Universidad Autónoma Metropolitana
Unidad Azcapotzalco

Maestría en Ciencias de la Computación

Idónea Comunicación de Resultados:

Metodología para la Comparación de Algoritmos de Aprendizaje Automático

Caso de estudio: Clasificación de Eventos Académicos

Presenta

Ing. Ariadna Gutiérrez Rosales
2171800349

Directores

Dra. Maricela Claudia Bravo Contreras
Dr. José Alejandro Reyes Ortiz

Agosto 2019

Contenido

Lista de Tablas

Lista de Figuras

Resumen

En este trabajo de investigación se presenta una metodología para la comparación de algoritmos de aprendizaje automático que consta de cuatro etapas:

- Recopilación e integración del conjunto de datos de entrada.
- Procesamiento y transformación de los datos al modelo espacio vectorial.
- Selección e implementación de los algoritmos de aprendizaje automático.
- Evaluación y comparación de los resultados obtenidos.

Se comparan cuatro algoritmos de aprendizaje supervisado: Naïve Bayes, k-vecinos más cercanos, C4.5 y Máquinas de soporte vectorial.

La metodología que se propone en este trabajo de investigación considera como caso de estudio a la clasificación de eventos académicos. Se tienen cuatro categorías distintas: eventos de difusión, cursos académicos y de actualización, asesorías académicas a alumnos y eventos ambientales.

Los eventos se obtienen a partir de la lectura de sensores que mantienen el registro de los eventos en bitácoras. Estos eventos se caracterizan y se procesan para ser utilizados por los clasificadores. Se tienen tres tipos de características de un evento, que son: características nominales, textuales[1] y numéricas.

Dada la naturaleza de los algoritmos empleados en este trabajo de investigación, se requiere de una etapa de entrenamiento, que, a partir de un conjunto de 362 eventos académicos, se eligen aleatoriamente un 70% de ellos para el entrenamiento de los algoritmos y un 30% restante para las pruebas.

Se consideraron 24 experimentos durante la etapa de pruebas en las que se analizan y se comparan los resultados obtenidos por los clasificadores. Las medidas de evaluación que se utilizan para elegir al mejor algoritmo de aprendizaje automático son la precisión y cobertura y la medida F1, sin embargo, el criterio de selección es la medida F1 ya que es un promedio de la precisión y la cobertura.

De los resultados obtenidos en la experimentación, se puede observar que C4.5 es el algoritmo con mejores resultados de clasificación de eventos; gracias a que es un algoritmo basado en árboles de decisión y las bondades que ofrece al trabajar con características combinadas.

[1] Para este trabajo de investigación se consideran únicamente textos escritos en idioma español.

Abstract

This research paper presents a methodology for comparing machine learning algorithms that consists of four stages:

a) Collection and integration of the input data set.
b) Processing and transformation of data to the vector space model.
c) Selection and implementation of machine learning algorithms.
d) Evaluation and comparison of the results obtained.

Four supervised learning algorithms are compared: Naïve Bayes, nearest k-neighbors, C4.5 and Vector Support Machines (VSM).

The methodology proposed in this research paper considers the classification of academic events as a case study. There are four different categories: dissemination events, academic and refresher courses, academic advice to students and environmental events.

The events are obtained from the reading of sensors that keep the log of the events in logbooks. These events are characterized and processed to be used by the classifiers. There are three types of characteristics of an event, which are: nominal, textual and numerical characteristics.

Given the nature of the algorithms used in this research work, a training stage is required, which, from a set of 362 academic events, 70% of them are randomly chosen for the training of the algorithms and 30 % remaining for tests.

24 experiments were considered during the test stage in which the results obtained by the classifiers are analyzed and compared. The evaluation measures that are used to choose the best machine learning algorithm are precision and coverage and the F1 measurement, however, the selection criterion is the F1 measurement since it is an average of the accuracy and coverage.

From the results obtained in the experimentation, it can be seen that C4.5 is the algorithm with the best event classification results; because it is an algorithm based on decision trees and the benefits it offers when working with combined characteristics.

I. Introducción

Con la evolución de las tecnologías de redes inalámbricas, de los dispositivos móviles y del Internet de las Cosas (del inglés *Internet of Things*, *IoT*) han surgido nuevas infraestructuras de hardware y software que tienen como propósito la implementación y manejo de espacios inteligentes. Uno de los requerimientos relevantes del procesamiento de la información dentro de un espacio inteligente es lograr la detección y clasificación automática de los eventos que ocurren en éste. Para lograr la identificación o clasificación automática de eventos que ocurren en espacios inteligentes es necesario incorporar algoritmos de aprendizaje automático. Sin embargo, existen muchos algoritmos de aprendizaje propuestos en la literatura, los cuales se han probado con muchas fuentes de datos diversas. Por ello es necesario desarrollar una metodología que permita evaluar un conjunto dado de algoritmos y que incluya medidas de evaluación mediante los cuales se pueda tomar una mejor decisión.

En este trabajo de investigación se propone una metodología compuesta de varias etapas. Se utiliza el término "metodología" tomando como referencia la definición propuesta por el Instituto de Ingeniería Eléctrica y Electrónica, conocido por sus siglas en inglés, *IEEE* (*Institute of Electrical and Electronics Engineers*, *IEEE*).

El Instituto de Ingeniería Eléctrica y Electrónica define a una metodología como "*una serie integrada y exhaustiva de técnicas o métodos que crean una teoría de sistemas general de cómo una clase de trabajo de pensamiento intensivo debe ser realizado*" [1].

De acuerdo con la definición anterior, se puede observar que una metodología está compuesta de métodos y técnicas. El IEEE, define a un método como un "*conjunto de procesos ordenados utilizados en la ingeniería de un producto o transformación de un servicio"* [2].

Con base en las definiciones citadas anteriormente, se presentan las etapas que constituyen a la metodología expuesta en este trabajo de investigación.

- **Etapa 1.** Recopilación e integración del conjunto de datos de entrada.

En esta etapa se reúnen y seleccionan las fuentes de datos que contienen la información necesaria para su análisis. Dicha información se extrae y se organiza integrando el conjunto de datos de entrada.

- **Etapa 2.** Procesamiento y transformación de los datos al modelo espacio vectorial.

En esta etapa se realiza la traducción (transformación) de los datos a un formato común, en el cual, los datos sean unificados con el fin de facilitar su procesamiento por el algoritmo de aprendizaje automático. La transformación de los datos consiste en una lista de características mediante el modelo espacio vectorial.

- **Etapa 3.** Selección e implementación de los algoritmos de aprendizaje

automático.

En esta etapa se decide cuáles algoritmos de aprendizaje automático se van a comparar de acuerdo con las características del problema que se desea resolver.

- **Etapa 4. Evaluación y comparación de los resultados obtenidos.**

En esta etapa se evalúan y comparan los resultados obtenidos por cada algoritmo de aprendizaje automático. Las medidas de evaluación que se utilizan son: precisión, cobertura y medida F_1.

El criterio para determinar cuál es el mejor algoritmo es aquel que presenta un mayor rendimiento en la medida F_1 por ser el promedio de la precisión y la cobertura.

En cada una de estas etapas se utilizaron distintas técnicas para resolver de la mejor manera los problemas presentes en cada etapa.

Los algoritmos de aprendizaje automático que se comparan son: Naïve Bayes, k-vecinos más cercanos, C4.5 y Máquinas de soporte vectorial (del inglés, *Support Vector Machine*, *SVM*). El entrenamiento de estos algoritmos se realiza mediante la utilización de características de los eventos, las cuales son: características textuales, características nominales y características numéricas.

La evaluación y comparación de los algoritmos se hace mediante las medidas de precisión, cobertura y medida F_1 y, se selecciona al algoritmo que presenta mayor rendimiento en la medida F_1. La razón por la que se utiliza a la medida F_1 es porque representa un promedio de la precisión y cobertura.

1.1 Planteamiento del problema

Dado que en un ambiente académico dotado de una red de sensores pueden ocurrir una gran cantidad de eventos de manera simultánea y generarse bitácoras de eventos muy extensas por las diversas características de los eventos, es necesario realizar un procesamiento inteligente de los datos de los eventos registrados. Por lo anterior, se requiere la selección e implementación de un algoritmo de clasificación de eventos, de tal forma que se filtre y organice en clases la información relevante registrada durante el periodo de monitoreo. Formalmente, el problema se puede plantear de la siguiente manera:

Dado un conjunto de eventos E, donde, $e_i \in E$, $0 < i <$ n

$$E = \{e_1, e_2, e_3, \ldots, e_n\}$$

Donde cada evento e consta de las siguientes características:

$$e = \{clase, nStudent, nProfessor, nVisitor, nPersonEvento, hInEvento, hFinEvento, durationTime, tEspacio, tLugarEvento, variation, eventName, description\}$$

Donde:

clase: Clase del evento
nStudent: Número de estudiantes
nProfessor: Número de profesores
nVisitor: Número de visitantes
nPersonEvento: Total de participantes
hInEvento: Horario inicial del evento
hFinEvento: Horario final del evento
durationTime: Tiempo del evento
tEspacio: Tipo de espacio
tLugarEvento: Tipo de lugar
eventName: Nombre del evento
description: Descripción del evento
variation: Variación en eventos ambientales

Y dado el conjunto de clases de eventos C

$$C = \{Difusión, Cursos, Asesoría, Ambiental\}$$

Formado por:

Difusión: eventos de difusión
Cursos: cursos académicos y de actualización
Asesoría: asesorías académicas a alumnos
Ambiental: eventos ambientales

Problema 1: Determinar a qué clase C_i pertenece cada evento e_i.

Problema 2: Dado un conjunto de algoritmos de clasificación conocidos, determinar cual tiene mejor precisión, cobertura y medida F_1 para un conjunto de eventos E.

1.2 Justificación

El aprendizaje automático aplicado en la detección de eventos constituye un campo de interés creciente para el desarrollo de espacios inteligentes; la incorporación de algoritmos de aprendizaje automático tiene como propósito construir sistemas de información capaces de proporcionar los servicios adecuados a sus usuarios mediante la automatización de tareas comunes.

El establecimiento de un espacio inteligente en un entorno académico exige el análisis y procesamiento de grandes cantidades de datos registrados por sensores, de tal forma que sea posible determinar automáticamente el tipo de evento que está sucediendo en el ambiente.

Identificar automáticamente los eventos que ocurren en ambientes académicos no es una labor sencilla ya que pueden suceder una infinidad de ellos; principalmente aquellos eventos relacionados con la docencia, la investigación y difusión cultural.

Para identificar los eventos que ocurren, es necesario el uso de algoritmos que permitan identificarlos de la mejor manera, sin embargo, a pesar de que en la literatura especializada existe una amplia variedad de algoritmos de aprendizaje automático, para seleccionar al mejor de ellos, se necesita una metodología que permita evaluarlos y determinar mediante medidas de evaluación cual es el más apropiado. Para este trabajo de investigación, específicamente, determinar cuál es el mejor algoritmo para la clasificación de eventos académicos.

La importancia de aplicar la metodología que se propone en este trabajo de investigación, en un ambiente académico para identificar o clasificar los eventos que en él ocurren, consiste en proporcionar información relevante del ambiente a sus usuarios. Estos usuarios son, principalmente, alumnos, profesores o administrativos, con la finalidad de mejorar aspectos de gestión de recursos universitarios.

Dado que, en otros campos de estudio, no sólo en el ambiente académico, es necesario emplear algún algoritmo de aprendizaje automático, aplicar la metodología que este trabajo propone, garantiza un buen resultado en el momento de elegir el mejor algoritmo con mayor rendimiento de acuerdo con sus medidas de evaluación. Esto beneficia a investigadores y analistas de eventos de tal forma que el tiempo invertido sea menor y las actividades manuales sean eliminadas.

1.3 Objetivos

Objetivo General

- Dada una colección de eventos que ocurren en un ambiente académico, un conjunto específico de algoritmos de clasificación, el objetivo es desarrollar una metodología para comparar y determinar el mejor algoritmo utilizando medidas de evaluación.

Objetivos específicos

- Diseñar e implementar un método para la extracción y análisis de las descripciones textuales de los eventos utilizando técnicas de Procesamiento de Lenguaje Natural.

- Implementar y comparar cuatro métodos de clasificación: Naïve Bayes, k - vecinos más cercanos, C4.5 y Máquinas de soporte vectorial utilizando las características de los eventos para clasificarlos por el tipo de evento.

- Evaluar los métodos de clasificación utilizando medidas de precisión, cobertura y medida F_1 para determinar cuál es el que obtiene mayores resultados.

1.4 Estructura del trabajo

Este trabajo de investigación se encuentra organizado de la siguiente forma:

En el Capítulo I, se presenta una breve introducción a este trabajo de investigación en la que se enumeran las características y razones que motivaron su desarrollo.

En el Capítulo II, se presenta el marco teórico en el que se exponen los conceptos más importantes que se utilizan en este trabajo para facilitar su comprensión a lo largo del documento.

En el Capítulo III, se hace un análisis de los trabajos desarrollados en el área de clasificación de eventos en los últimos años.

En el Capítulo IV, se muestra la metodología propuesta para la comparación de algoritmos de aprendizaje automático que este trabajo de investigación plantea.

El Capítulo V, presenta la implementación del caso de estudio en el que se aplica la metodología propuesta en la clasificación de eventos académicos.

El Capítulo VI, se presenta la experimentación y los resultados obtenidos por los algoritmos aplicados al caso de estudio empleado en este trabajo de investigación.

El Capítulo VII, presenta el análisis y discusión de los resultados.

Finalmente se presentan las conclusiones, trabajos a futuro, referencias, anexos y artículos publicados.

II. Marco teórico

En este capítulo se presentan los conceptos y fundamentos utilizados a lo largo de este trabajo. En primer lugar, se expone el aprendizaje automático y sus tipos; además se describen brevemente los métodos de aprendizaje automático, posteriormente, se muestra la definición de evento y finalmente el concepto de ambientes inteligentes.

2.1 Aprendizaje automático

A pesar de notables avances y mejoras que han tenido los métodos de aprendizaje automático (también conocido como aprendizaje de máquina), la manera en la que aprende un humano dista mucho de lo que una máquina podría hacerlo; un humano es capaz de asimilar conceptos nuevos con al menos un ejemplo, en cambio, un algoritmo de aprendizaje automático requiere de al menos unas decenas de ejemplos para asimilar un concepto al que nunca se había enfrentado antes.

Por lo tanto, a continuación, se citan algunas definiciones formales del aprendizaje automático.

Arthur Samuel, científico informático pionero en el estudio de la Inteligencia Artificial, dijo que el aprendizaje automático es "*el estudio que le da a las computadoras la capacidad de aprender sin estar explícitamente programadas*"[3].

Por otro lado, Tom Mitchell, define el aprendizaje automático como sigue: "*Se puede decir que un programa aprende de la experiencia E con respecto a algunas clases de tareas T y el rendimiento miden P, si su desempeño en tareas en T, medido por P, mejora con la experiencia E*"[4].

Ambos autores coinciden en que una computadora aprende con base en experiencias pasadas de ahí que el aprendizaje automático generaliza una regla desconocida a partir de un conjunto de datos de entrenamiento. En resumen, el aprendizaje automático es el estudio y diseño de programas que aprenden de los datos y utilizan la experiencia pasada para tomar decisiones futuras.

En el aprendizaje automático pueden requerirse grandes cantidades de datos para su análisis lo que hace imposible su análisis de manera manual. Por lo tanto, se necesitan técnicas de minería de datos que permitan el análisis de estos datos.

Fayyad, et al. [5] aportan la definición de minería de datos más utilizada en la literatura en la que afirman que la *‘‘minería de datos es un proceso no trivial de identificación de patrones de datos válidos, nuevos, potencialmente usables y comprensibles"*. En la minería de datos se identifican patrones

Las técnicas de aprendizaje automático en la minería de datos [6], se dividen en dos categorías: supervisadas o predictivas y no supervisadas o descriptivas. Estas categorías se presentan a continuación.

2.1.1 Técnicas de aprendizaje no supervisadas o descriptivas

Las técnicas no supervisadas o métodos de aprendizaje no supervisado basan su proceso de aprendizaje en un conjunto de datos sin etiquetas o clases previamente definidas. En este tipo de técnicas no se conoce el valor objetivo o de clase, su objetivo es encontrar grupos similares en el conjunto de datos [6]. Están dedicados a las tareas de agrupamiento o segmentación.

2.1.2 Técnicas de aprendizaje supervisadas o predictivas

Las técnicas supervisadas o métodos de aprendizaje supervisado se basan en un proceso de aprendizaje que, a partir de un conjunto de datos de entrenamiento previamente etiquetado, de un cierto dominio *D*, éstas construyen criterios para determinar el valor del atributo clase en un elemento cualquiera del dominio [6]. Los algoritmos que representan a esta categoría son los algoritmos de regresión y clasificación. Es este último, donde se enfoca esta investigación.

Clasificación. La clasificación es una de las principales tareas del aprendizaje automático que consiste en asignar instancias de un dominio determinado, descritas por un conjunto de atributos discretos o de valor continuo, a un conjunto de clases [6]. La función de clasificación puede verse como se muestra en la ecuación (1).

$$c : X \rightarrow C \qquad (1)$$

Dónde:

- c*:* representa la función de clasificación.
- *X:* representa el conjunto de atributos que forman una instancia.
- *C:* representa la etiqueta de clase de dicha instancia.

El aprendizaje automático divide a la clasificación en tareas binarias, multi-clase y jerárquicas.

Al tipo de clasificación en el que se requiere que un clasificador decida si un elemento de un dominio pertenece o no a una categoría, se le conoce como clasificación simple o binaria. En el caso de la clasificación multi–clase se encuentran aquellos elementos pertenecientes a un dominio que pueden pertenecer a múltiples categorías, como en el caso de las publicaciones científicas. Por otro lado, en la clasificación jerárquica, cuenta con todos los elementos de un dominio. Sin embargo, existe la posibilidad de que surjan nuevas categorías, ejemplo de ello son la clasificación de páginas web.

En este trabajo, se utiliza la clasificación multi-clase, además se abordan cuatro técnicas de aprendizaje supervisado: Naïve Bayes, k-vecinos más cercanos, C4.5 y SVM, que a continuación se describen:

2.1.3 Clasificación bayesiana

Los clasificadores bayesianos son clasificadores estadísticos, que pueden predecir tanto las

probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. La clasificación Bayesiana se basa en el teorema de Bayes [7]. Clasificadores como Naïve Bayes [7] permiten simplificar el costo computacional del modelo probabilístico, sin pérdida de expresividad por parte del mismo demostrando una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos. La teoría de la probabilidad y los métodos bayesianos son uno de los principales enfoques utilizados en el aprendizaje automático y la minería de datos; las razones por las que estos métodos resultan importantes son:

- Los métodos bayesianos permiten hacer inferencias a partir de los datos, formular hipótesis sobre nuevos valores, además, permiten calcular explícitamente la probabilidad asociada a cada una de las hipótesis posibles.

- Facilitan el trabajo para el análisis de numerosas técnicas de aprendizaje y minería de datos que no trabajan explícitamente con probabilidades.

- Naïve Bayes [7] es el modelo más simple de clasificación en redes bayesianas. Se basa en el concepto de probabilidad condicional. Su principal característica es que supone que todos los atributos son independientes, esto da lugar a un modelo gráfico probabilístico en el que existe un único nodo raíz (clase), en la que todos los atributos son nodos hoja en donde el único nodo padre es la clase. Gráficamente se tendría la estructura mostrada en la Figura 1.

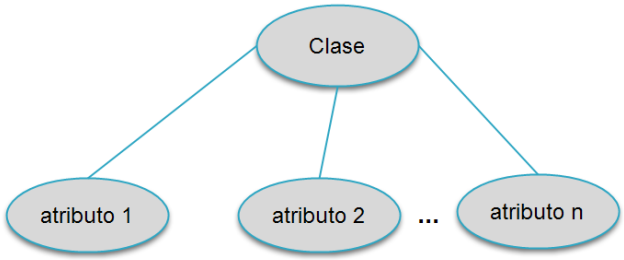

Figura 1. Representación gráfica del clasificador Naïve Bayes.

El clasificador Naïve Bayes [7], es utilizado cuando se quiere clasificar un ejemplo descrito por un conjunto de atributos (*xi's*) en un conjunto finito de clases (*c*) de acuerdo con el valor más probable, dados los valores de sus atributos [8] por lo tanto, el objetivo de este clasificador es encontrar la clase óptima para un determinado evento $d= \{d_1,d_2,d_3,\ldots,d_n\}$, calculando la clase que da la probabilidad posterior máxima a partir de los ejemplos del conjunto de entrenamiento. Dado que el problema de clasificación es el que se desea resolver, se tiene una variable de clase (C) y un conjunto de variables predictoras o atributos $\{d_1,\ldots, d_n\}$, el teorema de Bayes se ve de la siguiente forma:

$$arg\ max = P\ (c\ |\ d_1, d_2\ ,\quad \ldots, d_n\ ) \qquad (2)$$

$$= arg\,max\ \frac{P(d_1,\ d_2, d_{3,}\ \ldots, d_n \mid c)P(c)}{P(d_1, d_{2,}\ d_{3,}\ldots, d_n\ )} \tag{3}$$

$$\approx arg\,max\ \ P\ (c)\ \ \Pi(i,m)\ P\ (d_i\ |c) \tag{4}$$

Donde *p(c)* y $p(d_i|c)$ se estiman a partir del conjunto de entrenamiento utilizando las frecuencias relativas. Es decir, el número de casos favorables dividido por el número de casos totales, a esta técnica se le conoce como estimación por máxima verosimilitud.

2.1.4 Aprendizaje basado en ejemplos

La clasificación basada en ejemplos se realiza por medio de una función que mide la proximidad o parecido con los ejemplos existentes. Una métrica de distancia y los ejemplos más cercanos son utilizados para asignar la clase a la nueva instancia [9]. En la ecuación (5) se muestra formalmente que para un conjunto de elementos *X*, se considera distancia a toda función:

$$F: X \times X\ \rightarrow \mathcal{R} \tag{5}$$

Que cumpla las siguientes condiciones [6]:

- **No negatividad**. La distancia entre dos puntos siempre debe ser positiva, esto se muestra en la ecuación (6).

$$d(x,y) \geq 0\ \forall\ x,y \in X \tag{6}$$

- **Simetría**. La distancia entre un punto *a* y un punto *b* debe ser igual a la distancia entre el punto *b* y el punto *a.* Esto se muestra en la ecuación (7).

$$d(x,y) = d(y,x)\ \forall\ x,y \in X \tag{7}$$

- **Desigualdad triangular**. La distancia de un lado de un triángulo es menor que la suma de las longitudes de los otros dos lados. Esto se muestra en la ecuación (8).

$$d(x,z) \leq d(x,y) + d(y,z)\ \forall\ x,y,z\ \in X \tag{8}$$

La métrica de distancia euclideana es de las más utilizadas. Su expresión matemática se muestra en la ecuación (9):

$$de(x,d_i) = \sqrt{\sum_{j=1}^{m}(x_j - d_{ij})^2} \tag{9}$$

El clasificador *k* - vecinos más cercanos (del inglés, *k-Nearest Neighbor*, *KNN*) [10], es un

método de aprendizaje perezoso basado en ejemplos, que se basa en el modelo de espacio vectorial, el cual representa un conjunto de vectores de la forma ($a_1(x)$, $a_2(x)$,..., $a_n(x)$) donde $a_r(x)$ es el valor de la instancia para el atributo a_r. *KNN* compara todos los ejemplos de entrenamiento y la similitud entre sus vectores de características. Para encontrar los *k* ejemplos de entrenamiento más cercanos y el ejemplar desconocido es designado a los *k* vecinos más cercanos con mayor valor de clasificación. La principal ventaja de este algoritmo es su facilidad de implementación, sin embargo, un gran número de instancias en la etapa de entrenamiento, genera un alto costo computacional.

La similitud y el valor de *k* determinan la bondad del método. Para establecer el valor de la variable *k*, suele elegirse un número impar o primo, esto evitará que existan empates en el momento de decidir la clase a la que pertenezca una nueva instancia, si esto llegara a ocurrir, es necesario decidir cómo clasificar a la instancia. Algunas alternativas son: no dar la predicción o dar la clase más frecuente, en el conjunto de entrenamiento de las clases que han generado el empate.

Elegir un valor alto para la variable *k* significa un alto costo computacional además de que se consideran instancias que no son tan próximas para determinar la clase de una instancia provocando un error de generalización.

Su funcionamiento se detalla en el algoritmo 1 y se describe a continuación:

- Se determina el valor de *k*, comúnmente se elige un número pequeño impar o primo.
- Se selecciona una instancia *x* del conjunto de datos D_{test}
- El algoritmo selecciona las *k* instancias del conjunto de datos D_{train} más cercanas de acuerdo con la métrica de similitud utilizada
- Finalmente, la instancia *x* es asignada a la clase más frecuente de entre las *k* instancias seleccionadas como más cercanas

Algoritmo 1. Pseudocódigo de k-NN

k- NN (k, x)
para todo $d_i \epsilon D$ *hacer*
    calcular la distancia $d(d_i, x)$
fin para
$I \leftarrow C_i$
regresar I

2.1.5 Árboles de decisión

Un árbol de decisiones es una estructura en árbol, donde cada nodo representa un atributo a ser probado; las ramas representan la salida de la prueba y los nodos finales (hojas) representan la clasificación. Por lo tanto, un árbol de decisiones es una secuencia de condiciones del tipo *if- then*.

