, unigramas y bigramas, y partes de la oración. El resultado obtenido mostró una exactitud más grande que 80% con el uso de Naïve Bayes, Entropía Máxima y SVM, y con unigramas como característica.

En (Park, A. y P. Paroubek, 2010), se muestra el uso de un corpus con datos de Twitter, construyendo un clasificador de sentimientos capaz de determinar la polaridad positiva, negativa y neutral de datos escritos en idioma inglés. Para la clasificación utilizaron la presencia de unigramas, bigramas y trigramas como características binarias.

La investigación llevada a cabo en (Pimpalkar, A.P., 2013), menciona el desarrollo de un

[9] Se define como la diferencia de la asociación de la oración con las palabras de la semilla positiva y negativa (como medida para la información mutua punto a punto): SO(p) = PMI(p, excellent) - PMI(p, poor)

[10] Los emoticones son remplazados por palabras que presentan lo que el emoticon expresa

Sistema de análisis de sentimientos para comentarios de clientes en internet. Utilizaron SentiWordNet[11], también utilizaron el enfoque basado en reglas de medidas difusas. Encontraron algunos beneficios en esta investigación; tales como retroalimentación de usuarios en tiempo real, inteligencia de Mercado accionable basada en retroalimentación directa de usuario y retroalimentación, y tiempo de reacción mejorado para el servicio y calidad para el Mercado.

En otra solución presentada en (Hercig, T. et al., 2016), el autor presenta un Sistema con una aproximación basada sobre el clasificador de máxima entropía. El algoritmo de base atraviesa las tuplas al nivel de opiniones predichas de la misma categoría y cuenta las respectivas etiquetas de polaridad (positiva, negativa o neutral). La etiqueta de polaridad con la más alta frecuencia se asigna a la categoría a nivel texto, si no son tuplas de nivel de opinión de la misma categoría de la etiqueta polaridad, se determina sobre la base de todas las tuplas independientemente de la categoría. Para sus experimentos utilizaron Brainy (Máquina de aprendizaje automático), tanto como árboles de análisis de sintaxis, lematización y etiquetas POS de StandfordCoreNLP v3.6. Los resultados presentados muestran una exactitud de 86.9% en el dominio de laptops y 89.6% en el dominio de restaurantes en inglés. En el caso de (Mulay, S.A., et al., 2016), utilizaron tweets como reseñas escritas de usuario para una película particular, utilizaron hashtags sociales para la transmisión de datos de twitter. Utilizaron el Sistema de archivos Hadoop HDFS junto con la técnica MapReduce para manejar grandes cantidades de datos no estructurados, utilizaron Procesamiento de Lenguaje Natural (PLN) para el represamiento de tweets. Seleccionaron Naïve Bayes para clasificar tweets en positivo y negativo. Después del análisis de sentimientos se asignan los pesos para varios factores como el número de vistas en cada película, número de pantallas en donde la película es reproducida, número de tweets, etc. el peso de todos los factores son combinados para predecir el éxito general de la película en taquilla.

En (Mohammed y Moreno, 2017) describen SiTAKA un sistema para el análisis de sentimientos en twitter en idioma Inglés y Árabe, el sistema propone la representación de tweets utilizando un conjunto de características nuevo, el cual incluye una bolsa de palabras negadas e información proporcionada por algunos léxicos. La polaridad de los tweets la determinan por el clasificador Máquina de Soporte Vectorial. Tal sistema obtuvo el octavo lugar al evaluar tweets en idioma Inglés y el segundo lugar al evaluar tweets en idioma Árabe. En (Siordia, O.S. et al., 2015) presentaron los resultados de la tarea 1 de TASS 2015: Clasificación global de cinco niveles de polaridad para un conjunto de tweets en español. Utilizaron Máquina de Soporte Vectorial para clasificar incorporando los modelos de datos *LSI* y *TF-IDF*, el resultado fue una mejor clasificación, obteniendo 0.404 de F_1-*score*.

Hay muchos trabajos reportados en la literatura asociados con la minería de opiniones, en este documento se evitará ser exhaustivo en mencionar todos esos trabajos y se procederá a describir los enfoques empleados en nuestros experimentos.

