Textual Data Analysis in University Surveys to Understand and Reduce Student Dropouts

Herrera, M.[1,2] [0009-0006-2125-2970], Romagnano, M. [1,2] [0000-0002-8194-6994] and Ruiz, S. [1] [0009-0005-9817-9997]

[1] Department of Mathematics and Chemistry, Faculty of Exact, Physical and Natural Sciences, National University of San Juan, San Juan City, Argentina

[2] Informatics Institute, Faculty of Exact, Physical and Natural Sciences, National University of San Juan, San Juan City, Argentina
myiramhrrr@gmail.com – maritaroma@gmail.com – sbruizr@yahoo.com.ar

Abstract. In Argentine universities, the management of student data is a critical issue that needs to be addressed immediately.
These educational institutions collect a variety and quantity of data, such as the total number of students enrolled, the most chosen career/s, the dropout rate, among others. However, the retrieval, recording and analysis of these data is often inefficient and disorganized because many of them are in free textual content format and come from diverse information sources. This abundance of data, while valuable, presents a significant challenge due to its unstructured and heterogeneous nature. That is, how to process textual Big Data to obtain information and then acquire knowledge that can help us make valuable decisions?
In the educational domain, Text Analytics provides valuable information. This paper presents the Textual Data Analysis, collected from student surveys of two careers of the Faculty of Exact, Physical and Natural Sciences of the National University of San Juan. For this purpose, the ALCESTE method (Lexical Analysis of Cooccurrences in Simple Sentences of a Text) and other methods of the textual domain, such as word glossaries, concordances and the selection of the most specific vocabulary of each text, have been combined in order to provide a comparative tool.
As a result, it is shown how the study of the distribution of the lexicon used in a text allows us to detect the structuring of the meanings present in it.

Keywords: Data Management, Student Dropout, Text Analysis, Educational Surveys.

Análisis de Datos Textuales en Encuestas Universitarias para Comprender y Disminuir la Deserción Estudiantil

Resumen. En las universidades argentinas la gestión de los datos estudiantiles es una problemática crítica que requiere ser atendida de inmediato.
Estas instituciones educativas recopilan una variedad y cantidad de datos tales como el total de estudiantes matriculados, la/s carrera/s más elegida/s, la tasa de deserción, entre otros. Sin embargo, la recuperación, registro y el análisis de estos datos, a menudo, es ineficaz y desorganizada debido a que muchos de ellos se encuentran en formato de contenido textual libre y provienen de diversas fuentes de información. Esta abundancia de datos, aunque valiosa, presenta un desafío significativo, debido a su naturaleza desestructurada y heterogénea. Es decir, ¿cómo procesar Big Data textual para obtener información y luego adquirir conocimiento que pueda ayudarnos a tomar valiosas decisiones?
En el ámbito educativo, el Análisis de Texto proporciona información valiosa. En este trabajo se presenta el Análisis de Datos Textuales, relevados a partir de las encuestas estudiantiles de dos carreras de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan. Para ello se ha combinado el método ALCESTE (Análisis Lexical de Coocurrencias en Enunciados Simples de un Texto) y otros métodos del dominio textual, tales como los glosarios de palabras, las concordancias y la selección del vocabulario más específico de cada texto, para así proveer una herramienta comparativa.
Como resultado se muestra cómo el estudio de la distribución del léxico empleado en un texto permite detectar la estructuración de los significados presentes en el mismo.

Palabras clave: Gestión de Datos, Deserción Estudiantil, Análisis de Texto, Encuestas Educativas.

1 Introducción

En las universidades argentinas la gestión de los datos estudiantiles es una problemática crítica que requiere ser atendida de inmediato.

Estas instituciones educativas recopilan una variedad y cantidad de datos, tales como el total de estudiantes matriculados, la/s carrera/s más elegida/s, la tasa de deserción, entre otros. Sin embargo, la recuperación, registro y el análisis de estos datos, a menudo, es ineficaz y desorganizada, debido a que muchos de ellos se encuentran en formato de contenido textual libre, y provienen de diversas fuentes de información.

Esta abundancia de datos, aunque valiosa, presenta un desafío significativo, debido a su naturaleza desestructurada y heterogénea. Es decir, ¿cómo procesar Big Data textual para obtener información y luego adquirir conocimiento que pueda ayudarnos a tomar valiosas decisiones?

El Análisis de Texto es un método que permite transformar este cúmulo de datos en información útil y estructurada, facilitando su manipulación e interpretación. Entre las

técnicas más utilizadas en este campo se incluyen el análisis de sentimientos, la detección de temas y la extracción de palabras clave, que permiten medir opiniones y recolectar comentarios, favoreciendo a la pronta resolución de una situación problemática.

Particularmente, en el ámbito educativo, el Análisis de Texto proporciona información valiosa que puede llevar a la implementación de cambios positivos ante la deserción estudiantil. A través del uso de técnicas de procesamiento de lenguaje natural (NLP) y el análisis cualitativo, se pueden identificar tendencias que proporcionen conocimiento valioso para revisar y mejorar políticas académicas que tiendan a disminuir el abandono de los estudiantes de nivel superior.

Por lo tanto, el Análisis de Texto en encuestas de estudiantes universitarios es una herramienta esencial para comprender mejor las opiniones, experiencias y necesidades de los estudiantes. Este método implica la recolección y evaluación de respuestas abiertas en encuestas, permitiendo extraer patrones y percepciones significativas.

Nos podemos preguntar, ¿cómo el análisis de datos textuales en encuestas abiertas permite identificar percepciones y experiencias de los estudiantes de primer año de carreras LSI y LCC, relevantes para el diseño de estrategias preventivas de la deserción?