La profundidad máxima de un árbol de decisión es el número de condiciones que es

necesario resolver para llegar a una hoja. Un árbol de decisión se llama completo o puro si es posible subdividir el espacio de datos en subconjuntos más pequeños en donde cada subconjunto contenga solo elementos de una misma clase.

El algoritmo de árboles de decisión posee dos fases principales: en la primera llamada fase de crecimiento del árbol, el algoritmo inicia con todo el conjunto de datos como nodos raíz. Los datos son divididos en subconjuntos utilizando algún criterio de división. En la segunda fase, etapa de poda del árbol, el árbol total formado se poda para prevenir el exceso de ajuste (*over-fitting*) del árbol a los datos de entrenamiento. Es decir, se eliminan aquellos subárboles que no mejoran el árbol durante el proceso de creación. Los criterios para construir un árbol de decisión son los siguientes [6]:

- Criterio de parada (*Cp*). Determina en qué momento deja de subdividir nodos. Generalmente se usa un árbol completo.
- Criterio de selección (*Cs*). Determina qué nodo se particiona en dos o más subconjuntos. Normalmente es aquel que contiene más elementos de diferentes clases.
- Criterio de clasificación (*Cc*). Determina la clase que se asigna a un nodo hoja. Normalmente se trata de aquella que minimiza el error de clasificación.
- Criterio de división (*Cd*). Determina cómo se particiona un nodo en dos o más subnodos. Normalmente se utilizan árboles binarios.

El algoritmo para construir un árbol de decisión se muestra a continuación.

Algoritmo 2. Pseudocódigo para construir un árbol de decisión

```
DT(D)
  T ← { D }
  Etiquetar:  T ← Cc
  P ← { T }
  Mientras no se deba parar hacer
   Seleccionar: q ← Cs
    Si es posible particionar q entonces
      Particionar q en q1, ...qP
      Etiquetar: q1, ..., qP ← Cc
      Añadir: P ← q1, ..., qP
      Sustituir: q en T por un nodo interno
    fin si
    Eliminar: q de P
  fin mientras
 regresar T
```

Existen diversos algoritmos para construir árboles de decisión entre ellos ID3 [11], C4.5 [12], SPRINT [13], SLIQ [14] y PUBLIC [15]. El utilizado en este trabajo es el algoritmo C4.5 [12] que incluye diversos métodos para trabajar con atributos numéricos, valores ausentes, datos con ruidos y para generar reglas a partir de árboles de decisión.

En cuanto a métodos de aprendizaje, existen muchas razones por las que los árboles de decisión son de los más utilizados:

- Son sencillos de entender y los más adecuados para las tareas de clasificación, es decir, de manera más sencilla podemos determinar a qué clase pertenece algún objeto y específicamente en este trabajo a qué clase pertenece un determinado evento.
- Las posibles opciones, a partir de una determinada condición son excluyentes por lo que no pueden existir dos instancias que pertenezcan a dos clases distintas al mismo tiempo.
- Pueden combinar variables numéricas y categóricas en el mismo modelo.

2.1.6 Máquinas de soporte vectorial (SVM)

Máquinas de soporte vectorial [16] es un método de aprendizaje supervisado que, según la literatura, tiene un buen rendimiento en tareas de clasificación. Está basado en la clasificación lineal separando los datos en dos clases. El algoritmo pretende encontrar el hiperplano que maximiza el margen entre los vectores de soporte que define la posición del hiperplano ideal.

El margen es la zona de separación entre los distintos grupos a clasificar y se define como el doble de la distancia entre el hiperplano y el punto más cercano a este, cuanto más alejados estén los puntos del hiperplano será menos probable clasificar incorrectamente a una nueva instancia. En la Figura 2 se ilustra un ejemplo y se muestra el margen.

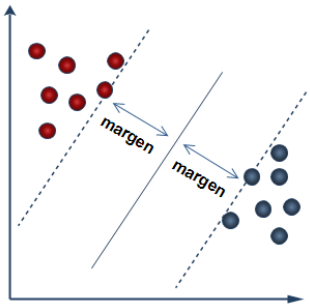

Figura 2. Función SVM y margen.

Existe una cantidad infinita de hiperplanos que separan ambas clases, pero es necesario utilizar aquel que minimice el error por exceso de ajuste para esto utiliza los vectores de soporte que sirven para anclar el hiperplano óptimo que pasa por su centro.

En caso de que no sea posible dividir las clases por medio de un separador lineal, SVM mapea los vectores a otro espacio en donde sí sea posible por medio de una función no lineal conocida como función Kernel.

- La ventaja de utilizar este método es su buen desempeño cuando se cuenta con un gran número de características y también cuando se tienen pocos elementos de entrenamiento en tareas de múltiples clases [16].

2.2 Evento

Dado que la definición de evento es muy amplia, es necesario conocer de manera general la definición formal de un evento. Posteriormente, la definición para un Ambiente Inteligente (del inglés *AMbient Intelligence, AmI*).

- Miller et al. [17] definen a un evento como un suceso que implica un cambio de estado, aspectos locativos, temporales y causales.

- Allen et al. [18] definen a los eventos como la forma en que los agentes clasifican ciertos patrones de cambio útil y relevante. Los eventos involucran a objetos, intervalos de tiempo y un cambio de estado.

- Galton et al. [19] afirman que todos los eventos están dados de acuerdo a intervalos e instantes de tiempo y que involucran una causalidad.

- Sowa [20] afirma que un evento es una entidad que puede involucrarse en la causalidad y que puede ser identificado por su ubicación en una región del espacio-tiempo.

- Zwaan et al. [21] afirman que los eventos se organizan en cinco índices: especialidad, temporalidad, protagonistas, causalidad e intencionalidad, cada uno está organizado sobre un marco temporal en el cual ocurren, una región espacial donde suceden, los protagonistas involucrados, su estado causal con respecto a eventos anteriores y su relación con el objetivo de los protagonistas.

- Para Koohzadi [22] los eventos pueden ser definidos como ocurrencias del mundo real que se desarrollan en el espacio y el tiempo. Un evento tiene una duración, ocurre en un lugar específico e implica cierto cambio de estado.

- Para Wasserkrug [23] en Inteligencia Ambiental (del inglés *AMbient Intelligence, AmI)* se refiere a un evento cuando se cumplen ciertas condiciones o cuando se alcanzan ciertos estados.

Tomando en consideración las definiciones anteriores, se puede concluir que un evento es un hecho que sucede en un espacio físico o lugar, en un momento determinado de tiempo, en el cual se ven involucrados participantes (agentes o personas). Asimismo, se puede observar que un evento tiene causas o motivos que lo originan y los posibles efectos. En este trabajo de investigación, se estudian los tres primeros aspectos de un evento: espacio, tiempo y participantes. En particular, se estudian cuatro clases de eventos que ocurren en ambientes académicos, éstas se describen a continuación.

a) **Eventos de difusión**. Evento cuyo objetivo es difundir los avances de los proyectos de investigación, de los proyectos de docencia y todo tipo de eventos

culturales.

b) **Cursos académicos y de actualización**. Esta clase de eventos tiene como objetivo la formación académica y profesional de estudiantes, profesores y público en general.

c) **Asesorías académicas a alumnos**. Este tipo de evento consiste en la atención personalizada que brinda un profesor a un estudiante para resolver cuestiones y dudas sobre temas específicos.

d) **Eventos ambientales**. Este tipo de evento puede definirse como un fenómeno ambiental que sucede al interior de un espacio académico (salón, oficina, auditorio, jardín, etc.) el cual es detectado mediante sensores que miden variaciones en las variables físicas (luz, humedad, temperatura, etc.).

En todos estos tipos de eventos se detectan dos aspectos fundamentales, la presencia de personas y el tiempo en el que ocurren.

2.3 Ambiente inteligente

En 2001 la Comisión Europea trazó por primera vez el camino para la investigación de la inteligencia ambiental (del inglés *AMbient Intelligence, AmI)* [24]. Un ambiente inteligente es aquel entorno enriquecido con sensores y dispositivos interconectados a través de una red. Los ambientes inteligentes son sensibles y adaptables al contexto.

Con el uso de los métodos de aprendizaje automático, es posible construir un sistema tal que actúe como mayordomo electrónico capaz de identificar a sus usuarios y características, los eventos que suceden en él, de esta manera proporcionales los servicios adecuados. A pesar de que el término ambiente inteligente es utilizado principalmente en “casas inteligentes” es posible extender su aplicación de estudio a los espacios académicos como en este trabajo.

III. Estado del arte

Este capítulo describe el trabajo realizado en el área de clasificación de eventos basada en información no estructurada como sus descripciones textuales. Además, se explora el uso de algoritmos de aprendizaje automático para dicha tarea. Con respecto a la clasificación de eventos utilizando textos, existen trabajos que han utilizado las redes sociales como su fuente de información.

En [25] se propone un enfoque de clasificación de mensajes de texto corto a través de utilizar un conjunto de características específicas extraídas del perfil y el texto del autor. El objetivo de este trabajo es clasificar de manera automática tweets entrantes en un conjunto predefinido de clases genéricas (noticias, eventos, opiniones, ofertas y mensajes privados) con la intención de que los usuarios no sean atestados por los datos en bruto, particularmente aquellas personas que acceden a Twitter a través de un dispositivo móvil. Los autores de este trabajo utilizan un conjunto de siete características binarias (presencia de acortamiento de palabras y el lenguaje utilizado, frases de eventos de tiempo, palabras con opinión, énfasis en palabras, signos de moneda y porcentaje, @nombre de usuario, al principio del tweet y nombre de usuario dentro del tweet) y una característica nominal (autor). Para la parte experimental obtienen una colección de tweets recientes de usuarios aleatorios de los cuales son eliminados aquellos tweets que no se encuentran en el idioma inglés y no cuenten con un mínimo de tres palabras excluyendo el saludo. Estos tweets se etiquetaron manualmente y, posteriormente eliminan palabras vacías. Para determinar la existencia de un evento, extraen la fecha, hora y frases de eventos de tiempo, además, realizan los experimentos con la implementación disponible del clasificador Naïve Bayes en WEKA utilizando una validación cruzada cinco veces.

En [26] analizan el flujo de mensajes de Twitter para distinguir entre mensajes sobre eventos del mundo real y mensajes que no son eventos. Identifican cada evento y sus mensajes de Twitter asociados mediante una técnica de agrupación en línea de tweets tópicamente similares. Los autores proponen un algoritmo de agrupación incremental con un parámetro de umbral que se sintonice empíricamente durante una fase de entrenamiento. Dicho algoritmo considera cada mensaje y determina la clase adecuada en función de la similitud del mensaje con los grupos existentes. Para identificar la categoría de los eventos en el flujo, calculan varias categorías de características que describen diferentes aspectos de las agrupaciones que desean modelar para determinar a cuál corresponden. Consideran características temporales, sociales, tópicas y centradas en Twitter que ayudan a detectar los clústeres que están asociados con eventos usando estadísticas del conjunto de mensajes. Estas características son utilizadas para entrenar a un clasificador para distinguir eventos de los que no son eventos.

En [27] describen “TwiCal” un sistema para la extracción y categorización de eventos de dominio abierto para Twitter. Proponen un enfoque no supervisado para descubrir automáticamente categorías de eventos importantes y clasificar eventos extraídos en función de variables latentes. Los eventos descubiertos de manera automática se inspeccionan

posteriormente para filtrar los que son incoherentes y el resto se anota con etiquetas informativas, algunas como: finanzas, educación, religión, deportes y política. El sistema aquí descrito extrae eventos representados como tuplas de cuatro elementos (una entidad nombrada, frase de evento, fecha del calendario y tipo de evento); obtiene el nombre de entidades en asociación con frases de eventos, fechas involucradas con eventos y se extraen frases, se resuelven expresiones temporales y los eventos se clasifican en tipos. El conjunto de eventos resultante de clases de eventos posteriormente es utilizado para categorizar cientos de millones de eventos reales extraídos de manera automática.

En [28], proponen un enfoque general no supervisado para explorar eventos a partir de tweets constituido por dos etapas: 1) filtrado y 2) extracción y clasificación. En la etapa de filtrado el ruido de los tweets es eliminado, se construye un léxico de palabras clave a partir de artículos de noticias publicados en el mismo periodo que los tweets y se extraen solo aquellos que contienen palabras que pueden encontrarse en el léxico; estos se utilizan para filtrar los no relacionados con eventos, además, emplean un segundo enfoque que percibe el filtrado de tweets como un problema de clasificación binaria. Las características que ellos utilizan para entrenar el clasificador binario y determinar la ocurrencia de un evento son: la presencia de entidades con nombre, la ubicación y la información de tiempo. Los eventos son representados como una tupla de cuatro elementos, compuesta por una entidad con nombre diferente a una ubicación, una fecha, una ubicación y palabras clave relacionadas con eventos (<*y, d, l, k*>). En la etapa de extracción y clasificación de eventos, los autores realizan un preprocesamiento, extracción y categorización utilizando el modelo de evento y categoría latentes (*LECM*). Para realizar la extracción de entidades nombradas, se utiliza un etiquetador de entidades con nombre y conservan sólo aquellas palabras etiquetadas como verbos, adjetivos y sustantivos.

En el dominio de la bioinformática, la clasificación de eventos a partir de textos médicos ha sido de gran ayuda para la identificación y extracción automática de eventos adversos.

En el Procesamiento Biomédico del Lenguaje Natural (BioNLP) intentan capturar fenómenos biomédicos de los textos mediante la extracción de relaciones entre entidades biomédicas como eventos complejos (es decir, proteínas y genes) por lo que en [29] proponen un sistema para la extracción automática de eventos complejos. El objetivo de este sistema es identificar eventos junto con sus tipos, desencadenadores principales (tokens que representan eventos), tema principal y argumentos de causa. Este sistema sigue un proceso de detección de disparos, detección de bordes y detección de eventos complejos. El sistema resuelve un problema de clasificación multi-clase y multi-etiqueta, además, construye un modelo nuevo. Sus principales contribuciones radican en proponer un método efectivo de detección de eventos biológicos utilizando el aprendizaje automático, la provisión de un sistema de extracción de eventos de alto rendimiento y la ejecución de un análisis de error cuantitativo.

La relación entre los datos sociales y los eventos que representan incrementan la tendencia de controlar los recursos de las redes sociales para comprender los eventos sociales. Dado que muchos contenidos de las redes sociales de hoy en día vienen con metadatos (marcas de tiempo, etiquetas, etiquetas geográficas, ID del cargador de metadatos, entre otras) y reflejan

parcialmente la información contenida, en [30] proponen explotar la etiqueta y el título como un problema de clasificación de texto supervisado en el que introducen un método para la detección de eventos basado en el procesamiento de lenguaje natural para detección de eventos sociales. Los autores de este trabajo, realizan un análisis de características específicas del lenguaje natural en las redes sociales y proponen un método basado en las características de las redes sociales en idioma inglés breve, ambiguo y no estándar. Seleccionan las características del lenguaje natural de las redes sociales luego extraen características usando herramientas comunes de procesamiento de lenguaje natural. Posteriormente, con el modelo bolsa de palabras simplifican la representación de datos de un contenido multimedia y finalmente emplean SVM como método de aprendizaje supervisado.

En [31] analizan eventos de riesgo en tiempo real (terremotos, erupciones volcánicas, deslizamientos de tierra, inundaciones o la aurora), como proyecto del GeoSocial mediante tweets de localización geográfica a través de un filtrado por palabras clave. Los autores examinan dos técnicas de clasificación automática de eventos de riesgo basada en tweets sobre la ‘‘Aurora’’. El filtrado de los tweets se realiza a través de palabras clave relacionadas con la aurora, eliminando las palabras vacías y seleccionando las escritas en inglés, los autores de este trabajo clasifican a los eventos en dos categorías ‘‘aurora-evento’’ o ‘‘no-aurora-evento’’ mediante dos técnicas de minería de datos: SVM y algoritmos de redes neuronales de convolución profunda (CNN). El resultado obtenido por estos clasificadores empleó un conjunto de entrenamiento conformado por un total de 1200 tweets.

En [32] se propone un método de clasificación textual (CLIM) para la identificación de eventos delictivos. Para ello se realiza un análisis semántico de noticias periodísticas en español. El proceso de clasificación se compone de tres subprocesos: extracción de información, proceso de clasificación y proceso de selección de las clases. Detectan cuatro tipos de delitos (homicidio, asalto, secuestro y abuso sexual) mediante técnicas de aprendizaje supervisado y proponen una metodología que consta de tres partes: el modelo de clasificación múltiple que consiste en la identificación de frases verbales y frases de nombre, luego, a través de una adaptación del modelo TF-IDF, se calcula el peso de cada palabra clave seleccionada. Los autores proponen un modelo para seleccionar automáticamente términos o palabras clave que proporciona al clasificador un mayor grado de autonomía y permite un autoaprendizaje eficiente para actualizar los perfiles en el cuerpo de CAD (Corpus Anotado de Delitos), utilizando el grado de dispersión de los términos según sus ponderaciones. Los resultados de este modelo no fueron tan altos como se esperaba ya que a veces coincidían las predicciones erróneas de los dos clasificadores convencionales. Las pruebas de eficiencia sobre este modelo se realizaron a través de la simulación de diferentes escenarios utilizando las noticias de CAD y se compararon la eficiencia lograda por clasificadores como SVM y NB, utilizados por su alto uso en tareas de clasificación textual.

En [33] tratan a la detección de eventos como un problema de clasificación binaria a nivel de oraciones que describen eventos. Esta tarea es de utilidad para aplicaciones de lenguaje natural como la respuesta a preguntas y resumen de textos. Los autores llevan a cabo la tarea de detección de eventos comparando el rendimiento de enfoques discriminativos frente a generativos a través de la evaluación de las técnicas de aprendizaje automático: SVM y

modelado de lenguaje. Los autores de este trabajo, construyen un clasificador binario para cada tipo de evento, de modo que una oración que pertenece a la recopilación de datos de ese tipo, se asigna a una oración en un evento u oración fuera del evento, es decir, una oración en un evento es una oración que contiene una o más instancias del tipo de evento objetivo y una oración fuera de evento es aquella que no contiene ninguna instancia del objetivo tipo de evento.

Los eventos que identifican son seis diferentes: muerte, ataques, lesiones, manifestaciones, transporte y acusaciones de cargo. Para ello usan SVM a fin de determinar qué función del núcleo es más adecuada para la clasificación. Crearon tres versiones de la SVM. El primero fue construido usando un núcleo lineal, el segundo con un núcleo polinomial, y el tercero usando una función de núcleo RBF. Cada variación fue evaluada utilizando los datos de IBC *(*artículos de *Newswire)*.

En [34] se propone un enfoque supervisado (utilizan: NB, KNN, SVM, DT, TF, NN, QDA y Adaboost) para identificación de usuarios en casas inteligentes a partir del análisis de patrones de comportamiento, mientras desempeñan sus actividades diarias. Los datos son obtenidos de sensores ambientales desplegados en los hogares inteligentes. En [34] la secuencia de eventos es etiquetada como un patrón de comportamiento del usuario. Basados en el modelo "bolsa de palabras", propone el modelo "bolsa de eventos" en el que las características se representan como una secuencia de actividades realizadas por un usuario. Utilizan dos categorías de características: eventos secuenciales (la actividad de cada persona es una secuencia de eventos del sensor) y bolsa de eventos de sensores para resolver el problema de variaciones en el comportamiento de los usuarios cuando desempeñan sus actividades pues algunas personas realizan sus actividades en diferente orden. Aquí cada estado de un sensor es considerado como una característica.

Los trabajos analizados se centran en las siguientes características:

a) **Algoritmos utilizados**. En la actualidad gran parte de la investigación hace referencia a los algoritmos de clasificación Naïve Bayes y SVM, esto sugiere fuertemente que son dos candidatos importantes para el análisis en este trabajo de investigación. Sin embargo, se puede observar que algoritmos como k - vecinos más cercanos y C4.5 son algunos ejemplos de algoritmos poco utilizados en las investigaciones recientes.

b) **Tipo de aprendizaje**. Hablar de aprendizaje automático está estrechamente relacionado con el tipo de aprendizaje que emplea un algoritmo de aprendizaje automático, lo que hace notar que el principal enfoque de investigación empleado en las investigaciones recientes, es un enfoque supervisado.

c) **Tipo de evento**. Desarrollar una metodología para la comparación de algoritmos de aprendizaje automático y tomar como caso de estudio a la clasificación de eventos académicos requiere de analizar distintos tipos de eventos. El estudio de eventos en las investigaciones recientes se inclina por eventos de interés general

entre los que destacan noticias y eventos sociales, sin embargo, no se ha presentado un análisis de eventos académicos, esto representa un área de oportunidad para este trabajo de investigación.

d) **Fuente de datos**. Este trabajo de investigación resalta la importancia de analizar la fuente de información de los eventos y se observa que la mayor parte de los trabajos previos en este campo ha utilizado textos cortos obtenidos de las redes sociales.

e) **Tipo de característica**. El estudio comparativo de los diferentes trabajos de investigación presentados en el Capítulo III, muestra que la mayor parte de los autores utiliza características textuales, y en solo uno de los casos utilizan características nominales y binarias. Por el contrario, en este trabajo de investigación se abordan características: textuales, numéricas y nominales.

f) **Idioma**. En el estudio realizado de los diferentes trabajos analizados anteriormente indica que el idioma inglés es el más explorado por los investigadores debido a la sencillez de su gramática en contraste con la del idioma español que se caracteriza por ser más compleja. Este proyecto de investigación resalta la importancia de evaluar distintas características textuales en idioma español.

Es interesante mencionar que en los trabajos relacionados se han abordado problemas de aprendizaje automático siguiendo procedimientos similares; sin embargo, ninguno ha presentado una metodología genérica para comparar y evaluar algoritmos de clasificación como lo hace este trabajo de investigación.

En la Tabla 1 se presenta un resumen de las características de cada trabajo de investigación analizado, las principales similitudes y las diferencias existentes entre éstos y el presente proyecto de investigación.

Tabla 1. *Tabla comparativa de los trabajos relacionados*

Autor	Algoritmos utilizados	Tipo de Aprendizaje	Tipo de evento	Fuente de datos	Tipo de característica	Idioma
[25]	Naïve Bayes	Supervisado	Noticias, eventos, opiniones, ofertas, mensajes privados	Twitter	Binarias Nominales	Inglés
[26]	Algoritmo de agrupamiento	No supervisado	Eventos del mundo real, mensajes que no son eventos	Twitter	Textuales	Inglés
[27]	Modelos bayesianos: Modelo de evento y categoría latentes (LECM)	No supervisado	Finanzas, educación religión, deportes, política	Twitter	Textuales	Inglés
[28]	Modelos de variables latentes	No supervisado	Eventos de dominio abierto	Twitter	Binarias	Inglés
[29]	SVM	Supervisado	Eventos biológicos	Textos médicos	Textuales	No Aplica
[30]	SVM	Supervisado	Eventos sociales	Textos obtenidos de las redes sociales	Textuales	Inglés
[31]	SVM CNN	Supervisado	Eventos de peligro	Twitter	Textuales	Inglés
[32]	SVM Naïve Bayes Hibrido CLIM	Supervisado	Eventos delictivos	Corpus criminológico (CAD)	Textuales	Español
[33]	SVM Naïve Bayes	Supervisado	Muerte, ataque lesionar, manifestación, transporte, acusaciones de cargo	El Corpus Multilingüe ACE 2005 Base de datos de IBC	Textuales	Inglés
[34]	Naïve Bayes KNN DT SVM RF Redes Neuronales Adaboost QDA	Supervisado	Identificación de usuarios basada en sus hábitos	Bolsa de eventos de sensores	Eventos secuenciales Bolsa de eventos de sensores Bolsa de actividades	No aplica
AGR	Naïve Bayes KNN C4.5 SVM	Supervisado	Eventos académicos	Bitácoras de eventos	Textuales Nominales Numéricas	Español

IV. Metodología propuesta

En este capítulo se describe la metodología desarrollada para la comparación de algoritmos de clasificación, utilizando como datos de entrada un conjunto de eventos que fueron registrados en un ambiente académico.

La metodología para la comparación de algoritmos (ver Figura 3) consiste en cuatro etapas:

1. Recopilación e integración del conjunto de datos de entrada.
2. Procesamiento y transformación de los datos al modelo espacio vectorial.
3. Selección e implementación de los algoritmos de aprendizaje automático.
4. Evaluación y comparación de los resultados obtenidos por cada algoritmo de aprendizaje automático utilizando las medidas de precisión, cobertura y medida F_1.