3. ENFOQUES PROPUESTOS

En este documento se proponen dos enfoques para determinar la polaridad de las opiniones de usuarios. Las opiniones analizadas fueron proporcionadas por los organizadores de la tarea 5 de SemEval-2016 y están almacenadas en dos conjuntos de datos: los datos de entrenamiento se utilizan para entrenar al clasificador y el conjunto de datos de prueba para predecir la categoría y/o polaridad de la opinión (Pontiki, M., et al., 2016). Primero, se aplica un pre-procesamiento al conjunto de datos de entrenamiento para obtener los vectores de representación para cada tupla {aspecto, polaridad}. Aplicamos el mismo proceso para el conjunto de datos de prueba, para aplicar los enfoques entre esos dos conjuntos de datos y para determinar la polaridad de cada elemento del conjunto de datos de prueba. El rendimiento de los enfoques se obtiene al comparar los resultados con los reportados en el *gold standar*.

[11] Herramienta que asigna pesos a los sentimientos de cada palabra encontrada en los comentarios.

A continuación, se describe el propósito de cada enfoque para la subtarea 2 de la tarea 5 de SemEval 2016.

3.1 Enfoque basado en el modelo espacio vectorial

En el primer enfoque, se propone un algoritmo que utiliza un modelo espacio vectorial con 1-*gram*, 2-*grams*, 3-*grams*. Este enfoque realiza los siguientes pasos:

1. Pre-procesamiento:

- Extracción de opiniones del documento XML.
- Eliminación de palabras vacías, símbolos de puntuación, caracteres aislados y clasificación de términos.
- Tokenización de opiniones por palabras y su ordenamiento para obtener el vocabulario del conjunto de datos.
- Reducción de vocabulario mediante la aplicación de *stemming*, el cual es un procedimiento computacional, que reduce todas las palabras a la raíz de cada una de ellas, usualmente eliminando cada palabra de sus sufijos derivacionales (Lovins, 1968). Se utiliza *Snowball* como algoritmo de *stemming* por su implementación en el lenguaje de programación.

2. Extracción de características:

- Frecuencia de término (*tf*): El número de veces que aparece un término en un documeto *d* o conjunto de datos, se representa con la ecuación 1.

$$tf(t,d) = f(t,d) \quad (1)$$

- Frecuencia Inversa del Documento (*idf*): Se calcula el número de documentos (|D|) en los que aparece un término *t*. Esta medida permite determinar la discriminación de un término. Los términos raros son más discriminados que los términos comunes (ver Ecuación 2).

$$idf(t,D) = log \frac{|D|}{d \in D : t \in d} \quad (2)$$

- Frecuencia de término y Frecuencia Inversa del Documento (*tf-idf*): Los valores *tf* e *idf* son combinados y producen la Ecuación 3, (Lavin Villa, M.E., 2010).

$$tf - idf = (t,d,D) = tf(t,d)\ x\ idf(t,D) \quad (3)$$

- *N-grams*: Para cada matriz de pesado, ambos, *tf* y *tf-idf* son calculados por diferentes secuencias de palabras llamadas *n-grams*. Esas cadenas de texto son el resultado de agrupar una secuencia de palabras de un texto dado, previo al paso de pre-procesamiento. En el enfoque, consideramos n=1; 2; 3; es decir, unigramas, bigramas y trigramas de palabras.

3. Vector representativo para cada categoría-polaridad de los datos de entrenamiento:

- Basado en las matrices de pesado generadas con el conjunto de entrenamiento, se procede a crear un vector representativo para cada categoría (entidad-atributo) se considera una polaridad asociada, por ejemplo, para la entidad ambiente, el atributo general y este corresponde a un tipo de polaridad (positiva, negativa, neutral y conflicto), se obtienen cuatro diferentes vectores representativos:

{*AMBIENCE#GENERAL*; *positivo*}
{*AMBIENCE#GENERAL*; *negativo*}
{*AMBIENCE#GENERAL*; *neutral*}
{*AMBIENCE#GENERAL*; *conflicto*}

4. Detección por medio de la medida de similitud coseno:

- Aplicamos la medida de similitud coseno, ver Ecuación (4), para determinar la similitud entre dos vectores de pesado (o ponderación), uno del conjunto de datos de entrenamiento (*dj*) y el otro del conjunto de prueba (*q*). La opinión debe ser asignada con la polaridad de acuerdo al valor obtenido por la medida de coseno (la más alta).

$$sim(dj,q) = \frac{d_j,q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}} \quad (4)$$

5. Evaluación de polaridad: Para determinar el desempeño del enfoque empleamos

exactitud como medida de evaluación. Esta medida evalúa el promedio de las predicciones correctas, es decir de las categorías y polaridades asignadas correctamente en las opiniones, la fórmula para obtener exactitud se puede ver en la ecuación 5.

$$Exactitud = \frac{CantidadDeCasosClasificadosCorrectamente}{TotalDeCasos} \quad (5)$$

3.2 Enfoque basado en un clasificador automático

En esta aproximación usamos un clasificador automático, primero se detecta la categoría y enseguida la polaridad. A continuación se explica el enfoque.

