RECIBIDO EL 22 DE JULIO DE 2021 - ACEPTADO EL 23 DE OCTUBRE DE 2021

APLICACIÓN DEL PROCESAMIENTO DEL LENGUAJE NATURAL COMO TÉCNICA DE ANÁLISIS EN LA PRODUCCIÓN TEXTUAL, CASO ESTUDIANTES DE INGENIERÍA DE SISTEMAS UFPSO

APPLICATION OF NATURAL LANGUAGE PROCESSING AS AN ANALYSIS TECHNIQUE IN TEXTUAL PRODUCTION, CASE OF UFPSO SYSTEMS ENGINEERING STUDENTS

Claudia Marcela Durán Chinchilla[1]

Carmen Liceth García Quintero[2]

Alveiro Alonso Rosado Gómez[3]**.**

Universidad Francisco de Paula Santander, Ocaña

1 Doctora en Educación, Magister en Pedagogía, Especialista en Práctica Pedagógica, Licenciada en lingüística. Docente Universidad Francisco de Paula Santander Ocaña .Investigadora asociada y directora del grupo de investigación GIFEAH cmduranc@ufpso.edu.co. https://orcid.org/0000-0001-9291-7841. https://scholar.google.es/citations?user=XszWRgQAAAAJ&hl=es

2 Ms. Practica Pedagógica, Ms en Zoología, especialista en docencia Universitaria, Zootecnista. Docente Universidad Francisco de Paula Santander Ocaña, investigadora del grupo GIADS. clgarciaq@ufpso.edu.co. https://orcid.org/0000-0001-9314-8714. https://scholar.google.com/citations?user=_dE97_oAAAAJ&hl=es

3 Ms en Gestión, Aplicación y Desarrollo de Software, especialista en gestión de proyectos informáticos, investigador del grupo GITYD. aarosadog@ufpso.edu.co https://orcid.org/0000-0003-2932-3383 https://scholar.google.com/citations?user=hLiV3CgAAAAJ&hl=es

RESUMEN

Esta investigación utilizó las palabras, las frases, oraciones y párrafos con la intención de identificar la capacidad que los estudiantes poseen para transmitir información verbal. Para lograr alcanzar este objetivo, se tuvo en cuenta la investigación cuantitativa descriptiva. Como técnica de recolección de información, se aplicó un taller de producción textual, en el cual, los alumnos escribieron un pequeño texto, tipo resumen, a partir de la lectura de un escrito. Una vez, realizado el ejercicio se procedió al análisis de la construcción gramatical, utilizando el procesamiento del lenguaje natural, permitiendo separar, las palabras en conjuntos adecuados para evaluar su intencionalidad. Los resultados obtenidos, sugieren que gran parte de los estudiantes, presentan dificultad para construir escritos y que la forma en que intentan expresar sus ideas tiene una estructura gramatical sencilla, impidiendo que la comunicación cumpla con la intencionalidad y el mensaje sea efectivo.

PALABRAS CLAVE

Análisis semántico, Análisis de texto, Procesamiento del lenguaje natural, Minería de texto

ABSTRACT

This research used words, phrases, sentences and paragraphs with the intention of identifying the ability that students have to transmit verbal information. To achieve this objective, quantitative descriptive research was taken into account. As an information gathering technique, a textual production workshop was applied; in which, the students wrote a short text, summary type, from the reading of a writing. Once the exercise was carried out, the analysis of the grammatical construction was carried out, using natural language processing, allowing the words to be separated into adequate sets to evaluate their intentionality. The results obtained suggest that a large part of the students have difficulty in constructing writings and that the way in which they try to express their ideas has a simple grammatical structure, preventing the communication from fulfilling the intention and the message being effective.

KEYWORDS

Semantic Analysis, Text Analysis, Natural Language Processing, Text Mining.