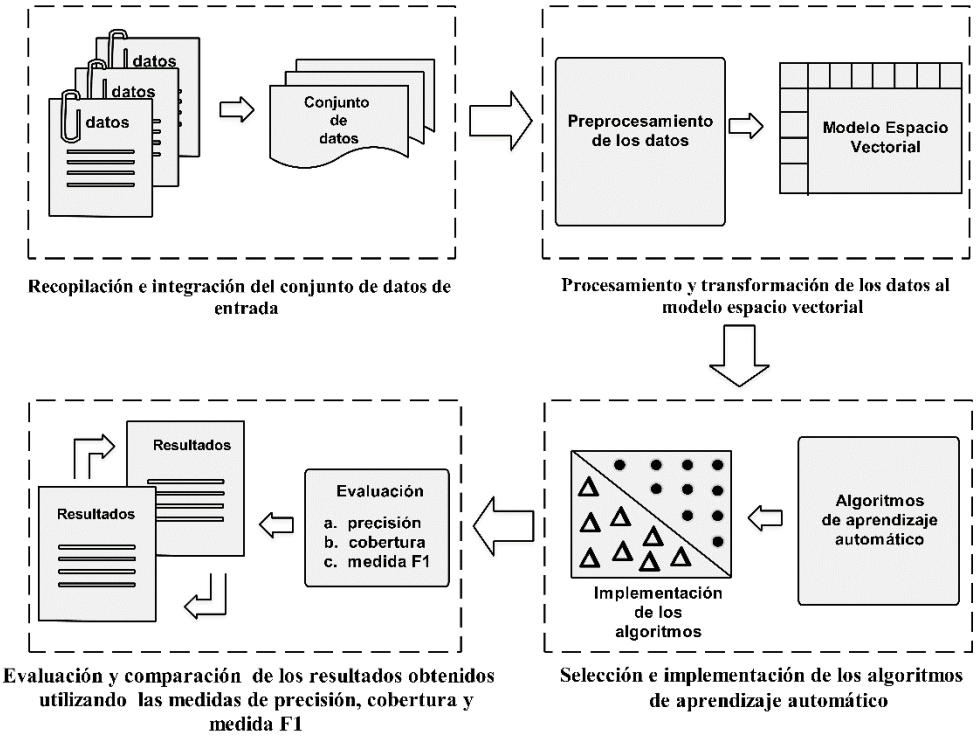

Figura 3. Metodología propuesta

A continuación, se explica la metodología propuesta en este trabajo de investigación.

4.1 Etapa de recopilación e integración del conjunto de datos de entrada

En la etapa de recopilación e integración del conjunto de datos de entrada se reúnen y seleccionan las fuentes de datos que contienen la información necesaria y que sirve como punto de partida para el análisis. Esta información se extrae y se organiza integrando así el

conjunto de datos de entrada.

4.2 Etapa de procesamiento y transformación de los datos al modelo espacio vectorial

En esta etapa se realiza la traducción de los datos, es decir, que el conjunto de datos obtenido en la etapa anterior pasa una transformación a un formato común en el cual los datos sean unificados con el fin de que sea posible su procesamiento por el algoritmo de aprendizaje automático.

La transformación de los datos consiste en una lista de características mediante el modelo espacio vectorial. En esta etapa se detectan y eliminan inconsistencias, tal como se describe en la sección 5.2.1.

4.3 Etapa de selección e implementación de los algoritmos de aprendizaje automático

En esta etapa se seleccionan los algoritmos de aprendizaje automático que se van a comparar de acuerdo con las características del problema que se desea resolver.

4.4 Etapa de evaluación y comparación de los resultados obtenidos

En la etapa de evaluación y comparación de resultados se comparan los resultados obtenidos por cada algoritmo de aprendizaje automático y se utilizan las medidas de evaluación: precisión, cobertura y medida F_1; y se decide cuál de ellos presenta mayor rendimiento. El criterio para seleccionar al mejor algoritmo es la medida F_1que representa un promedio de las dos medidas de evaluación anteriores.

V. Implementación del caso de estudio

En este capítulo se describe detalladamente la implementación de la metodología para la comparación de los cuatro algoritmos de clasificación aplicados en la identificación de los tipos de eventos que suceden en un ambiente académico.

5.1 Recopilación e integración del conjunto de datos de entrada

En esta etapa utiliza la información almacenada en una bitácora que, mantiene el registro histórico de 362 eventos académicos detectados por una red de sensores en un espacio académico.

La estructura presente en esta bitácora de eventos se compone de dos tipos de datos, datos estructurados y datos no estructurados, esto se puede observar en la Tabla 2, que muestra un ejemplo de los datos que almacena una bitácora de eventos.

Tabla 2. *Estructura de una bitácora de eventos*

Bitácora de eventos									
clase	nStudent	nProfessor	nvisitor	hInEvento	hFinEvento	tLugarEvento	variation	eventName	description
…								…	
Datos estructurados								Datos no estructurados	

En cuanto al tipo de datos estructurados se tienen:

- *Clase*. Indica la categoría a la que pertenece un evento
- *nStudent.* Indica número de estudiantes
- *nProfessor*. Indica el número de profesores
- *nVisitor*. Indica el número de personas externas a un espacio académico
- *hInEvento*. Indica la hora en la que se registra el inicio de un evento
- *hFinEvento.* Indica la hora en la que se registra el término de un evento
- *tLugarEvento*. Indica el espacio físico en el que se registra un evento
- *variation*. Indica los cambios registrados en las lecturas de los sensores ambientales

En cuanto al tipo de datos no estructurados se tienen:

- *eventName*. Indica el nombre del evento
- *description*. Indica la descripción en idioma español que describe al evento

Estas bitácoras se obtienen utilizando un programa simulador[2] de eventos para un espacio académico.

Los elementos que caracterizan a un evento son: el nombre, la descripción del evento, los participantes del evento (alumnos, profesores y visitantes), el lugar donde ocurrió el evento, el horario en el que sucedió cada evento y los datos sobre las variaciones en las lecturas de los sensores en caso de que se trate de eventos ambientales.

Los datos que almacenan estas bitácoras son la base para el análisis que se realiza, pues a partir de ellos, se obtienen algunos otros datos que pueden ser de utilidad para la clasificación de los eventos

Los eventos se extraen y se almacenan en archivos de texto plano y se obtienen aquellos datos faltantes, como son el total de participantes en un evento, el tiempo total que tarda en realizarse un evento y el tipo de espacio, ya sea un espacio abierto o cerrado (en interiores o al aire libre). Toda esta información constituye el conjunto de datos de entrada, y, en una etapa siguiente son procesados y se representan como un conjunto de características de cada evento.

5.2 Procesamiento y transformación de los datos al modelo espacio vectorial

El procesamiento y transformación depende de la recopilación de datos pues se realiza la traducción (transformación) de los eventos almacenados en las bitácoras de eventos a la lista de características y valores que caracterizan a un evento. Por lo anterior, es necesario un preprocesamiento de los datos que posteriormente serán utilizados en la etapa de clasificación. El preprocesamiento de eventos se realiza en dos etapas independientes: extracción de características textuales y extracción de características numéricas y nominales.

5.2.1 Extracción de características textuales

El término “lexicografía” hace referencia a la rama de la Lingüística encargada de la elaboración de diccionarios, por lo tanto, el papel de la lexicografía es recopilar las unidades léxicas (palabras) de un idioma y descubrir el sentido y empleo de cada una de ellas. En el léxico se incluye la información morfológica, la categoría gramatical, irregularidades sintácticas y representación del significado. Con base en lo anterior, en esta etapa se realiza un análisis morfológico del nombre y descripción de cada evento, cuyo objetivo es obtener los verbos, adjetivos y sustantivos presentes en los eventos, como parte de este análisis se hizo uso de técnicas de Procesamiento de Lenguaje Natural.

En la Tabla 3 se puede observar al nombre y la descripción del evento como características de texto las cuales posteriormente se van a representar mediante características lexicográficas (verbos, adjetivos y sustantivos).

[2] Para referencia ver el anexo A

Tabla 3. *Tabla de características de texo*

Tipo de característica	Característica de texto	Descripción	Posibles valores
Característica de texto	Nombre del evento	Denominación verbal en español que se le asigna a un evento	Valor nominal = {cadena}
Característica de texto	Descripción del evento	Narración de extensión corta, en español que se hace sobre un evento	Valor nominal = {cadena}

La extracción de características textuales consta de las siguientes tareas: segmentación, limpieza, lematización y la obtención de la lista de características, que se explican a continuación.

a) **Segmentación**. Es la primer tarea en la extracción de características lexicográficas; para realizar la segmentación es necesario obtener todas las cadenas delimitadas por un espacio en blanco que de acuerdo con Jackson y Moulinier [35], son conocidos como *tokens*. Si una cadena contiene o se encuentra escrita con letras mayúsculas, éstas se convierten a minúsculas. La Tabla 4 muestra un ejemplo de esta tarea.

Tabla 4. *Ejemplo de segmentación*

Nombre y descripción del evento	Evento segmentado
Sistemas multiagente de robots móviles. En este seminario se presentan los resultados en el área de coordinación de movimiento para sistemas multiagente.	sistemas, multiagente, de, robots, moviles, . , en, este, seminario, se, presentan, los, resultados, en, el, área, de, coordinación, de, movimiento, para, sistemas, multiagente, . ,

Todas las cadenas son palabras excepto el punto. Los caracteres especiales también son segmentados.

b) **Limpieza.** Una vez que se ha obtenido el conjunto de segmentos (*tokens*) que conforman a cada evento, se descartan aquellos datos que no aportan información relevante al proceso de clasificación. Por lo tanto, esta tarea se conforma de dos sub-tareas:

    1. **Eliminación de números, signos de puntuación y caracteres especiales.** En esta sub-tarea se identifican datos numéricos, signos de puntuación y cualquier otro carácter especial tanto en el nombre como la descripción del evento. En la Tabla 5, se muestra el resultado de eliminar números, signos de puntuación y caracteres especiales del ejemplo anterior.

Tabla 5. *Eliminación de números, signos de puntuación y caracteres especiales*

Segmentos del evento	Eliminación de números, signos de puntuación y caracteres especiales
sistemas, multiagente, de, robots, móviles, . , en, este, seminario, se, presentan, los, resultados, en, el, área, de, coordinación, de, movimiento, para, sistemas, multiagente, . ,	sistemas, multiagente, de, robots, móviles, en, este, seminario, se, presentan, los, resultados, en, el, área, de, coordinación, de, movimiento, para, sistemas, multiagente.

En este ejemplo se identificó al "punto". Este signo de puntuación es eliminado.

2. **Eliminación de palabras vacías (*stop words*).** Esta subtarea consiste en eliminar aquellas palabras que carecen de un significado por sí solas, denominadas palabras vacías (*stop words*), algunos ejemplos son: artículos, preposiciones y conjunciones como puede observarse en la Tabla 6.

Tabla 6. *Ejemplo de eliminación de palabras vacías*

Segmentos del evento	Eliminación de palabras vacías (stop words)
sistemas, multiagente, de, robots, móviles, en, este, seminario, se, presentan, los, resultados, en, el, área, de, coordinación, de, movimiento, para, sistemas, multiagente.	sistemas, multiagente, robots, móviles, seminario, presentan, resultados, área, coordinación, movimiento, para, sistemas, multiagente, todos, resultados, validados, través, simulaciones

Las palabras vacías o stop words identificadas son: 'de', 'en', 'este', 'se' 'los', 'en', 'el', 'de', 'de'. Todas estas palabras son eliminadas.

c) **Lematización**. En esta tarea se eliminan las partes no esenciales de una palabra para obtener su forma base.

La lematización empleada en este trabajo de investigación, requirió de implementar un módulo de lematización para idioma español con el cual fuera posible identificar sin ambigüedades al conjunto de características lexicográficas que componen un evento (verbos, adjetivos y sustantivos). Para poder identificar las características lexicográficas se realizó un análisis morfológico de las palabras resultantes de la etapa anterior y se tomaron en cuenta las variaciones que una palabra pueda presentar. A la variación que sufre una palabra dependiendo de su género, número o tamaño se le conoce como flexión, en español forman flexión nominal los adjetivos, sustantivos y pronombres con los morfemas flexivos de género y número (masculino, femenino y singular o plural respectivamente), en cuanto a los verbos, lo hacen con la conjugación.

Considerando que el módulo *pattern.es* para español de CLiPS (*Computational Linguistics & Psycholinguistics Research Center*) [36] contiene un etiquetador gramatical y distintas herramientas para la conjugación de verbos, la singularización y pluralización de adjetivos y sustantivos. El módulo lematizador desarrollado en este trabajo se adaptó

a partir de *pattern* para identificar la categoría gramatical de cada palabra.

A continuación, se explica brevemente la implementación del funcionamiento del módulo lematizador.

Módulo lematizador

El módulo lematizador desarrollado en este trabajo de investigación emplea el etiquetador gramatical de CLiPS. Estas etiquetas son: "VB" para identificar a los verbos, "VBG" para identificar a los verbos conjugados; "JJ" los adjetivos, "NN" para sustantivos y NNS para sustantivos en plural.

Verbos

Los verbos son aquellas palabras que indican una acción y pueden presentarse en sus distintos tiempos verbales. Para identificar a un verbo y sus distintos tiempos verbales, se utilizaron las etiquetas VB y VBG del etiquetador gramatical de CLiPS para diferenciar a un verbo en infinitivo y a un verbo conjugado. Como primer paso se identificaron a los verbos en infinitivo y posteriormente se encontraron a los verbos conjugados que mediante las herramientas de conjugación de verbos que ofrece *Pattern*, se transformaron a su forma base, es decir, se trasladaron al infinitivo (los verbos en infinitivo tienen las terminaciones *ar*, *er*, *ir*). En la Tabla 7, se ilustran algunos ejemplos de los verbos identificados en este trabajo.

Tabla 7. *Ejemplo de lematizar de verbos*

Palabra original (verbos)	**Lema**
construyendo	construir
disminuyó	disminuir
presentará	presentar

Adjetivos

Los adjetivos son aquellas palabras que indican cualidades o propiedades de un sustantivo. La identificación de adjetivos se realizó a través de sus morfemas flexivos de género y número (*–o -a,-os, -as, -as o –es*). Según sea el género de los adjetivos identificados, se obtuvo su forma base y posteriormente se transforman al singular. Para identificar los adjetivos, se utilizó la etiqueta "JJ" del etiquetador gramatical de CLiPS. La Tabla 8, presenta la lematización de algunos adjetivos en este trabajo de investigación.

Tabla 8. *Ejemplo de lematizar de adjetivos*

Palabra original (adjetivos)	**Lema**
matemáticos	matemático
universitaria	universitario
básicas	básico
presenciales	presencial

Sustantivos

Los sustantivos son aquellas palabras que identifican a una persona, una entidad, un lugar o un concepto. Para identificar los sustantivos presentes en un evento, se utilizó la etiqueta "NN" del etiquetador gramatical de CLiPS la cual permite identificar los sustantivos en singular, sin embargo, también es posible que se encuentren en plural por lo que, en este trabajo de investigación, los sustantivos se identifican por sus morfemas de número (-s, -es) en plural para garantizar que sustantivos en plural como en singular fueran tomados en cuenta. La identificación de sustantivos en plural, se realizó con la etiqueta "NNS" del etiquetador gramatical de CLiPS, posteriormente se transformaron al singular.

Dado que muchos sustantivos son invariables, es decir, no cambian de género, son masculinos o femeninos, los sustantivos derivados de verbos se descartan de esta categoría gramatical y son asignados a la categoría gramatical "verbos". La Tabla 9, muestra la lematización de algunos sustantivos identificados en este trabajo de investigación.

Tabla 9. *Ejemplo de lematizar sustantivos*

Palabra original (sustantivos)	**Lema**
segunda	segundo
conferencias	conferencia
condiciones	condición

Para el desarrollo del módulo de lematización de este trabajo se empleó el lenguaje de programación denominado *Python* v2.7, además de las herramientas de *NLTK* de *Scikit-learn* [37].

d) **Lista de características.** Una vez concluido el proceso de lematización, basados en el modelo BoW (del inglés, *Bag of Words*), se obtiene una lista de las características lexicográficas de los eventos (verbos, adjetivos y sustantivos) y se eliminan aquellos términos repetidos A continuación, en la Tabla 10 se muestra un resumen de las características lexicográficas obtenidas en este trabajo de investigación.

Tabla 10. *Características lexicográficas*

Tipo de característica	**Característica lexicográfica**	**Total**	**Descripción**
Características de texto	Verbo	306	Valor numérico que indica el total de verbos en los eventos
	Adjetivo	294	Valor numérico que indica el total de adjetivos en los eventos
	Sustantivo	1233	Valor numérico que indica el total de sustantivos en los eventos

5.2.2 Extracción de características numéricas y nominales

El valor nominal es el título que se les asigna a las características de los eventos relacionadas con el horario y el espacio en el que se llevó a cabo dicho evento. Además, se tienen

características numéricas como son la variación que presentan los eventos ambientales y el total de participantes en un evento. De acuerdo con la entidad que representan las características numéricas y nominales, se dividen en: características de agente, características de tiempo, características de espacio y características ambientales.

El procesamiento de las bitácoras de eventos para la extracción de características numéricas y nominales, empleado en este trabajo, dispone de un análisis distinto al proceso para la extracción de características lexicográficas descrito anteriormente ya que estas características son de naturaleza distinta. A continuación, se explica brevemente el proceso destinado para obtener estas características.

a) **Características de agente**. Las características de agente son aquellas características que indican la cantidad de participantes en un evento.

   Las características de agente son el número de alumnos, número de profesores y número de visitantes involucrados en el evento.

   Para obtenerlas simplemente se toman de las bitácoras de eventos que mantienen un registro del número tanto de alumnos, como de profesores y visitantes.

   El total de participantes se obtiene mediante la suma del total de alumnos más el número total de profesores y el total de visitantes.

b) **Características de tiempo**. Las características de tiempo son aquellas que identifican el horario en el que se desenvuelve un evento. Se tienen dos tipos, nominales y numéricas.

   Estas características nominales definen el horario inicial y el horario final del evento, las cuales se obtienen a partir de rangos de tiempo en los que se registra un evento.

   Estos rangos se definen con base en cuatro turnos: matutino, vespertino, intermedio y nocturno, como se muestra en la Tabla 11.

Tabla 11. *Características de tiempo*

Turno	Rango de tiempo
Turno matutino	08:00:00 am – 12:00:00 pm
Turno vespertino	12:01:00 pm – 16:00:00 pm
Turno intermedio	16:01:00 pm – 20:00:00 pm
Turno nocturno	20:01:00 pm – 22:00:00 pm

Como resultado de aplicar estos rangos, se obtienen los valores nominales para asignar el título a la característica.

Por otro lado, características numéricas en cuanto al tiempo son la duración de

un evento expresado en minutos. Estas características se obtienen de la diferencia que existe entre la hora final del evento menos la hora inicial del evento.

c) **Características de espacio**. Las características de espacio representan el lugar en el cual se registra un evento. Se tienen dos características nominales: el tipo de espacio y el tipo de lugar.

El tipo de lugar se refiere al espacio físico en el que se desarrolla un evento, estos son salones (aulas de clase), oficinas de profesores, laboratorios, auditorios, plazas y jardines.

A estos lugares, se les asigna un valor nominal para asignar el título a la característica como se muestra en la Tabla 12:

Tabla 12. *Características de espacio*

Lugar	Valor nominal
salón	1
oficina de un profesor	2
laboratorio	3
auditorio	4
plaza	5
jardín	6

Conocer el lugar en el cual se desarrolla un evento, contribuye a conocer el tipo de espacio. El tipo de espacio, puede ser un espacio abierto (en exteriores o al aire libre) o cerrado (en interiores).

Por lo tanto, para conocer el tipo de espacio, en el cual ocurre un evento, se identifica el tipo de lugar mediante reglas:

1. Si el tipo de lugar es un salón, oficina de un profesor, laboratorio o auditorio el tipo de espacio es cerrado.
2. Si el tipo de lugar es una plaza o un jardín, el tipo de espacio es abierto.

Como consecuencia de aplicar estas reglas, se obtiene el valor nominal para asignar el título a la característica. Como se muestra en la Tabla 13.

Tabla 13. *Tipo de espacio*

Tipo de espacio	Valor nominal
abierto	0
cerrado	1

d) **Características ambientales.** Las características ambientales están determinadas por la diferencia que existe entre dos valores registrados por los sensores ambientales en un tiempo t_1 y t_2.

Estos valores son de tipo numérico que registran cambios en la iluminación, la humedad, la temperatura o la presencia de una persona en un área determinada.

Clase. Los datos fueron etiquetados manualmente de acuerdo al tipo de evento que sucede en el ambiente académico.

Se tienen cuatro clases principales: evento de difusión, cursos académicos, asesoría y evento ambiental. Las etiquetas de clase se muestran en la Tabla 14.

Tabla 14. *Etiquetas de clase*

Clase	Valor nominal
Evento de difusión	0
Cursos académicos	1
Asesorías académicas	2
Evento ambiental	3

En consecuencia, en la Tabla 15, se muestran las características nominales y numéricas de un evento, su descripción y sus posibles valores.

Tabla 15. *Tabla de características nominales*

Tipo de característica	Característica	Descripción	Posibles valores
Características de Agente	Número de estudiantes	Total de alumnos participantes en un evento	Valor numérico que indica la cantidad de estudiantes en un evento
	Número de profesores	Total de profesores participantes en un evento	Valor numérico que indica la cantidad de profesores en un evento
	Número de visitantes	Total de participantes externos a un espacio académico	Valor numérico que indica la cantidad de visitantes en un evento
	Total de participantes	Total de participantes en un evento	Valor numérico que indica la cantidad de estudiantes, profesores y visitantes en un evento
Características de tiempo	Horario inicial del evento	Tiempo en que inicia un evento	Valor nominal = {turno matutino = 1, turno vespertino = 2, turno intermedio = 3, turno nocturno = 4 }
	Horario final del evento	Tiempo en que termina un evento	Valor nominal = {turno matutino = 1, turno vespertino = 2, turno intermedio = 3, turno nocturno = 4}
	Tiempo del evento	Tiempo que tarda en efectuarse un evento	Valor numérico que indica la duración en minutos de un evento
Características de espacio	Tipo de espacio	Interior o exterior (al aire libre) en el que sucede un evento	Valor nominal que indica si un espacio es abierto o cerrado = {abierto = 0, cerrado = 1}
	Tipo de lugar	Espacio físico en el que se efectúa un evento	Valor nominal = {salón = 1, oficina de un profesor = 2, laboratorio = 3, auditorio = 4, plaza = 5, jardín = 6 }
Características ambientales	Variación ambiental	Cambio en luminosidad, temperatura o humedad	Valor numérico que indica la variación en eventos de luminosidad, temperatura y humedad

5.3 Transformación de los eventos

Una vez que se ha extraído la lista de características, los eventos se representan como un vector de características por pares del tipo EVENT: FEATURE (evento: característica). Las

características para cada EVENT: FEATURE se representan utilizando el modelo espacio vectorial. Las filas representan los eventos y las columnas f_i representan las características de cada evento.

Las características f_i de cada evento se representan por el conjunto de todas las características como se muestra en (10):

$$F = \{f_1, f_2, f_3, \dots, f_n\} \tag{10}$$

Los eventos son el conjunto de todos los eventos (11):

$$E = \{e_1, e_2, e_3, \dots, e_n\} \tag{11}$$

El modelo espacio vectorial en el ambiente académico se define como un modelo matemático y algebraico para transformar y representar a los eventos académicos como vectores numéricos. Las dimensiones para cada vector será el número total de características distintas de todos los eventos. En este trabajo, el modelo espacio vectorial está compuesto por las características nominales, características numéricas y características textuales (características lexicográficas).

En un solo evento puede aparecer más de una vez una característica determinada, esta característica es importante porque permite distinguir a este evento del resto, además, alguna característica puede considerarse con mayor grado de importancia que otra, por lo que es importante normalizar los vectores. Existen distintos métodos para calcular el grado de importancia de cada término en el vector que representa a cada evento.

La relación existente entre un evento y sus características está determinada por la función de peso (12):

$$W(e_i, f_i) \tag{12}$$

En este trabajo se utilizan tres de los esquemas de pesado más utilizados en la literatura:

a) **Frecuencia del Término (TF)** [38]. Se refiere a la frecuencia absoluta del término dentro del evento. Se representa con la fórmula (13):

$$TF_{ij} = f_{ij} \tag{13}$$

Donde f_{ij} es la frecuencia de la característica *i* en el evento *j*.

b) **Frecuencia del Término - Frecuencia Inversa del Término (TF - IDF)** es la unión del esquema de pesado TF (Frecuencia del Término) [38] con IDF (Frecuencia Inversa del Término) [39]. La fórmula empleada para el cálculo de la medida de la Frecuencia Inversa del Término se muestra en (14):

$$IDF(i) = log\left(\frac{N}{df}\right) \tag{14}$$

Donde *N* es el número total de eventos y *df* el número de eventos en donde aparece el término *i*.

En TF-IDF [40] cada vector está conformado por los pesos que representan la relevancia que tiene una característica en un evento. De acuerdo con [40] aquellas características que ocurren con menor frecuencia en una colección de eventos se consideran más importantes que aquellas que ocurren con mayor frecuencia. Su fórmula se muestra en la ecuación (15):

$$tf - idf_{ij} = f_{ij} * log\left(\frac{N}{df}\right) \quad (15)$$

Donde f_{ij} es la frecuencia de la característica *i* en el evento *j*, *N* es el número de eventos y *df* es el número de descripciones en donde aparece el término *i*.

c) **Pesado booleano.** Consiste en asignar un 1 a a_{ij} si existe la característica *i* en un evento *j* o un 0 si no existe, como se muestra a continuación en la ecuación (16):

$$a_{ij} = \begin{cases} 1 & si\ f_{ij} > 0 \\ 0 & en\ otro\ caso \end{cases} \quad (16)$$

Donde f_{ij} es la frecuencia de la característica *i* en el evento *j*.

