Herramienta para mejorar la toma de decisiones en la Dirección de Posgrado de la UCI

Dainys Gainza Reyes [1*], Henry Raúl González Brito [2]
[1] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km 2 ½, reparto Torrens, municipio Boyeros, La Habana, Cuba. dgainza@uci.cu
[2] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km 2 ½, reparto Torrens, municipio Boyeros, La Habana, Cuba. henryraul@uci.cu
[*] Autor para correspondencia: dgainza@uci.cu

Resumen
La colaboración entre la educación superior y los gobiernos para impulsar el desarrollo local mediante sistemas de innovación en diferentes comunidades del país es de vital importancia. En este contexto, la Universidad de Ciencias Informáticas (UCI) ha actualizado su misión y ha ofrecido cursos de posgrado en diferentes formatos para complementar, profundizar o actualizar la formación profesional de los profesionales cubanos. En respuesta a la pandemia de COVID-19, la UCI ha desarrollado las Escuelas de Posgrado a Distancia, eventos de carácter internacional que han tenido un gran impacto en todo el país. Estos cursos de posgrado se ofrecen en diferentes formatos, incluyendo modalidades presenciales, semipresenciales y a distancia, las mismas han sido exitosas al atraer a más de 7000 profesionales cubanos en más de 13 eventos. Sin embargo, el procesamiento de las encuestas de satisfacción realizadas por los cursistas se realiza manualmente, lo que limita la capacidad de la Dirección de Educación de Posgrado (DEP) de tomar decisiones importantes relacionadas con la calidad de los cursos y la evaluación de su apertura. Por lo que se hace necesario proporcionar a la DEP de una herramienta que le permita mejorar la toma de decisiones en temas relacionados con las escuelas de posgrado a distancia según los datos recopilados en las encuestas de satisfacción utilizando técnicas de machine learning como el análisis de sentimiento y el análisis de tópicos

Palabras clave: Machine learning, aprendizaje automático, análisis de sentimientos, análisis de tópico

Temática: Competencias digitales en la Educación Superior

Introducción (Times New Roman, negritas, 12 puntos)

La educación superior y los gobiernos pueden colaborar para impulsar el desarrollo local mediante la implementación de sistemas de innovación en provincias, municipios y comunidades a lo largo del país (Díaz-Canel y González, 2020). Con el objetivo de fortalecer y mejorar su conexión con la agenda digital cubana, la UCI ha actualizado su misión. Una de sus estrategias principales ha sido la oferta de cursos de posgrado en formatos, como presenciales, semipresenciales y a distancia. Los mismos tienen como finalidad complementar, profundizar o actualizar la formación profesional a través de temáticas interesantes basadas en resultados de investigación relevantes y elementos esenciales para mejorar el desempeño laboral.

Desde el año 2013, la UCI ha organizado Escuelas Internacionales de Invierno y Verano, que ofrecen actividades de posgrado enfocadas principalmente a profesionales de la informática y disciplinas relacionadas. Hasta marzo de 2020, estas Escuelas se llevaban a cabo en modalidad presencial. El contexto generado por la pandemia de COVID-19, aunque limitó las actividades presenciales en las universidades cubanas y destacó la brecha digital que afecta al sistema educativo, impulsó una mayor demanda de actividades de capacitación virtual en Cuba. Este desafío se convirtió en una oportunidad para crear o renovar formas de capacitación para el cuerpo docente, los graduados y los profesionales de la informática y disciplinas afines. Una respuesta efectiva fue el desarrollo de las Escuelas de Posgrado a Distancia, eventos de carácter internacional que han tenido un gran impacto en todo el país al haber atraído a más de 7000 profesionales cubanos en más de 13 eventos.

Desde el 2020 hasta diciembre del 2024 se han desarrollado exitosamente 13 Escuelas de posgrado a distancia y dentro de ella 4 escuelas de verano y 4 escuelas de invierno. Cada una de ella cuenta con las encuestas de satisfacción realizadas por los cursistas dónde expresan su grado de satisfacción. Estas encuestas son de vitales importancias para la Dirección de Educación de Posgrado (DEP) de la UCI, pues le permite tomar decisiones importantes relacionadas con la calidad de los cursos y evaluar la apertura o no de algunos cursos teniendo en cuenta las respuestas obtenidas en las mismas.