En este trabajo se presenta el análisis de datos textuales, relevados a partir de las encuestas realizadas a los estudiantes de las carreras de Licenciatura en Sistemas de Información (LSI) y Licenciatura en Ciencias de la Computación (LCC), de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan, al finalizar el segundo semestre del 2024. Para ello se han combinado varios métodos de análisis de datos textuales, tales como el método ALCESTE (Alba, 2017), cuya utilidad e interés se observa en su rápida difusión y por su aplicación en todas las disciplinas involucradas con el análisis de materiales discursivos; y otros métodos del dominio textual, tales como los glosarios de palabras, las concordancias y la selección del vocabulario más específico de cada texto, para así proveer una herramienta comparativa.

Como resultado se muestra cómo el estudio de la distribución del léxico empleado en un texto permite detectar la estructuración de los significados presentes en el mismo.

2 Antecedentes

El Análisis de Datos Textuales es una aplicación de los métodos de Análisis de Datos en la perspectiva de la escuela francesa, es decir, métodos de análisis multidimensionales exploratorios. Las primeras aplicaciones fueron realizadas por (Benzécri, 1973), quien desarrolló el análisis de correspondencias. Posteriormente, (Lebart, 1982) continuó los desarrollos ante la necesidad de tratar preguntas abiertas con métodos más automáticos que la post codificación manual, que se hacía hasta entonces, y que en la mayoría de los casos aún se sigue realizando. A su vez, el desarrollo de un paquete

informático para el tratamiento de datos textuales, el SPAD.T (Centre international de statistique et d'informatique appliquées et al., 1989), se debe a (Bécue, 1992).

El Análisis de Datos Textuales consiste en aplicar estos métodos, en especial el análisis de correspondencias y la clasificación a tablas específicas, creadas a partir de los datos textuales. Estos métodos se completan con métodos propios del dominio textual como los glosarios de palabras, las concordancias y la selección del vocabulario más específico de cada texto, con lo que se tiene una herramienta comparativa de los mismos (Peralta et al., 2020).

Por otro lado, la Estadística Textual, en pleno desarrollo, se encuentra en la encrucijada de varias disciplinas: Estadística Clásica, Análisis del Discurso, Informática y Procesamiento de Encuestas. De hecho, los investigadores y profesionales de la actualidad tienen que afrontar un desarrollo dual. Por un lado, el de los textos de encuestas, entrevistas, archivos, bases de datos documentales y, por otro, el de las herramientas informáticas de entrada de datos y la gestión de textos. El Análisis Estadístico de Datos Textuales (AEDT) comprende una serie de herramientas que se enmarcan en el análisis estadístico multidimensional descriptivo, frecuentemente llamado "Análisis de datos". El AEDT tiene su origen en los análisis cuantitativos realizados sobre obras literarias, que iban dirigidos al recuento de palabras, el estudio de la distribución del vocabulario, la comparación del léxico empleado por distintos autores o por un mismo autor en diferentes períodos creativos. Las investigaciones realizadas por algunos autores, tales como: (Yule, 1944), (Zipf, 1946), (Guiraud, 1960), (Muller, 1968), y el posterior desarrollo y popularización de la informática se encuentran en la base de los métodos de la denominada estadística textual, que han acabado aplicándose al estudio de los datos textuales en muy diversos ámbitos: historia, literatura, sociología, educación, redes sociales, entre otros.

Según Maldonado et al. (2015), el AEDT surge "de la relación que se ha dado entre el estudio cuantitativo de los textos literarios y la corriente de la estadística moderna llamada análisis de datos" (Bécue et al., 1992); parte del número de ocurrencias de las palabras contenidas en el conjunto de textos provenientes de libros, artículos periodísticos o a partir de las respuestas a preguntas abiertas y su relación con características propias de los encuestados, entre otros. La cadena de tratamiento estadístico habitualmente sigue cuatro etapas: problema, datos, tratamiento e interpretación; este esquema marca las actividades que un estadístico debe seguir, aunque en la práctica las situaciones que se presentan se pueden enriquecer con un Análisis Multivariado (Lebart et al., 2000).

3 Metodología

En esta investigación, se utilizaron datos textuales tomados de las encuestas realizadas a los estudiantes de las carreras de Licenciatura en Sistemas de Información y Ciencias de la Computación, de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan, al finalizar el segundo semestre del 2024. Se muestra cómo el estudio de la distribución del léxico empleado en un texto permitió

detectar la estructuración de los significados presentes en el mismo. Para ello se siguió la metodología ALCESTE, que se inspira igualmente en los métodos de análisis de datos de la escuela francesa, basada en técnicas como la clasificación jerárquica descendente o el cálculo de Chi-cuadrado.

Nuestro interés se centró en los métodos de Análisis Factorial de Correspondencias (AFC) y Clasificación Automática, dos métodos exploratorios multivariantes complementarios, adecuados al tratamiento de datos cualitativos.

La propuesta aplicó estos métodos a tablas específicas, creadas a partir de los datos textuales. Estos se completan con métodos propios del dominio textual como los glosarios de palabras, las concordancias y la selección del vocabulario más específico de cada texto, para así proveer una herramienta comparativa.

También, cabe destacar que los métodos a aplicar, cuándo y cómo, depende del tipo de estudio. Este trabajo se centró en el análisis de respuestas abiertas (ARA) ya que el estudiante puede expresarse libremente o “con mayor soltura”.

4 Experimentación y Resultados

Para llevar a cabo esta etapa, se realizaron las siguientes tareas:

4.1 Definición del Corpus

El corpus es el conjunto de datos o textos científicos, literarios, informáticos, jurídicos, periodísticos, etc., que pueden servir de base a la investigación.

