ISSN 1870-4069

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.

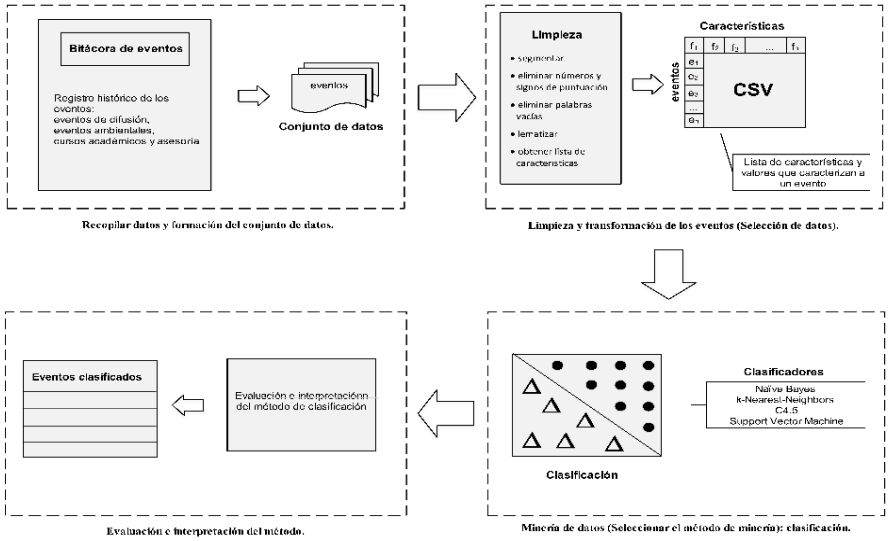

Fig. 1. Proceso de clasificación de eventos.

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:

$$F = \{f_1, f_2, f_3 \ldots f_n\}. \quad (1)$$

Los eventos son el conjunto de todos los eventos:

$$E = \{e_1, e_2, e_3 \ldots e_n\}. \quad (2)$$

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:

$$tf - idf_{ij} = f_{ij} * \log\left(\frac{N}{\mathrm{df}}\right), \tag{3}$$

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:

$$Precisión = \frac{Número\ de\ eventos\ clasificados\ correctamente}{Total\ de\ eventos}. \tag{4}$$

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.

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.

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

–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), pp. 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), pp. 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), pp. 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), pp. 263–272 (2013)
10. Molina, J., García, J.: Técnicas de Minería de Datos basadas en Aprendizaje Automático. Técnicas de Análisis de Datos, pp. 96-–66 (2008)
11. Friedman, N., Geiger, D., Goldszmidt M.: Bayesian Network Classifiers. Mach. Learn. 29, pp. 131–163 (1997)
12. Viera, A. F. G.: Técnicas de aprendizaje de Máquina Utilizadas para la Minería de Texto. Investigación bibliotecológica. 31(71), pp. 103–126 (2017)
13. Cover, T. M., Hart, P. E.: Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theor. 13, pp. 21–27 (1967)
14. Quinlan J. R.: Induction of Decision Trees, J. Mach. Learn. 1 (1), pp. 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), pp. 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), pp. 503–520 (2004)
22. Salton, G., Yang, C. S.: On the Specification of Term Values in Automatic Indexing. Journal of documentation. 29(4), pp. 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, pp. 2825–2830 (2011)
25. Smedt,T. D., Daelemans, W.: Pattern for Python. J. Mach. Learn. Res., 13, pp. 2063–2067 (2012)