Las encuestas de satisfacción actualmente están compuestas por 16 ítems agrupados en 3 categorías (aspectos académicos, recursos empleados y valoración general). Se trabajó con variables cuantitativas de tipo discretas que usan una escala de Linker en 15 de los ítems, por lo que para su procesamiento puede usarse un análisis descriptivo usando gráficos y tablas donde se muestren la distribución de frecuencia de ocurrencia, ya sea relativa o absoluta. Sin embargo, el ítem 16, da la

posibilidad al cursistas de expresar sus consideraciones sobre el curso, por lo que en estos momentos es considerada por parte de la DEP la variable más importante de la encuesta y de la que se pueden sacar más conclusiones.

Actualmente el procesamiento de estas encuestas se realiza manualmente por parte de un especialista de la DEP, por lo que la toma de decisiones se afecta, pues el tiempo que se dispone para este procesamiento es limitado y por lo general no se hacen análisis profundo del ítem 16 por lo complicado que resulta poder agrupar y clasificar de manera manual todas las opiniones, por lo que la DEP se ha visto en la necesidad de buscar soluciones automatizadas que te permitan procesar estas encuestas en menor tiempo de manera que posibilite la toma de decisiones oportunas por parte de la dirección.

Al estudiar la bibliografía se puedo constatar que existen métodos que pueden utilizarse para el tratamiento de las variables discretas como son: Análisis de frecuencias (American Psychological Association, 2020) y Análisis de regresión (Field, 2013). Para el caso de la variable cualitativa los autores recomiendan el uso de técnicas de aprendizaje automático o (Machine Learning) por sus siglas en inglés. Dentro de las técnicas más usadas tenemos: Análisis de sentimiento (Pang & Lee, 2008), la Clasificación de datos (Alpaydin, 2010), y el análisis de tópicos (Blei., Ng, & Jordan, 2003).

La utilización de técnicas de machine learning para el procesamiento de encuestas es cada vez más frecuente en la actualidad (Jordan & Mitchell, 2015). En el caso específico del procesamiento de encuestas, el machine learning puede ser utilizado para analizar grandes cantidades de datos de manera automatizada, lo que permite ahorrar tiempo y esfuerzo en comparación con métodos tradicionales de análisis manual.

Autores como (Pang & Lee., 2008; Blei, Ng, & Jordan, 2003) plantean que dentro de las principales y más usadas técnicas de machine learning para el análisis de opiniones tenemos el análisis de sentimiento y el análisis de tópicos. Estos análisis se llevan a cabo mediante la aplicación de técnicas avanzadas de procesamiento de lenguaje natural (PNL), aprendizaje automático e inteligencia artificial (IA). El análisis de sentimiento es un método que utiliza técnicas de aprendizaje supervisado. Puede ser útil para identificar patrones en las respuestas y para clasificar a los encuestados según su actitud hacia el tema de la encuesta. Por su parte el análisis de tópicos (Topic Analysis en inglés) es una técnica de aprendizaje no supervisado, que se utiliza para identificar los temas principales o tópicos que se tratan en un conjunto de documentos. Este análisis se realiza a partir de un modelo matemático que identifica patrones en el uso de palabras y frases en los documentos

Son muchas las técnicas usadas para realizar análisis de sentimientos y análisis de tópicos, continuación, se resume lo planteado por diferentes autores sobre el tema.

Técnicas de análisis de sentimientos:

Baccianella, Esuli, & Sebastiani (2010) plantean que una de las técnicas más usada es el Análisis léxico, esta técnica utiliza listas de palabras positivas y negativas para asignar una puntuación a un texto en función de la cantidad de palabras positivas y negativas que contiene. Dentro de las ventajas de usar dicha técnica tenemos que es una técnica sencilla y rápida de implementar y no requiere un conjunto de datos de entrenamiento. Mientras que dentro de los inconvenientes tenemos que puede no ser precisa en textos que contienen ironía o sarcasmo y puede dar resultados inexactos en idiomas diferentes al del léxico utilizado.