RESUMO

Esta pesquisa utilizou palavras, frases, sentenças e parágrafos com o intuito de identificar a habilidade que os alunos possuem para transmitir informações verbais. Para atingir esse objetivo, foi considerada a pesquisa quantitativa descritiva. Como técnica de coleta de informações, foi aplicada uma oficina de produção textual; em que os alunos escreveram um texto curto, do tipo resumo, a partir da leitura de uma carta. Terminado o exercício, procedeu-se à análise da construção gramatical, utilizando processamento de linguagem natural, permitindo a separação das palavras em conjuntos adequados para avaliar a sua intencionalidade. Os resultados obtidos sugerem que grande parte dos alunos tem dificuldade em construir a escrita e que a forma como procuram expressar suas ideias tem uma estrutura gramatical simples, impedindo que a comunicação cumpra a intenção e a mensagem seja efetiva.

PALAVRAS-CHAVE

Análise semântica, Análise de texto, Processamento de linguagem natural, Mineração de texto

1. INTRODUCCIÓN

La escritura es un proceso que incita el análisis crítico; esto lleva a pensar que el proceso de escritura se convierte en un elemento lingüístico que involucra un alto nivel de elaboración, rigor

sintáctico, semántico y morfológico. Saber redactar, es escribir un texto con claridad y coherencia, factores que sin duda alguna llevan al lector a comprender e interpretar lo que leen de la manera más precia. La mala utilización, de las palabras y de los signos de puntuación llevan a que el mensaje se distorsione y pierda sentido lo que el emisor desea transmitir (Aguilar, Albarrán, Errázuriz, & Lagos, 2016).

En la producción textual, se hace indispensable que quien escribe tenga claro elementos comunicativos básicos para que el propósito se cumpla; dentro de ellos está la coherencia y la cohesión (Briesmaster & Etchegaray, 2017), en tal sentido, la coherencia permite al escrito amalgamar proposiciones con sentido dentro de una unidad mayor, de tal manera, la coherencia se convierte en una propiedad semántica que da contextura al mensaje, relaciona palabras y frases (Cassany, 1995) y da significado a lo que se desea expresar.

Para Van Dijk (1980), la coherencia está dada en tres niveles: la primera coherencia local, la segunda coherencia lineal y la tercera coherencia global; en cuanto al primer nivel, consiste en la organización del texto a partir de estructuras gramaticales (sujeto, predicado y elementos como uso de verbos, artículos, pronombres); el segundo nivel, se refiere a la continuidad del discurso a partir de enlaces discursivos (conectores) y el tercer nivel, establece el significado temático (ideas principales). Los tres niveles permiten que el texto cobre valor comunicativo, si falla uno de ellos, es posible, la intención comunicativita se afecte y el mensaje se distorsiona.

Respecto a la cohesión, se refiere a la relación sintáctica que adquieren las palabras, frases u oraciones en su nivel lingüístico (Casado, 1995) así mismo, corresponde a los elementos paralingüísticos que ayudan a dar énfasis y aclarar las ideas (Kohan, 2010), por lo tanto, las palabras, las frases, las oraciones y los párrafos se relacionan entre sí a través de medios gramaticales, evitando la repeticiones innecesarias y redundantes; haciendo uso de elementos o enlaces de conexión; acudiendo a los signos de puntuación como elementos de significación; uso de tiempos verbales y recursos semánticos.

La comprensión de los discursos textuales depende en gran medida de la capacidad cognitiva del lector; sin embrago, cuando el discurso posee fallas en sus estructuras gramaticales, en el uso de elementos de coherencia y cohesión, el lector, tendrá serias dificultades para poder interpretar lo que se desea expresar. De la misma manera, cuando un texto no plasma ideas claras, es posible, el escritor no forme un proceso reflexivo de lo que desea transmitir, lo que trae como consecuencia, escritos carentes de sentido en los que lo único que se puede percibir son malas improvisaciones, desorden textual, aliteraciones, ambigüedades e incongruencias (Huerta, 2010).

Pedagógicamente, la enseñanza de la escritura, exige que el docente oriente al alumno en aspectos claves del proceso como lo son: tener claro a qué público va dirigido el escrito, la intención del texto, las normas gramaticales, semánticas y morfológicas; aspectos lingüísticos lexicales, aspectos cognitivos, entre otros; hasta que estos aspectos no sean apropiados por los estudiantes, será muy difícil conseguir que éstos produzcan textos con propósitos claros (Uribe & Camargo, 2011).