1. Pre-procesamiento:
   - Extracción de opiniones del documento XML.
   - Eliminación de palabras vacías, símbolos de puntuación, carácteres aislados y clasificación de términos.
   - Tokenización de opiniones por palabras y ordenamiento de las palabras resultantes para obtener el vocabulario del conjunto de datos.
   - Reducción de vocabulario mediante la aplicación de *stemming* a cada palabra. Este paso se realiza de la misma manera y con el mismo algoritmo al del enfoque anterior (3.1).

2. Extracción de características: Calculamos la característica *tf-idf* (ver Ecuación 3).

3. Fase de entrenamiento:
   - En esta fase, utilizamos el clasificador Naïve Bayes para entrenarlo dos veces con las características extraídas. La primera vez, las categorías son obtenidas. La segunda vez la polaridad es obtenida.

4. Fase de prueba: El conjunto de prueba es clasificado con el modelo obtenido en el paso anterior. Después de esta fase, cada opinión de los datos de prueba es etiquetada con categoría y polaridad.

5. Evaluación: Para determinar el desempeño del enfoque, aplicamos exactitud como medida de evaluación. Se puede ver en la ecuación 5, la fórmula para obtenerla.

En la siguiente sección, se reportan los resultados obtenidos con los dos enfoques.

4. RESULTADOS EXPERIMENTALES

En esta sección describimos los resultados obtenidos con los enfoques propuestos.

4.1 Conjuntos de datos

En los experimentos realizados, se utilizan los conjuntos de datos de entrenamiento y prueba, provenientes de SemEval 2016, tarea 5, subtarea 2[12]. El conjunto de prueba incluye las evaluaciones del *gold standard*, para ser posible medir la calidad del enfoque propuesto. Las opiniones de los usuarios son dadas para dos dominios: Restaurantes (escritos en Inglés y Español), y Laptops (escritas únicamente en Inglés). En la Tabla 1 se muestra el número de textos (opiniones) proporcionadas por SemEval 2016.

Tabla 1. Número de opiniones dadas para cada dominio en la subtarea 2 de la tarea 5.

Dominio	Entrenamiento	Prueba	Oro
Restaurante (Español)	627	268	268
Restaurante (Inglés)	335	90	90
Laptops (Inglés)	395	80	80

Es importante mencionar que las opiniones en ambos conjuntos, entrenamiento y prueba, pueden ser asignadas con más de una categoría y polaridad (tuplas).

En la tabla 2, se muestra el número de tuplas que las opiniones pueden tener asociado para cada dominio, es decir, las diferentes categorías y polaridades por opinión.

[12] http://alt.qcri.org/semeval2016/task5/index.php?id=data-and- tools

Tabla 2. Número de tuplas por dominio.

Dominio	Entrenamiento	Prueba	Oro
Restaurante (Español)	2,121	881	881
Restaurante (Inglés)	1,435	404	404
Laptops (Inglés)	2,082	545	545

4.2 Resultados usando el enfoque basado en el modelo de espacio vectorial

Con el enfoque de similitud coseno, las tuplas son primero evaluadas por categoría y después por polaridad. En la tabla 3 se presenta el dominio, el total de tuplas por dominio, la cantidad de muestras clasificadas empleando la representación de texto *tf* con unigramas (1-*grams*), bigramas (2-*grams*) y trigramas (3-*grams*), y lo mismo usando *tf-idf*. En la tabla 4, se reportan los mismos resultados expresados con el porcentaje obtenido de la medida exactitud. Los resultados obtenidos permiten determinar que *tf* reporta exactitud mayor que 50% para el dominio de Laptops, 66% para el dominio de Restaurantes en Español, mientras que *tf-idf* obtiene resultados aceptables cuando se usa unigramas de palabras.

Tabla 3. Resultados de opiniones clasificadas correctamente de acuerdo a polaridad por dominio.