5.4 Selección e implementación de los algoritmos de aprendizaje automático

En esta etapa se describe la implementación de los algoritmos presentados en este trabajo de investigación, además, se presentan las razones por las cuales se eligieron para la clasificación de eventos académicos.

La elección de un algoritmo, depende en gran manera del tamaño y el tipo de datos con los que se trabaja (datos estructurados o no estructurados) y las características del problema que se desea resolver.

Este trabajo de investigación pretende resolver un problema de clasificación, específicamente, la clasificación de eventos académicos. Se tienen eventos de cuatro categorías distintas.

Dado que se desea resolver un problema de clasificación se eligen cuatro (Naïve Bayes, k-vecinos más cercanos, C4.5 y SVM) de los algoritmos de clasificación más utilizados en la literatura especializada.

La implementación de estos algoritmos, se hizo en el lenguaje de programación *Python v2.7* con los algoritmos de clasificación incluidos en *Scikit-learn,* la biblioteca de aprendizaje automático de Software libre para *Python.*

Los parámetros de configuración que se emplearon para los algoritmos son los valores por defecto.

Los algoritmos de clasificación reciben como parámetros de entrada a una matriz de eventos.

Esta matriz de eventos es el conjunto de eventos representados en el modelo espacio vectorial con los tres diferentes ponderados (TF, TF-IDF y Booleano).

Los eventos se dividen en dos subconjuntos de eventos seleccionados aleatoriamente. El primero, con un 70% para el entrenamiento de los clasificadores y el segundo con un 30% para pruebas. Como salida se obtienen los eventos clasificados por los algoritmos.

5.5 Evaluación y comparación de los resultados obtenidos

En esta etapa se evalúan y comparan los resultados obtenidos por los modelos y se determina cuál presenta mejores resultados para la clasificación de eventos.

5.5.1 Medidas de evaluación de los algoritmos de clasificación

La precisión, el recuerdo o cobertura y la medida F_1 son las medidas más comunes en la evaluación de los algoritmos de clasificación [41].

Para evaluar los algoritmos que se utilizaron en este trabajo de investigación se aplican tres medidas de evaluación: precisión, cobertura y medida F_1.

La precisión indica qué tan exacta fue la clasificación de los eventos; se define como la relación entre los eventos clasificados correctamente sobre el total de eventos. La fórmula de la precisión se muestra en la ecuación (17):

$$precisión = \frac{Número\ de\ eventos\ clasificados\ correctamente}{Total\ de\ eventos} \tag{17}$$

La cobertura se define como la relación entre los eventos clasificados correctamente sobre el total de eventos que pertenecen a una clase *i*, es decir, da a conocer si los eventos que pertenecen a una clase *i*, se clasificaron dentro de esa clase. El cálculo de la cobertura se presenta en la ecuación (18):

$$cobertura = \frac{Número\ de\ eventos\ clasificados\ en\ la\ clase\ i}{Número\ de\ eventos\ que\ pertenecen\ a\ la\ clase\ i} \tag{18}$$

La medida F_1 es una combinación entre las medidas de precisión y cobertura que representa el porcentaje de las predicciones que son correctas. La medida F_1 se estima como se muestra en la ecuación (19):

$$F_1 = \frac{2\ *\ precisión * cobertura}{precisión\ + cobertura} \tag{19}$$

VI. Experimentación y resultados

El objetivo de este capítulo es presentar las pruebas y resultados obtenidos de los experimentos realizados con los cuatro algoritmos de clasificación y sus combinaciones con las diferentes características de los eventos.

6.1 Conjunto de datos

Durante la etapa de pruebas se utilizó un total de 362 eventos académicos. Los eventos son de cuatro tipos distintos (cursos académicos, evento de difusión eventos ambientales y asesoría) entre los cuales hay subtipos.

En la Tabla 16 se presentan los cuatro tipos de eventos, así como su descripción y el subtipo de evento para las clases: cursos académicos, eventos de difusión y eventos ambientales a excepción de las asesorías que no tienen subtipo.

Tabla 16. *Descripción de eventos*

Tipo	Descripción de evento	Subtipo de evento	Total de eventos
Asesoría	Consulta que brinda un profesor a un estudiante para resolver cuestiones sobre temas que domina		20
Cursos académicos	Su objetivo es la formación académica y profesional de estudiantes y profesores	Cursos de licenciatura, posgrado y actualización al personal académico	60
Evento de difusión	Evento cuyo objetivo es difundir temas relacionados con la investigación y la cultura	Congreso, panel de discusión, taller, seminario y presentación	122
Ambiental	Evento en el cual se encuentran involucradas las variables del ambiente	Presencia, luminosidad, temperatura y humedad	160

6.2 Diseño de experimentos

Para comparar los algoritmos aplicados en la clasificación de eventos académicos se utilizaron 24 experimentos. Estos experimentos consisten en realizar ocho combinaciones de características textuales, nominales y numéricas con tres esquemas de pesado: TF, TF-IDF y Booleano.

Las pruebas se realizaron en cuatro pasos:

- En primer lugar, se emplean características textuales de manera individual, es decir, se aplican los algoritmos al conjunto de verbos, adjetivos y sustantivos de manera independiente.
- En un siguiente paso se aplican los algoritmos de clasificación sobre el conjunto de todas las características textuales, es decir, se emplean verbos, adjetivos y sustantivos.

- ❖ Finalmente, se aplican los algoritmos de clasificación sobre el conjunto de características nominales y numéricas combinadas con verbos, con adjetivos y con sustantivos de manera individual.
- ❖ En un siguiente paso, se aplican los algoritmos sobre las características nominales, numéricas y textuales combinadas con todas las características texturales, es decir, se realizan pruebas sobre el conjunto conformado por las características nominales. verbos, adjetivos y sustantivos.

El diseño de los experimentos para este trabajo de investigación se muestra en la Tabla 17.

Tabla 17. *Diseño de experimentos*

Prueba	**NB, KNN, SVM y C4.5 (pesado TF)**	**NB, KNN, SVM y C4.5 (pesado TF-IDF)**	**NB, KNN, SVM y C4.5 (pesado Booleano)**
1	Verbos	Verbos	Verbos
2	Adjetivos	Adjetivos	Adjetivos
3	Sustantivos	Sustantivos	Sustantivos
4	verbos, adjetivos y sustantivos	verbos, adjetivos y sustantivos	verbos, adjetivos y sustantivos
5	nominales, numéricas y verbos	nominales, numéricas y verbos	nominales, numéricas y verbos
6	nominales, numéricas y adjetivos	nominales, numéricas y adjetivos	nominales, numéricas y adjetivos
7	nominales, numéricas y sustantivos	nominales, numéricas y sustantivos	nominales, numéricas y sustantivos
8	nominales, numéricas, verbos, adjetivos y sustantivos	nominales, numéricas, verbos, adjetivos y sustantivos	nominales, numéricas, verbos, adjetivos y sustantivos

6.2.1 Pesado frecuencia de término (TF)

Se realizaron ocho pruebas con el esquema de pesado frecuencia de término y se comparó el rendimiento de los algoritmos de clasificación Naïve Bayes, k-vecinos más cercanos, C4.5 y SVM. A continuación, se muestran los experimentos realizados y el análisis de cada uno en el cual se exponen a los dos mejores algoritmos y al algoritmo que obtiene los resultados más bajos durante las pruebas.

Verbos

Los resultados obtenidos de la experimentación sobre el conjunto de verbos para las características individuales se presentan en la Tabla 18.

Tabla 18. *Resultados de la experimentación con verbos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	62.16	69.7	65.71
	Cursos	46.67	31.82	37.84
	Asesoría	30.77	66.67	42.11
	Ambiental	90.91	83.33	86.96
	Promedio	**57.63**	**62.88**	**58.15**
KNN	Difusión	60	9.09	15.79
	Cursos	90	40.91	56.25
	Asesoría	100	16.67	28.57
	Ambiental	50.54	97.92	66.67
	Promedio	**75.13**	**41.15**	**41.82**
C4.5	Difusión	57.58	57.58	57.58
	Cursos	85	77.27	80.95
	Asesoría	100	16.67	28.57
	Ambiental	67.27	77.08	71.84
	Promedio	**77.46**	**57.15**	**59.74**
SVM	Difusión	50	87.88	63.74
	Cursos	100	22.73	37.04
	Asesoría	0	0	0
	Ambiental	80.43	77.08	78.72
	Promedio	**57.61**	**46.92**	**44.87**

El análisis de resultados obtenido para el conjunto de verbos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 77.46% de precisión, cobertura del 57.15% y medida F1 con un 59.74%. En contraste, SVM muestra un 57.61% de precisión, 46.92% de cobertura y 44.87% para la medida F_1, siendo estos los resultados más bajos. KNN también muestra un desempeño similar con un 75.13% de precisión, 41.15% de cobertura y un 41.82% para la medida F_1.

Aunque la medida F_1 obtenida por KNN es menor que la medida F_1 obtenida por SVM, se puede observar que la precisión de KNN es mayor, ya que SVM no logró clasificar a ningún evento del tipo asesoría.

Esto, debido a que el número de eventos en esta categoría es inferior al número de eventos presentes en las demás categorías. Por lo tanto, El número de elementos necesarios para que un algoritmo clasifique correctamente a un evento está relacionado con el número de instancias en una cada clase.

Adjetivos

Los resultados obtenidos de la experimentación sobre el conjunto de adjetivos para las características individuales se presentan en la Tabla 19.

Tabla 19. *Resultados de la experimentación con adjetivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	65.38	51.52	57.63
	Cursos	50	40.91	45
	Asesoría	2.5	16.67	4.35
	Ambiental	64	33.33	43.84
	Promedio	**45.47**	**35.61**	**37.7**
KNN	Difusión	87.5	21.21	34.15
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	46.94	95.83	63.01
	Promedio	**33.61**	**29.26**	**24.29**
C4.5	Difusión	70	42.42	52.83
	Cursos	57.14	18.18	27.59
	Asesoría	0	0	0
	Ambiental	56.41	91.67	69.84
	Promedio	**45.89**	**38.07**	**37.56**
SVM	Difusión	54.9	84.85	66.67
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	72.41	87.5	79.25
	Promedio	**31.83**	**43.09**	**36.48**

El análisis de resultados obtenido para el conjunto de adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 y NB obtienen los mejores resultados de clasificación con un 45.89% y 45.47% de precisión, cobertura de 38.07% y 35.61%. La medida F_1 muestra resultados del 37.56% y 37.7%. Si bien NB tiene un valor más alto en la medida F_1, se considera mejor clasificador a C4.5 puesto que la precisión de C4.5 es mayor que la obtenida por NB.

Por otro lado, SVM muestra un 31.83% de precisión, 43.09% de cobertura y un 36.48% para la medida F_1. Se puede notar que los bajos resultados obtenidos por SVM nuevamente están relacionados con el número de eventos presentes en las categorías: Cursos con 60 eventos y Asesoría con 20 eventos.

Por otro lado, el rendimiento de los algoritmos disminuyó considerablemente en comparación con los resultados obtenidos por los verbos, pues, se tienen menos adjetivos (294 adjetivos) que verbos.

Sustantivos

Los resultados obtenidos de la experimentación sobre el conjunto de sustantivos para las características individuales, se presentan en la Tabla 20.

Tabla 20. *Resultados de la experimentación con sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	65.96	93.94	77.5
	Cursos	90.91	45.45	60.61
	Asesoría	100	83.33	90.91
	Ambiental	95.65	91.67	93.62
	Promedio	**88.13**	**78.6**	**80.66**
KNN	Difusión	100	24.24	39.02
	Cursos	44.44	18.18	25.81
	Asesoría	100	50	66.67
	Ambiental	53.93	100	70.07
	Promedio	**74.59**	**48.11**	**50.39**
C4.5	Difusión	80	96.97	87.67
	Cursos	100	72.73	84.21
	Asesoría	100	100	100
	Ambiental	95.74	93.75	94.74
	Promedio	**93.94**	**90.86**	**91.65**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para el conjunto de sustantivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 93.94% de precisión, cobertura del 90.86% y medida F_1 con un 91.65%; NB también ofrece buenos resultados de clasificación con 88.13% de precisión, 78.60% para la cobertura y un 80.66% para la medida F_1.

SVM muestra los peores resultados de clasificación con resultados menores al 50% para la precisión, cobertura y medida F_1.

El total de sustantivos supera al número de verbos y al número de adjetivos mejorando así el rendimiento de los clasificadores. Como se puede observar, todos los clasificadores mejoraron sus resultados comparados contra los obtenidos durante las pruebas con adjetivos, con excepción de SVM que disminuyó.

Verbos, adjetivos y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de todas las características lexicográficas se muestran en la Tabla 21.

Tabla 21. *Resultados de la experimentación con verbos, adjetivos y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	62.75	96.97	76.19
	Cursos	90	40.91	56.25
	Asesoría	0	0	0
	Ambiental	93.75	93.75	93.75
	Promedio	**61.62**	**57.91**	**56.55**
KNN	Difusión	100	3.03	5.88
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.86	100	61.94
	Promedio	**36.21**	**25.76**	**16.95**
C4.5	Difusión	78.95	90.91	84.51
	Cursos	88.89	72.73	80
	Asesoría	100	100	100
	Ambiental	95.74	93.75	94.74
	Promedio	**90.90**	**89.35**	**89.81**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para la combinación de características lexicográficas (verbos, adjetivos y sustantivos) mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 90.90% de precisión, cobertura 89.35% y medida F_1 de 89.81%. Mientras que SVM muestra resultados por debajo del 50% con un 11.01% de precisión, 25% de cobertura y 15.29% para la medida F_1.

Se puede observar que SVM se encuentra sujeto al número de eventos en las clases puesto que la única clase en la cual arroja resultados es en la clase ambiental que cuenta con 160 ejemplos.

A pesar de que se tienen más elementos en el entrenamiento de los algoritmos, en esta prueba se observa que C4.5 tuvo un ligero cambio en cuanto a la precisión con la que clasifica a los eventos. En la clasificación de sustantivos, C4.5 tiene un 100% de precisión para eventos de tipo asesoría y cursos; en la combinación de verbos, adjetivos y sustantivos: la precisión para

cursos tiene un 88.89%, una cobertura del 72.73% y una medida F_1 de 89.81%. Con esto se observa que aunque la precisión de C4.5 es menor en la clasificación de cursos académicos sigue siendo el que mejores resultados ofrece en la medida F_1.

Características numéricas, nominales y verbos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y verbos se muestran en la Tabla 22.

Tabla 22. *Resultados de la experimentación con características numéricas, nominales y verbos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	67.5	81.82	73.97
	Cursos	62.5	45.45	52.63
	Asesoría	100	100	100
	Ambiental	100	97.92	98.95
	Promedio	**82.50**	**81.30**	**81.39**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	91.67	100	95.65
	Cursos	100	86.36	92.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**97.92**	**96.59**	**97.08**
SVM	Difusión	44	100	61.11
	Cursos	0	0	0
	Asesoría	100	33.33	50
	Ambiental	100	66.67	80
	Promedio	**61**	**50**	**47.78**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y verbos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 97.92% de precisión, cobertura del 96.59% y medida F_1 con un 97.08% a diferencia de la evaluación realizada al conjunto de verbos, KNN también ofrece buenos resultados de clasificación con un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1. SVM muestra un menor rendimiento con un 61% de precisión, 50% de cobertura y 47.78% para la medida F_1.

Características numéricas, nominales y adjetivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y adjetivos se muestran en la Tabla 23.

Tabla 23. *Resultados de la experimentación con características numéricas, nominales y adjetivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	72.22	78.79	75.36
	Cursos	62.5	68.18	65.22
	Asesoría	33.33	16.67	22.22
	Ambiental	100	95.83	97.87
	Promedio	**67.01**	**64.87**	**65.17**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	81.58	93.94	87.32
	Cursos	88.24	68.18	76.92
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**92.45**	**90.53**	**91.06**
SVM	Difusión	45.21	100	62.26
	Cursos	0	0	0
	Asesoría	100	16.67	28.57
	Ambiental	100	72.92	84.34
	Promedio	**61.3**	**47.4**	**43.79**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 92.45% de precisión, cobertura del 90.53% y medida F_1 de 91.06%, en este caso KNN también ofrece buenos resultados de clasificación con un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1. SVM muestra un menor rendimiento con un 61.3% de precisión, 47.4% de cobertura y 43.79% para la medida F_1.

Características numéricas, nominales y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y sustantivos se muestran en la Tabla 24.

Tabla 24. *Resultados de la experimentación con características numéricas, nominales y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	68.09	96.97	80
	Cursos	83.33	45.45	58.82
	Asesoría	100	83.33	90.91
	Ambiental	100	93.75	96.77
	Promedio	**87.85**	**79.88**	**81.63**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	81.08	90.91	85.71
	Cursos	83.33	68.18	75
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**91.1**	**89.77**	**90.18**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para el conjunto de características nominales más adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que KNN muestra mayor rendimiento con un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1 y C4.5 ofrece un buen desempeño, sin embargo, en un grado menor que KNN con 91.1% de precisión, 89.77% de cobertura y 90.18% para la medida F_1 a diferencia del mostrado en la evaluación individual para los sustantivos. SVM muestra rendimiento muy bajo con un 11.01% de precisión, 25% de cobertura y 15.29% para la medida F_1.

Características numéricas, nominales, verbos, adjetivos y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales, verbos, adjetivos y sustantivos se muestran en la Tabla 25.

Tabla 25. *Resultados de la experimentación con características numéricas, nominales, verbos, adjetivos y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	63.46	100	77.65
	Cursos	90	40.91	56.25
	Asesoría	100	16.67	28.57
	Ambiental	100	95.83	97.87
	Promedio	**88.37**	**63.35**	**65.09**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	88.57	93.94	91.18
	Cursos	90	81.82	85.71
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**94.64**	**93.94**	**94.22**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para la combinación de características numéricas, nominales y características lexicográficas mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 94.64% de precisión, cobertura 93.94% y medida F_1 de 94.22%. KNN también muestra un buen desempeño con un 93.33% de precisión, 90.15% de cobertura y 90.84% para la medida F_1, mientras que SVM muestra un bajo grado de confianza con 11.01% de precisión, 25% de cobertura y un 15.29% para F_1.

Las pruebas realizadas a la combinación de características lexicográficas, numéricas y nominales indican que C4.5 presenta un rendimiento superior en la mayoría de los casos. Únicamente para el caso de combinar sustantivos y características nominales KNN muestra superioridad. Por último, la evaluación de combinar todas las características lexicográficas con características numéricas y nominales demuestra que C4.5 logra un mayor rendimiento. Pues C4.5 dispone de métodos capaces de trabajar con datos nominales, numéricos y

textuales.

6.2.2 Pesado frecuencia del término - frecuencia inversa del término (TF-IDF)

Se realizaron ocho pruebas con el esquema de pesado frecuencia de término-frecuencia inversa del término y se comparó el rendimiento de los algoritmos de clasificación Naïve Bayes, k-vecinos más cercanos, C4.5 y SVM. A continuación se muestran los experimentos realizados y el análisis de cada uno en el cual se exponen a los dos mejores algoritmos y al algoritmo que obtiene los resultados más bajos durante las pruebas.

Verbos

Los resultados obtenidos de la experimentación sobre el conjunto de verbos para las características individuales se presentan en la Tabla 26.

Tabla 26. *Resultados de la experimentación con verbos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	62.16	69.7	65.71
	Cursos	46.67	31.82	37.84
	Asesoría	30.77	66.67	42.11
	Ambiental	90.91	83.33	86.96
	Promedio	**57.63**	**62.88**	**58.15**
KNN	Difusión	75	9.09	16.22
	Cursos	83.33	22.73	35.71
	Asesoría	100	16.67	28.57
	Ambiental	47.96	97.92	64.38
	Promedio	**76.57**	**36.6**	**36.22**
C4.5	Difusión	57.58	57.58	57.58
	Cursos	85	77.27	80.95
	Asesoría	100	16.67	28.57
	Ambiental	67.27	77.08	71.84
	Promedio	**77.46**	**57.15**	**59.74**
SVM	Difusión	44.78	90.91	60
	Cursos	100	4.55	8.7
	Asesoría	0	0	0
	Ambiental	85.37	72.92	78.65
	Promedio	**57.54**	**42.09**	**36.84**

El análisis de resultados obtenido para el conjunto de verbos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 77.46% de precisión, cobertura del 57.15% y medida F_1 con un 59.74%. NB presenta un rendimiento similar al de C4.5 con un 57.63% de precisión, un 62.88% de cobertura y 58.15% para la medida F_1. SVM por su parte, tiene una precisión

aún menor del 57.54%, una cobertura de 42.09% y una medida F_1de 36.84%. A pesar de que los resultados mostrados para F_1 son inferiores al 70%, C4.5 indica un mayor porcentaje en su rendimiento.

Adjetivos

Los resultados obtenidos de la experimentación sobre el conjunto de adjetivos para las características individuales, se presentan en la Tabla 27.

Tabla 27. *Resultados de la experimentación con adjetivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	62.96	51.52	56.67
	Cursos	50	40.91	45
	Asesoría	2.56	16.67	4.44
	Ambiental	64	33.33	43.84
	Promedio	**44.88**	**35.61**	**37.49**
KNN	Difusión	75	9.09	16.22
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	45.1	95.83	61.33
	Promedio	**30.02**	**26.23**	**19.39**
C4.5	Difusión	70	42.42	52.83
	Cursos	57.14	18.18	27.59
	Asesoría	0	0	0
	Ambiental	56.41	91.67	69.84
	Promedio	**45.89**	**38.07**	**37.56**
SVM	Difusión	50	93.94	65.26
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	84.44	79.17	81.72
	Promedio	**33.61**	**43.28**	**36.75**

El análisis de resultados obtenido para el conjunto de adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 y NB obtienen mejores resultados de clasificación con un 45.89% y 44.88% de precisión, cobertura de 38.07% y 35.61%. La medida F_1 muestra resultados del 37.56% y 37.49%. Los resultados obtenidos muestran un porcentaje menor al 70% deseable, sin embargo, el rendimiento de C4.5 sobresale del resto.

Sustantivos

Los resultados obtenidos de la experimentación sobre el conjunto de sustantivos para las características individuales, se presentan en la Tabla 28.

Tabla 28. *Resultados de la experimentación con sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	60.87	84.85	70.89
	Cursos	76.92	45.45	57.14
	Asesoría	50	16.67	25
	Ambiental	89.58	89.58	89.58
	Promedio	**69.34**	**59.14**	**60.65**
KNN	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.44	100	61.54
	Promedio	**11.11**	**25**	**15.38**
C4.5	Difusión	80	96.97	87.67
	Cursos	100	72.73	84.21
	Asesoría	100	100	100
	Ambiental	95.74	93.75	94.74
	Promedio	**93.94**	**90.86**	**91.65**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para el conjunto de sustantivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 93.94% de precisión, cobertura del 90.86% y medida F_1 con un 91.65%. SVM muestra un rendimiento deficiente con un 11.01% para la precisión, 25% para la cobertura y 15.29% para la medida F_1.

Verbos, adjetivos y sustantivos.

Los resultados obtenidos de la experimentación realizada a la combinación de todas las características lexicográficas se muestran en la Tabla 29.

Tabla 29. *Resultados de la experimentación con verbos, adjetivos y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	59.26	96.97	73.56
	Cursos	90	40.91	56.25
	Asesoría	0	0	0
	Ambiental	95.56	89.58	92.47
	Promedio	**61.20**	**56.87**	**55.57**
KNN	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**
C4.5	Difusión	78.95	90.91	84.51
	Cursos	88.89	72.73	80
	Asesoría	100	100	100
	Ambiental	95.74	93.75	94.74
	Promedio	**90.90**	**89.35**	**89.81**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para la combinación de características lexicográficas (verbos, adjetivos y sustantivos) mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 90.90% de precisión, cobertura 89.35% y medida F_1 de 89.81% mientras que SVM y KNN muestran resultados por debajo del 50% con un 11.01% de precisión, 25% de cobertura y 15.29% para la medida F_1.

Las pruebas individuales realizadas a cada conjunto de características lexicográficas indican que en todos los casos C4.5 presenta un mayor rendimiento. Y para el caso de combinar todas las características lexicográficas, C4.5 presenta un rendimiento superior en comparación con el resto de los algoritmos.

Características numéricas, nominales y verbos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y verbos se muestran en la Tabla 30.

Tabla 30. *Resultados de la experimentación con características numéricas, nominales y verbos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	65.79	75.76	70.42
	Cursos	55.56	45.45	50
	Asesoría	100	66.67	80
	Ambiental	95.92	97.92	96.91
	Promedio	**79.32**	**71.45**	**74.33**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	91.67	100	95.65
	Cursos	100	86.36	92.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**97.92**	**96.59**	**97.08**
SVM	Difusión	42.31	100	59.46
	Cursos	0	0	0
	Asesoría	100	16.67	28.57
	Ambiental	96.67	60.42	74.36
	Promedio	**59.74**	**44.27**	**40.60**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y verbos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 97.92% de precisión, cobertura del 96.59% y medida F_1 con un 97.08%. NB por otro lado, muestra un 73.32% de precisión, 71.45% de cobertura y 74.33% para la medida F_1. SVM muestra un 59.74% de precisión, 44.27% de cobertura y 40.60% para la medida F_1, por lo tanto, C4.5 es el mejor clasificador para la combinación de características numéricas, nominales y verbos.