En sintonía con lo expresado anteriormente, algunos autores (Marchant, Lucchini, & Cuadrado, 2007), indican que existe una estrecha relación entre la capacidad léxica y el acervo lingüístico que se posee con la calidad de lo que se escribe, indican estos autores que para lograrlo se hace indispensable que el individuo se documente y lea permanentemente; logrado esto, es posible, se alcancen niveles

escriturales satisfactorios y en consecuencia se logre lo que denominamos comprensión lectora.

Evaluar, la capacidad y la habilidad escritural en los estudiantes, especialmente en la educación superior resulta muchas veces, complicado, especialmente en la educación en ingenierías, toda vez que tiende a ser técnica y se aleja un poco de lo humanístico; en primer lugar, porque ya se traen vicios difíciles de romper y, en segundo lugar, los niveles escriturales y de comprensión que se exigen son mayores a los que le puede pedir a un estudiante de educación básica; las actividades, talleres y ejercicios de producción textual debe estar claramente formulados para que lleven al estudiante a mejorar el proceso (Duran & Rosado, 2018); igualmente, las evaluaciones deben abordar cada uno de los elementos y factores gramaticales, sintácticos, lexicales, semánticos y morfológicos, tarea que no es nada fácil para el docente, pues cada elemento requiere, ser evaluado a partir de características propias de cada elemento, razón por la cual, usar una técnica como el procesamiento de lenguaje natural, facilita al docente la evaluación, además permite detectar de manera puntual, dónde están las fallas o dificultades en cuanto a producción escritural (Cornejo, Roble, Barrero, & Martín, 2012).

El procesamiento del lenguaje natural (PLN), es una disciplina ubicada dentro de la inteligencia artificial y la lingüística computacional, que tiene como propósito la exploración del lenguaje natural como forma de comunicación entre humanos y máquinas (Velásquez, Puentes, & Espinel, 2016); los propósitos del procesamiento natural están dimensionados en: el procesamiento de textos, es decir, extraer la información relevante e importante a través de la generación automática de síntesis; la traducción automática por medio de sistemas multilingües que traducen textos de un idioma a otro, y la generación de interfaces del lenguaje natural en la cual se dan ordenes en lenguaje natural y la máquina procesa y obedece (Joseph, Hlomani, Letsholo, Kaniwa, & Sedimo, 2016). Las tareas que más son usadas en el procesamiento natural son: extracción de información, reproducción de síntesis de manera maquinal, reconocimiento de conceptos, interpretación de discursos, generación de lenguaje natural en distintas lenguas, reconocimiento de voz, análisis semántico, morfológico y sintáctico y análisis de sentimientos para determinar la polaridad de contenidos (Eisenstein, 2019).

El PLN, siendo un lenguaje entre el ser humano y una máquina, esta última debe responder a la tarea asignada, para ello existen varios modelos para la recuperación de información que establecen diferencias entre una consulta y las respuestas de dicha consulta, pare ello se recurre a múltiples formas o modelos de búsqueda. En este caso se convierte en una herramienta importante que permite conocer sentimientos que expresan los textos a través de medios computacionales, lo cual permite que se haga un análisis del comportamiento de la estructura que se utilizó para exponer una idea y poder realizar el texto desde diferentes enfoques que permite PLN. Desde este enfoque, esta investigación intenta encontrar recursos dentro de PLN que permitan realizar la evaluación de un texto determinando que tan cerca está de un resumen ideal de un documento.

2. METODOLOGÍA

La investigación estuvo amparada en el paradigma cuantitativo descriptivo, dado que este tipo de investigación permite explicar o predecir fenómenos o situaciones a partir de cifras (Gay, Mills, & Airasian, 2009). En el presente estudio participaron 34 estudiantes de ingeniería de sistemas primer semestre, a los cuales se les aplicó un taller con una única actividad, la cual correspondía en elaborar un resumen de un texto conformado por cuatro párrafos; los estudiantes debían redactar un resumen tipo texto y subirlo a la plataforma

Moodle; el resumen fue evaluado a partir de criterios establecidos en una rúbrica la cual, así mismo, fue validada por expertos, y en la que se tenía en cuenta factores claves: ortografía, vocabulario, cohesión y coherencia.

