DOI: http://dx.doi.org/10.23857/dc.v8i1.2646

Ciencias Técnicas y Aplicadas
Artículo de Investigación

Detección de tópicos de textos en español usando machine learning, caso discursos Guillermo Lasso Presidente de Ecuador

Spanish text topic detection using machine learning, case of speeches by Guillermo Lasso President of Ecuador

Detecção de tópicos em textos em espanhol usando aprendizado de máquina, caso de discursos Guillermo Lasso Presidente do Equador

César Humberto Espin-Riofrio [I]
cesar.espinr@ug.edu.ec
https://orcid.org/0000-0001-8864-756X

Tania Jeesenia Peralta-Guaraca [II]
tania.peraltag@ug.edu.ec
https://orcid.org/0000-0002-4879-6824

Luis Merino-Salcedo [III]
luis.merinos@ug.edu.ec
https://orcid.org/0000-0002-9082-3240

Gerardo Parra-Barrezueta [IV]
gerardo.parrab@ug.edu.ec
https://orcid.org/0000-0001-6878-8306

Correspondencia: cesar.espinr@ug.edu.ec

*Recibido: 25 de febrero del 2022 *Aceptado: 25 de marzo de 2022 * Publicado: 01 de abril de 2022

I. Magister en Sistemas de Información Gerencial, Universidad de Guayaquil, Guayaquil, Ecuador.
II. Magister en Ingeniería de Software y Sistemas Informáticos, Universidad de Guayaquil, Guayaquil, Ecuador.
III. Universidad de Guayaquil, Guayaquil, Ecuador.
IV. Universidad de Guayaquil, Guayaquil, Ecuador.

Resumen

El presente artículo tiene como objetivo centrarse en dos de las principales técnicas para Procesamiento de Lenguaje Natural de machine learning para el modelado y detección de tópicos, se trata de los algoritmos Non-negative Matrix Factorization and Latent Dirichlet Allocation que se usarán para experimentar y verificar en corpus de textos en el idioma español, basados en el estado de arte de la atribución de autoría relacionado a la detección de tópicos mediante el análisis de artículos científicos de relevancia sobre el tema, además se investigan los distintos modelos destinados a la detección de tópicos resaltando cuales son los más utilizados, también se busca evaluar el comportamiento y resultados de los dos modelos escogidos. La experimentación se realiza sobre los discursos políticos pasados a texto del Sr. Guillermo Lasso Presidente del Ecuador, se identifican los diferentes tópicos o temas sobre los que trata el corpus de textos formado por los discursos con el fin de conocer directamente su contenido o para dónde estos están apuntando de manera preliminar sin necesidad de leer el contenido en su totalidad, los resultados se presentan comparando los modelos, así se logra determinar con cuál de los dos algoritmos se obtienen resultados más acertados.

Palabras Clave: Aprendizaje automático; LDA; NMF; Detección de tópicos; Procesamiento de Lenguaje Natural.

Abstract

This paper aims to focus on two of the main techniques for machine learning Natural Language Processing for topic modeling and detection, namely the Non-negative Matrix Factorization and Latent Dirichlet Allocation algorithms that will be used to experiment and verify in a corpus of texts in the Spanish language, based on the state of the art of authorship attribution related to topic detection through the analysis of relevant scientific articles on the subject, in addition, the different models for topic detection are investigated, highlighting which are the most used, and the behavior and results of the two chosen models are also evaluated. The experimentation is carried out on the political speeches of Mr. Guillermo Lasso President of Ecuador, the different topics or themes on which the corpus of texts formed by the speeches deals with are identified in order to know directly their content or where they are pointing to in a preliminary way without the need to read the content in its entirety, the results are presented comparing the models, thus determining with which of the two algorithms the most accurate results are obtained.

Keywords: Machine Learning; LDA; NMF; Topical Detection; Natural Language Processing.