Dominio	Prueba	*tf*			*tf-idf*		
		1-*grams*	2-*grams*	3-*grams*	1-*grams*	2-*grams*	3-*grams*
Restaurante (Español)	881	589	506	459	535	570	466
Restaurante (Inglés)	404	288	248	249	268	248	248
Laptops (Inglés)	545	290	276	295	324	302	296

Tabla 4. Resultados de exactitud por dominio.

Dominio	*tf*			*tf-idf*		
	1-*gram*	2-*gram*	3-*gram*	1-*gram*	2-*gram*	3-*gram*
Restaurante(Español)	66.85	57.43	52.09	60.72	64.69	52.89
Restaurante (Inglés)	71.28	61.38	61.63	66.33	61.38	61.38
Laptops (Inglés)	53.21	50.64	54.12	59.44	55.49	54.31

4.3 Resultados usando el enfoque basado en clasificador automático

En la tabla 5, se presenta el dominio, el total de tuplas por dominio, la cantidad de muestras clasificadas empleando la representación *tf-idf* de texto. En la tabla 6, los mismos resultados son reportados con el promedio de exactitud para cada dominio. Los resultados obtenidos determinan que la clasificación de categoría y polaridad reportan exactitud mayor al 56% para el dominio de Laptops, mientras que categoría + polaridad (C+P) obtienen resultados aceptables para el dominio de Restaurantes

Tabla 5. Resultados de opiniones clasificadas correctamente de categoría, polaridad y categoría y polaridad por dominio.

Dominio	Categoría	Polaridad	Categoría + Polaridad (C + P)
Restaurante (Español)	686	639	526
Restaurante (Inglés)	302	285	218
Laptops (Inglés)	308	302	191

Tabla 6. Resultados de exactitud para categoría, polaridad y categoría con polaridad por dominio.

Opiniones	Categoría	Polaridad	Categoría + Polaridad (C + P)
Restaurante (Español)	77.8	72.5	59.7
Restaurante (Inglés)	74.7	70.5	53.9
Laptops (Inglés)	56.5	55.4	35.0

En la tabla 7 se representa el dominio, la exactitud obtenida por el primer enfoque (Modelo vectorial) y el segundo enfoque (Naïve Bayes). De acuerdo a los resultados, el primer enfoque logra resultados más altos al 50% en el dominio de restaurantes que en el segundo enfoque

Tabla 7. Comparación entre los dos enfoques que detectan categoría y polaridad

Opiniones	Exactitud	
	Coseno-TF/1-gram	Naïve Bayes
Restaurante (Español)	66.8	59.7
Restaurante (Inglés)	71.2	53.9
Laptops (Inglés)	53.2	35.0

4.4 Resultados del mejor lugar calificado en el dominio de restaurantes

Considerando los resultados del segundo enfoque y únicamente los resultados de polaridad sobre el dominio de Restaurantes, el enfoque propuesto puede predecir con un 77% de exactitud la cantidad de opiniones positivas que el restaurante recibe de sus clientes. En la tabla 8, se muestra que en los datos oro el restaurante Barceloneta[13] obtiene 69 opiniones positivas y el enfoque propuesto obtiene exitosamente 56, es decir el Sistema implementado obtiene un 81.1% de exactitud para predecir el número de opiniones positivas. En el caso del dominio de restaurantes en Inglés, Blue Ribbon Shushi[14], tiene 71 opiniones positivas en los datos de oro y el enfoque obtiene 55, es decir, un 77.4% de exactitud.

Tabla 8. Resultados de los mejores lugares calificados por polaridad.

Dominio	Nombre del restaurante	Enfoque	Oro	Exactitud
Restaurante (Español)	Barceloneta	69	56	81.1
Restaurante (Inglés)	Blue Ribbon Sushi	71	55	77.4

5. CONCLUSIONES

En la investigación de análisis de sentimientos en opiniones se llevó acabo en dos dominios diferentes: restaurantes y laptops, se trabajó con dos idiomas diferentes: Español e Inglés. Específicamente, el dominio de restaurantes fue analizado para el idioma Español e Inglés y para el dominio de laptops únicamente en Inglés. Se presentaron diferentes soluciones para la clasificación de polaridad y categoría, así como la prueba de esos métodos con la medida de exactitud usando el conjunto de datos de entrenamiento y prueba. Los clasificadores automáticos que se utilizaron en algunas soluciones se probaron en el entorno Python.