Pang & Lee (2008) plantean que otra técnica muy usada es el aprendizaje supervisado, esta técnica utiliza un conjunto de datos de entrenamiento que consiste en textos etiquetados con su polaridad (positiva, negativa o neutra) para entrenar un modelo de aprendizaje automático, como un clasificador de Naive Bayes. Dentro de las ventajas del uso de esta técnica tenemos que puede ser muy precisa en textos de dominio específico y puede manejar ironía y sarcasmo con mayor precisión que el análisis léxico, sin embargo, como elementos negativos se puede mencionar que requiere un conjunto de datos de entrenamiento etiquetado que puede ser costoso y difícil de obtener y puede ser sensible a la calidad y cantidad del conjunto de datos de entrenamiento.

Por su parte, Tang, Qin, & Liu (2015) plantean que pueden usarse también las redes neuronales, las cuales son modelos de aprendizaje profundo que se utilizan para el análisis de sentimientos. Estos modelos se entrenan con un conjunto de datos etiquetados y utilizan una arquitectura de red para aprender patrones en los datos y predecir la polaridad de nuevos textos. Dentro de las ventajas que presentan se tiene que puede manejar texto no estructurado y de diferentes dominios con alta precisión y puede aprender patrones complejos en los datos de entrenamiento. Dentro de los inconvenientes se tienen que requiere un conjunto de datos de entrenamiento etiquetado que puede ser costoso y difícil de obtener y puede ser computacionalmente intensivo y requiere hardware especializado.

Técnicas de análisis de tópicos: Varios autores como (Blei, Ng, & Jordan, 2003; Ramage, Dumais, & Liebling2010; Pang & Lee 2008) concuerdan que dentro de las técnicas más reconocidas para el análisis de tópico tenemos: El Análisis de frecuencia de palabras, esta técnica consiste en identificar las palabras más frecuentes en un texto y agruparlas en temas. Dentro de las ventajas que ofrece podemos decir que es una técnica simple y fácil de usar y utiliza el aprendizaje no supervisado, sin embargo, no considera el contexto y puede ser engañosa debido a las palabras comunes que no tienen relación con el tema.

El análisis de co-ocurrencia de palabras es una técnica que se basa en la identificación de las palabras que aparecen juntas con mayor frecuencia en un texto y agruparlas en temas. Es una técnica más precisa que la anterior, ya que considera la relación entre palabras, pero también puede ser engañosa debido a la frecuencia de las palabras irrelevantes.
El Análisis de Latent Dirichlet Allocation (LDA) es un modelo probabilístico que identifica los temas en un texto basándose en la distribución de palabras en el texto. Es una técnica precisa y ampliamente utilizada, pero requiere una gran cantidad de datos para entrenar el modelo y puede ser complicada de implementar.
El Análisis de Redes Semánticas (SNA) se basa en la identificación de las palabras clave y su relación en un texto para crear una red semántica. Es una técnica útil para identificar la relación entre temas, pero puede ser limitada debido a la falta de datos para crear una red semántica completa. Por lo general estás técnicas de análisis de tópicos están basadas en el uso de aprendizaje no supervisado para su funcionamiento.
Para el procesamiento de las variables discretas los autores (García, 2018 y Martínez, 2019) proponen como métodos más usados los de Análisis de frecuencia y el Análisis de regresión. El análisis de frecuencia se utiliza para describir la distribución de frecuencia de una variable, es decir, para contar cuántas veces ocurre cada valor posible de la variable en un conjunto de datos. Por su parte el análisis de regresión se utiliza para analizar la relación entre una variable discreta y otra variable continua o discreta. El objetivo es encontrar una función matemática que describa la relación entre las variables y que pueda utilizarse para predecir los valores de la variable discreta en función de los valores de la variable independiente.
Teniendo en cuenta lo antes mencionado por los autores para la realización de este trabajo se decidió utilizar dentro de las técnicas de machine learning las de análisis de sentimiento y de tópico por las ventajas que ofrecen y ser las que más se adaptan al entorno de trabajo de la DEP. Para realizar el análisis de sentimiento se utilizará la técnica de análisis léxico por las ventajas descritas y teniendo en cuenta que la base de conocimiento es limitada y para el análisis de tópicos se usará el LDA teniendo en cuenta que es de los métodos más usados a nivel mundial y es muy precisa. Para las variables discretas se utilizará el análisis de frecuencia teniendo en cuenta que se busca saber la frecuencia de ocurrencia de un determinado evento, teniendo como objetivo proporcionar a la Dirección de Educación de Posgrado de la UCI una herramienta que le permita mejorar la toma de decisiones en los temas referentes a las escuelas de posgrados a distancia de verano e invierno que se ofertan como parte de las actividades de posgrado de la universidad. y que pueda predecir algunos comportamientos de acuerdo a los datos recogidos en las encuestas de satisfacción realizada a los cursistas utilizando técnicas de machine learning.