Resumo

O objetivo deste artigo é focar em duas das principais técnicas de Processamento de Linguagem Natural de aprendizado de máquina para a modelagem e detecção de tópicos, são eles os algoritmos de Fatoração de Matrizes Não Negativas e Alocação de Dirichlet Latente que serão usados para experimentar e verificar em corpus de textos em língua espanhola, com base no estado da arte de atribuição de autoria relacionada à detecção de tópicos por meio da análise de artigos científicos relevantes sobre o assunto, além disso, os diferentes modelos destinados à detecção de tópicos são investigados, destacando quais são os mais utilizados, busca também avaliar o comportamento e os resultados dos dois modelos escolhidos. A experimentação é realizada nos discursos políticos passados a texto pelo Sr. Guillermo Lasso, Presidente do Equador, identificando os diferentes tópicos ou temas sobre os quais trata o corpus de textos formado pelos discursos para conhecer diretamente seu conteúdo ou onde esses estão apontando de forma preliminar sem a necessidade de ler o conteúdo na íntegra, os resultados são apresentados comparando os modelos, assim é possível determinar com qual dos dois algoritmos os resultados mais precisos são obtidos.

Palavras-chave: Aprendizado de máquina; LDA; NMF; Detecção tópica; Processamento de linguagem natural.

Introducción

En la actualidad, en pleno siglo XXI existe una gran variedad de estudios referente a la inteligencia artificial (IA) y a sus campos, (Bodem, 2017) en su libro menciona que la inteligencia no es una dimensión única, sino un espacio profusamente estructurado de capacidades diversas para procesar la información. Del mismo modo, la IA utiliza muchas técnicas diferentes para resolver una gran variedad de tareas.

Entre los campos de la IA está el machine learning (ML) (Ethem Alpaydin, 2021). Según Murphy (2012) el machine learning o aprendizaje automático se define como un conjunto de métodos capaces de detectar automáticamente patrones en los datos. Por otro lado, en el trabajo de (Naga, y Murphy, 2015) afirman que el ML es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circulante (p.3-11), es meramente dificultoso obtener un solo concepto de ML pues es un campo extenso que alberga

diversos escenarios. Dentro del campo del aprendizaje automático existen diversas técnicas enfocadas en el modelado y detección de tópicos que se complementan con la atribución de autoría para el procesamiento de datos. Estas técnicas de ML permiten procesar una inmensa cantidad de datos, llamados en un lenguaje técnico como corpus, y con ello ayudará al usuario a entender de manera más simplificada grandes volúmenes de textos. El modelado de tópicos tiene como fin encontrar a través de algoritmos estadísticos, los principales temas de colecciones de documentos (Hernández et al., 2015).

Partiendo de los conceptos mencionados, este trabajo se centra en el análisis y comparación de los resultados luego de aplicar la Factorización Matricial No Negativa (Non-negative Matrix Factorization, NMF) (Lee & Seung, n.d.) y Análisis Discriminante Lineal (Latent Dirichlet Allocation, LDA) (Blei et al., 2003). Estos dos métodos estadísticos pertenecientes al aprendizaje automático, la cual es una raíz de la inteligencia artificial, permitirán que grandes cantidades de textos se puedan simplificar y sean más sencillos de entender para los lectores. También estos algoritmos serán complementados por la atribución de autoría, la cual realiza un perfil lingüístico guiado por un protocolo metodológico.

Para la ejecución de los dos algoritmos se hará uso de los discursos dados por el Presidente del Ecuador Sr. Guillermo Lasso, los tópicos serán identificados de manera automática pues estos métodos usan patrones que se encuentran analizando una colección llamada corpus, en la cual básicamente se reunirán las palabras claves importantes que se localizan en los discursos, una vez ejecutados estos métodos se obtendrán los resultados de manera ordenada y categorizada, estos datos permitirán la comparación experimental entre los dos métodos, previo a esto se realizará un análisis de los dos algoritmos, la cual será complementada con los resultados de los mismos y así con ello obtener una mejor perspectiva comparativa de los dos métodos.