Características numéricas, nominales y adjetivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y adjetivos se muestran en la Tabla 31.

Tabla 31. *Resultados de la experimentación con características numéricas, nominales y adjetivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	69.44	75.76	72.46
	Cursos	60.87	63.64	62.22
	Asesoría	33.33	16.67	22.22
	Ambiental	97.87	95.83	96.84
	Promedio	**65.38**	**62.97**	**63.44**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	81.58	93.94	87.32
	Cursos	88.24	68.18	76.92
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**92.45**	**90.53**	**91.06**
SVM	Difusión	38.82	100	55.93
	Cursos	0	0	0
	Asesoría	100	16.67	28.57
	Ambiental	100	47.92	64.79
	Promedio	**59.71**	**41.15**	**37.32**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 92.45% de precisión, cobertura del 90.53% y medida F_1 de 91.06%, en este caso KNN también ofrece buenos resultados de clasificación con un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1. SVM muestra el menor rendimiento con un 57.71% de precisión, 41.15% de cobertura y 37.32% para la medida F_1.

Características numéricas, nominales y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y sustantivos se muestran en la Tabla 32.

Tabla 32. *Resultados de la experimentación con las características numéricas, nominales y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	64.58	93.94	76.54
	Cursos	76.92	45.45	57.14
	Asesoría	66.67	33.33	44.44
	Ambiental	100	93.75	96.77
	Promedio	**77.04**	**66.62**	**68.73**
KNN	Difusión	80	96.97	87.67
	Cursos	92.86	59.09	72.22
	Asesoría	0	0	0
	Ambiental	87.27	100	93.2
	Promedio	**65.03**	**64.02**	**63.27**
C4.5	Difusión	81.08	90.91	85.71
	Cursos	83.33	68.18	75
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**91.10**	**89.77**	**90.18**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y sustantivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 ofrece un buen desempeño con un 91.1% de precisión, 89.77% de cobertura y 90.18% para la medida F_1. NB presenta un 77.04% de precisión, 66.62% de cobertura y un 68.73% para la medida F_1. SVM muestra el rendimiento más bajo con un 11.01% de precisión, 25% de cobertura y 15.29% para la medida F_1.

Características numéricas, nominales, verbos, adjetivos y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales, verbos, adjetivos y sustantivos se muestran en la Tabla 33.

Tabla 33. *Resultados de la experimentación con las características numéricas, nominales, verbos, adjetivos y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	62.26	100	76.74
	Cursos	90	40.91	56.25
	Asesoría	0	0	0
	Ambiental	97.83	93.75	95.74
	Promedio	**62.52**	**58.66**	**57.18**
KNN	Difusión	80	96.97	87.67
	Cursos	92.86	59.09	72.22
	Asesoría	0	0	0
	Ambiental	87.27	100	93.2
	Promedio	**65.03**	**64.02**	**63.27**
C4.5	Difusión	88.57	93.94	91.18
	Cursos	90	81.82	85.71
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**94.64**	**93.94**	**94.22**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para la combinación de características numéricas, nominales y características lexicográficas mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 94.64% de precisión, cobertura 93.94% y medida F_1 de 94.22%. KNN muestra un desempeño con 65.03% de precisión, 64.02% de cobertura y 63.27% para la medida F_1, mientras que SVM muestra un bajo rendimiento con 11.01% de precisión, 25% de cobertura y un 15.29% para F_1.

Las pruebas realizadas a la combinación de características lexicográficas y características nominales indican que C4.5 presenta un rendimiento superior en todos los casos. Por último, la evaluación de combinar todas las características lexicográficas y características nominales demuestra nuevamente que C4.5 dispone de un mayor rendimiento.

6.2.3 Pesado booleano

Se realizaron ocho pruebas con pesos booleanos y se comparó el rendimiento de los algoritmos de clasificación Naïve Bayes, k-vecinos más cercanos, C4.5 y SVM. A continuación, se muestran los experimentos realizados y el análisis de cada uno en el cual se exponen a los dos mejores algoritmos y al algoritmo que obtiene los resultados más bajos durante las pruebas.

Verbos

Los resultados obtenidos de la experimentación sobre el conjunto de verbos para las características individuales se presentan en la Tabla 34.

Tabla 34. *Resultados de la experimentación con verbos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	62.16	69.7	65.71
	Cursos	46.67	31.82	37.84
	Asesoría	30.77	66.67	42.11
	Ambiental	90.91	83.33	86.96
	Promedio	**57.63**	**62.88**	**58.15**
KNN	Difusión	80	12.12	21.05
	Cursos	75	13.64	23.08
	Asesoría	0	0	0
	Ambiental	47	97.92	63.51
	Promedio	**50.50**	**30.92**	**26.91**
C4.5	Difusión	58.97	69.7	63.89
	Cursos	90	81.82	85.71
	Asesoría	50	16.67	25
	Ambiental	70.83	70.83	70.83
	Promedio	**67.45**	**59.75**	**61.36**
SVM	Difusión	61.7	87.88	72.5
	Cursos	100	31.82	48.28
	Asesoría	0	0	0
	Ambiental	83.64	95.83	89.32
	Promedio	**61.33**	**53.88**	**52.52**

El análisis de resultados obtenido para el conjunto de verbos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 67.45% de precisión, cobertura de 59.75% y medida F_1 de un 61.36%. NB presenta un rendimiento menor con un 57.63% de precisión, un 62.88% de cobertura y 58.15% para la medida F_1. KNN, muestra una precisión menor del 50.5%, una cobertura

de 30.92% y una medida F_1 de 26.91%.

Adjetivos

Los resultados obtenidos de la experimentación sobre el conjunto de adjetivos para las características individuales, se presentan en la Tabla 35.

Tabla 35. *Resultados de la experimentación con adjetivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	65.38	51.52	57.63
	Cursos	56.25	40.91	47.37
	Asesoría	2.5	16.67	4.35
	Ambiental	59.26	33.33	42.67
	Promedio	**45.85**	**35.61**	**38**
KNN	Difusión	50	6.06	10.81
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.66	95.83	60.93
	Promedio	**23.67**	**25.47**	**17.93**
C4.5	Difusión	70.83	51.52	59.65
	Cursos	60	13.64	22.22
	Asesoría	0	0	0
	Ambiental	57.89	91.67	70.97
	Promedio	**47.18**	**39.20**	**38.21**
SVM	Difusión	59.52	75.76	66.67
	Cursos	100	4.55	8.7
	Asesoría	0	0	0
	Ambiental	71.21	97.92	82.46
	Promedio	**57.68**	**44.55**	**39.45**

El análisis de resultados obtenido para el conjunto de adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que SVM presenta un 57.68% de precisión, 44.55% de cobertura y 39.45% para la medida F_1. C4.5 presenta un 47.18% de precisión, cobertura de 39.20% y 38.21% para el caso de la medida F_1. El clasificador KNN indica un rendimiento inferior con un 23.67% de precisión, 25.47% de cobertura y 17.93% para la medida F_1. El análisis indica que todos los algoritmos aplicados a este conjunto de datos presentan resultados por debajo del 50%.

Sustantivos

Los resultados obtenidos de la experimentación sobre el conjunto de sustantivos para las características individuales, se presentan en la Tabla 36.

Tabla 36. *Resultados de la experimentación con sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	66.67	90.91	76.92
	Cursos	81.82	40.91	54.55
	Asesoría	100	83.33	90.91
	Ambiental	95.83	95.83	95.83
	Promedio	**86.08**	**77.75**	**79.55**
KNN	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**
C4.5	Difusión	84.21	96.97	90.14
	Cursos	100	72.73	84.21
	Asesoría	100	100	100
	Ambiental	95.92	97.92	96.91
	Promedio	**95.03**	**91.90**	**92.81**
SVM	Difusión	46.38	96.97	62.75
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	90	75	81.82
	Promedio	**34.09**	**42.99**	**36.14**

El análisis de resultados obtenido para el conjunto de sustantivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 95.03% de precisión, cobertura de 91.90% y medida F_1 con un 92.81%; NB también ofrece buenos resultados de clasificación con 86.08% de precisión, 77.75% para la cobertura y un 78.55% para la medida F_1. En este caso, KNN indica un menor rendimiento comparado con el resto de algoritmos con un 11.01% de precisión, 25% de cobertura y un 15.29% para la medida F_1.

Verbos, adjetivos y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de todas las características lexicográficas se muestran en la Tabla 37.

Tabla 37. *Resultados de la experimentación con verbos, adjetivos y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	64.58	93.94	76.54
	Cursos	77.78	31.82	45.16
	Asesoría	100	33.33	50
	Ambiental	92	95.83	93.88
	Promedio	**83.59**	**63.73**	**66.4**
KNN	Difusión	0	0	0
	Cursos	100	4.55	8.7
	Asesoría	0	0	0
	Ambiental	44.44	100	61.54
	Promedio	**36.11**	**26.14**	**17.56**
C4.5	Difusión	82.86	87.88	85.29
	Cursos	94.12	72.73	82.05
	Asesoría	100	100	100
	Ambiental	92.16	97.92	94.95
	Promedio	**92.28**	**89.63**	**90.57**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para la combinación de características lexicográficas (verbos, adjetivos y sustantivos) mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 92.28% de precisión, cobertura del 89.63% y medida F_1 de 90.57%. NB presenta un buen rendimiento con 83.59% de precisión, 63.73% de cobertura y 66.40% para la medida F_1. SVM muestra un rendimiento deficiente con un 11.01% para la precisión, 25% para la cobertura y 15.29% para la medida F_1.

Las pruebas individuales realizadas a cada conjunto de características lexicográficas indican que C4.5 presenta un mayor rendimiento en la mayoría de los casos, de igual forma, se observa que la evaluación para la combinación de características lexicográficas muestra que C4.5 indica un rendimiento superior en comparación con el resto de los algoritmos.

Características numéricas, nominales y verbos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y verbos se muestran en la Tabla 38.

Tabla 38. *Resultados de la experimentación con características numéricas, nominales y verbos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	67.5	81.82	73.97
	Cursos	62.5	45.45	52.63
	Asesoría	100	100	100
	Ambiental	100	97.92	98.95
	Promedio	**82.50**	**81.30**	**81.39**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	91.67	100	95.65
	Cursos	100	86.36	92.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**97.92**	**96.59**	**97.08**
SVM	Difusión	45.83	100	62.86
	Cursos	0	0	0
	Asesoría	100	33.33	50
	Ambiental	100	72.92	84.34
	Promedio	**61.46**	**51.56**	**49.30**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y verbos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 97.92% de precisión, cobertura del 96.59% y medida F_1 con un 97.08%. KNN también presenta buenos resultados de clasificación con un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1. SVM muestra un menor rendimiento con un 61.46% de precisión, 51.56% de cobertura y 49.30% para la medida F_1.

Características numéricas, nominales y adjetivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y adjetivos se muestran en la Tabla 39.

Tabla 39. *Resultados de la experimentación con características nominales y adjetivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	72.22	78.79	75.36
	Cursos	62.5	68.18	65.22
	Asesoría	33.33	16.67	22.22
	Ambiental	100	95.83	97.87
	Promedio	**67.01**	**64.87**	**65.17**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	81.58	93.94	87.32
	Cursos	88.24	68.18	76.92
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**92.45**	**90.53**	**91.06**
SVM	Difusión	49.25	100	66
	Cursos	0	0	0
	Asesoría	100	50	66.67
	Ambiental	100	81.25	89.66
	Promedio	**62.31**	**57.81**	**55.58**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales y adjetivos mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 92.45% de precisión, cobertura del 90.53% y medida F_1 de 91.06%, en este caso KNN también ofrece buenos resultados de clasificación con un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1. SVM muestra un bajo rendimiento con un 62.31% de precisión, 57.81% de cobertura y 55.58% para la medida F_1.

Características numéricas, nominales y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales y sustantivos se muestran en la Tabla 40.

Tabla 40. *Resultados de la experimentación con las características numéricas, nominales y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	67.39	93.94	78.48
	Cursos	75	40.91	52.94
	Asesoría	100	83.33	90.91
	Ambiental	100	95.83	97.87
	Promedio	**85.60**	**78.50**	**80.05**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	81.08	90.91	85.71
	Cursos	83.33	68.18	75
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**91.10**	**89.77**	**90.18**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para el conjunto de características numéricas, nominales más sustantivos mostró en la evaluación global promedio de cada algoritmo de clasificación que KNN presenta un 93.33% de precisión, 90.15% de cobertura y un 90.84% para la medida F_1 y C4.5 ofrece un buen desempeño con un 91.1% de precisión, 89.77% de cobertura y 90.18% para la medida F_1. SVM muestra el rendimiento más bajo con un 11.01% de precisión, 25% de cobertura y 15.29% para la medida F_1.

Características numéricas, nominales, verbos, adjetivo y sustantivos

Los resultados obtenidos de la experimentación realizada a la combinación de las características numéricas, nominales, verbos, adjetivos y sustantivos se muestran en la Tabla 41.

Tabla 41. *Resultados de la experimentación con las características nominales, verbos, adjetivos y sustantivos*

Algoritmo	Clase	Precisión	Cobertura	F_1
NB	Difusión	64	96.97	77.11
	Cursos	77.78	31.82	45.16
	Asesoría	100	66.67	80
	Ambiental	100	95.83	97.87
	Promedio	**85.44**	**72.82**	**75.04**
KNN	Difusión	80	96.97	87.67
	Cursos	93.33	63.64	75.68
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**93.33**	**90.15**	**90.84**
C4.5	Difusión	88.57	93.94	91.18
	Cursos	90	81.82	85.71
	Asesoría	100	100	100
	Ambiental	100	100	100
	Promedio	**94.64**	**93.94**	**94.22**
SVM	Difusión	0	0	0
	Cursos	0	0	0
	Asesoría	0	0	0
	Ambiental	44.04	100	61.15
	Promedio	**11.01**	**25**	**15.29**

El análisis de resultados obtenido para la combinación de características numéricas nominales y características lexicográficas mostró en la evaluación global promedio de cada algoritmo de clasificación que C4.5 obtiene mejores resultados de clasificación con un 94.64% de precisión, cobertura 93.94% y medida F_1 de 94.22%.

KNN muestra un desempeño con 93.33% de precisión, 90.15% de cobertura y 90.84% para la medida F_1, mientras que SVM muestra un bajo rendimiento con 11.01% de precisión, 25% de cobertura y un 15.29% para F_1.

Las pruebas indican que C4.5 presenta un rendimiento superior en la mayor parte de los casos, únicamente para el caso de combinar a sustantivos y características nominales KNN indica un valor más alto, gracias a que el rendimiento de KNN mejora con el número de características que posee un evento. Cuanto mayor sea el número de características de un

evento, mayor será su precisión, por lo tanto mejora la medida $F_{1.}$

Por último, combinar todas las características lexicográficas y características nominales demuestra que C4.5 dispone de un mayor rendimiento por su capacidad de trabajar con datos numéricos, nominales y textuales.

VII. Análisis y discusión de resultados

En esta sección se presentan los resultados obtenidos de realizar pruebas exhaustivas a 24 conjuntos de datos conformados por diversas características (lexicográficas y nominales). Los resultados que se presentan toman como referencia a la medida F_1 y el esquema de pesado TF, TF-IDF y Booleano.

7.1 Análisis de resultados con el esquema de pesos basados en frecuencia del término (TF)

En las pruebas realizadas individualmente sobre cada conjunto de características lexicográficas (verbos, adjetivos y sustantivos) aplicando el pesado TF se puede observar que C4.5 muestra el mejor rendimiento, sin embargo, los resultados mostrados al evaluar a verbos y adjetivos indican que todos los clasificadores estiman un bajo desempeño, no así para los sustantivos con 91.65%.

La combinación de todas las características lexicográficas (verbos, adjetivos y sustantivos) nuevamente demuestra que C4.5 se encuentra por arriba de los clasificadores utilizados con un 89.81%, resultado que mejora con la combinación de características nominales más las características lexicográficas (nominal más verbos más adjetivos más sustantivos) a un 94.22%.

En cuanto a las pruebas realizadas individualmente sobre cada conjunto de características lexicográficas (verbos, adjetivos y sustantivos), NB es el segundo mejor algoritmo. El rendimiento más alto fue obtenido por los sustantivos con 80.66%, del mismo modo, para la combinación de todas las características lexicográficas (verbos, adjetivos y sustantivos) NB obtiene los resultados más altos con un 56.55%.

La combinación de todas las características nominales más características lexicográficas (nominal más verbos más adjetivos más sustantivos) mejoran los resultados obtenidos para la combinación de características nominales más verbos, adjetivos y sustantivos un 90.84% con KNN. Esto debido a que se tienen más características de los eventos útiles para el entrenamiento del algoritmo KNN.

Por otro lado, las pruebas realizadas con SVM demuestran que este algoritmo se encuentra sujeto al número de eventos para su entrenamiento, dado que, los resultados más bajos obtenidos por este algoritmo se ven reflejados en la categorización de aquellos eventos con menor número de elementos, mientras que los eventos con mayor número de instancias, si muestran resultados. A pesar de ello, estos resultados se consideran bajos ya que se encuentran por debajo del 50%, lo que quiere decir que se requieren de más instancias para su entrenamiento.

7.2 Análisis de resultados con el esquema de pesos basados en frecuencia del término-frecuencia inversa del término (TF – IDF)

Se puede observar que en las pruebas realizadas individualmente sobre cada conjunto de

características lexicográficas (verbos, adjetivos y sustantivos) aplicando el pesado TF-IDF, C4.5 presenta un rendimiento superior siendo del 91.65% de rendimiento obtenido por los sustantivos, caso contrario a los obtenidos en la evaluación de verbos y adjetivos que al igual que en el pesado TF se mantienen por debajo del 70%. C4.5 presenta obtiene el valor más alto con un 59.74% para verbos y un 37.56% para adjetivos. Resultado de combinar todas las características lexicográficas (verbos, adjetivos y sustantivos) muestra a C4.5 con un 89.81% como el mejor clasificador, este resultado mejora después de combinar características lexicográficas con características nominales donde los verbos indicaron el valor más alto con un a un 97.08%. La combinación de todas las características nominales más las características lexicográficas (nominal más verbos más adjetivos más sustantivos) mostraron que C4.5 estima un rendimiento del 94.22%.

En cuanto a las pruebas realizadas individualmente sobre cada conjunto de características lexicográficas (verbos, adjetivos y sustantivos), NB es el segundo mejor algoritmo con un rendimiento del 60.65% obtenido por los sustantivos, del mismo modo para la combinación de todas las características lexicográficas (verbos, adjetivos y sustantivos) con un 55.57% y a la combinación de características nominales y lexicográficas de manera individual (verbos, adjetivos y sustantivos) NB estima un rendimiento mayor para la combinación de características nominales más verbos con un 74.33% y 68.73% para la combinación de características nominales más sustantivos. La combinación de características nominales y adjetivos indican para este caso a KNN como el segundo mejor algoritmo de clasificación con un 90.84% para este conjunto de datos. La combinación de todas las características nominales más características lexicográficas (nominal más verbos más adjetivos más sustantivos) presentan a KNN como el segundo mejor algoritmo de clasificación a pesar de que presenta un rendimiento menor al 70%. Esto se debe a que se tienen más características para el entrenamiento del algoritmo KNN. Por lo tanto, al aumentar el número de características en las pruebas, el algoritmo KNN obtiene mejores resultados.

Por otro lado, los resultados presentados por SVM muestran un rendimiento menor al del 50% en la mayor parte de los casos. La razón por la que ocurre este comportamiento es porque SVM es sensible a los datos con los cuales trabaja. Las pruebas demostraron que en todos los casos los resultados obtenidos están estrechamente relacionados con la cantidad de eventos en cada clase. La manera más sencilla de verlo es que el algoritmo favorece los casos de clasificación para eventos ambientales que poseen más instancias en comparación con los eventos de asesoría donde sólo se tienen 20 eventos.

7.3 Análisis de resultados con el esquema de peso booleano

Se puede observar que en las pruebas realizadas individualmente sobre cada conjunto de características lexicográficas (verbos, adjetivos y sustantivos) aplicando los pesos booleanos, C4.5 presenta un rendimiento superior al resto de clasificadores en la mayoría de los casos. Nuevamente C4.5 para verbos y adjetivos presentan un resultado menor al 70%; a pesar de que en ambos casos C4.5 presenta un bajo rendimiento, nuevamente C4.5 indica un rendimiento superior.

En la combinación de características lexicográficas (verbos, adjetivos y sustantivos) C4.5 presenta un resultado superior a todos los clasificadores con el 90.57%.) La combinación

de todas las características nominales más las características lexicográficas (nominal más verbos más adjetivos más sustantivos) indican a C4.5 como el mejor clasificador con un rendimiento del 97.08%. Finalmente, al resultado de combinar todas las características nominales más las características lexicográficas (nominal más verbos más adjetivos más sustantivos) demuestran que C4.5 tienen el mejor rendimiento con un 94.22%.

En cuanto al rendimiento observado en las pruebas realizadas sobre cada conjunto de características lexicográficas, el segundo mejor clasificador es NB pues el rendimiento observado en los sustantivos es de 79.55%. En la combinación de todas las características lexicográficas (verbos, adjetivos y sustantivos) NB presenta un rendimiento del 66.44%; en la combinación de las características nominales y lexicográficas de manera individual (verbos, adjetivos y sustantivos) KNN presenta un rendimiento del 90.84% para la combinación de características nominales más verbos y para la combinación de características nominales y adjetivos respectivamente.

Finalmente, al resultado de combinar las características nominales más las características lexicográficas (nominal más verbos más adjetivos más sustantivos) KNN presenta un rendimiento del 90.84%, mientras que SVM presenta el rendimiento más bajo en todos los casos con un 15.29% de rendimiento.

Con lo anterior se puede decir que, los algoritmos C4.5 y KNN mejoran considerablemente su rendimiento con la combinación de características nominales, textuales y numéricas. Si se tienen más características, un algoritmo de clasificación mejora su precisión y como consecuencia su rendimiento en la medida F1 mejora

Conclusiones

En este trabajo de investigación se presentó una metodología para realizar la comparación de algoritmos de aprendizaje automático. En particular, se aplicó la metodología en un caso de estudio relacionado con la clasificación de eventos en un ambiente académico. Los algoritmos de clasificación que se compararon fueron: Naïve Bayes (NB), k vecinos más cercanos (KNN), C4.5 y SVM; estos algoritmos fueron seleccionados por ser los más utilizados en la revisión de la literatura especializada. Los algoritmos usados son de tipo supervisado, por lo tanto, requieren de una etapa de entrenamiento, para lo cual se utilizaron un 70% de los datos procesados con características de texto y características nominales.

La segunda etapa de la metodología consiste en el preprocesamiento de los datos de entrada mediante el cual se extraen las características de cada evento, las cuales son de dos tipos: características textuales y características nominales. A partir de las características de texto (el nombre y descripción de un evento), se obtuvieron conjuntos de características lexicográficas divididas en tres grupos: verbos, adjetivos y sustantivos; además un conjunto de características nominales. Todos estos conjuntos de características se utilizaron en la etapa de entrenamiento de los modelos de clasificación.

En este trabajo de investigación se implementó un módulo lematizador con la finalidad de identificar de manera adecuada y sin ambigüedades las características lexicográficas más representativas de un evento mediante un análisis morfológico del nombre y la descripción de cada evento.

Es importante destacar que para la implementación del módulo lematizador se utilizaron técnicas de Procesamiento de Lenguaje Natural de tal forma que se logra la desambiguación de términos similares. Este lematizador es mejor que otros lematizadores utilizados en las pruebas para la obtención de características. Este módulo lematizador está basado en el análisis morfológico destinado al análisis de las características de texto, el cual consiste en la segmentación de las cadenas presentes en un texto también llamada tokenización, una limpieza de las cadenas obtenidas, es decir se eliminan los caracteres especiales, así como signos de puntuación y números, además se identificaron y eliminaron las denominadas stop words para finalmente obtener la lista de características más representativas de un evento.

Otra característica del módulo lematizador es que está enfocado hacia el idioma español resaltando su importancia en comparación con el resto de los sistemas lematizadores utilizados en la revisión del estado del arte, los cuales están enfocados principalmente al idioma inglés, además, es relevante destacar que el idioma español es más complejo, esto realza las aportaciones de este trabajo de investigación.

Se utilizaron tres esquemas de ponderación de las características morfológicas y se utilizaron 24 configuraciones de experimentos, las cuales se basaron en el modelo espacio vectorial para su representación y fácil manejo por los algoritmos clasificadores. Los esquemas de ponderación utilizados en este trabajo de investigación son: pesos basados en

la frecuencia del término, la frecuencia de termino-frecuencia inversa del término y booleano.