Desarrollo, Materiales y métodos o Metodología (Times New Roman, negritas, 12 puntos)

Se utilizó Python, un lenguaje de programación, el cual ofrece una amplia variedad de librerías específicas para el procesamiento de texto y la ciencia de datos. Para implementar el dashboard, se seleccionó PyCharm como el entorno de desarrollo integrado. En cuanto a los elementos fundamentales de la investigación, como el análisis de sentimientos y la detección de tópicos todos fueron implementados en Python. Para desarrollar aplicaciones web personalizadas para ciencia de datos y aprendizaje automático, se utilizó Streamlit, una biblioteca Python de código abierto. Streamlit permite a los científicos de datos crear y compartir aplicaciones web de manera rápida y sencilla sin necesidad de aprender tecnologías web adicionales como HTML, CSS y JS. Una de las principales ventajas de Streamlit es su naturaleza pura de Python, lo que lo hace más accesible para aquellos que no están familiarizados con los marcos de aplicaciones GUI tradicionales. Aunque Streamlit funciona de manera diferente a una interfaz de usuario tradicional al ejecutar el script de arriba a abajo cada vez que se realiza una acción de la interfaz de usuario, su almacenamiento en caché agresivo a través del decorador @cached permite una ejecución eficiente al reejecutar solo el código necesario en cada cambio, siempre y cuando los argumentos sean hashables.
Para dar inicio a la investigación, el primer paso fue recolectar los datos necesarios. En este caso, la DEP de la universidad contaba con todas las encuestas de satisfacción de los cursos ofertados en las últimas 11 escuelas de posgrado a distancia, incluyendo dos escuelas de verano y dos de invierno. Luego de recopilar los datos, se creó el dataset en formato cvs, que contenía el nombre del curso, su identificador y una columna para cada pregunta de la encuesta.La siguiente etapa consistió en la minería de texto, aplicando herramientas del procesamiento del lenguaje natural. El preprocesamiento de los datos, que implica la limpieza de los mismos, fue una tarea fundamental en este proceso. Es importante destacar que la limpieza y el preprocesamiento de los datos suelen representar el 80% del tiempo invertido en un proceso de ciencia de datos, según el principio de Pareto.
Para las variables discretas se utilizó el análisis de frecuencia para brindar a la DEP gráficos y tablas que le posibilitaran observar mejor el comportamiento de dichas variables. De esta manera es más fácil para la DEP poder sacar conclusiones sobre dichas variables. Para el caso de la pregunta 16 se realiza análisis de tópico y sentimientos.

Resultados y discusión (Times New Roman, negritas, 12 puntos)

Para la implementación del dashboard lo primero fue la identificación de temas en la encuesta, para ello se utilizó el método de aprendizaje no supervisado topicmodels, basado en el algoritmo matemático LDA. Este algoritmo se basa en la idea de que cada documento está compuesto por varios temas, y que cada tema está representado por un conjunto de palabras que co-ocurren juntas en la estructura de bag of words. En la generación de temas, se sigue una lógica inversa, planteando que no son las palabras las que determinan los temas, sino al revés. Cada palabra tiene un peso distinto en los diferentes temas, lo que implica que pueden ser más relevantes en algunos temas que en otros. Cada palabra es el resultado de un encadenamiento de distribuciones y se realiza la inferencia hacia atrás para calcular la distribución más probable, dadas las palabras y los documentos.