Metodología

Como es notorio la cantidad de información que se encuentra disponible en internet en la actualidad es gigantesca y más aún crece día a día de forma exponencial, esta información que encuentra en textos en como redes sociales, revistas, periódicos, blogs, emails, foros, etc. Teniendo esta apreciación en cuenta surge la necesidad de poder analizar esta información, pero se ve dificultada por la carencia de las herramientas adecuadas. (Stamatatos, 2009) menciona que en la atribución de autoría (AA) se tiene como fin el poder diseñar un algoritmo que sea capaz de reconocer o identificar

a los autores aprendiendo de su estilo de escritura, mencionando algunas de las problemáticas relacionadas con la AA se tiene la atribución de mensajes de terroristas, acoso e intimidación, spam, detección de plagio, disputas de derechos intelectuales, etc. Con esto se puede determinar la importancia de la AA en algunos trabajos usando clasificación supervisada, más aún cuando se atribuyen la propiedad a decenas de autores (Pavelec et al., 2008).

El estudio de la atribución de autoría tiene una extensa historia (Sarwar & Nutanong, 2016) y esta ha ido evolucionando progresivamente desde su comienzo a finales del siglo pasado, consecuentemente su aplicación se ha encontrado en un continuo avance y a su vez abarcando diferentes áreas de las cuales se ha referenciado en este proyecto. El primer intento en el campo de la AA fue proclamado por (Mendenhall, 1887) con su trabajo en estilometria para lograr identificar la autoría en base al análisis del estilo de su escritura en obras de estilo literario como es en la obra de Shakespeare. Esencialmente en el trabajo de los autores (Mosteller & Wallace, 1963) este inició con unos estudios de atribución de autoría no tradicionales, a diferencia de los basado en métodos expertos tradicionales. Desde entonces y hasta después de 1990 las investigaciones de atribución de autoría fueron dominadas por intentos en definir características para cuantificar el estilo de escritura como lo fue esencial en la investigación de (Holmes & Holmes, 1998) quien formuló un conjunto de características para contabilizar los estilos de escritura de los autores los cuales son conocidos como estilometría, el estudio de la estilometría está comprendido en análisis estadísticos de variaciones de los estilos literarios del autor representado como un conjunto de características, lo cual los mantiene relativamente sin cambios en diferentes documentos.

(Carleo et al., 2019) menciona que el machine learning abarca una amplia gama de algoritmos y herramientas de modelado utilizado para una gran variedad de tareas de procesamiento de datos que han entrado en la materia de las disciplinas científicas en los últimos años. Dicho esto, para el desarrollo de esta investigación se aplicó dos de las técnicas del ML las cuales serán descritas y analizadas para determinar cuál de las dos es más funcional para la detección de tópicos junto a su atribución de autoría.

Según (Vayansky et al., 2020), el modelado de tópicos es una herramienta analítica popular para evaluar datos (p.1). Se han desarrollado numerosos métodos de modelado para temas que consideran muchos tipos de relaciones y restricciones dentro de los conjuntos de datos. Este modelado permitirá que los lectores entiendan grandes volúmenes de textos de manera más simplificada y sencilla. Este método será acompañado por la atribución de autoría. En los estudios de atribución de autor, la

medición del uso diferencial de palabras funcionales es el procedimiento más común, aunque a menudo se utilizan estadísticas léxicas. Rara vez se ha empleado el análisis de contenido, (Martindale & P McKenzie, 1995). Estas técnicas serán aplicadas en dos de los algoritmos de ML.

Los algoritmos LDA y NMF son modelos de aprendizaje automático no supervisado, cuyo funcionamiento consiste en identificar las palabras claves de los diversos temas. LDA es un modelado probabilístico, se basa en la suposición de que los documentos se componen de varios temas y no precisamente de palabras, donde un tema es una distribución multinomial en un vocabulario fijo. LDA es considerado como un modelo mixto que es capaz de capturar la intercambiabilidad de palabras y documentos. La suposición de intercambiabilidad por palabra en un documento significa que el orden de las palabras en un documento no es importante, y del mismo modo para la ordenación de documentos en un corpus, (Mifrah, 2020).

NMF es considerado un algoritmo de aprendizaje automático. El método NMF es considerado también como un modelo de pertenencia mixta, ya que los documentos procesados no se asocian a un único tema. En su lugar, a cada documento se le asignan porcentajes de todos los temas. Así, un documento puede ser un 25 % sobre desempleo, un 10 % sobre inflación, etc. (Hansen, n.d.).