Para determinar cuál clasificador es el más adecuado para el ambiente académico, se comparó el rendimiento de cada clasificador y se tomó en cuenta la precisión, la cobertura y la medida F_1 de cada uno, sin embargo, se utilizó la medida F_1 de conservar un equilibrio entre la precisión y la cobertura.

De los 24 de experimentos se observó que C4.5 y Naïve Bayes presentan los mejores resultados para verbos con un 59.74% en la medida F_1 para los pesados TF y TF-IDF en el caso de C4.5 y un 58.15% para los esquemas TF-IDF y booleano para Naïve Bayes.

En cuanto a los adjetivos C4.5 y Naïve Bayes también presentan los resultados más altos con un 59.74% de rendimiento para C4.5 en el esquema de pesado TF; Naïve Bayes presenta un porcentaje del 38.21% para los pesos booleanos, pese a que Naïve Bayes y C4.5 muestran los resultados más altos, en esta prueba su desempeño es menor al 70%.

Las pruebas realizadas a sustantivos muestran una mejora en los resultados obtenidos durante la prueba anterior.C4.5 muestra un rendimiento del 92.81% para la medida F_1 con los pesos booleanos mientras que Naïve Bayes en el esquema de pesado TF muestra un rendimiento del 80.66%. Por otro lado, la combinación de todas las características lexicográficas (verbos, adjetivos y sustantivos), C4.5 con los pesos booleanos presenta un rendimiento del 92.81%, Naïve Bayes en este esquema de pesado muestra un resultado del 66.44%.

La combinación de características nominales y verbos C4.5 presenta los resultados más altos en los tres esquemas de pesado con un rendimiento del 97.08%; en esta configuración de experimento KNN se posiciona como el segundo mejor método con un 90.84% para el pesado TF pesos booleanos.

En cuanto a la combinación de las características nominales y adjetivos C4.5 nuevamente presenta un rendimiento elevado en todos los esquemas de pesado, además KNN también presenta un muy buen rendimiento en todos los esquemas de pesado. C4.5 presenta un rendimiento del 91.06% en la medida F_1 mientras que KNN presenta un rendimiento del 90.84% en la medida F_1.

Las pruebas realizadas a la combinación de características nominales más sustantivos indican que el rendimiento de KNN es de 90.84% para el pesado TF y los pesos booleanos, en la configuración de este experimento, C4.5 es el segundo mejor algoritmo con un rendimiento del 90.18 en la medida F_1 para los pesados TF y booleano.

Por ultimo las pruebas realizadas al conjunto de todas las características nominales y lexicográficas (nominales, verbos, adjetivos y sustantivos) presentan a C4.5 con un porcentaje del 94.22% para todos los esquemas y a KNN con un 90.84% para los esquemas de pesado TF y booleano.

En todas las configuraciones de experimentos analizados se pudo observar que SVM presenta los resultados más bajos a pesar de ser el algoritmo más utilizado en la literatura especializada. A manera de resumen en la Tabla 42 se presentan los dos mejores algoritmos en la tarea de clasificación de eventos académicos correspondientes a la medida F_1 obtenidos en cada experimento.

Tabla 42. *Tabla resumen de los resultados más altos en la clasificación de eventos*

Configuración del experimento	Pesado TF %	Pesado TF-IDF %	Pesado Booleano %
Verbos	C4.5	C4.5	C4.5
	59.74	59.74	61.36
	KNN	NB	NB
	41.82	58.15	58.15
Adjetivos	C4.5	C4.5	SVM
	59.74	37.56	39.45
	NB	NB	C4.5
	37.7	37.49	38.21
Sustantivos	C4.5	C4.5	C4.5
	91.65	91.65	92.81
	NB	NB	NB
	80.66	60.65	79.55
Verbos, adjetivos y sustantivos	C4.5	C4.5	C4.5
	89.81	89.81	92.81
	NB	NB	NB
	56.55	55.57	66.44
Nominales, numéricas y verbos	C4.5	C4.5	C4.5
	97.08	97.08	97.08
	KNN	NB	KNN
	90.84	74.33	90.84
Nominales, numéricas y adjetivos	C4.5	C4.5	C4.5
	91.06	91.06	91.06
	KNN	KNN	KNN
	90.84	90.84	90.84
Nominales, numéricas y sustantivos	KNN	C4.5	KNN
	90.84	90.18	90.84
	C4.5	NB	C4.5
	90.18	68.73	90.18
Nominales, numéricas, verbos, adjetivos y sustantivos	C4.5	C4.5	C4.5
	94.22	94.22	94.22
	KNN	KNN	KNN
	90.84	63.27	90.84

Aportaciones de este trabajo de investigación

La principal contribución de este trabajo de investigación es la metodología propuesta para comparar algoritmos de aprendizaje automático la cual facilita el trabajo a investigadores y personas que deseen iniciarse en este campo de investigación.

Cabe mencionar que la metodología aquí propuesta fue aplicada particularmente en la clasificación de eventos académicos logrando determinar al mejor algoritmo de clasificación de eventos.

El uso de las características nominales y textuales resalta la importancia de la presente investigación en comparada con la literatura analizada en el estado del arte, donde, la mayor parte de los autores se han centrado en características de texto.

Otra aportación importante en este trabajo de investigación es la implementación de un módulo lematizador para la extracción de características textuales. Este lematizador permite identificar sin ambigüedades los términos que otros lematizadores tienden a confundir.

Por último, se tiene que el algoritmo C4.5 es el método que ofrece mejores resultados en la clasificación de eventos con distintas clases de características de tipo textual y nominal y numéricas, sin embargo, se pudo notar que SVM ofrece los resultados más bajos a pesar de ser el más utilizado en literatura. La razón del comportamiento de estos algoritmos está relacionada principalmente con la naturaleza de los algoritmos. C4.5 es un algoritmo basado en árboles de decisión lo le permite ir tomando decisiones, además de que cuenta con métodos que le permiten trabajar con datos nominales, textuales y numéricos como en el caso de este trabajo de investigación.

También, durante las pruebas realizadas se puede notar que SVM está sujeto al número de instancias con las que cuenta para su entrenamiento, ya que, SVM no logró determinar la clase aquellos eventos que cuentan con pocas instancias, cabe señalar que, en la literatura SVM es de los algoritmos más populares en la clasificación de textos, sin embargo, en las pruebas con las características textuales empleadas en este trabajo de investigación presentó un bajo rendimiento, este rendimiento mejora con la incorporación de características nominales y textuales con los pesos booleanos.

Los resultados anteriormente mencionados sugieren que todavía existe un campo muy extenso por explorar en futuras investigaciones.

Trabajos a futuro

El presente trabajo de investigación sugiere que, en futuras investigaciones, la metodología aquí propuesta sea aplicada en otros campos de investigación con el fin de elegir al mejor algoritmo de aprendizaje automático para aplicarlo en áreas como la medicina en la detección automática de enfermedades, así como en la mercadotecnia para simplificar el proceso de venta ya que elegir el algoritmo más adecuado permite a las empresas simplificar el proceso de venta a sus clientes en función de sus necesidades. En la seguridad informática sería de utilidad tener una metodología que permita elegir al mejor algoritmo para la detección de acciones fraudulentas que comprometa la seguridad de datos confidenciales en una empresa.

Una vez identificado al algoritmo C4.5 como el mejor algoritmo de clasificación para este caso de estudio se propone desarrollar un sistema de clasificación de eventos con comunicación directa con usuarios.

La extracción de características textuales, como alternativa, se presenta el uso de n-gramas como parte de la tarea de reconocimiento de patrones. Otro posible trabajo a futuro es la mejora del módulo lematizador implementado en este proyecto incorporando reglas que faciliten la detección de un número mayor de términos.

Referencias

[1] IEEE (1990) Standard Glossary of Data Management Terminology. IEEE Std 610.5

[2] IEEE (1990b) Standard Glossary of Software Engineering Terminology. IEEE Std 610.12 (Revision and redesignation of IEEE Std 792-1983).

[3] Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 3(3), 210-229.

[4] Mitchell T. Machine Learning, New York, USA, McGraw-Hill

[5] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[6] Gironés, J. Minería de datos: modelos y algoritmos, Editorial UOC, 2017. ProQuest Ebook Central.

[7] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2-3), 131-163.

[8] Cover, T. M. (1967). Nearest neighbour pattern classification. IEEE Transactions in information Theory, 13(1), 21-27.

[9] Viera, A. F. G. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investigación bibliotecológica, 31(71), 103-126.

[10] Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1): 21-27.

[11] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[12] Quinlan, J. R. (1993). C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco.C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco.

[13] Shafer, J., Agrawal, R., & Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining. In VLDB (Vol. 96, pp. 544-555).

[14] Mehta, M., Agrawal, R., & Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. In International conference on extending database technology (pp. 18-32). Springer, Berlin, Heidelberg.

[15] Rastogi, R., & Shim, K. (1998). Public: A decision tree classifer that integrates building and pruning. In Proc. 1998 Int. Conf. Very Large Data Bases (pp. 404-415).

[16] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

[17] Miller, G. A., & Johnson-Laird, P. N. (1976). Language and perception. Belknap

Press.

[18] Allen, J. F., & Ferguson, G. (1994). Actions and events in interval temporal logic. Journal of logic and computation, 4(5), 531-579.

[19] Galton, A., & Augusto, J. C. (2002, September). Two approaches to event definition. In International Conference on Database and Expert Systems Applications (pp. 547-556). Springer, Berlin, Heidelberg.

[20] Sowa, J. F. (2000). Knowledge representation: logical, philosophical, and computational foundations. PWS.

[21] Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model.Psychological science, 6(5), 292-297.

[22] Koohzadi, M., & Keyvanpour, M. R. (2014). An analytical framework for event mining in video data. Artificial Intelligence Review, 41(3), 401-413.

[23] Wasserkrug, Segev, (2016), Event Prediction in Encyclopedia of Database Systems (Pp. 1-5). Springer New York.

[24] I.A. Group, Scenarios for Ambient Intelligence In 2010, 2001

[25] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 841-842). ACM.

[26] Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In Fifth international AAAI conference on weblogs and social media.

[27] Ritter, A., Etzioni, O., & Clark, S. (2012). Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1104-1112). ACM.

[28] Zhou, D., Chen, L., & He, Y. (2015). An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

[29] Miwa, M., Sætre, R., Kim, J. D., & Tsujii, J. I. (2010). Event extraction with complex event classification using rich features. Journal of bioinformatics and computational biology, 8(01), 131-146.

[30] Nguyen, D. D., Dao, M. S., & Nguyen, T. V. T. (2015). Natural language processing for social event classification. In Knowledge and Systems Engineering (pp. 79-91). Springer, Cham.

[31] Filgueira, R., Bee, E. J., Diaz-Doce, D., Poole Sr, J., & Singh, A. (2017). Applying

machine-learning techniques to Twitter data for automatic hazard-event classification. In AGU Fall Meeting Abstracts.

[32] Moreno-Jiménez, L. G., Torres-Moreno, J. M., Castro-Sánchez, N. A., Nava-Zea, A., & Sierra, G. (2017, October). Criminal events detection in news stories using intuitive classification. In Mexican International Conference on Artificial Intelligence (pp. 120-132). Springer, Cham.

[33] Naughton, M., Stokes, N., & Carthy, J. (2010). Sentence-level event classification in unstructured texts. Information retrieval, 13(2), 132-156.

[34] Lesani, F. S., Ghazvini, F. F., & Amirkhani, H. (2017). Smart home user identification using bag of events approach. In 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 379-383). IEEE.

[35] Jackson, P., & Moulinier, I. (2002). Natural language processing for online applications: Text retrieval. Extraction and Categorization.

[36] Smedt, T. D., & Daelemans, W. (2012). Pattern for python. Journal of Machine Learning Research, 13 2063-2067.

[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12, 2825-2830.

[38] Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval. New York, NY: McGrawHill.

[39] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation, 60(5), 503-520.

[40] Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. Journal of documentation, 29(4), 351-372.

[41] Jackson, P., & Moulinier, I. (2007). Natural language processing for online applications: Text retrieval, extraction and categorization (Vol. 5). John Benjamins Publishing.

Anexo A

El programa simulador de eventos “Agentes de software concurrentes para la generación de datos sobre eventos en un espacio académico inteligente” es un sistema multi-agente para la generación automática de eventos implementado en el lenguaje de programación Java que simula la generación de eventos detectados por una red de sensores que monitorea un espacio académico.

Los eventos simulados por este sistema son eventos de cuatro categorías:

a) Eventos de difusión, los eventos que destacan en esta categoría son congresos, seminarios, talleres y eventos culturales en general.
b) Eventos de cursos académicos a nivel licenciatura, posgrado y de actualización al personal académico.
c) Eventos ambientales como son de temperatura, humedad, luminosidad y presencia.
d) Asesorías académicas en las que un profesor atiende a un alumno sobre temas relacionados a su campo de estudio.

En todos eventos generados, se define a los participantes, el lugar en el cual se llevó a cabo el evento; el horario de inicio y fin en el que se registra el evento y el lugar en el que registra un evento. Adicionalmente, un evento tiene un nombre y una descripción, ambas se encuentran escritas en el idioma español y son asignadas automáticamente de una BD de descripciones generadas por expertos a fin de utilizar un lenguaje más familiar para los usuarios del ambiente académico.

Estos eventos se mantienen en bitácoras de eventos que poseen una estructura definida. Las bitácoras registran datos estructurados: información referente a los participantes, el espacio físico en el que se detecta un evento, el tiempo y las variaciones detectadas por los sensores ambientales. En cuanto a los datos no estructurados se tienen el nombre y descripción de cada evento.

Artículos publicados

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 141 (2018) 175–182

Procedia Computer Science

www.elsevier.com/locate/procedia

The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018)

Ontology for Academic Context Reasoning

Maricela Bravo*, José A. Reyes-Ortiz, Isabel Cruz-Ruiz, Ariadna Gutiérrez-Rosales, and Josué Padilla-Cuevas

Autonomous Metropolitan University, San Pablo 180, Azcapotzalco, CDMX, México

Abstract

Ontologies have gained popularity in the scientific community as representational mechanisms to support intelligent reasoning and execute inferences. In this paper we describe an ontology designed specifically to represent academic contexts at a public university. This model consists of a collection of ontologies designed to represent persons, physical space, sensor networks, events, etc. among other entities that exist in the academic environment. In particular, we describe the design requirements that guided the construction of the ontologies. The resulting ontology model is evaluated considering the competency of the ontology, and the concept domain coverage. Results are promising and the set of competency questions are translated to queries showing that the ontology model adheres to the requirements.

Keywords: Context ontology; Ontology design; Ontology competency

1. Introduction

The myriad of daily activities occurring at universities generate a high demand for information services, which causes the requirement to process large volumes of data and offer information to users in a fast, concise, and pertinent manner. These information requirements vary according to the context of the user, the reasons why he / she is in a given building of the institution, and the time in which an event happens. Commonly, users who demand information services at the university are teachers, researchers, students, employees and visitors. The type of information requested ranges from course schedules, location of facilities, location of professors, etc. Therefore, the

* Corresponding author. Tel.: +52 55 53 18 95 32 ext. 149.
E-mail address: mcbc@correo.azc.uam.mx

10.1016/j.procs.2018.10.165

processing of users contexts must consider various situations and characteristics: a) public universities count with wireless network technologies through which they offer Internet access and diverse applications for users at the university; b) most people at the university have mobile devices capable of wireless network connection; and c) universities count with sufficient computing resources to support the user's duties. Given these conditions, it is possible to develop the computational infrastructure to facilitate smart interconnection of devices and data processing, that is, to implement Internet of Things (IoT) technology. However, the implementation of the IoT presents a series of challenges that must be solved so that the interconnection of objects and people in a university can be efficiently exploited. Among the most relevant problems to solve are:

a) Context management and service personalization. Context management consists of context acquisition, context processing, and context reasoning. Service personalization is the intelligent outcome of context management.
b) Event detection, management, and activation. Event management requires the representation of events at the academic environment in such a way that it is possible to know *What* happened? *When* happened? *Where*? and *Who* was involved? Handling of events requires the detection, registration and activation of actions in response to events.
c) The efficient interaction between intelligent information systems and hardware. This interaction requires interoperability between communication protocols, and the dynamic interconnection of mobile devices with wireless networks and sensors/actuators networks.

There are other relevant and derived problems. However, in this article we present an ontology solution approach for the representation and management of contexts and events in the university environment addressing particularly academic issues. Ontologies were selected as the formal representational mechanism as they are very popular in IoT based applications. Ontologies facilitate reusability, knowledge sharing, and execution of formal reasoning tasks such as satisfiability of concepts, consistency checking, classification and inference.

2. Related Work

The revision of the related work was done emphasizing the use of ontologies for context representation, management and reasoning. In 2003 Chen, Finin, and Joshi [1] described CoBrA, a context broker agent architecture that is capable of managing a shared model of the context and reasoning support for context-aware applications. Later in 2004 authors detailed the SOUPA ontology [2] which consists of vocabularies for describing person contact information, beliefs, desires, and intentions of an agent, actions, policies, time, space, and events. Razmerita, Angehrn, and Maedche [3] presented in 2003 OntobUM, a generic ontology-based user modeling architecture. This architecture integrates three ontologies: the user ontology, the domain ontology, and the log ontology. Later in 2007 [4] authors augmented their OntobUM model by representing the behavior of the user, such as: level of activity, type of activity, level of knowledge sharing, etc. Wang et al. [5] described in 2004 CONON, an ontology for modeling context in pervasive computing environments divided into upper ontology and specific ontology. The upper ontology model defines computational entity, location, person and activity as the most important entities of a context model. Later in 2004 [6] authors presented SOCAM, a Service-Oriented Context Aware Middleware architecture to support the construction of context-aware services in intelligent environments. SOCAM architecture incorporates CONON ontology. Preuveneers et al. [7] presented in 2004 CoDAMoS, an extensible context ontology for ambient intelligence, which describes four main concepts: user, environment, platform, and service. Authors described the requirements for ambient intelligence: application adaptability, resource awareness, mobile services, semantic service discovery, code generation, and context-aware user interfaces. In 2007 Ejigu et al. [8] presented an ontology-based context model to facilitate context reasoning by providing structure for contexts, rules and their semantics. This work takes the basic elements of a pervasive computing environment characterized by their dynamicity, heterogeneity, and ubiquity of users, devices and resources, ad hoc connection between devices; and the existence of logical and physical sensors. The sources of context information are implemented as classes: User, Device, Application, Physical Environment, Resource, Network, Location, and Activity; all these classes are subclasses of a general class Context. In 2010 Zainol and Nakata [9] presented a Generic Context Ontology Model to represent context information in general. The aim is to facilitate the common context representation, context matching and context reasoning. This approach is based on the idea that a well-defined context model will minimize the

complexity of context-aware systems enhancing their maintainability. This model consists of a formal specification of the semantics of context identifiers; allowing sharing knowledge among different resources. In 2010 Poveda-Villalón et al. [10] presented mIO! an ontology network for a mobile environment. mIO! Ontology consists of eleven modular ontologies: user, role, environment, location, time, service, provider, device, interface, source, and network. This ontology covers a wide range of concepts related with context representation. Skillen et al. [11] presented in 2012 a user profile model for context-aware application personalization; authors concentrated on concepts to model a dynamic context: user time, user location, user activity, and user context. In 2013 Guermah et al. [12] described an architecture for the development of context-aware services based on ontologies. This architecture is composed of three main elements: a meta-model of context, an ontology for the meta-model, and a reasoning engine. The context is modeled based on a meta-model that defines the context and sub-contexts. Context properties are gathered from sensors, each property has a context validity and context specification. In 2014 Nadozeva and Kiritsis [13] presented an ontology-based context model to capture general concepts about users and business. The aim of this work is to propose an ontology-based model and rules to classify the context of users and business. This ontology describes general concepts: space, matter, object, event and action. Domain specific extensions specify the vocabulary and properties related to a generic domain by specializing terms in the upper ontology. This model is composed of three sub-models: user context, business context, and information feature model. In 2015 Kayes, A. Han, J, Colman A. [14] presented Ontology-based Context-Aware Access Control (OntCAAC) a generic framework that models dynamic con-texts and access control policies. The aim is to use a policy model for specifying and enforcing context-aware access control. The OntCAAC provides the capability to control access for software services and resources by taking into account the context information. This ontology defines the general context entity classes: User, Role, Source, Owner, Relationships, Place, EnvPerson and Device. It also defines the different types of dynamic context information: ContextInfo, LocationInfo, TemporalInfo. The place ontology identifies different buildings, departments and rooms of the hospital. The Relationship ontology represents the relations between users: Person-Centric or Location-Centric. The former is used for represent relationship between users, the latter represents that the concerned people are collocated. The Status Information Ontology stores the health status, the current personal status, the current location status. Miraoui et al. [15] proposed in 2015 an ontology based on modeling a smart living room environment and their contextual information for enabling a common understanding of context and enhancing its sharing. Authors propose a definition of context for service oriented systems as any information that triggers a service or changes the quality of a service if its values change. Based on this concept modeling a smart room starts by specifying the services that each equipment can provide and the set of information that triggers the service.

None of revised works fully achieve the requirements specification, majority of ontologies include information about person or users, but do not consider IoT-based identification such as RFID tags or MAC address, for instance [2], [3], [5], [6], and [7]. Majority of revised related ontologies consider the geographical localization, but do no correlate localization with physical spaces and persons [3] and [7] not allowing the automatic identification and localization of persons and objects. Another important requirement was the incorporation of networks (computer-based networks and sensor networks), which was not considered. And finally, events or activities management is important, few reported works considered. All these requirements are necessary for context reasoning.

3. Ontology Design Methodology

In this section, the methodology that was defined and executed for design, construction and evaluation of the ontology is described. An initial set of competency questions were used for term elicitation and for final competency evaluation. Ontology design encompasses three stages: specification of ontology requirements, ontology construction, and ontology evaluation.

3.1. Specification of Ontology Requirements

In order to specify a set of initial ontology requirements, we reviewed the concepts of context and events management. Abowd et al. [16] compared and analyzed different definitions of context, and presented their

definition as follows: "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and the application themselves". From this definition we consider that the entities (or objects) present at an academic environment are: ***Person*** (for example professors, students, staff, etc.), ***Physical space*** (classroom, building, laboratory, bathroom, etc.), and computing ***Devices*** and ***Networks***, among others. Schilit et al. [17] stated that the tree important aspects of context are: where you are, who you are with, and what resources are nearby. From this approach, a context should include information about the geographical localization of objects and persons; together with the accurate identification of persons and objects. Hanzal et al. [18] distinguish between objects and events (both important concepts related with the concept of context [16] [17]). They clarify the distinction between objects and events from a philosophical perspective stating that objects are *continuants* (they exist and persist through time) and events are *ocurrents* (they happen or take place at some point). This clarification of the concept ***event*** emphasizes the need to represent not only objects or entities, but events occurring at a point, this means that the ontology model should include the concept of ***Time***. From this initial analysis we have defined the concept coverage requirements of the ontology and defined the main objective of the ontology, which is to facilitate intelligent context processing in the academic environment.

3.1.1. Concept coverage

Based on the definitions of context and event given in [16] and [17] the ontology model should include the following concepts:

1. Person profile information to represent the user data that is possible to gather from public networks, public Web pages, or public data bases available such as DBLP.
2. Data for the identification of persons such as id credentials, RFID tags, MAC address, and passwords provided.
3. Data for the identification of objects (RFID Tags, MAC address, etc.)
4. Data for geographical localization of persons, objects or physical locations. In order to enable localization services, this data should consist of longitude, latitude and height.
5. Physical space to represent the buildings organization with their precise geographical coordinates.
6. Sensor networks to represent the intranet and private nets organization as they are currently arranged into the institution.
7. Environmental data to represent the physical sensor measures located at the different physical locations.
8. Device to represent any kind of hardware device available including personal computers, microcontrollers, cellular and any device capable of data processing.
9. Events that may occur at an academic environment which result of interest for the users at the university. Events should be correlated with time, physical space, the location, and the person involved.

4. Ontology Construction

The ontology was incrementally implemented using the Protégé ontology editor, and represented using the standard Web Ontology language (OWL). The resulting ontology model consists of a set of modular ontologies among which a set of semantic relations are defined to support intelligent context reasoning.

4.1. Person Ontology

Person ontology represents all possible human users that may be present at the university in a moment or period of time, such as: visitor, professor, student, employee, etc. Additionally, this ontology included the concepts of ***Department*** and ***AcademicTitle*** to improve the context descriptions with useful attributes for ***Academics***. Figure 1 shows the class hierarchy of the ***Person*** ontology. An important characteristic of this ontology was to define a unique identifier for every type of person that would be present inside the sensor-enabled context. The concept ***Person*** is defined as an equivalence through the *hasName* and *hasGender* data properties, indicating that every person should provide his name and gender. The concept ***Employee*** is defined as a sub class of a ***Person*** that

hasEconomicNumber data property. Whereas the concept ***Student*** is defined as a sub class of ***Person*** that *hasStudentId*. An important concept is a ***Professor*** which is an ***Academic***, is an ***Employee*** and is a ***Person*** that *hasCategory*, *hasDepartment*, and *hasEmail*; and inherits the data property of an ***Academic*** *hasProject*. The class hierarchy of the ***Person*** ontology shows the sub-classification of the class ***Student*** into ***RegularStudent*** and ***AssistantStudent***. This classification addresses a particular need to represent the two types of students that exist in the university where the ***AssistantStudent*** is an ***Academic***, is an ***Employee*** and a ***Student***.