Este corpus puede ser, por ejemplo:

- El conjunto de transcripciones de entrevistas realizadas en una investigación.
- Noticias que aparecieron en diferentes diarios sobre una misma temática, o
- Las respuestas abiertas registradas en cuestionarios sobre una misma temática, realizados a estudiantes de diferentes asignaturas, etc.

Para que el análisis que se vaya a realizar tenga sentido, es necesario que el conjunto textual esté centrado en una temática principal o el objeto conceptual de investigación.

El corpus de texto está constituido por un conjunto de textos. La definición de cada uno de estas unidades dependerá de la naturaleza de la investigación.

Retomando los ejemplos anteriores sobre el corpus de textos, se encontró que:

- En un estudio documental sobre noticias, cada una de ellas corresponde a un texto.
- En un estudio con entrevistas, la transcripción de cada una de las entrevistas realizadas a diferentes sujetos de investigación corresponde a un texto. En (Ghiglione et al, 1989) recomiendan entre 20 y 30 textos, siendo 20 textos para cada grupo si se plantean estudios comparativos (Camargo et al., 2013).

4.2 Reducción de las unidades del corpus textual

La reducción de las unidades en segmentos de texto, denominadas Unidades de Contexto Elementales (UCE) representa un texto de dos o tres líneas. El tamaño de las UCE varía de acuerdo al tamaño del corpus. El objetivo del análisis es proponer una clasificación de estas UCE.

4.3 Creación de unidades de contexto (UC)

Una Unidad de Contexto (UC) es un conjunto de Unidades de Contexto Específicas (UCE) que contiene un número mínimo de "formas activas" (verbos, sustantivos, adverbios y adjetivos), en contraste con las "formas suplementarias" o "palabras-herramienta" (preposiciones, pronombres, adjetivos posesivos y algunos verbos y adverbios comunes). Cada UC debe incluir al menos dos formas activas. Para su análisis, se elaboran dos tablas de contingencia: una con las UC en las filas y las formas en las columnas, diseñada para cada número mínimo requerido.

4.4 Lematización

De manera predeterminada, las palabras son sometidas a lematización automáticamente, lo que implica que cada término es reemplazado por su forma canónica o raíz (Lemaire, 2008).

4.5 Construir la matriz de datos

Se trabaja con matrices que suelen mencionarse como “matrices de individuos x variables” dado que las filas de la misma representan a los individuos, personas u objetos bajo estudio (I) y las columnas representan a las variables que se estudian sobre cada uno (J). En el cuerpo de la tabla aparecen los valores numéricos de esas variables o los códigos de las modalidades si se trata de variables nominales (Fig. 1).

En el análisis de datos textuales, las palabras o segmentos cumplen el papel de las modalidades de una variable nominal, la totalidad de las palabras contenidas en las respuestas aparecen como columnas y cada fila corresponde a una persona. En el cuerpo de la tabla aparecen las frecuencias con que cada individuo utilizó cada palabra en su respuesta libre.

Esta matriz permite realizar un estudio de la distribución del vocabulario.

## clase muy	buena	atención	consultas	virtuales
## UC1	1	1	0	1	0
## UC2	0	1	1	0	1
## UC3	1	1	1	0	1

Fig. 1. Ejemplo de matriz binaria de presencia / ausencia del término en el documento. Fuente: Elaboración propia.

4.6 Clasificación jerárquica descendente de Max Reinert

En cada etapa, las UC se particionan en dos grupos con el propósito de maximizar la inercia intragrupos (variabilidad o dispersión de los datos dentro de cada grupo, en otras palabras, es la homogeneidad o heterogeneidad de los grupos.). Este proceso se aplica a tablas de contingencia creadas a partir de las unidades de contexto.

Un proceso aproximado para particionar es:

1. Se lleva a cabo un Análisis Factorial de Correspondencias (AFC) sobre la primera tabla de contingencia, posteriormente las líneas se ordenan según sus aportaciones sobre el primer factor.
2. Se busca a lo largo de este primer factor, la partición en 2 clases que maximiza la inercia (distancia chi-cuadrado) inter-clase.
3. Por último, un algoritmo de intercambio permuta cada línea de una clase a la otra y verifica la variación de inercia inter-clase.

El mismo paso es aplicado sobre otra tabla de contingencia, y así sucesivamente (Marín Riveros y Melo Carrasco, 2020).

Se puede decir que es un procedimiento de clúster jerárquico descendiente, lo cual implica partir de un solo grupo o clase, proporcionado, en este caso, por el análisis factorial de correspondencias, ya que incluye a todos los individuos, y en cada etapa se llevan a cabo las subdivisiones, que llegaría, al final, a representar a cada individuo.

4.7 Análisis de Similitud

Para realizar este análisis se toma como base a la teoría de grafos, donde un grafo es un conjunto de vértices (las palabras o formas) y aristas (la relación entre ellas). Para esto se usó ipysigma (Medium, 2024), que es una biblioteca de Python para renderizar visualizaciones gráficas.

El propósito es el estudio de la proximidad y la relación entre los elementos de un conjunto (formas-lemas), pero reduciendo el número de enlaces hasta llegar a "un gráfico conectado sin ciclo", es decir un camino cerrado en el que no se repite ningún vértice a excepción del primero que aparece dos veces como principio y fin del camino), tal como se aprecia en la Fig 2.

Como se puede observar Fig 2, a la izquierda se muestran todas las posibles conexiones entre cada elemento (formas/palabras). A partir de estos enlaces, se buscará el "árbol de máximo", que se crea a partir de los bordes más fuertes (mayor similitud, peso, asociación etc.) de los gráficos, que es el que presenta la imagen de la derecha. Este es el árbol más simple que se puede lograr, pero es también el más rico en información.

El análisis nos proporciona todos esos árboles mínimos considerando el corpus de texto.