Con lo mencionado anteriormente, lo que busca esta investigación es realizar una comparación experimental de los algoritmos LDA y NMF y, junto a ello, identificar el algoritmo más adecuado y certero para la detección de tópicos empleando metodologías con un perfil lingüísticas de atribución de autor tomando como objeto de estudio cinco discursos del Sr. Guillermo Lasso Presidente del Ecuador.

Objeto de prueba

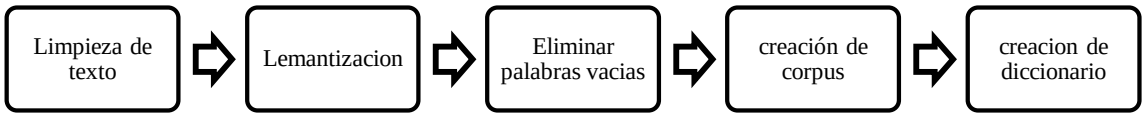
El objeto de prueba propuesto para este artículo son los cinco discursos dados por el Sr. Guillermo Lasso, presidente del Ecuador mostrados en la siguiente figura.

index	
0	Cumbre sobre sistemas alimentarios.txt
1	Medidas de seguridad en los centros penitenciarios.txt
2	decreto para mejoras en el sistema educativo.txt
3	Reformatoria a la ley de creacion Uni Amawtay Wasi.txt
4	Avances en el proceso de vacunación segunda dosis al presidente.txt

Discursos políticos del Sr. Guillermo Lasso presidente del Ecuador

Preprocesamiento de datos

En esta etapa se debe cumplir un proceso de preparación para los datos que se van a evaluar esto comprende la eliminación de caracteres especiales seguido de la lematización que implica estandarizar las palabras, aunque, además tenemos que eliminar aquellas palabras que no son relevantes para la detección de tópicos para el proceso final que comprende la creación de diccionario y de corpus, este proceso se puede apreciar mejor en la siguiente figura.



Proceso de preparación de datos.

Modelado de tópicos

El modelado de tópicos es una metodología donde se procede a procesar grandes cantidades de datos que son los que forman el corpus, estos datos contienen características útiles. Existen varios algoritmos para detectar tópicos y en este artículo nos centraremos en dos modelos que son muy usados como LDA y NMF.

Latent Dirichlet Allocation (LDA)

El modelo LDA es un modelo probabilístico que se basa en la suposición del modelo de bolsa de palabras más conocido como bow. Los datos usados para este modelo pueden comprender múltiples temas y cada tema una distribución de probabilidad por palabra, es decir que se segmenta por cada documento bajo un conjunto de reglas, esto supone una gran variedad de tópicos probables. Un ejemplo claro para poder entender cómo funciona este algoritmo es que puede tener tópicos clasificados en “relacionados a gatos” y “relacionados a perros”. Un tópico tiene probabilidades de generar varias palabras como leche, miau y gatito, las cuales pueden ser clasificadas e interpretadas como “relacionadas a gatos”. Naturalmente, la palabra gato en sí misma tendrá una alta probabilidad dado este tópico. Por otro lado, los tópicos “relacionados a perros” tienen probabilidades altas para las siguientes palabras: cachorro, ladrar y hueso.(Hammoe, 2018).

Non-negative Matrix Factorization (NMF)

El modelo NMF es un algoritmo de algebra línea que comprende técnicas no supervisadas para el modelado de tópicos que nos permite reducir de forma considerable conjuntos de datos en atributos representativos, de tal forma que si tenemos una cantidad determinada de documentos esta nos devolverá una representación más pequeña reduciendo la dimensionalidad de los datos. A diferencia de otros modelos NMF es un algoritmo que relaciona sus matrices llegando a representar de forma única el corpus obteniendo tópicos de documentos latentes. La esencia de este algoritmo radica en que dada la matriz de documentos X, NMF nos dará dos matrices una matriz W con temas por palabras y la matriz de coeficientes H con documentos por temas. (*Modelado de Temas Con NMF Para Clasificación de Reseñas de Usuarios*, 2020)