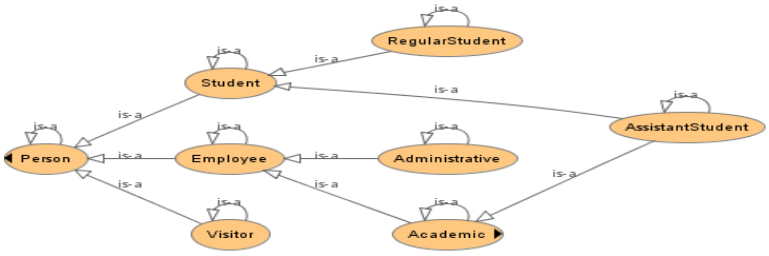

Fig. 1. Class hierarchy of the ***Person*** Ontology.

4.2. Physical Space Ontology

The ***PhysicalSpace*** ontology represents any kind of physical location such as buildings, cubicles, classrooms, offices, parking lots, plazas, green areas, etc. The class hierarchy of the ontology shown in Figure 2 contains as a main concept the ***PhysicalSpace*** class to conceptualize any kind of physical space which is divided into two main sub-class definitions: ***Internal*** and ***External***. An internal physical space is used to represent closed rooms; whereas external physical space is used to represent open spaces such as: hallway, green areas, parking, etc. The set of data properties that were defined are: *hasName*, *hasDoorState*, *hasAirConditioner*, *hasLampsNumber*, *hasLevel*, *hasPeopleCapacity*, *hasCarCapacity*, *hasProjector*, *hasService*, *hasWindow*, *isOpen*, *has Area*, among other properties. Two important relationships between concepts (object properties) defined for this class are: *isLocatedInto*, this property is used to specify that a physical space is part of another physical space, enabling physical objects composed of physical objects. For instance a classroom *isLocatedInto* a building; and the property *isBesideOf*, which is used to specify neighboring localities, for instance a cubicle A *isBesideOf* cubicle B and cubicle B *isBesideOf* bathroom 2.

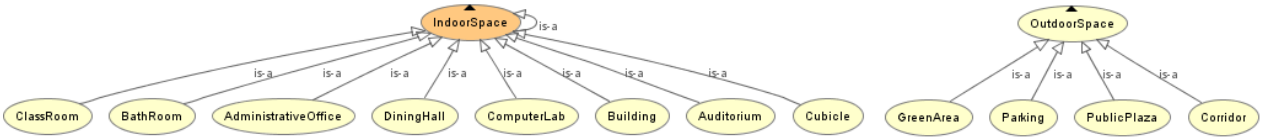

Fig. 2. Class hierarchy of the ***PhysicalSpace*** class.

4.3. Sensor Network Ontology

The ***SensorNetwork*** ontology represents a collection of physical sensors, mobile sensors and actuators. The objective of the network sensor is to obtain data from the physical context, user context and eventually activate some actuators. This ontology aims at representing environmental data such as temperature, lighting, humidity, and presence of humans into the environment. Another important objective of this model is the possible identification of the users and the data generated by user interaction with the environment. The following types of sensors are considered: *environmental sensors*, which are used to obtain data room temperature, humidity, luminosity, and presence of persons; *mobile and wearable sensors*, such as: accelerometer, gyroscope, magnetometer, proximity sensor, light sensor, barometer, thermometer, pedometer, heart rate monitor, fingerprint sensors, etc.; *identification*

sensors carried by the user, which will serve both the user ID to the acquisition of additional information; *actuators* represent the hardware devices through which actions are activated in order to achieve an ideal state.

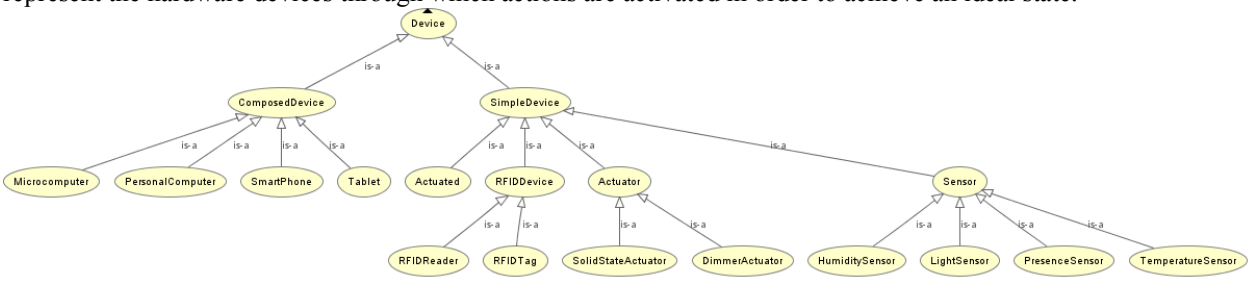

Fig. 3. Taxonomy of the ***Device*** class from the ***SensorNetwork*** Ontology.

4.4. Time Ontology

Time representation is an important requirement to manage events in the academic context. Figure 4 shows the structure of this ontology which consists of a class *TemporalEntity* sub divides into the concepts of *Instant* and *Interval* as follows: an Instant is defined as an individual that has the data attributes of *hasYear*, *hasMonth*, *hasDay*, *hasHour* and *hasMinute*; whereas *Interval* is defined with the object properties *hasBeginning* with range *Instant* and *hasEnd* with range *Instant*.

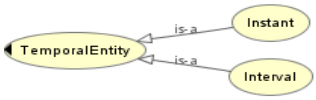

Fig. 4. Time Ontology.

5. Context Reasoning

Ontology reasoning consists of executing a program to infer logical consequences from a set of asserted facts or axioms. In order to realize the intelligent context reasoning the set of ontologies described above were imported and integrated into the ***IntelligentEnvironment*** ontology (see Fig. 5). This ontology was completed with additional class hierarchies, data properties and object properties. An important class hierarchy that was included into the ontology is the ***Event*** class to define a variety of events that may occur at the academic environment. A ***Presence*** event is sub divided into entering and leaving a physical space. These concepts are of great importance for the identification of persons into the academic intelligent environment. A Presence Event is defined thorough the data and object properties: *hasDescription*, *hasPersonInvolved* (*Person* class), *hasTime* (*TemporalEntity* class), and *happensIn* (*PhysicalSpace* class). All these semantic relationships are useful to answer any question regarding events occurring in the academic environment. What happened? When? Who was involved? At what time occurred the event?

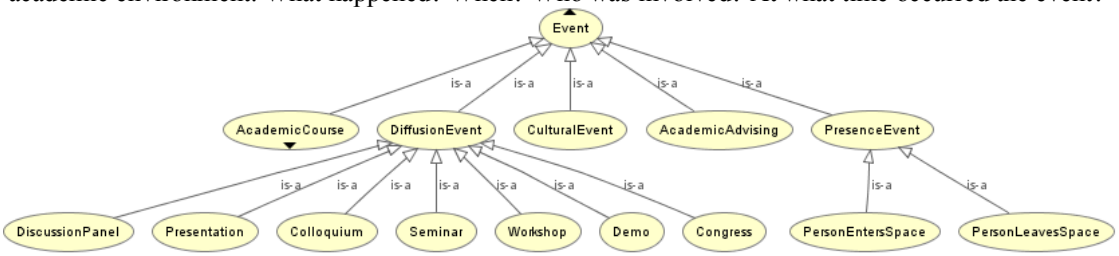

Fig. 5. The Event class hierarchy.

6. Ontology Evaluation

Ontology Evaluation [19] concerns the correct building of the ontology, ensuring that its definitions correctly implement the ontology requirements and competency questions. For evaluation two important aspects are considered: the ***competence of the ontology***, that is, if it is able to respond to a set of competency questions; and the verification of ***requirements compliance***. The following competency questions were correctly answered:

1. Who is present in the classroom E313?

   ```
   Person and (personLocatedIn value classRoomE313)
   ```

2. Where is the cubicle of professor Alejandro Reyes?

   ```
   Cubicle and (hasEmployeeAssigned some (hasName value "REYES ORTIZ JOSE ALEJANDRO"))
   ```

3. Where is professor Maricela?

   ```
   PhysicalSpace and (hasPersonDetected some (Professor and hasName value “BRAVO CONTRERAS MARICELA CLAUDIA”))
   ```

4. At what time did Professor Alejandro Reyes Ortiz leave the Babbage computing laboratory?

   ```
   Instant and (instantPersonLeaves some (hasPersonInvolved  some (hasName value "REYES ORTIZ JOSE ALEJANDRO")))
   ```

5. Who is the professor responsible of Distributed Systems course?

   ```
   Professor and (isProfessorResponsibleOfCourse some (hasEventName value "Curso Sistemas Distribuidos"))
   ```

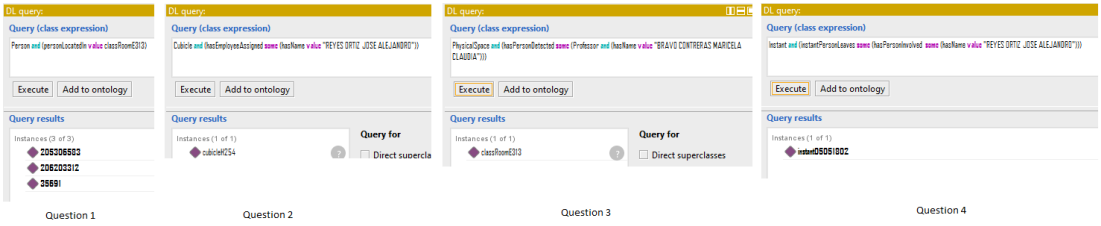

Fig. 6. Answers to competency questions using DL Query tool.

6. How many publications does the professor with economical number “14233” has?

   ```
   Professor(?prof1) ^ hasPublish(?prof1, ?pub) ^ hasEconomicNumber(?prof1, ?e) ^ swrlb:equal(?e, "14233") -> sqwrl:count(?pub)
   ```

7. With whom the professor identified with economical number “14233” collaborates with?

   ```
   Professor(?p1) ^ collaborateWith(?p1, ?p2) ^ hasEconomicNumber(?p1, ?e) ^ swrlb:equal(?e, "14233") -> sqwrl:select(?p1, ?p2)
   ```

8. How many women professors from Systems Department have published at least one article?

   ```
   Professor(?prof) ^ hasGender(?prof, ?gen) ^ swrlb:equal(?gen, "FEMENINO") ^ hasDepartment(?prof, ?dep) ^ swrlb:equal(?dep, "SISTEMAS") ^ hasPublish(?prof, ?pub) -> sqwrl:count(?prof)
   ```

6.1. Evaluation of the Requirements

Ontology requirements were fully attended as follows: Person profile information and data for the automatic identification of persons were included in the ***Person*** ontology. The particular properties *hasRFIDTag* with domain ***Employee*** and ***Student*** and range ***RFIDTag***, *hasMACAddress*, and *hasIPAddress* allow the automatic identification

of persons and objects at the intelligent environment. Geographical localization of persons, objects or physical locations is included in the ***PhysicalSpace*** ontology by means of the following properties: *isLocatedInto*, *isBesideOf*, *hasLatitude*, *hasLongitude*, and *hasAltitude*. Network and Device is represented at the ***SensorNetwork*** ontology. Environmental data is represented at the ***PhysicalMeasure*** ontology. Events are represented at the ***IntelligentEnvironment*** ontology, in this ontology events are correlated with time, physical space, and persons.

7. Conclusions

The ontology model reported in this paper is envisioned for a wireless networked environment, where users may be identified by their mobile device mac address or by an RFID card. Such an environment may be an office or laboratory into an academic institution or university, where users enter and leave the environment freely. The ontology was constructed from scratch, because the full list of requirements was not included in any of the reported works. For evaluation, a set of competency questions were translated to queries and executed to show that the ontology is complete and is capable of context reasoning. Results and the evaluation of the ontology model show promising advances towards the construction of an integral IoT platform.

References

[1] Chen, H., Finin, T., & Joshi, A. An ontology for context-aware pervasive computing environments. The knowledge engineering review, 18 (03), 197-207, (2003).

[2] Chen, H., Finin, T., & Joshi, A. (2005). The SOUPA ontology for pervasive computing. In Ontologies for agents: Theory and experiences (pp. 233-258). Birkhäuser Basel.

[3] Razmerita, L., Angehrn, A., & Maedche, A. Ontology-based user modeling for knowledge management systems. In International Conference on User Modeling (pp. 213-217). Springer Berlin Heidelberg, (2003).

[4] Razmerita, L. Ontology-based user modeling. In Ontologies (pp. 635-664). Springer US, (2007).

[5] Wang, X. H., Zhang, D. Q., Gu, T., & Pung, H. K. Ontology based context modelling and reasoning using OWL. In Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on (pp. 18-22). IEEE, (2004).

[6] Gu, T., Wang, X. H., Pung, H. K., & Zhang, D. Q. An ontology-based context model in intelligent environments. In Proceedings of communication networks and distributed systems modeling and simulation conference (Vol. 2004, pp. 270-275), (2004).

[7] Preuveneers, D., Van den Bergh, J., Wagelaar, D., Georges, A., Rigole, P., Clerckx, T., ... & De Bosschere, K. Towards an extensible context ontology for ambient intelligence. In European Symposium on Ambient Intelligence (pp. 148-159). Springer Berlin Heidelberg, 2004.

[8] D. Ejigu, M. Scuturici and L. Brunie, An Ontology-Based Approach to Context Modeling and Reasoning in Pervasive Computing, Pervasive Computing and Communications Workshops, 2007. PerCom Workshops '07. Fifth Annual IEEE International Conference on, White Plains, NY, pp. 14-19, (2007).

[9] Zainol, A., Nakata, K., Generic Context Ontology Modelling: A review and Framework, on 2nd International Conference on Computer Technology and Development (ICCTD), pp 126-130, (2010).

[10] Poveda Villalon, M., Suárez-Figueroa, M. C., García-Castro, R., & Gómez-Pérez. A context ontology for mobile environments. (2010).

[11] Skillen, K. L., Chen, L., Nugent, C. D., Donnelly, M. P., Burns, W., & Solheim, I. (2012, December). Ontological user profile modeling for context-aware application personalization. In International Conference on Ubiquitous Computing and Ambient Intelligence (pp. 261-268). Springer Berlin Heidelberg.

[12] Guermah, H., Fissaa, T., Hafiddi, H., Nassar, M., & Kriouile, A, Context modeling and reasoning for building context aware services, In Computer Systems and Applications (AICCSA), 2013 ACS International Conference on, pp. 1-7, (2013).

[13] Nadoveza A., Kiritsis D., Ontology-Based Approach for Context Modeling in Enterprises, Computer in Industry, pp. 1218-1231, (2014).

[14] Kayes, A., Han J., Colman, A. OntCAAS: An Ontology-Based Approach to Context-Aware Access Control for Software Services, The Computer Journal, v. 58, No 11, (2015).

[15] Miraoui, M., El-etriby, S.,Tadj, Ch., Zaid Abid, A., Ontology-Based Context Modeling for a Smart Living Room, Proceedings of the World Congress on Engineering and Computer Science 2015 Vol I WCECS 2015, October 21-23, 2015, San Francisco, USA

[16] Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggles, P. (1999, September). Towards a better understanding of context and context-awareness. In International Symposium on Handheld and Ubiquitous Computing (pp. 304-307). Springer, Berlin, Heidelberg.

[17] Schilit, B., Adams, N., & Want, R. (1994, December). Context-aware computing applications. In Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on (pp. 85-90). IEEE.

[18] Hanzal, T., Svátek, V., & Vacura, M. (2016, July). Event Categories on the Semantic Web and Their Relationship/Object Distinction. In FOIS (pp. 183-196).

[19] Gómez-Pérez, A. Towards a framework to verify knowledge sharing technology. Expert Systems with Applications, 11(4), 519-529, (1996).

COMIA 2019

La Sociedad Mexicana de Inteligencia Artificial (SMIA), la Unidad de Transferencia Tecnológica Tepic, del Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE-UT³) y la Universidad Autónoma de Nayarit (UAN)

OTORGAN ESTE CERTIFICADO A

Ariadna Gutiérrez-Rosales, José A. Reyes-Ortiz and Maricela Bravo

POR LA PRESENTACIÓN DEL ARTÍCULO TITULADO

Clasificación de Eventos Académicos a partir de su Descripción Textual

En el 11° Congreso Mexicano de Inteligencia Artificial – COMIA 2019, Tepic, Nayarit, México, del 4 al 7 de junio de 2019

DR. FÉLIX CASTRO ESPINOZA
PRESIDENTE SMIA

DR. JUAN MARTÍNEZ MIRANDA
COMITÉ LOCAL COMIA

COMIA · SMIA · CICESE UT³ · UNIVERSIDAD AUTÓNOMA DE NAYARIT · INTELIGENCIA ARTIFICIAL

Clasificación de eventos académicos a partir de su descripción textual

Ariadna Gutiérrez-Rosales, José A. Reyes-Ortiz, Maricela Bravo

Universidad Autónoma Metropolitana, División de Ciencias Básicas e Ingeniería,
Departamento de Sistemas,
Azcapotzalco, Ciudad de México
México

ariadna.gtzr08@gmail.com, {jaro, mcbc}@azc.uam.mx

Resumen. El mundo se ha introducido cada vez más en la era inteligente y con ello surgen los espacios inteligentes, donde predecir los eventos que suceden en él no es una labor sencilla para lo cual se necesita brindarle inteligencia al medio para que éste sea capaz de anticipar eventos por suceder y proporcionar los servicios adecuados al usuario de acuerdo con sus necesidades. La clasificación de eventos académicos es deseable ya que permite predecir acontecimientos, por mencionar algunos ejemplos: visitas, reuniones, seminarios, cursos académicos y asesorías. En los espacios académicos, los métodos de clasificación pueden considerar eventos pasados y el tiempo en que ocurrieron, tales modelos dotarían al espacio académico de cierto grado de inteligencia para actuar sobre algunas situaciones o decisiones a futuro. Los eventos a clasificar en este artículo están relacionados con docencia, investigación y difusión de la cultura, pertenecientes a cuatro clases: evento de difusión, evento ambiental, evento de cursos académicos y evento de asesoría. Esta clasificación tiene como objetivo determinar el tipo de evento que sucede dentro del espacio académico, para ello se evalúan cuatro de los principales modelos de clasificación más utilizados en la literatura (Naïve Bayes (NB), K-Nearest-Neighbors (KNN), C4.5 y Support Vector Machine (SVM)) y se expone cuál es el más adecuado para un espacio académico.

Palabras clave: eventos académicos, clasificación, aprendizaje automático, procesamiento de lenguaje natural.

Academic Event Classification from Textual Descriptions

Abstract. The world has been introduced more and more into the intelligent age and with it intelligent spaces arise, where to predict the events that take place in it is not a simple task for which it is necessary to provide intelligence to the environment so that it is capable of anticipating events to happen and provide the appropriate services to the user according to their needs. The classification of academic events is desirable since it allows predicting events, to mention a few

examples: visits, meetings, seminars, academic courses and consultancies. In academic spaces, classification methods can consider past events and the time they occurred, such models would give the academic space a certain degree of intelligence to act on some situations or decisions in the future. The events to be classified in this article are related to teaching, research and dissemination of culture, belonging to four classes: dissemination events, environmental events, academic courses and advice. This classification aims to determine the type of event that happens within the academic space, for which four of the main classification models most used in the literature are evaluated (Naïve Bayes (NB), K-Nearest-Neighbors (KNN), C4.5 and Support Vector Machine (SVM)) and which is best suited for an academic space.

Keywords: academic events, classification, machine learning, natural language processing

1 Introducción

El mundo se ha introducido cada vez más en la era inteligente y con ello surgen los espacios inteligentes, donde identificar los eventos que suceden en él no es una labor sencilla, motivo por el cual se necesitan enfoques computacionales para que éste sea capaz de anticipar eventos por suceder y proporcionar los servicios adecuados al usuario de acuerdo con sus necesidades. A pesar de que el término ambiente inteligente es utilizado principalmente en “casas inteligentes” es posible extender su aplicación de estudio a los espacios académicos.

La clasificación, es una de las principales tareas del aprendizaje automático, ofrece información que puede ser utilizada para la toma de decisiones, y eliminación de tareas manuales y repetitivas. Algunos ejemplos en los que se utiliza la clasificación son: la medicina, detección de fraude y seguridad, sistemas de recomendación, identificación de correo no deseado, y ambientes inteligentes. Es este último, donde se enfoca este trabajo.

La clasificación de eventos académicos es deseable ya que permite predecir acontecimientos, por mencionar algunos ejemplos: visitas, reuniones, seminarios, cursos académicos y asesorías. En los espacios académicos, los métodos de clasificación pueden considerar eventos pasados y el tiempo en que ocurrieron, tales modelos dotarían al espacio académico de cierto grado de inteligencia para actuar sobre algunas situaciones o decisiones a futuro. Los eventos a clasificar en este artículo están relacionados con docencia, investigación y difusión de la cultura, pertenecientes a cuatro clases: evento de difusión, evento ambiental, evento de cursos académicos, y evento de asesoría. Esta clasificación tiene como objetivo determinar el tipo de evento que sucede dentro del espacio académico, para ello se evalúan cuatro de los principales modelos de clasificación más utilizados en la literatura (Naïve Bayes (NB), K-Nearest-Neighbors (KNN), C4.5 y Support Vector Machine (SVM)) y se expone cuál es el más adecuado para un espacio académico. El resto del trabajo se organiza como sigue. En la Sección 2 se presentan los trabajos relacionados con la clasificación de eventos basado en sus descripciones textuales. La Sección 3 expone el enfoque utilizado para la clasificación de

eventos que incluye los algoritmos de aprendizaje automático. La Sección 4, presenta los resultados obtenidos con los diversos algoritmos y combinando las características textuales extraídas. Finalmente, las conclusiones y el trabajo a futuro son presentado en la Sección 5.

2 Trabajos relacionados

En esta sección se describe el trabajo realizado en el área de clasificación de eventos basada en información no estructurada como sus descripciones textuales. Además, se explora el uso de algoritmos de aprendizaje automático para dicha tarea, así como el uso de recursos externos como las ontologías.

Con respecto a la clasificación de eventos utilizando textos, existen trabajos que han utilizado las redes sociales como su fuente de información. En [1] se presenta un enfoque tradicional basado en "Bolsa de palabras" para la clasificación de mensajes de Twitter en diversas categorías entre las que destacan los eventos y noticias. [2] expone un enfoque para clasificar mensajes de la red social en dos categorías mensajes sobre eventos del mundo real y mensajes que no son eventos; los autores utilizan la técnica de clasificación en línea y agrupamiento basado en tópicos junto con características textuales.

Un sistema para la extracción y clasificación de eventos en un dominio abierto a partir de Twitter es presentado en [3]. Los autores proponen un enfoque basado en modelos de variables latentes que descubren un conjunto apropiado de tipos de eventos que coinciden con los datos. Los eventos descubiertos automáticamente se inspeccionan posteriormente para filtrar los que son incoherentes y el resto se anota con etiquetas informativas, algunas como: finanzas, educación, religión, deportes y política. El conjunto resultante de clases de eventos se aplica luego para categorizar cientos de millones de eventos reales extraídos de manera automática.

En [4] se propone un enfoque no supervisado para explorar eventos a partir de Twitter, el cual consiste en un proceso de filtrado, extracción y categorización de eventos. En la etapa de filtrado el ruido de los tweets es eliminado, mientras que para la extracción se utiliza un lexicón para separar los tweets de aquellos que no son relevantes. Finalmente, para la categorización, los tweets son representados en vectores de características textuales y un modelo Bayesiano es utilizado para clasificar los eventos sin el uso de datos etiquetados.

En el dominio de la bioinformática, la clasificación de eventos a partir de textos médicos ha sido de gran ayuda para la identificación y extracción automática de eventos adversos, como en [5], donde se utilizan un método de aprendizaje automático para la detección efectiva de eventos en biomedicina; en [6] que extraen las relaciones entre medicamentos y efectos adversos como eventos a partir de literatura médica. En [7] se presenta un sistema que extrae seis tipos de eventos (pruebas, problema, tipo de diagnóstico, tratamiento, evidencias y ocurrencias) a partir de notas médicas, utilizando características semánticas como nombres de medicamentos, tratamientos, enfermedades, síntomas y regiones anatómicas extraídas del conjunto de datos utilizado como entrenamiento.

Finalmente, el uso de ontologías para apoyar la minería de textos en biomedicina, se presenta en [8], donde exponen un enfoque basado en reglas de decisión para la extracción y clasificación de eventos y hechos. Las ontologías ayudan a la identificación de características semánticas como el reconocimiento de entidades nombradas.

Con la revisión de los trabajos relacionados, se puede notar que la mayoría de los esfuerzos se centran en dominios como la medicina y utilizando textos en inglés extraídos de redes sociales y literatura científica. Con ello, es evidente la necesidad de un enfoque para la clasificación de eventos académicos utilizando textos en español, como lo presenta este trabajo de investigación.

3 Clasificación de eventos

El proceso de clasificación de eventos, involucra una serie de etapas que a continuación se enumeran en la Figura 1.