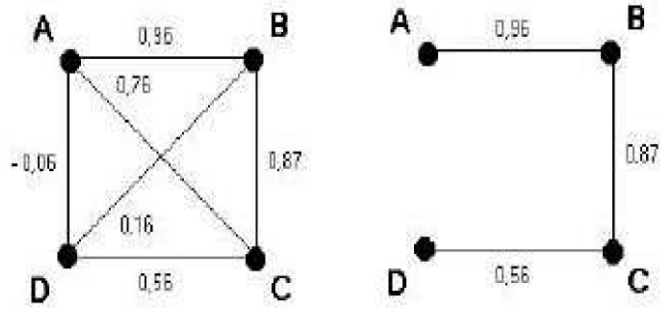

Fig. 2. Análisis de las similitudes. Fuente: Marchand et al., 2012.

4.8 Análisis de discursos producidos por los estudiantes

A continuación, se presenta la aplicación de la metodología a los discursos producidos por 56 estudiantes de LSI y LCC, bajo una encuesta libre, anónima y no estructurada.

Se intentó obtener información sobre el rendimiento y la posible deserción de estudiantes del primer año de las carreras, pero la idea fue que el estudiante se expresara con sus palabras, es decir si estaba o no de acuerdo con la materia. Es por eso que se le realizo una encuesta, en las asignaturas del área Matemática, de primer año, que contaba en un principio con preguntas tales como:

¿Alcanzó a estudiar los contenidos a evaluar en el parcial 1 de las asignaturas? En caso de responder que no, explique por qué no pudo hacerlo. ¿Realizó en su totalidad todas las guías de ejercitación a ser evaluadas en el parcial 1? En caso de responder que no, explique por qué no pudo hacerlo. ¿Qué temas, a ser evaluados en el Parcial 1, no comprendió? ¿Asistió clases de consulta? Si asistió a las clases de consulta indique de qué manera le resultaron beneficiosas. En caso de responder que no, explique por qué no pudo hacerlo, entre otras. Al finalizar se le pidió que expresara todo lo que pensaba de la materia tanto con respecto a los materiales de trabajo, la modalidad de enseñanza, su relación con sus compañeros. ¿Usted dejó de cursar esta asignatura? A continuación, por favor, sólo responder si dejaste de cursar la asignatura ¿Cuál/es fueron los motivos que lo llevaron a dejar de cursar? ¿Ha seguido cursando alguna otra asignatura?

De este modo se pudo definir la Matriz de Dato, cuyas columnas fueron establecidas por elementos del vocabulario y filas por las unidades de contexto.

Para realizar el análisis se usó el software libre IRaMuTeQ (Fig. 3) (Camargo et al., 2013).

Fig. 3. Entorno de trabajo de IRaMuTeQ. Fuente: Elaboración propia.

En la Fig. 4 se puede observar la preparación del material del texto. Entre las variables que se tuvieron en cuenta, se mencionan:

- Carrera que cursan (2 categorías o modalidades).
- Área de conocimiento del curso (2 categorías o modalidades),
- Lugar de procedencia del estudiante (3 categorías o modalidades).

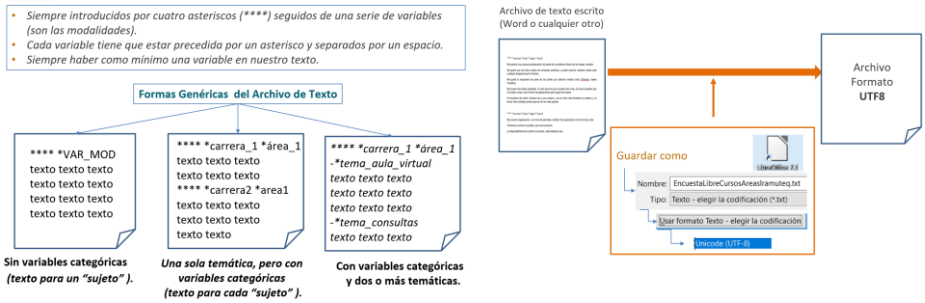

Fig. 4. Preparación del material del texto. Fuente: Elaboración propia.

Luego, en la Fig. 5 se presenta la primera carilla al trabajar con la base de datos: Base_de_Datos_Encuesta_Estudiantes para las asignaturas de primer año.

```
**** *carrera_1 *area_1 *lugar_1 *sexo_1
Me pareció muy buena predisposición de parte de la profesora titular dar las
clases virtuales.
Me gustó que nos diera clases de consultas prácticas y pueda resolver nuestras
dudas ante cualquier pregunta que le hicimos.
Me gustó la respuesta de parte de las profes por distintos medios (mail, Edmodo,
clases virtuales).
Muy bueno las clases grabadas, no solo para los que cursaban día a día, sino
para aquellos que no podían cursar, pero tenían las grabaciones para seguir las
clases.
Al momento de rendir, tomaron de a una unidad, y así se hizo más llevadera la
materia y no tomar más unidades juntas que por ahí es más pesado.

**** *carrera_1 *area_1 *lugar_1 *sexo_2
Muy buena organización, a la hora de parciales, también de organizarse con los
temas a dar.
Teníamos muchas consultas, eso era buenísimo.
La disponibilidad de la profe muy buena, cabe destacar eso.
El apoyo hacia los alumnos muy bueno también.
Me pasaba que quizá veía algo que era difícil y la profe lo explicaba muy bien,
la versatilidad para hacernos comprender muy buena.
Podíamos consultar ya sea por video llamada o por correo, otro punto sobre la
versatilidad de la materia.
La empatía también, la profe consideraba siempre en pasar parciales a otras
fechas para que podamos llegar estudiando ya que sabía que quizá en otras
materias teníamos parciales o trabajos por presentar o ya sea porque no nos
sentíamos
A pesar de no tener una presencialidad, se encargaron de que todo sea muy
eficaz, se valora mucho porque creo que nadie estaba acostumbrado a la
virtualidad y a pesar de todo se las arreglaron muy bien para hacernos aprender.

**** *carrera_1 *area_1 *lugar_1 *sexo_1
La organización del curso fue óptima.
El tiempo entre parciales permitía a los alumnos poder estudiar de forma tal que
no generaba exceso de carga horaria.
Los docentes daban la posibilidad de consultas varios días a la semana.
Facilitaban todos los materiales necesarios.
Los docentes se supieron adaptar a la situación.
```

Fig. 5. Base de datos Encuesta de Estudiantes. Fuente: Elaboración propia.