El análisis de la información se realizó a través de PLN; el texto inicialmente fue convertido a toquen para trabajarlo como un vector de palabras para poder posteriormente aplicar conteo de palabras y generar acciones de similitud dentro del texto. Luego de tener el texto separado y adaptado, se aplicaron varias herramientas de PLN que permiten hacer conteos y relacionar similitud entre textos, con el ánimo de describir las diferencias que existen entre los resultados mejor y menor valorados (Bird, Klein, & Loper, 2009).

3. RESULTADOS Y DISCUSIÓN

Una de los primeros interrogantes que se planteó es el hecho que los estudiantes con mejor calificación podrían ser aquellos que hicieron una exposición de palabras mayor que los estudiantes de menor calificación, en la tabla 1 se hace un resumen de los conteos de las palabras divida entre los estudiantes que aprobaron y los que reprobaron el taller. Como se puede observar en los diferentes valores que se muestran en la tabla, existen diferencias entre las medidas y algunos casos los textos que fueron aprobados tienen valores mayores que los reprobados; sin embargo, al mirar las medidas en su conjunto no existe una variación significativa entre el número de palabras que permita afirmar que a mayor numero mejor calificación.

Tabla 1. Conteo de palabras

Media	Aprobado	Reprobado
Total	23	11
Promedio	507	477
Desviación	171	215
Mínimo	149	174
25%	386	319
50%	486	496
75%	600	607
Máximo	872	815

Fuente. Elaboración propia

Un análisis de la estructura interna (sintáctico) de los textos puede dar pautas sobre la forma y uso de las palabras, para lograrlo se recomienda el uso de n-gramas que produce mejores resultados que hacer un análisis de caracteres individuales (Takase, Suzuki, & Nagata, 2019). En la tabla 2, se muestra, la comparación de los cinco tri-gramas, más comunes en los textos; en la parte izquierda se muestran las frecuencias de los resúmenes que aprobaron y en la parte derecha, los que fueron reprobados; estos resultados sugieren, que el comportamiento entre los dos grupos de estudiantes en términos generales es similar, tiendo una única diferencia que es la mención que se hacen los que aprobaron sobre la carta de las naciones unidas y los que reprobaron Naciones Unidas utilizan. Con el ánimo de organizar la información que se muestra en la tabla se hace la siguiente convención:

F1: octubre 24 1945

F2: Organización Naciones Unidas

F3: Carta Naciones Unidas

F4: Progreso económico social

F5: Fundada octubre 24

F6: Naciones Unidas utilizan

Tabla 2. Análisis tri-gramas

Palabras	Frecuencia	Palabra	Frecuencia
F1	20	F1	8
F2	16	F6	7
F3	13	F2	7
F4	12	F5	6
F5	11	F4	6

Fuente. Elaboración propia

Otro tipo de análisis que se le puede hacer al texto es el discursivo; para buscar la similitud que pueden tener dos textos (Mota, Da Cunha, & López-Escobedo, 2016). Una técnica utilizada corresponde a la búsqueda de cadena más larga (longest common subsequence, LCS), que consiste en una subsecuencia de todas las cadenas de entrada que están presentes en otra cadena con la cual se compara (Djukanovic, Berger, Raidl, & Blum, 2020). Otra técnica que también se utiliza para esta clase de análisis es la similitud coseno (cosine similarity, CS), la cual consiste en convertir en vectores las cadenas que se quieren comparar y el Angulo que exista entre ellas es la similitud que tienen (Singh, Maurya, Tripathi, Narula, & Srivastav, 2020). Para las dos formas expuestas la forma de interpretar sus resultados es la misma; los valores están comprendidos entre 0 y 1, entre más cercano a uno sea el valor existe mayor similitud.