La clasificación de categoría y polaridad con la propuesta de similitud coseno obtiene los mejores resultados en el dominio de restaurantes en Inglés con las características *tf* y 1-*grams*, obteniendo exactitud de 71.28%. Igualmente, exactitud de 59.7% fue obtenida con el clasificador Naïve Bayes para la clasificación de categoría y polaridad. Es importante mencionar que cuando se clasifica únicamente categoría, el clasificador obtiene una exactitud de 77.8% en el dominio de restaurantes en Español. Los resultados obtenidos con estas propuestas nos llevan a una investigación más profunda, en la cual se considera necesario el uso de otras

[13] www.restaurantbarceloneta.com
[14] www.blueribbonrestaurants.com

características o herramientas tales como diccionarios de pesado, como SentiWordNet o alguna otra forma de clasificación automática.

6. AGRADECIMIENTOS

Esta investigación es parcialmente apoyada por el proyecto PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, por el proyecto ID 00478 VIEP-BUAP. Apoyado por el Fondo Sectorial de Investigación para la Educación, proyecto Conacyt CB-257357.

7. REFERENCIAS

Go, A., Bhayani R. and Huang L. 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1,12.

Hercig, T., Brychcín T., Svoboda L. and Konkol M. 2016. Uwb at semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, Association for Computational Linguistics: 354-361.

Jabreel, M., y Moreno, A. 2017. A. SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter Based on a Rich Set of Features.

Jiménez, S.M., E. Martínez, M.T.Martín y A. Ureña. 2014. Desafíos del análisis de sentimientos. In: SINAI – Sistemas Inteligentes de Acceso a la Información: 15-18.

Lavin Villa, M.E. 2010. Construcción automática de diccionarios semánticos usando la similitud distribucional. Master's thesis, Centro de Investigación en Computación, México.

Lovins, J. B. 1968. Development of a stemming algorithm. Mech. Translat. & Comp. Linguistics: 22-31.

Mulay, S.A., Joshi S.J., Shaha M.R., Vibhute H.V. and Panaskar M.P. 2016. Sentiment analysis and opinion mining with social networking for predicting box office collection of movie. International Journal of Emerging Research in Management & Technology 5:74-79.

Pak, A. and Paroubek P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In: LREc.10:1320-1326.

Pimpalkar, A.P. 2013. A sentimental analysis of movie reviews involving fuzzy rulebased. International Journal of Artificial Intelligence and Knowledge Discovery 3: 9-14.

Pontiki, M., Galanis D., Papageorgiou H., Androutsopoulos I., Manandhar S., AL-Smadi M., Al-Ayyoub M., Zhao Y., Qin B., De Clercq O., Hoste V., Apidianaki M., Tannier X., Loukachevitch N., Kotelnikov E., Bel N., Jiménez-Zafra S.M. y G. Eryi_git. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, Association for Computational Linguistics: 19-30

Rothfels, J. and Tibshirani J. 2010. Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items. CS224N-Final Project.

Sangeetha Suresh Harikantra, R.F. 2016. Opinion mining on twitter data. International Journal of Innovative Research in Science, Engineering and Technology 5: 205-209.

Siordia, O. S., Moctezuma, D., Graff, M., Miranda-Jimenez, S., Téllez, E. S., y Villaseñor, E. A. 2015. Sentiment Analysis for Twitter: TASS 2015. In *TASS@ SEPLN* (pp. 65-70).

Yadav, V 2016. Thecerealkiller at semeval-2016 task 4: Deep learning based system for classifying sentiment of tweets on two point scale. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California, Association for Computational Linguistics: 100-102

Zagibalov, T. and Carroll J. 2008. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd International Conference on Computational Linguistics, Association for Computational Linguistics 1: 1073-1080.

SECRETARÍA DE EDUCACIÓN PÚBLICA

SEP

Instituto Tecnológico de Cd. Victoria

División de Estudios de Posgrado e Investigación

Maestría en

CIENCIAS EN BIOLOGÍA

PADRÓN NACIONAL DE POSGRADO DE CALIDAD (SEP-CONACYT)

Especialidad:
Manejo y Conservación de Recursos Naturales (Terrestres o Acuáticos)

Becas Disponibles

Maestría en Ciencias en Biología

PERFIL

El programa está diseñado para egresados de la carrera de biología o afines como médicos veterinarios, ingenieros agrónomos, ingenieros ambientales e ingenieros forestales. Podrán participar egresados de otras carreras con la aprobación del consejo de posgrado.