Posteriormente, una vez cargada la base de datos y configurado el algoritmo para el procesamiento, se presenta una descripción general del corpus, tal como se observa en la Fig. 6.

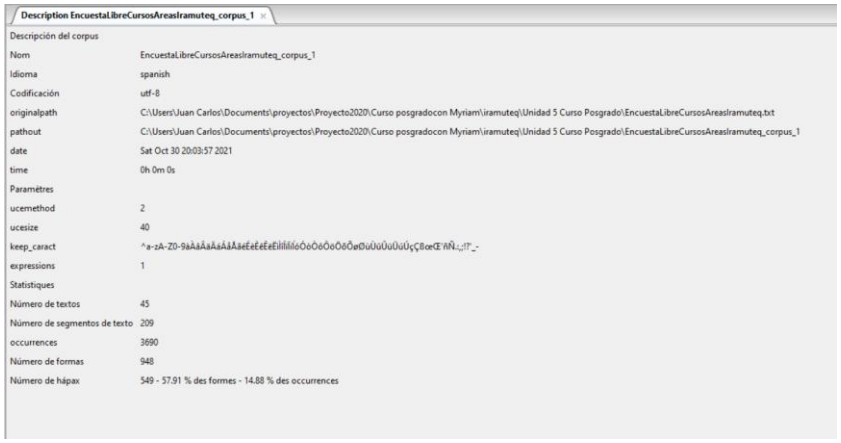

Fig. 6. Descripción del corpus. Fuente: Elaboración propia.

A partir de este momento, el software ofrece la posibilidad de realizar diferentes análisis (Fig. 7).

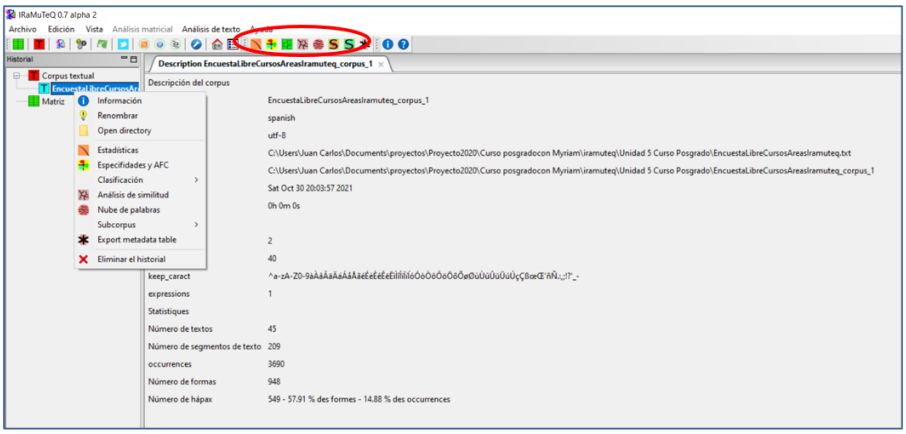

Fig. 7. Posibilidades de análisis. Fuente: Elaboración propia.

Si bien el equipo de investigación realizó un análisis completo, en este trabajo solo se muestra una parte de ellos (Fig. 8, 9 y 10).

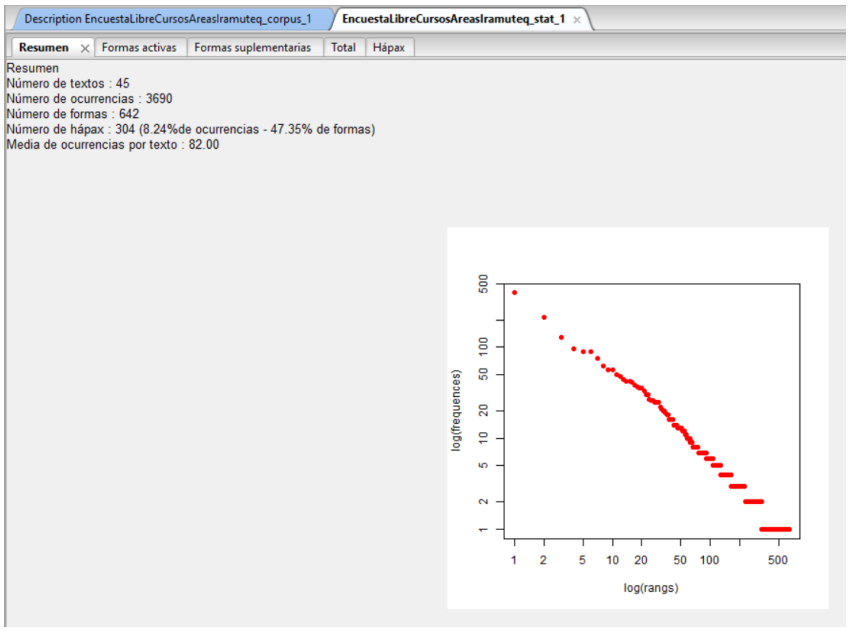

Fig. 8. Análisis de frecuencia. Fuente: Elaboración propia.