Para aplicar las herramientas mencionadas, se redactó un texto que contendría la forma adecuada de resumir la lectura asignada, este texto se comparó con los textos redactados por los estudiantes para revisar cual, de las dos formas, muestra mejor resultado. En la tabla 3, se muestran las diez mejores calificación emitida por el docente (valores entre 0 y 5), el resultado de LCS y CS, si se toman estos valores y se llevan a una escala entre cero y cinco, se puede indicar que la similitud coseno es la más precisa y emite valores cercanos a los utilizados por el docente; sin embargo, no en todos los casos el valor fue tan preciso, esto se puede deber a no utilizar procesos como el de lematización, por que como lo señalan Diaz-Mendivelso & Suarez-Baron (2019), producen mayor precisión en esta clase de comparaciones.

Tabla 3. Comparación de texto discursivo

Calificación	LCS	CS
4.0	0.04	0.73
4.0	0.13	0.80
4.0	0.06	0.81
4.0	0.09	0.73
3.9	0.07	0.73
3.9	0.05	0.71
3.8	0.03	0.85
3.8	0.17	0.75
3.8	0.09	0.82
3.8	0.03	0.75

Fuente. Elaboración propia

Si bien, CS, dio mejores resultados para conseguir la similitud entre dos cadenas de texto, existen algunos problemas para que permita delegar en esta herramienta confianza como para realizar procesos automáticos de validación de similitud.

Dentro del análisis de subsecuencias entre las palabras presentes en los textos, en esta investigación se encontró que este tipo de análisis no permitió hacer deducciones que diferenciaran el comportamiento de los dos grupos de estudiantes, siendo muy similares sus asociaciones, esto sugiere una necesidad de tener acciones más centradas en el texto y en el contexto de la aplicación de PLN, teniendo que profundizar en el conocimiento semántico de un dominio restringido, para hacer una estructura especifica que permita encontrar diferencias entre los textos de los participantes (Hurtado, Costa, Segarra, Garcıa-Granada, & Sanchis, 2016).

En esta investigación se indago por la forma en que se podría llegar a realizar un proceso de identificación de comportamientos de escritura, acompañados por herramientas tecnológicas pero automatizado, para que la intervención humana permita depurar más profundamente los comportamientos (Cornejo, Roble, Barrero, & Martín, 2012). Sin embargo, estas diferencias produjeron una nueva revisión en la rúbrica de evaluación que estaba siendo rigurosa en algunos aspectos que afectaban la calificación final del estudiante.

Si bien PLN, es robusto para adelantar procesos de detección automática de patrones de redacción, de comparación de cadenas y polaridad de los sentimientos presentes en las mismas, aun se tienen algunos limitantes que permitan que soluciones como la planteada tengan que realizar procesos específicos de creación de un corpus estricto para el contexto de estudio, porque elementos como las negaciones y el sarcasmo necesitan mayor nivel de especificidad (Martí, y otros, 2016).

3. CONCLUSIONES

En esta investigación se abordaron elementos que brinda el procesamiento del lenguaje natural en conjunto con el lenguaje de programación Python, para poder extraer información de texto, específicamente de resúmenes realizados sobre un documento guía, lo cual permite que las actividades cotidianas de los docentes puedan de alguna manera agilizarse para encontrar de forma más rápida patrones diferenciadores en la forma de escribir y apropiar la transmisión de ideas escritas por los estudiantes. Para lograrlo, se siguieron acciones comunes en la literatura

que ayudan a realizar una exploración del conjunto de textos con los que se cuenta para ir paulatinamente, definiendo afirmaciones sobre ellos e ir separando el comportamiento correcto del que no lo es.

Como se mostró en esta investigación, automatizar la evaluación e interpretación de la escritura es un proceso en el que inicialmente se debe tener un conjunto de prácticas o experiencia del docente en la identificación de patrones de comportamiento frecuentes en cuanto al mejor y menor rendimiento; esto permite guiar la intervención del procesamiento del lenguaje natural, delimitando la búsqueda en los estudiantes con problemas y sus rasgos de redacción característicos, para posteriormente generar un modelo que tenga la capacidad de sugerir un rendimiento a partir de un texto guía.