REQUISITOS DE INGRESO Y DOCUMENTACIÓN

- Carta de exposición de motivos indicando porque desea cursar una maestría y porque desea ingresar a este programa, Maestría en Ciencias en Biología-ITCV.
- Copia (s) de título profesional, certificado de calificaciones, diploma (s) y constancias de otros estudios.
- Constancia de promedio mínimo de 8 (ocho) en estudios de licenciatura.
- Currículum vitae con documentos probatorios adjuntos.
- Comprender el idioma inglés y aprobar examen de inglés del programa de MCB-ITCV.
- Dos fotografías tamaño credencial.
- Aprobar examen de admisión.
- Carta compromiso indicando que terminará su programa de maestría en dos años.
- Disposición para desarrollar e integrarse en proyectos de investigación.
- Entrevista con el comité de posgrado.
- Ser estudiante de tiempo completo.

PLAN DE ESTUDIOS

El programa está diseñado para concluirse en dos años y consta de cinco materias básicas, seis optativas y presentación de tesis de grado.

Áreas disponibles actualmente para investigación y desarrollo de tesis:

Malacología, Entomología, Micología, Mastozoología, Ciencias Forestales (Biodiversidad, Sistemática, Ecología y Fisiología).

PLANTA DOCENTE

Almaguer Sierra Pedro, Dr. UANL. Agua-Suelos, Agrometeorología e Hidroponia.

Azuara Domínguez Ausencio. Dr. Colegio de Posgraduados. Manejo Integrado de Plagas.

Barrientos Lozano Ludivina, Ph.D. **Universidad de Gales, College of Cardiff**. Reino Unido. Entomología Aplicada. Ecología y Sistemática de **Orthoptera**.

Correa Sandoval Alfonso, Dr. UNAM. Malacología y Ecología Marina.

Flores Gracia Juan, Dr. UANL. Genética y Biotecnología.

García Jiménez Jesús, Dr. UANL. Micología y Parasitología Forestal.

González Gaona Othón Javier. Dr. ITESM. Toxicología.

Guevara Guerrero Gonzalo, Dr. UANL. Biotecnología y Micología.

Horta Vega Jorge V., Dr. CINVESTAV-IPN Neurociencias y Entomología.

Navar Cháidez José de Jesús. Ph.D. Dr. Manejo sustentable de recursos naturales.

Rangel Lucio José Antonio. Dr. Colegio de Posgraduados. Edafología.

Venegas Barrera Crystian Sadiel. Dr. CIBNOR. Manejo y Preservación de Recursos Naturales (Ecología).

INFORMES

INSTITUTO TECNOLÓGICO DE CD. VICTORIA
División de Estudios de Posgrado e Investigación

Blvd. Emilio Portes Gil No. 1301 Cd. Victoria, Tam. C.P. 87010 Apdo. Postal 175
Tel. (834) 153 2000 Ext. 325

http:///www.postgradositcv.com
http://www.itvictoria.edu.mx
E-mail: jhortavega@yahoo.com.mx
E-mail:almagavetec@hotmail.com

SECRETARÍA DE EDUCACIÓN PÚBLICA

SEP

Instituto Tecnológico de Cd. Victoria

División de Estudios de Posgrado e Investigación

Doctorado en

CIENCIAS EN BIOLOGÍA

PADRÓN NACIONAL DE POSGRADO DE CALIDAD (SEP-CONACYT)

Convocatoria: 2018

Recepción de solicitudes: enero-marzo de 2018

Líneas de investigación

- Biodiversidad y Ecología
- Manejo y Conservación de Recursos Naturales
- Procesos Biotecnológicos