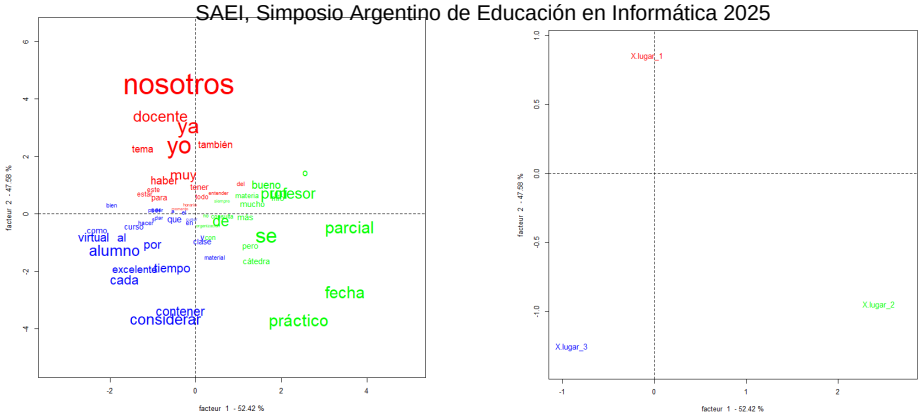

Fig. 9. Análisis de factorial de correspondencia. Fuente: Elaboración propia.

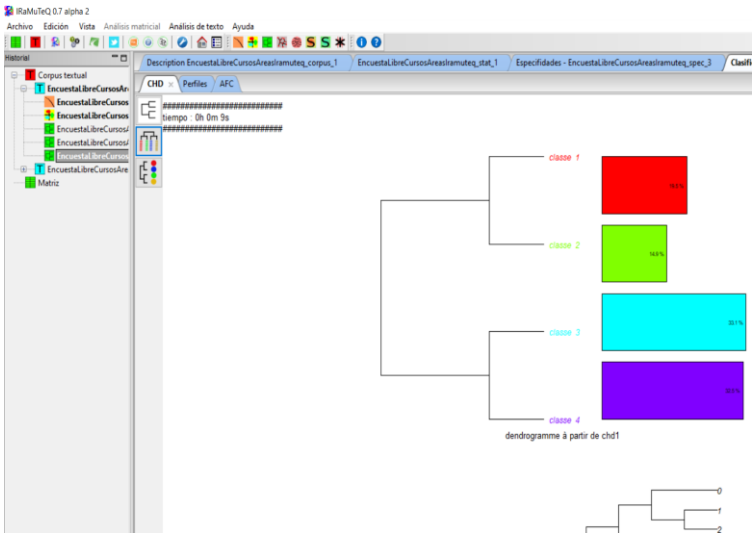

Fig. 10. Dendograma formas específicas de las clases. Fuente: Elaboración propia.

Las formas (unidades léxicas) sirven de orientación para hacer una primera valoración sobre el contenido lexical de cada clase. El tamaño de cada una de las formas orienta sobre la significatividad estadística de la forma dentro de ese mundo léxico (Fig. 11).

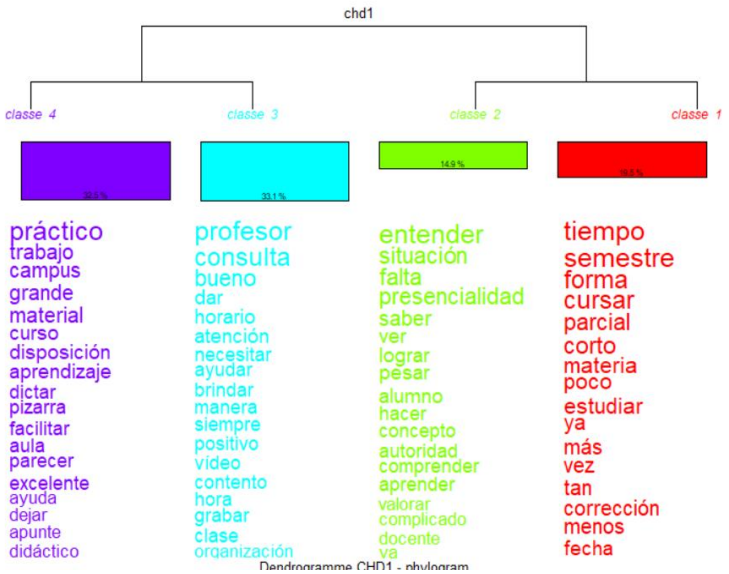

Fig. 11. Formas Léxicas. Fuente: Elaboración propia.

Sin embargo, para un análisis más detallado sobre esta cuestión, es necesario acudir a la segunda pestaña, denominada “perfiles” (Fig. 12).

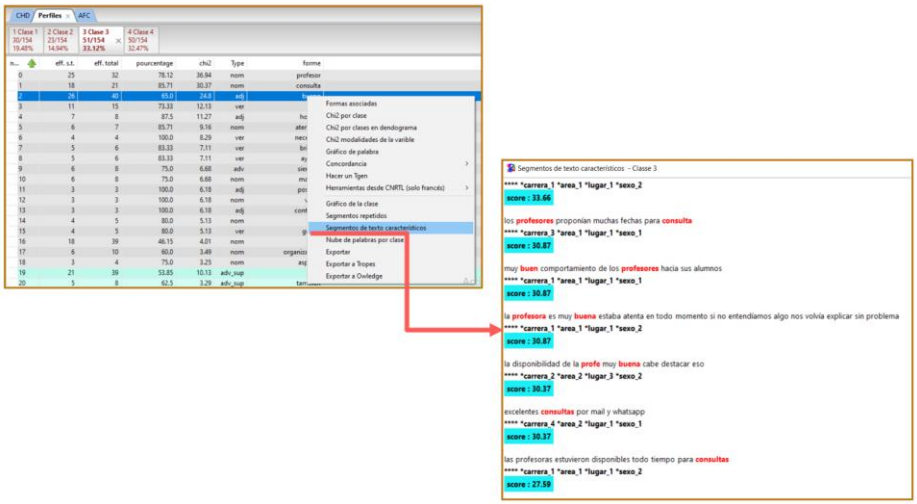

Fig. 12. Análisis de perfiles. Fuente: Elaboración propia.