REFERENCIAS BIBLIOGRÁFICAS

Aguilar, P., Albarrán, P., Errázuriz, M., & Lagos, C. (2016). Teorías implícitas sobre los procesos de escritura: Relación de las concepciones de estudiantes de Pedagogía Básica con la calidad de sus textos. Estudios Pedagógicos, 7-26.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. Sebastopol: O'Reill.

Briesmaster, M., & Etchegaray, P. (2017). Pédagogique basée sur la métacognition. Íkala, Revista de Lenguaje y Cultura, 183-202.

Casado, B. (1995). Poder y Escritura en la Edad Media. Espacio, Tiempo y Forma, 143-168.

Cassany, D. (1995). La cocina de la escritura. Barcelona: Anagrama.

Cornejo, J., Roble, M., Barrero, C., & Martín, A. (2012). Hábitos de lectura en alumnos universitarios de carreras de ciencia y de tecnología. Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 155-163.

Diaz-Mendivelso, J., & Suarez-Baron, M. (2019). Análisis social aplicando técnicas de lenguaje natural a información extraída de Twitter. Scientia et Technica, 496-503.

Djukanovic, M., Berger, C., Raidl, G., & Blum, C. (2020). An A∗ search algorithm for the constrained longest common subsequence problem. Information Processing Letters, 1-12.

Duran, C., & Rosado, A. (2018). La Comprensión Lectora y el Rendimiento Académico en Estudiantes de Ingeniería. Revista Colombiana de Tecnologías de Avanzada, 9-15.

Eisenstein, J. (2019). Introduction to Natural Language Processing. Cambridge: The MIT Press.

Gay, L., Mills, G., & Airasian, P. (2009). Educational Research: Competencies for Analysis and Applications. Londres: Pearson Education.

Huerta, S. (2010). Coherencia y cohesión. Herencia, 76-80.

Hurtado, L., Costa, I., Segarra, E., Garcıa-Granada, F., & Sanchis, E. (2016). Traducción Automática usando conocimiento semántico en un dominio restringido. Procesamiento del Lenguaje Natural, 101-108.

Joseph, S., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. (2016). Natural Language Processing: A Review. International Journal of Research in Engineering and Applied Sciences, 207-217.

Kohan, S. (2010). Gramática para escritores y no escritores. Madrid: Alba.

Marchant, T., Lucchini, G., & Cuadrado, B. (2007). ¿Por qué Leer Bien es Importante? Asociación del Dominio Lector con Otros Aprendizajes. Psykhe, 3-16.

Martí, M., Taulé, M., Nofre, M., Marsó, L., Martín-Valdivia, M., & Jiménez-Zafra, S. (2016). La negación en español: análisis y tipología de patrones de negación. Procesamiento del Lenguaje Natural, 41-48.

Mota, M., Da Cunha, I., & López-Escobedo, F. (2016). Un Corpus de Paráfrasis en Español: Metodología, Elaboración y Análisis. Revista de lingüística teórica y aplicada, 85-112.

Singh, R., Maurya, S., Tripathi, T., Narula, T., & Srivastav, G. (2020). Movie Recommendation System using Cosine Similarity and KNN. International Journal of Engineering and Advanced Technology (IJEAT), 556-559.

Takase, S., Suzuki, J., & Nagata, M. (2019). Character n-Gram Embeddings to Improve RNN Language Models. AAAI Conference on Artificial Intelligence (págs. 5074-5082). Honolulu: Association for the Advancement of Artificial Intelligence.

Uribe, G., & Camargo, Z. (2011). Prácticas de lectura y escritura académicas en la universidad colombiana. Magis, Revista Internacional De Investigación En Educación, 317-341.

Van Dijk, T. (1980). Estructuras y funciones del discurso. Coyoacan: Siglo veintiuno.

Velásquez, T., Puentes, A., & Espinel, E. (2016). Marco Ontológico Para La Estructuración Semántica y la Recuperación de Recursos Bibliográficos Empleando Procesamiento del Lenguaje Natural. Revista Colombiana de Tecnologías de Avanzada, 9-15.