Requisitos y antecedentes académicos de ingreso de los candidatos

- Contar con grado de Maestría (indispensable estar titulado) en un programa experimental o de investigación en el área de las Ciencias Biológicas.
- Promedio igual o superior a 8 (80 de 100) en estudios de maestría.
- Disponer de tiempo completo para cumplir con el programa doctoral.
- Aprobar el examen de conocimientos que aplica el programa o acreditar con al menos un 75% en conocimientos básicos y un 60% en habilidades de investigación en el EXANI-III del CENEVAL.
- Acreditar el examen de Inglés TOEFL, al ingresar al programa, mínimo 500 puntos. O bien acreditarlo este examen antes de egresar del programa, ya que este es un requisito para sustentar examen de grado y poder titularse.
- Presentar dos cartas académicas de recomendación expedidas por profesionistas reconocidos.
- Carta de exposición de motivos para el ingreso al doctorado, no mayor de una cuartilla, con fecha y firma.
- Visto bueno en entrevista con miembros del Claustro Doctoral.
- Presentar por escrito protocolo de investigación (3-5 cuartillas) para evaluar aptitudes y habilidades de experiencia previa, en el área de ciencias naturales.
- Carta de aceptación de uno de los miembros del Claustro Doctoral.

PLANTA DOCENTE

Almaguer Sierra Pedro, Dr. UANL. Agua- Suelos, Agrometeorología e Hidroponia.

Azuara Domínguez Ausencio. Dr. Colegio de Posgraduados. Manejo Integrado de Plagas.

Barrientos Lozano Ludivina, Ph.D. Universidad de Gales, Cardiff. Reino Unido. Entomología Aplicada. Ecología y Sistemática de Orthoptera.

Correa Sandoval Alfonso, Dr. UNAM Malacología y Ecología Marina.

Flores Gracia Juan, Dr. UANL. Genética y Biotecnología.

García Jiménez Jesús. Dr. UANL. Ciencias Forestales y Micología.

González Gaona Othón Javier. Dr. ITESM. Toxicología.

Guevara Guerrero Gonzalo, Dr. UANL. Biotecnología y Micología.

Horta Vega Jorge V., Dr. CINVESTAV-IPN Neurociencias y Entomología.

Navar Cháidez José de Jesús. Ph.D. Manejo sustentable de recursos naturales.

Rangel Lucio José Antonio. Dr. Colegio de Posgraduados. Edafología.

Venegas Barrera Crystian Sadiel. Dr. CIBNOR. Manejo y Preservación de Recursos Naturales (Ecología).

INFORMES

INSTITUTO TECNOLÓGICO DE CD. VICTORIA. División de Estudios de Posgrado e Investigación.

Blvd. Emilio Portes Gil No. 1301 Cd. Victoria, Tam. C.P. 87010 Apdo. Postal 175.
Tel. (834) 153 2000, Ext. 325

http:///www.postgradositcv.com
http://www.itvictoria.edu.mx
E-mail: jhortavega@yahoo.com.mx
E-mail: almagavetec@hotmail.com

CONVOCATORIA PARA PUBLICAR EN TecnoINTELECTO

TÍTULO CON MAYÚSCULAS, DEBIDAMENTE ACENTUADAS, EN NEGRITAS, CENTRADO, ARIAL 10, INTERLINEADO SENCILLO

Autor(es) Arial 10 puntos, itálica, centrado, interlineado sencillo; nombre (s) completo y apellidos completos, separados por un guión, sin grado académico, más de un autor separado por comas e indicador numérico para los datos siguientes: Institución(es) en 10 Arial, en itálica y centrado, interlineado sencillo, correo electrónico de los autores centrado, interlineado sencillo

RESUMEN: Deberá ser lo más general y significativo posible, de manera que en pocas palabras exprese la aportación más relevante del artículo. Letra tipo Arial de 10 puntos, interlineado sencillo y espaciado anterior de 8 puntos y posterior de 6, iniciando con la palabra **RESUMEN** en negritas. Texto con alineación ajustada en todo el artículo. Si el artículo está en español, adjuntar el resumen inglés.

PALABRAS CLAVE: Colocar las palabras (tres a cinco) más significativas en el artículo, no repetir palabras del título, fuente de 10 puntos, dejando un espacio entre el párrafo anterior.

ABSTRACT: The abstract shall be as general and substantial as possible, in such a way that provides in a few words a clear idea of the paper's contribution. Please use Arial font 10 points, single space, space above 8 points and below 6 points, begin text with the word **ABSTRACT** in bold face. All text through the paper must be aligned to fit page. If paper is in Spanish abstract shall be in English.

KEY WORDS: Please use the most (three to five) significant words, font of 10 points, leaving a space between the preceding paragraphs.