Es importante destacar que el preprocesamiento es fundamental en todos los problemas de aprendizaje automático. En el caso del modelado de temas, el preprocesamiento implica la eliminación de elementos que pueden interferir con la identificación de los temas, como las entidades nombradas, las palabras irrelevantes y los caracteres especiales. En el caso específico de la investigación, se desea eliminar varios tipos de entidades nombradas y caracteres espaciales. Para esto, se utilizó una función que elimina el ruido de los documentos de texto, utilizando las librerías gensim y nltk, se Tokenizó el texto en palabras para permitir la vectorización. Se usó la librería syuzhet, que opera con un diccionario de términos en español como parte del NRC Word-Emotion Association Lexicon, compuesto por una lista de palabras y sus asociaciones.

El último paso en el preprocesamiento de documentos fue usar spacy para realizar la lematización. La lematización es un proceso de normalización de palabras que consiste en reducir las palabras a su forma base o lema, con el objetivo de agrupar las palabras que tienen una raíz común. En otras palabras, la lematización busca reducir las palabras a su forma esencial sin perder su significado. La lematización se utiliza comúnmente en el procesamiento del lenguaje natural, especialmente en tareas como el modelado de temas, la clasificación de texto y la recuperación de información. Al igual que otras técnicas de preprocesamiento, la lematización puede ayudar a mejorar la calidad de los datos y reducir el ruido, lo que a su vez puede mejorar la precisión y eficiencia en el análisis de datos.

Luego se crea y entrena el modelo de tema utilizando la librería Gensim. Para la visualización, nos basamos en el artículo Visualización de modelado de temas: ¿Cómo presentar los resultados de los modelos LDA?, específicamente para las visualizaciones de resultados del modelo de temas. La visualización más compleja realizada fue la de nubes de palabras y dado que ya había un paquete de Python para hacer precisamente eso, la tarea resultó muy sencilla. Se muestran varios gráficos a la DEP para que puedan tomar decisiones.

El análisis de sentimiento se realizó se usando la librería Textblob que opera con un diccionario de términos en español como parte del NRC Word-Emotion Association Lexicon, compuesto por una lista de palabras y sus asociaciones con ocho emociones (ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y disgusto) y tres sentimientos (positivo, negativo y neutro) lo que posibilita la clasificación de los temas. La realización de este análisis es importante ya que permite saber cuál es la satisfacción de una persona en un tema determinado, además permite realizar un análisis de riesgo y oportunidades sobre los temas planteados. Esto permitió clasificar las opiniones.

Conclusiones (Times New Roman, negritas, 12 puntos)

Con el presente trabajo la DEP cuenta con una herramienta útil para la toma de decisiones en sus actividades de posgrado a distancia, de manera que se potencie la transformación digital de la sociedad, teniendo en cuenta que se potencian los cursos más solicitados y mejores puntuados.

Referencias bibliográficas (Times New Roman, negritas, 12 puntos)

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). MIT Press.

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). https://doi.org/10.1037/0000165-000

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. LREC 2010, 2200-2204.

Díaz-Canel Bermúdez, M. y Fernández González, A. (2020). Gestión de gobierno, educación superior, ciencia, innovación y desarrollo local. Retos de la Dirección 14 (2), 5-32

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

García, J. R. (2018). Análisis de frecuencia de variables discretas. Revista de Estadística Aplicada, 54(2), 145-158. https://doi.org/10.1016/j.resap.2018.06.002

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260. Kelleher, J. D., & Tierney, B. (2018). Data science: An introduction (1st ed.). CRC Press. Murphy, K. P. (2012). Machine learning: A probabilistic perspective (1st ed.). MIT Press.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135. https://doi.org/10.1561/1500000011
Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. In Proceedings of the fourth international AAAI conference on weblogs and social media (pp. 130-137).
Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1422-1432.