La Fig. 13 muestra un análisis de similitudes, donde las formas que se encuentran en los nodos de la gráfica y las aristas/enlaces representan la coocurrencia entre ellos. A mayor frecuencia de las palabras, mayor tamaño de las mismas en el gráfico. A mayor coocurrencia entre palabras, más grueso se representa el enlace entre ellas.

Para dar interpretaciones a las relaciones entre las formas, teniendo en cuenta el/los objetivos de investigación cabe destacar: Los estudiantes piensan que el profesor dio buenas clases, entienden lo dado en clase, necesitan tiempo y más consultas.

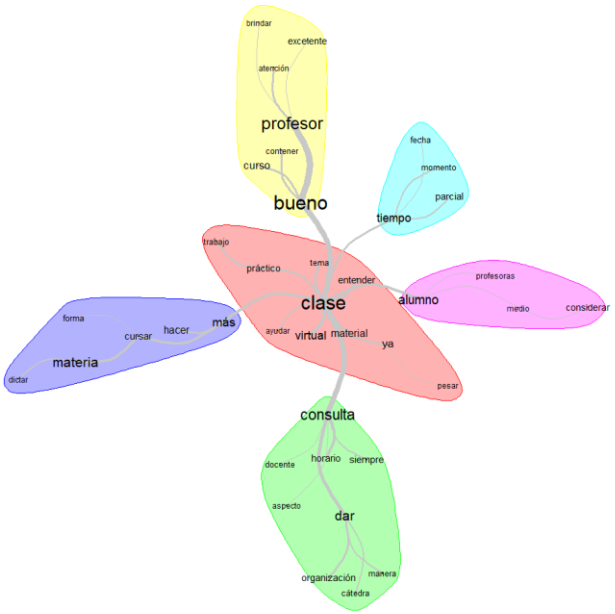

Fig. 13. Análisis de similitudes. Fuente: Elaboración propia.

En la Fig. 14 se presenta el último análisis, la Nube de palabras, en función de la frecuencia de aparición en el Corpus.

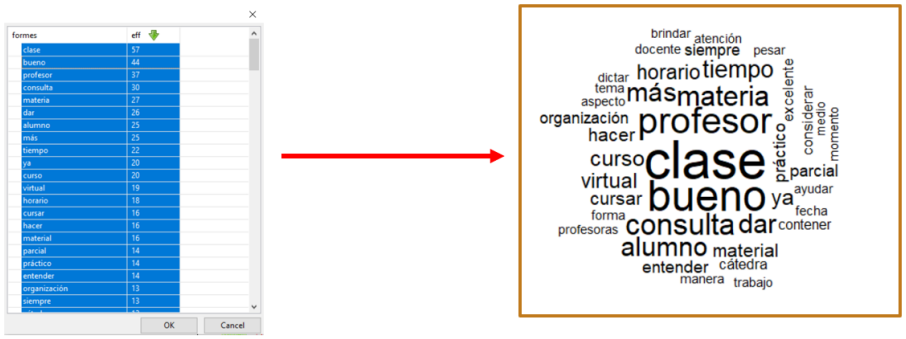

Fig. 14. Nube de palabras. Fuente: Elaboración propia.

5 Conclusión

En este trabajo, se estudió la distribución del léxico empleado en encuestas a estudiantes universitarios para identificar la estructuración de los significados presentes en sus respuestas. El objetivo principal fue explorar las percepciones de los estudiantes respecto a sus estudios y los posibles factores que contribuyen a la deserción, a través del análisis de datos textuales libres.

Se ha evidenciado una problemática persistente en las universidades argentinas, donde la gestión de los datos estudiantiles, crucial para comprender y mitigar la deserción, es a menudo ineficiente debido a la naturaleza desestructurada y heterogénea de la información textual recolectada en encuestas. Si bien se realizan encuestas estructuradas, la presente investigación subraya la necesidad de comprender las experiencias y razones directas de los estudiantes a través de sus propias palabras.

A través de la aplicación de métodos de análisis de datos textuales, incluyendo el método ALCESTE y otras técnicas como glosarios de palabras, concordancias y selección de vocabulario específico, se pudo procesar y transformar este cúmulo de datos en información útil. El estudio se centró en las respuestas abiertas de 56 estudiantes de las carreras de Licenciatura en Sistemas de Información y Licenciatura en Ciencias de la Computación de la Universidad Nacional de San Juan.

El análisis de frecuencia (Fig. 8), la nube de palabras (Fig. 14), y la clasificación jerárquica (Fig. 10 y 12) fueron los principales instrumentos para la interpretación. Específicamente:

Necesidad de Tiempo y Organización para Parciales: Los análisis textuales, incluyendo la nube de palabras (Fig. 14) y las formas léxicas más significativas (Fig. 11 y 12), mostraron consistentemente que los estudiantes “hacen hincapié en el tiempo que necesitan para llegar a rendir los parciales”. Esto se corrobora con la alta frecuencia de términos relacionados con “tiempo” y “organización” en el corpus textual analiza-

do, lo que sugiere que la gestión del tiempo y la preparación para las evaluaciones son desafíos significativos para ellos.