Resultados

Para evaluar la eficiencia de los modelos LDA y NMF se realizó un proceso de experimentación, con la herramienta de desarrollo Google Colab y el lenguaje de programación Python, los discursos políticos fueron extraídos de fuentes confiables y se procesaron para esta evaluación con el modelo de LDA usando Gensim[1] y NMF usando Scikit Learn[2], debido a que los textos evaluados fueron cinco las variaciones en los tiempos de ejecución no difirieron, finalmente se generaron los tópicos donde se podía medir el nivel de coherencia con relación a los discursos identificando el modelo con los mejores tópicos.

```
Discurso #0
['escuela', 'hoy', 'país', 'educativo', 'niño', 'encontrar', 'educación']
Discurso #1
['educación', 'libre', 'presidente', 'ecuatoriano', 'señor', 'joven', 'universidad']
Discurso #2
['policía', 'libertar', 'humano', 'derecho', 'carcelario', 'centro', 'señor']
Discurso #3
['gobernar', 'alimentación', 'alimentario', 'agroalimentario', 'importante', 'año', 'país']
Discurso #4
['segundar', 'vacunación', 'manera', 'hacer', 'vacunar', 'dosis', 'ecuatoriano']
```

Tópicos obtenidos por el modelo NMF

[1] https://radimrehurek.com/gensim/models/ldamodel.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html#sklearn.decomposition.NMF

```
[(0,
  '0.021*"educación" + 0.019*"niño" + 0.012*"joven" + 0.010*"excelencia" + 0.009*"llamar" + 0.009*"conocimiento" + 0.008*"profesional"'),
 (1,
  '0.017*"alimentario" + 0.017*"gobernar" + 0.016*"agroalimentario" + 0.016*"millón" + 0.015*"alimentación" + 0.011*"sociedad" + 0.011*"agricultura"'),
 (2,
  '0.037*"educación" + 0.023*"joven" + 0.019*"universidad" + 0.018*"niño" + 0.017*"libre" + 0.016*"millón" + 0.015*"manera"'),
 (3,
  '0.037*"derecho" + 0.037*"carcelario" + 0.028*"ley" + 0.028*"paz" + 0.028*"centro_carcelario" + 0.020*"director" + 0.020*"vez"'),
 (4,
  '0.005*"niño" + 0.005*"decir" + 0.005*"ley" + 0.005*"abrazar" + 0.005*"alimentación" + 0.005*"recurso" + 0.005*"agroalimentario"')]
```

Tópicos obtenidos por el modelo LDA

Como se puede observar al momento de obtener los tópicos para los modelos existe diferencia, y es que el modelo LDA al generar un diccionario y corpus con menores términos obtiene los resultados menos relevantes, esto debido a la reducción de la distribución probabilística, a diferencia de NMF que obtuvo tópicos más comprensibles.

Discusión

Mediante las evaluaciones realizadas a los modelos LDA y NMF se demostró la diferencia en cuanto a resultados de los algoritmos bajo librerías que están orientadas a estos fines como lo son Gensim y Scikit-learn, es importante el hecho de haber realizado una comparación de estos algoritmos en lenguaje en español. Es necesario discutir sobre ciertas condiciones que nos proveerán de mejores o peores resultados y que se ven afectados por la cantidad de documentos y los parámetros usados, podemos indicar también que para efectos de este análisis se determinaron estos campos por lo que cambiarlos alteraría los resultados, aunque se hayan condicionado para este análisis en específico, esto es de carácter general y se aplica a los diferentes estudios encontrados, Por otro parte, se toma en cuenta que los discursos fueron tomados de forma aleatoria para su análisis con los modelos LDA y MNF y estos están en lenguaje español, pero qué pasaría si se analizan estos documentos con un idioma diferente o que de igual manera en los mismos documentos se encuentren palabras que aporten sustancialmente al tema en otro idioma, existirían muchos inconvenientes ya que el modelo fue preparado para un lenguaje en específico por lo tanto se debería tomar en cuenta para futuras evaluaciones.