1. Recopilar datos y formación del conjunto de datos.
2. Limpieza y transformación de los eventos (Selección de datos).
3. Minería de datos (Seleccionar el método de minería): Clasificación.
4. Evaluación e interpretación del método.

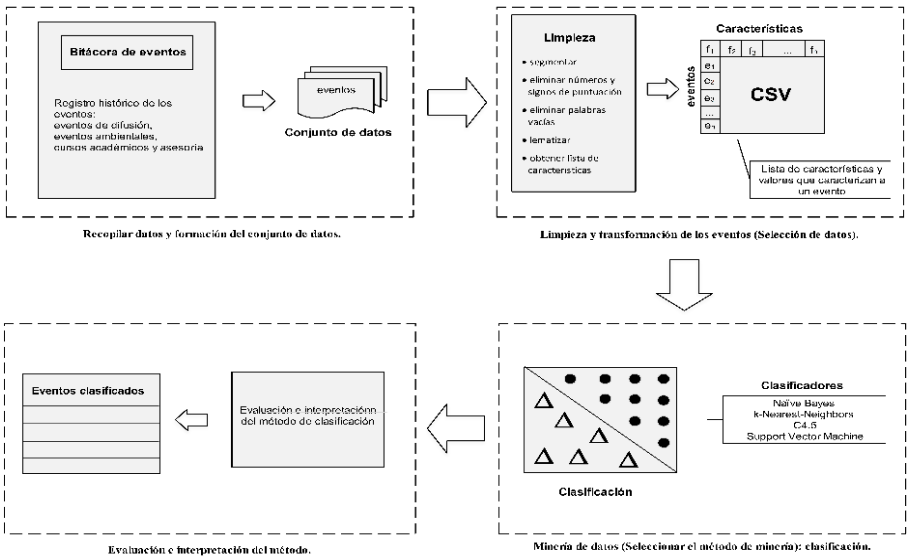

Fig. 1. Proceso de clasificación de eventos.

3.1 Recopilar datos y formación del conjunto de datos

La etapa de recopilación de datos parte de una serie de bitácoras las cuales tienen como objetivo mantener el registro histórico de los eventos ocurridos en el espacio académico: eventos de difusión, cursos académicos, asesorías y eventos ambientales. Estas

bitácoras son obtenidas a partir de la implementación de un programa que, de manera automática realiza la simulación de eventos aleatorios ocurridos en un espacio académico. Los datos que se almacenan en las bitácoras de eventos son: el tipo participante, el horario y el lugar en el que se registra el evento, así como el nombre y descripción del evento.

Se describe la implementación de un método para la extracción automática y el análisis de estos eventos, que se almacenan en archivos de texto plano y forman el conjunto de datos de entrada, en una etapa siguiente, serán representados como un conjunto de características morfológicas de cada evento.

3.2 Limpieza y transformación de los eventos (selección de datos)

La limpieza y transformación depende de la recopilación de datos pues se realiza la traducción (transformación) de los eventos almacenados en las bitácoras de eventos a la lista de características y valores que caracterizan a un evento para ello es necesario un proceso de limpieza de los datos. Los eventos se representan por pares del tipo EVENT: FEATURE (evento: característica). Las características para cada EVENT: FEATURE se representan utilizando el modelo espacio vectorial como en [9], donde las características se representan numéricamente.

En este modelo es muy común representar a los elementos en una tabla, las filas representan los eventos y las columnas f_i representan las características de cada evento.

Las características f_i de cada evento se representan por el conjunto de todas las características:

$$F = \{f_1, f_2, f_3 \ldots f_n\} \tag{1}$$

Los eventos son el conjunto de todos los eventos:

$$\mathrm{E} = \{e_1, e_2, e_3 \ldots e_n\} \tag{2}$$

3.3 Minería de datos (seleccionar el método de minería): Clasificación

Las técnicas de minería de datos se clasifican en dos categorías: supervisadas o predictivas y no supervisadas o descriptivas [10]. En esta fase es donde se decide cuál es la tarea (clasificación) a realizar y las técnicas descriptivas o predictivas a utilizar (seleccionar el método de clasificación). A continuación, se describen las utilizadas en este trabajo.

3.3.1 Clasificación bayesiana

Los clasificadores Bayesianos son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. La clasificación Bayesiana se basa en el teorema de Bayes [10].

Clasificadores como Naïve Bayes [11] permiten simplificar el coste computacional del modelo probabilístico, sin pérdida de expresividad por parte del mismo demostrando una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos.

La teoría de la probabilidad y los métodos bayesianos son uno de los principales enfoques utilizados en el aprendizaje automático y la minería de datos; las razones por las que estos métodos resultan importantes son:

- Los métodos bayesianos permiten hacer inferencias a partir de los datos, formular hipótesis sobre nuevos valores, y además permiten calcular explícitamente la probabilidad asociada a cada una de las hipótesis posibles.
- Facilitan el trabajo para el análisis de numerosas técnicas de aprendizaje y minería de datos que no trabajan explícitamente con probabilidades.

Naïve Bayes [11] es el modelo más simple de clasificación en redes bayesianas. Su principal característica es que supone que todos los atributos son independientes esto da lugar a un modelo gráfico probabilístico en el que existe un único nodo raíz (clase), y en la que todos los atributos son nodos hoja en donde el único nodo padre es la clase.

El clasificador Naïve (ingenuo) Bayes [11], es utilizado cuando se quiere clasificar un ejemplo descrito por un conjunto de atributos (x_i's) en un conjunto finito de clases (c) de acuerdo con el valor más probable dados los valores de sus atributos [10] por lo tanto el objetivo de este clasificador es encontrar la clase óptima para un determinado evento, calculando la clase que da la probabilidad posterior máxima.

3.3.2 Aprendizaje basado en ejemplos

La clasificación basada en ejemplos se realiza por medio de una función que mide la proximidad o parecido con los ejemplos existentes usando una métrica de distancia y los ejemplos más próximos son utilizados para asignar la clase a la nueva instancia [12].

El clasificador *k* vecinos más próximos [13], es un método de aprendizaje perezoso basado en ejemplares, se basa en el modelo de espacio vectorial, el cual representa un conjunto de vectores de la forma ($a_1(x)$, $a_2(x)$,..., $a_n(x)$) en la que $a_r(x)$ es el valor de la instancia para el atributo a_r. El algoritmo KNN [13] procura por todos los ejemplos de entrenamiento comparar la similitud entre sus vectores de características, para encontrar los *k* ejemplos de entrenamiento más próximos y el ejemplar desconocido es designado a los *k* vecinos más próximos con mayor valor de clasificación.

La principal ventaja de este algoritmo es su facilidad de implementación, pero su costo computacional es alto cuando el tamaño de las instancias usadas en el entrenamiento crece. El mejor valor de *k* depende del conjunto de datos y del dominio de la aplicación.

3.3.3 Árboles de decisión

El árbol de decisión es una estructura en árbol, donde cada nodo representa un atributo a ser probado; las ramas representan la salida de la prueba y los nodos finales (hojas) representan la clasificación.

El algoritmo de árboles de decisión posee dos fases principales: en la primera llamada fase de crecimiento del árbol, el algoritmo inicia con todo el conjunto de datos como nodos raíz. Los datos son divididos en subconjuntos utilizando algún criterio de división. En la segunda fase, etapa de poda del árbol, el árbol total formado se poda para prevenir el exceso de ajuste (*over-fitting*) del árbol a los datos de entrenamiento.

Existen diversos algoritmos para construir árboles de decisión entre ellos ID3 [14], C4.5 [15], SPRINT [16], SLIQ [17] y PUBLIC [18]. El utilizado en este trabajo es el algoritmo C4.5 [15] que incluye diversos métodos para trabajar con atributos numéricos, valores ausentes, datos con ruidos y para generar reglas a partir de árboles de decisión.

3.3.4 Máquinas de soporte vectorial (SVM)

Máquinas de Soporte Vectorial [19] es un método de aprendizaje supervisado con un alto grado de clasificación, su funcionamiento está basado en la clasificación lineal separando los datos en dos clases. El algoritmo pretende encontrar el hiperplano que maximiza el margen entre los vectores de soporte que define la posición del hiperplano ideal. La ventaja de utilizar este método es su buen desempeño cuando se cuenta con un gran número de características y también cuando se tienen pocos elementos de entrenamiento en tareas de múltiples clases [19].

3.4 Evaluación e interpretación del método

En esta etapa se evalúa y se validan las conclusiones obtenidas comparando los modelos y se determina cuál ofrece mejores resultados de clasificación, para lograr esto se realizó la representación de eventos basada en el modelo vectorial y el esquema de pesado TF-IDF (Frecuencia del Término - Frecuencia Inversa del Término). TF - IDF es la unión del esquema de pesado TF (Frecuencia del Término) [20] con IDF (Frecuencia Inversa del Término) [21]. En TF-IDF [22] cada vector está conformado por los pesos que representan la relevancia que tiene una característica en un evento. De acuerdo con [22] aquellas características que ocurren con menor frecuencia se consideran más importantes que aquellas que ocurren con mayor frecuencia. Su fórmula se muestra en la ecuación 3.

$$tf - idf_{ij} = f_{ij} * \log(\frac{N}{\mathrm{df}}) \qquad (3)$$

Donde f_{ij} es la frecuencia de la característica *i* en el evento *j*, *N* es el número de descripciones y *df* es el número de descripciones en donde aparece el término *i*.

La precisión, el recuerdo y la exactitud son las métricas de evaluación más comunes en la evaluación de los algoritmos de clasificación, en este trabajo se utiliza a la *precisión*. La *precisión* indica qué tan exacta fue la clasificación de los eventos, mientras que el recuerdo da a conocer si los eventos que pertenecen a una clase *i*, se clasificaron dentro de esa clase; la *exactitud* representa el porcentaje de las predicciones que son correctas. La fórmula de la *precisión* se muestra en la ecuación 4.

$$Precisión = \frac{Número\ de\ eventos\ clasificados\ correctamente}{Total\ de\ eventos} \quad (4)$$

4 Experimentación y resultados

En esta sección se presentan la experimentación y los resultados obtenidos con los algoritmos de aprendizaje supervisado presentados anteriormente. Además, del conjunto de datos utilizado para esta experimentación y su transformación para lograr la clasificación de eventos.

4.1 Conjunto de datos

En la etapa de recopilación de datos y formación del conjunto de datos, a partir de una serie de bitácoras se obtuvieron 363 eventos académicos de cuatro clases de eventos: eventos de asesoría, cursos académicos, eventos de difusión y eventos ambientales. De estos eventos son de interés los participantes del evento, el lugar y el horario en el cual se llevó a cabo cada evento y su variación en caso de que se trate de eventos ambientales, así como el nombre del evento y su descripción. En la Tabla 1 se muestran los cuatro tipos de eventos considerados, algunos de los cuales incluyen subtipos de eventos y sus descripciones correspondientes.

Tabla 1. Descripción de eventos

Tipo	Descripción de evento	Tipos de eventos
Asesoría	Consulta que brinda un profesor a un estudiante para resolver cuestiones sobre temas que domina	
Cursos académicos	Su objetivo es la formación académica y profesional de estudiantes y profesores	Licenciatura, posgrado y actualización
Evento de difusión	Evento cuyo objetivo es difundir temas relacionados con la investigación y la cultura	Congreso, panel de discusión, taller, seminario y presentación
Ambiental	Evento en el cual se encuentran involucradas las variables del ambiente	Presencia, luminosidad, temperatura y humedad

4.2 Extracción de características

En esta etapa se realizó la representación de los eventos y características mediante el modelo espacio vectorial propuesto por Salton [23], para cada par se extraen trece características que se dividen de acuerdo con la información morfológica del evento: la(s)

persona(s) participante(s) en el evento, tiempo en el que ocurre el evento, espacio en el cual ocurre dicho evento, y en el caso de los eventos ambientales, la variación, así como la clase de evento, además con el modelo "Bolsa de palabras", se obtiene el conjunto de características lexicográficas que componen a un evento: verbos, adjetivos y sustantivos. Estas características se describen en la tabla 2.

Tabla 2. Características morfológicas de un evento

Característica	Descripción
Características de agente	Definen al participante que inicia o participa en un evento
Características de tiempo	Describen el tiempo en el que ocurren los eventos
Características de espacio	Describen el espacio físico en el cual ocurren los eventos
Características de magnitud	Describen variables de ambiente en un evento ambiental
Características de clase	Describen el tipo de evento
Características lexicográficas	Definen las palabras más representativas de un evento

La obtención de las características se realizó mediante un pre-procesado de las descripciones de los eventos. Este pre-procesado fue realizado con la herramienta NLTK de Scikit-learn [24] y el módulo Pattern desarrollado por el Centro de investigación CLiPs (Computational Linguistics & Psycholinguistics) [25] en el lenguaje de programación denominado Python.

- **Pre-procesado.** En este paso se seleccionan los datos que serán utilizados en la clasificación. El pre-procesado en este trabajo consiste en segmentar, limpiar, eliminar palabras vacías, lematizar y obtener la lista de características (bolsa de palabras) de un evento.
- **Segmentar.** Tarea que consiste en obtener las cadenas delimitadas por un espacio en blanco.
- **Limpieza.** Proceso que descarta aquellos datos que no aportan información relevante al proceso de clasificación, estos datos son: números, signos de puntuación, caracteres especiales y aquellas palabras que carecen de un significado por si solas, denominadas palabras vacías (stop words), algunos ejemplos son: artículos, preposiciones y conjunciones. Este trabajo utiliza el módulo de stop words de *NLTK* para español.
- **Lematización.** Proceso mediante el cual se eliminan partes no esenciales de una palabra para obtener su forma base, este proceso implica un análisis morfológico de cada palabra en el que se identifica a través de un etiquetado automático (POS tagging) su categoría gramatical. Existen muchas herramientas que permiten realizar esta labor. Sin embargo, las pruebas realizadas a los eventos académicos mostraron ambigüedades en la asociación de una palabra con su categoría gramatical razón por la que se optó por la implementación de un módulo de lematización automático para español. Este módulo se desarrolló en *Python* 2.7 como lenguaje de programación y la identificación de la categoría gramatical se realiza con la ayuda de *Pattern.es* [25] en su versión

para el español. A la variación que sufre una palabra dependiendo de su género, número o tamaño se le conoce como flexión, en español forman flexión nominal los adjetivos, sustantivos y pronombres con los morfemas flexivos de género y número (masculino, femenino y singular o plural respectivamente), los verbos lo hacen con la conjugación. La asignación de categorías gramaticales de un verbo se realiza identificando sus diferentes formas verbales que dependiendo de esta se trasladan al infinitivo. La identificación de adjetivos se realiza a través de sus morfemas flexivos de género y número *(-o,-a,-os,-as,-as o -es*), según sea el género se obtiene su forma base y se singularizan. En el caso de los sustantivos se identifican sus morfemas de número *(-s, -es*), en plural para posteriormente singularizar, en su mayoría, los sustantivos son invariables (no cambian de género) son masculinos o femeninos, de estos, se descartan los sustantivos derivados de verbos (expresan acciones, eventos o procesos).

- **Bolsa de palabras.** El modelo "bolsa de palabras" (del inglés, Bag of Words) está compuesto por el conjunto de características lexicográficas obtenidas durante el proceso de lematización, cabe mencionar que en este modelo no se admiten términos repetidos. En la tabla 3 se muestra el listado de características utilizadas para la clasificación de eventos junto con su descripción.

Tabla 3. Conjunto de características

Característica	Descripción	Posibles valores
Número de estudiantes	Total de alumnos participantes en un evento	Valor numérico que indica la cantidad de estudiantes en un evento
Número de profesores	Total de profesores participantes en un evento	Valor numérico que indica la cantidad de profesores en un evento
Número de visitantes	Total de participantes externos a un espacio académico	Valor numérico que indica la cantidad de visitantes en un evento
Total de participantes	Total de participantes en un evento	Valor numérico que indica la cantidad de estudiantes, profesores y visitantes en un evento
Horario inicial del evento	Rango de tiempo en el que sucede un evento	Valor nominal = {turno matutino = 1, turno vespertino = 2, turno intermedio = 3, turno nocturno = 4 }
Horario final del evento	Rango de tiempo en el que sucede un evento	Valor nominal = { turno matutino = 1, turno vespertino = 2, turno intermedio = 3, turno nocturno = 4 }
Tiempo del evento	Tiempo que tarda en desarrollarse un evento	Valor numérico que indica la duración en minutos de un evento
Tipo de espacio	Lugar en el que sucede un evento	Valor booleano que indica si es un espacio interior o al aire libre = {si = 0, no =1}

Tipo de lugar	Espacio físico en el que se desarrolla un evento	Valor nominal = {salón = 1, oficina de un profesor = 2, laboratorio = 3, auditorio = 4, plaza = 5, jardín = 6 }
Variación ambiental	Cambio en luminosidad, temperatura o humedad	Valor numérico que indica la variación en eventos de luminosidad, temperatura y humedad
Clase del evento	Clase a la que pertenece un evento	Valor nominal = {Evento de difusión = 0, Cursos académicos = 1, Asesoría =2, Evento ambiental = 3}
Nombre del evento	Denominación verbal en español que se le asigna a un evento	Valor nominal = {cadena}
Descripción del evento	Narración de corta extensión, en español que se hace sobre un evento	Valor nominal = {cadena}
Verbo	306	Valor numérico que indica el total de verbos en los eventos
Adjetivo	294	Valor numérico que indica el total de adjetivos en los eventos
Sustantivo	1233	Valor numérico que indica el total de sustantivos en los eventos

4.3 Clasificación de eventos

En esta sección se describen las pruebas realizadas con cuatro de los métodos de clasificación más usados en la literatura: Naïve Bayes, KNN, Árboles de decisión (C4.5) y Support vector Machine (SVM). Posteriormente se muestra el score obtenido para cada uno.

Para la realización de las diferentes pruebas se utilizó el esquema de pesado TF – IDF sobre cada conjunto de datos y se dividió en dos pequeños subconjuntos de eventos seleccionados aleatoriamente, el primer grupo corresponde al de entrenamiento con el 70% de los eventos y el segundo con el 30% restante para su evaluación.

4.4 Resultados

Las pruebas se realizaron de manera individual para las características lexicográficas (verbos, adjetivos y sustantivos) y haciendo una combinación de estas, además se hizo la combinación de las características lexicográficas con sus características nominales. Por último, la combinación de todos, es decir, características lexicográficas: verbos, adjetivos, sustantivos y las características nominales de cada evento.

En el conjunto de pruebas llevadas a cabo se puede observar que Naïve Bayes y C4.5 muestran mejores resultados para el conjunto de verbos con una precisión del 67% mientras que C4.5 y SVM obtienen mayor precisión para los adjetivos con un 56% y un 63% respectivamente, Naïve Bayes y C4.5 nuevamente ofrecen mejores resultados

para los sustantivos. En el caso de las pruebas con combinaciones de características, se tiene que combinando verbos, adjetivos, sustantivos y características nominales se ha logrado un 94 % de precisión con el algoritmo C4.5. En el total de pruebas realizadas, se ha concluido que el algoritmo C4.5 obtiene los mejores resultados en la clasificación de eventos académicos como se observa en la Tabla 4 que expone los resultados de precisión obtenidos de la evaluación de cada uno de los algoritmos aplicados con los diferentes conjuntos de características.

Tabla 4. Resultados de precisión

Conjunto de datos	NB	KNN	C4.5	SVM
Verbo	0.67	0.51	0.67	0.60
Adjetivo	0.39	0.44	0.56	0.63
Sustantivo	0.75	0.44	0.90	0.44
Verbo + Adjetivo + Sustantivo	0.77	0.44	0.88	0.44
Nominal + Verbo	0.78	0.91	0.87	0.57
Nominal + Adjetivo	0.78	0.91	0.91	0.52
Nominal + Sustantivo	0.80	0.85	0.90	0.44
Nominal + Verbo + Adjetivo + Sustantivo	0.79	0.85	0.94	0.67

5 Conclusiones

En este artículo se ha presentado un enfoque para la clasificación de eventos académicos utilizando algoritmos de aprendizaje automático basado en sus descripciones textuales. El enfoque que presenta consiste de una etapa de entrenamiento de los modelos de clasificación, donde se utilizan características textuales como la frecuencia de palabras y se hace uso de información morfosintáctica, como la categoría de las palabras (verbos, sustantivos, adjetivos). Los cuatro algoritmos utilizados son Naïve Bayes (NB), *k* vecinos más próximos (KNN), C4.5 y máquinas de soporte vectorial (SVM).

Las principales aportaciones de este trabajo son a) el conjunto de datos sobre eventos académicos etiquetados en cuatro categorías; b) el enfoque para la clasificación automática de eventos académicos basada en sus descripciones textuales en español; c) la comparación de diversos clasificadores combinándolos con diversos tipos de características.

Con la experimentación y resultados, se hace notar que la mejor configuración de experimentos es utilizando el algoritmo de árboles de decisión (C4.5) y haciendo uso de todas las características: nominales, verbos, adjetivos y sustantivos. Esta configuración ha logrado un 94 % de precisión en la tarea de clasificación de eventos académicos.

Los resultados de este trabajo son de gran utilidad para los analistas de eventos académicos, debido a que ellos realizan un análisis y categorización de este tipo de eventos de manera manual. El enfoque propuesto en este artículo apoyaría en disminuir los tiempos de análisis de eventos desde que propone un razonamiento automático a partir de sus descripciones textuales.

Como trabajo a futuro se propone la experimentación con eventos de otros dominios como la medicina, la política y seguridad. Además, se propone el modelo de n-gramas y n-gramas sintácticos, por su simplicidad e independencia de idioma. Un sistema de clasificación automática de eventos con comunicación directa con los usuarios, sería de gran utilidad para la comunidad que se dedica al análisis de eventos.

Referencias

1. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: 33rd international Proceedings ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841-842. ACM, New York (2010)
2. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on twitter. In: Fifth AAAI International Conference on Weblogs and Social Media (2011)
3. Ritter, A., Etzioni, O., Clark, S.: Open Domain Event Extraction from Twitter. In: 18th ACM SIGKDD International conference on Knowledge discovery and data mining, pp. 1104-1112. ACM, New York, NY, USA (2012)
4. Zhou, D., Chen, L., He, Y.: An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
5. Miwa, M., Sætre, R., Kim, J. D., Tsujii, J. I.: Event Extraction with Complex Event Classification Using Rich Features. Journal of Bioinformatics and Computational Biology. 8(1), 131-146 (2010)
6. Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E. M., Kors, J. A.: Knowledge-Based Extraction of Adverse Drug Events from Biomedical Text. BMC bioinformatics. 15(1), 64 (2014)
7. Sohn, S., Wagholikar, K. B., Li, D., Jonnalagadda, S. R., Tao, C., Komandur Elayavilli, R., Liu, H.: Comprehensive Temporal Information Detection from Clinical Text: Medical Events, Time, and TLINK Identification. Journal of the American Medical Informatics Association. 20(5), 836-842 (2013)
8. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A.: Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. Briefings in Bioinformatics. 6(3), 239-251 (2005)
9. Reyes, J. A., Montes, A., González, J. G., Pinto, D. E.: Clasificación de Roles Semánticos Usando Características Sintácticas, Semánticas y Contextuales. J. Comp. y Sist. 17(2), 263-272 (2013)
10. Molina, J., García, J.: Técnicas de Minería de Datos basadas en Aprendizaje Automático. Técnicas de Análisis de Datos, 96-266 (2008)
11. Friedman, N., Geiger, D., Goldszmidt M.: Bayesian Network Classifiers. Mach. Learn. 29, 131-163 (1997)
12. Viera, A. F. G.: Técnicas de aprendizaje de Máquina Utilizadas para la Minería de Texto. Investigación bibliotecológica. 31(71), 103-126 (2017)
13. Cover, T. M., Hart, P. E.: Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theor. 13, 21-27 (1967)

14. Quinlan J. R.: Induction of Decision Trees, J. Mach. Learn. 1 (1), 81-106 (1986)
15. Ross, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
16. Shafer, J.C., Agrawal, R., Mehta, M.: SPRINT: A Scalable Parallel Classifier for Data Mining. In: 22th International Conference on Very Large Data Bases, pp. 544–555. Morgan Kaufmann, San Francisco (1996)
17. Mehta, M., Agrawal, R., Rissanen, J.: SLIQ: A Fast Scalable Classifier for Data Mining. In: 5th International Conference on Extending Database Technology: Advances in Database Technology, Springer-Verlag, London, UK, pp. 18-32. (1996)
18. Rastogi, R.,Shim, K.:PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning. In: 24th International Conference on Very Large Data Bases, pp. 24-27. Morgan Kaufmann, San Francisco (1998)
19. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., Scholkopf, B.: Support Vector Machines. IEEE Intelligent Systems. 13(4), 18-28 (1998)
20. Salton, G., McGill, M. J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1986)
21. Robertson, S.: Understanding Inverse Document Frequency: on Theoretical Arguments for IDF. Journal of documentation. 60(5), 503-520 (2004)
22. Salton, G., Yang, C. S.: On the Specification of Term Values in Automatic Indexing. Journal of documentation. 29(4), 351-372 (1973)
23. Salton, G.: Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Longman Publishing Co., Boston (1989)
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J.: Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825-2830. (2011)
25. Smedt,T. D., Daelemans, W.: Pattern for Python. J. Mach. Learn. Res., 13, 2063-2067 (2012)