Valoración de Clases y Consultas: Se observó que los estudiantes valoran positivamente “la forma de enseñanza del profesor durante las clases y en las consultas”. El análisis de similitud (Fig. 13) refuerza esta conclusión, al mostrar una fuerte co-ocurrencia entre términos como “profesor”, “clases” y “consultas”, y la presencia de términos como “buenas clases” y la indicación de que “entienden lo dado en clase”. La frecuencia y las conexiones entre estas palabras en los gráficos indican su relevancia en el discurso de los estudiantes, sugiriendo que perciben estos recursos como beneficiosos y necesarios.

En cuanto a la observación de que “los estudiantes universitarios no completan sus carreras”, si bien este trabajo no presenta datos estadísticos directos sobre tasas de finalización de carreras, la investigación se enmarca en la problemática general de la deserción estudiantil universitaria, que es un “desafío que aún no se ha logrado resolver”. El estudio fue diseñado para “intentar obtener información sobre el rendimiento y la posible deserción de estudiantes del primer año de las carreras”. Las preguntas de la encuesta, tales como “¿Usted dejó de cursar esta asignatura?” y “¿Cuál/es fueron los motivos que lo llevaron a dejar de cursar?”, apuntan directamente a recopilar las percepciones de los estudiantes sobre el abandono de asignaturas, lo cual es un indicador relevante de la problemática de la deserción a nivel de carrera. El análisis de estas respuestas textuales busca comprender los “porqué” de la deserción.

Por lo tanto, al prestar mayor atención a la forma de expresión libre de los estudiantes y utilizando metodologías de análisis de datos textuales, se logra una comprensión más profunda de sus opiniones, experiencias y necesidades. Esta aproximación proporciona conocimiento valioso que puede ser utilizado para revisar y mejorar políticas académicas, con el fin último de disminuir el abandono de los estudiantes de nivel superior.

Referencias

Alba Martha (2017). El Método Alceste y su aplicación al estudio de las representaciones sociales. This edition first published 2017© 2017 John Wiley & Sons Ltd.

Benzecri (1973). Análisis de Correspondencia, https://cms.dm.uba.ar/academico/materias/2docuat2017/sem_herr_avan/Analisis%20de%20Correspondencia.pdf, último acceso 15/03/25.

Bécue Bertaut, M. (1992) Análisis de Datos Textuales, Cisia. París, https://revistes.ub.edu/index.php/Anuario-psicologia/article/download/9263/11854/0, último acceso 1/02/25.

Camargo, B. V., Justo, A. M. (2013). IRAMUTEQ: un software libre para el análisis de datos textuales. Temas psicol. [online]. 2013, vol.21, n.2, pp.513-518. ISSN 1413-389X, https://diposit.ub.edu/dspace/bitstream/2445/113063/1/Trabajar_con_IRAMUTEQ_PAUTAS.pdf, último acceso 25/02/25.

Centre international de statistique et d’informatique appliquées, Lebart, L., Morineau, A., & Bécue Bertaut, M. (1989). SPAD. T: système portable pour l’analyse des données textuelles :manuel de l’utilisateur. CISIA.

Ghiglione, R. y Matalon, B. (1989). Las encuestas sociológicas. Teoría y práctica. México: Trillas.

Guiraud, P. (1960). ProblCmes et Méthodes de la Statistique Linguistique. Paris: PUF

Lebart, L. (1982). L' Analyse Statistique des Réponses Libres dans les EnquEtes Socio-économiques. Consornmation, I, pp. 39-62, Paris: Dunod.

Lebart, L., Salem, A., & Bécue M. (2000). Análisis Estadístico de Textos. España: Milenio.

Lemaire, G., Jeuffroy, M.H. and Gastal, F. (2008). Diagnosis Tool for Plant and Crop N Status in Vegetative Stage. Theory and Practices for Crop N Management. European Journal of Agronomy, 28, 614-624, http://dx.doi.org/10.1016/j.eja.2008.01.005, último acceso 20/02/25.

Marchand, P., y Ratinaud, P. (2012). L'analyse de similitude appliqueé aux corpus textueles: les primaires socialistes pour l'election présidentielle francaise. Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles (pp. 687-699).

Maldonado, M. L. H., Palmeros, H. D., & Fernández, A. O. J. (2015). Análisis estadístico de datos textuales aplicado al uso de redes sociales. Revista CPU-e, (21), 1-27, https://www.redalyc.org/pdf/2831/283140301002.pdf, último acceso 1/04/25.

Marín Riveros, J., y Melo Carrasco, D. (2020). El Mediterráneo como "comunidad retórica": Los paratextos prologales y la temprana historiografía árabo-islámica. Estudios filológicos, (65), 153-167. https://dx.doi.org/10.4067/S0071-17132020000100153, último acceso 1/02/25.

Medium (2024). Ipysigma. Easily visualize networks with thousands of nodes and edges in Python. https://medium.com/@msdatashift/ipysigma-easily-visualize-networks-with-thousands-of-nodes-and-edges-in-python-3ecdbe0321, último acceso 26/01/25.

Muller, C. (1968). Initiafion a la Statistique Linguistique. Paris: Larousse.

Peralta, N., Castellaro, M. y Santibáñez, C. (2020). El análisis de datos textuales como metodología para el abordaje de la argumentación: una investigación con estudiantes de pregrado en universidades chilenas. Íkala, Revista de Lenguaje y Cultura, vol. 25, núm. 1, pp. 209-227, 2020. Escuela de Idiomas, Universidad de Antioquia, https://www.redalyc.org/journal/2550/255066212012/html/, último acceso 20/02/25.

Yule, G.U. (1944). A Statistical Stzidy of Vocabulary. Cambridge University Press.

Zipf, G.K. (1935). The Psychobiology of Lunguage, an Zntroduction to DY~~~TZ~C Philology. Eloston: Houghton-Mifflin.