Conclusiones

Como se ha podido ver a lo largo del análisis se determina el estado del arte de la detección de tópicos y la atribución de autoría donde se investigó de fuentes confiables y contribuciones científicas de relevancia referentes al tema, también se logró identificar los algoritmos para el modelado de tópicos más usados de los cuales para este proceso se eligieron dos de estos para ser evaluados, los modelos LDA y NMF con el fin de obtener los tópicos de los discursos del Sr. Guillermo Lasso Presidente del Ecuador, este proceso se pudo llevar aplicando librerías de Python para la detección de tópicos para finalmente mostrar los resultados de los tópicos en los cuales se muestra de forma clara que el modelo en obtener los tópicos más comprensibles es el NMF a diferencia del LDA que no obtuvo tópicos acordes a los temas tratados, se puede determinar que este método tendría un mejor comportamiento teniendo como prueba un corpus de texto mucho mayor al tomado en este artículo.

Referencias

1. Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).
2. Boden, M. A. (2017). *Inteligencia artificial*. Turner.
3. Carleo, G., Cirac, I., Cramer, K., Daudet, L., & Schuld, M. (2019). El aprendizaje automático y las ciencias físicas. *Reseñas de Física Moderna , 91 (4), 045002.*
4. el Naqa, I., & Murphy, M. J. (2015). *¿Qué es el aprendizaje automático?* Aprendizaje automático en oncología radioterápica. Springer, Cham.
5. Ethem Alpaydin. (2021). *Machine learning*. MIT Press.
6. *Factorización matricial no negativa HistoriayFondo*. (n.d.). Retrieved March 7, 2022, from https://hmong.es/wiki/Non-negative_matrix_factorization
7. *Función Dirichlet - Función Dirichlet modificada, Otras propiedades, Continuidad e integrabilidad, Definición | KripKit*. (n.d.). Retrieved March 7, 2022, from https://kripkit.com/funcin-dirichlet/
8. Hammoe, L. (2018). *Detección de tópicos: utilizando el modelo LDA*. INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA.
9. Hansen, S. (n.d.). *APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO AL ANÁLISIS ECONÓMICO Y LA FORMULACIÓN DE POLÍTICAS*.

10. Hernández, A., Tomás, D., & Borja Navarro. (2015). Una aproximación a la recomendación de artículos científicos según su grado de especificidad. *Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos*.
11. Lee, D. D., & Seung, H. S. (n.d.). *Algorithms for Non-negative Matrix Factorization*.
12. Martindale, C., & P McKenzie, D. (1995). On the utility of content analysis in author attribution: The Federalist. *Computadoras y Humanidades , 29 (4), 259-270.*
13. Mifrah, S. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5756–5761. https://doi.org/10.30534/ijatcse/2020/231942020
14. Murphy, K. (2012). *Aprendizaje automático: una perspectiva probabilística*. Prensa del MIT.
15. Pavelec, D., Oliveira, L. S., Justino, E., & Batista, L. V. (2008). *Using Conjunctions and Adverbs for Author Verification*.
16. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556. https://doi.org/10.1002/ASI.21001
17. Vayansky, I., AP Kumar, S., & Sathish, A. K. (2020). Una revisión de los métodos de modelado de temas. *Sistemas de Información , 94 , 101582.*
18. Holmes, R. M., & Holmes, S. T. (1998). *Contemporary perspectives on serial murder*. 246.
19. Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science*, *9*(214), 237–246. https://doi.org/10.1126/SCIENCE.NS-9.214S.237
20. *Modelado de temas con NMF para clasificación de reseñas de usuarios*. (2020). ICHI.PRO. https://ichi.pro/es/modelado-de-temas-con-nmf-para-clasificacion-de-resenas-de-usuarios-111674468812030
21. Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association*, *58*(302), 275. https://doi.org/10.2307/2283270
22. Sarwar, R., & Nutanong, S. (2016). The Key Factors and Their Influence in Authorship Attribution. *Research in Computing Science*, *110*(1), 139–150. https://doi.org/10.13053/rcs-110-1-12