1. INTRDUCCIÓN

Los criterios para la revisión técnica son: importancia de la contribución a la divulgación científica, pertinencia de métodos empleados, correcta presentación de datos, soporte del manuscrito con literatura relevante y actualizada, discusión suficiente o necesaria. Además, figuras y tablas adecuadas. El manuscrito pasará al comité editorial, quien dictaminará si contiene el mínimo indispensable para ser publicado, lo cual se notificará vía electrónica en formato pdf.

2. CARACTERÍSTICAS

El cuerpo del artículo en dos columnas con 0.6 cm entre ellas y todos sus márgenes de 3 cm. Cada sección deberá contener un título numerado con formato de párrafo espaciado anterior de 12 y posterior de 6 puntos. La fuente de todo el manuscrito es Arial. En el cuerpo de 10 puntos, interlineado sencillo, con secciones numeradas con números arábigos.

2.1 Idioma Español o inglés.

2.2 Subsecciones

Las subsecciones en formato tipo título, negritas, interlineado sencillo y espaciado anterior y posterior de 6 puntos.

2.3. Las gráficas y tablas

Pueden ser **a color** o en ***escala de grises*** y se ajustarán de acuerdo a las características de ellas y al gusto del investigador. Deberán ser posicionadas de acuerdo a la necesidad del investigador y bajo su responsabilidad.

3. LINEAMIENTOS

Los artículos deberán ser inéditos. Cada trabajo deberá presentarse en un mínimo de 6 y un máximo de 12 páginas. De 6 páginas se considerarán artículos cortos y se publicarán a recomendación del comité editorial.

4. RESPONSABILIDADES

El investigador es responsable del contenido, la sintaxis y el envío de su artículo en Word a la coordinación editorial actual de TecnoINTELECTO: ludivinab@yahoo.com, almagavetec@hotmail.com. El Instituto Tecnológico de Cd. Victoria será responsable de la revisión y aceptación o rechazo de los manuscritos, la edición de la revista, el índice,

la impresión y distribución, apoyándose en el Comité Editorial y otras instituciones, si lo considera pertinente.

Los artículos que no se ajusten a las normas editoriales serán rechazados para su adecuación.

El máximo número de autores y/o coautores por artículo es de 5.

5. FECHAS IMPORTANTES

Recepción de artículos todo el año. **Publicación julio-agosto y diciembre-enero.**

6. LITERATURA CITADA

6.1 Referencias en texto

Sin numerar, solo citar apellido(s) según el caso y el año separado por una coma, si son más citas separar por punto y coma; dos autores se separan "y" y si son más de dos autores solo se pondrá el apellido(s) del primer autor seguido de "*et al.,,"*.

Al final, listar en orden alfabético sin numeración. Autor (es) iniciando con apellido (s) seguido por la inicial del nombre (s), si es el caso puede escribir los dos apellidos separados por un guion. Año. Título del artículo. Nombre de la Revista, Volumen y número de páginas, tipo Arial, 10 puntos, interlineado sencillo.

Artículo científico

Armenta, C. S., H. Bravo y R. Reyes. 1978. Estudios bioecológicos de *Epilachna varivestis* Mulsant, bajo condiciones de laboratorio y campo. Agrociencia, 34: 133-146.

Ávila-Valdez, J., L. Barrientos-Lozano y P. García-Salazar. 2006. Manejo Integrado de la Langosta centroamericana (*Schistocerca piceifrons piceifrons* Walker) (Orthoptera: Acrididae) en el sur de Tamaulipas. Entomología Mexicana, 5: 636-641.

Libro o Tesis

Jaffe, K., J. Lattke y E. Pérez. 1993. *El mundo de las hormigas.* Equinoccio Ediciones. Universidad Simón Bolívar, Venezuela. 196 pp. En el caso de tesis señalar después del título si es profesional o de grado.

Capítulo de libro:

Navarrete-Heredia, J. L. y A. F. Newton. 1996. Staphylinidae (Coleoptera). Pp. 369-380. *In:* J. E. Llorente-Bousquets, A. N. García-Aldrete y E. González-Soriano (Eds.). Biodiversidad, Taxonomía y Biogeografía de Artrópodos de México: Hacia una Síntesis de su Conocimiento. Instituto de Biología, UNAM, México, D. F.

Instituto Tecnológico de Cd. Victoria

División de Estudios de Posgrado e Investigación-Coordinación Editorial de TecnoINTELECTO.

Dra. Ludivina Barrientos Lozano:
ludivinab@yahoo.com,
almagavetec@hotmail.com