UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Comparación automática de Planes de Gobierno utilizando técnicas de Procesamiento de lenguaje natural: Caso de Estudio – Las Elecciones Presidenciales de Ecuador (Segunda Vuelta)

.

Mike Arthur Pinta Pacheco

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 06 de mayo de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Comparación automática de Planes de Gobierno utilizando técnicas de Procesamiento de lenguaje natural: Caso de Estudio –Las Elecciones Presidenciales de Ecuador (Segunda Vuelta)

Mike Arthur Pinta Pacheco

Nombre del profesor, Título académico **Daniel Riofrío, Ph. D.**

Quito, 06 de mayo de 2021

© DERECHOS DE AUTOR

Nombres y apellidos: Mike Arthur Pinta Pacheco

Código: 00125826

Cédula de identidad: 0704986504

Lugar y fecha: Quito, 06 de mayo de 2021

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Comparar planes de gobierno siempre ha sido complicado debido a las diferencias ideológicas, situaciones sociales del momento y trasfondo del candidato, pero como punto en común todos los candidatos tienen manifiestos que exponen sus propuestas y puntos de vista en diferentes áreas. Este documento explora una manera de poder comparar propuestas de campaña a través de sus propios manifiestos mediante técnicas de procesamiento natural de lenguaje usando el algoritmo de Doc2Vec. Como corpus lingüístico usamos todos los artículos en español de Wikipedia y usamos dos modelos de redes neuronales, DBOW (Distributed bag of words) y DM (Distributed memory model) y cada párrafo de los manifiestos son clasificados en 7 diferentes dominios acordes con el Manifiesto Project, de esta manera cada parte del manifiesto está organizado por temática y podemos comparar todo el documento.

Palabras clave: Minería de texto, Natural Language Processing, Doc2Vec, Elecciones, Manifiestos políticos, Manifesto Project

ABSTRACT

Comparing government policies has always been difficult due to the ideological differences, social situations of the moment and the background of the candidate, but as a common point, all candidates have manifestos that expose their proposals and points of view in different areas. This document explores a way to be able to compare campaign proposals through your own manifests using natural language processing techniques using the Doc2Vec algorithm. As a linguistic corpus we use all the articles written in spanish from Wikipedia and we use two models of neural networks, DBOW (Distributed bag of words) and DM (Distributed memory model) and each paragraph of the manifestos is classified into 7 different domains according to the Manifesto Project, in this way each part of the manifesto is organized by theme and we can compare the entire document.

Key words: Text mining, Natural Language Processing, Doc2Vec, Elections, Political manifest, Manifesto Project

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Todos los partidos políticos tienen su propio manifiesto político que les permite exponer sus ideas a sus votantes están divididos en múltiples temas como salud, empleo, economía, educación, vivienda, cultura y problemas relevantes para el país, con temas como el medio ambiente que se están volviendo muy discutidos.

Este proyecto nace de la necesidad de poder analizar programas electorales entre candidatos en el Ecuador evitando la parcialidad que pueda tener cualquier persona al analizar las propuestas de políticos proponemos modelos neuronales que logren detectar similitud entre manifiestos políticos y similitud entre diferentes temas basados en el sistema de dominios de Manifesto Project.

Usamos Doc2Vec que es una técnica de procesamiento de lenguajes naturales basada en Word2Vec que crea una red neuronal que puede aprender palabras que se asocien entre sí basada en un cuerpo lingüístico. Es un algoritmo ampliamente usado para comparar similitudes lingüísticas entre palabras de cualquier lenguaje (Zhu, Wang, & Zou, 2016) (Ramadhanti & Mariyah, 2019). Los vectores que se crean pueden ser usados con operaciones matemáticas como la similitud de coseno o la distancia entre vectores que nos indican la similitud semántica entre palabras; o en el caso de Doc2Vec, nos indica la similitud entre documentos, el modelo DBOW (Distributed bag of words) y el modelo DM (Distributed memory model). Estos dos modelos se generan de diferente manera, pero los resultados entre ellos son parecidos cuando hay relación entre las palabras. (Mikolov, Chen, Corrado, & Dean, 2013)

El Manifesto Project nació por parte del Manifesto Research on Political Representation (MARPOR) como una propuesta para poder analizar manifiestos políticos de las elecciones democráticas de todo el mundo creando así un sistema de dominios que abarca los principales temas importantes para todas las naciones en general, usamos este sistema para poder clasificar

partes de los planes de gobierno en temas como democracia, relaciones internacionales, cultura, salud, economía, grupos minoritarios y recientemente también temas ecológicos. Así poder comparar propuestas. El proceso de etiquetado se realizó manualmente bajo la observación de un experto siguiendo las definiciones descritas en el Manifesto Project (Volkens, y otros, 2020).

Ya se han usado algoritmos de word embedding y los dominios de Manifesto Project para comparar discursos políticos en el parlamento de Reino Unido y poder determinar una diferencia entre discursos, análisis de sentimiento y opinión, y preferencias políticas (Abercrombie, Nanni, Batista-Navarro, & Ponzetto, 2019). En Latinoamérica se ha usado Word2Vec para análisis del discurso socialista de los líderes del siglo 20 y 21, logrando así distinguir una similitud entre discursos políticos entre Fidel Castro, Ernesto Che Guevara, Hugo Chávez, Nicolás Maduro, Evo Morales, Rafael Correa y el Subcomandante Marcos (Zapata & Peignier, 2017). Sin embargo, con la relevancia política que tienen las elecciones en el Ecuador, este es el primer trabajo que incursiona en el ámbito de determinar similitudes a través del manifiesto político, comparando todo el documento y comparándolo por temas.

Los candidatos que analizamos son aquellos que pasaron a la segunda vuelta: Andrés Arauz y Guillermo Lasso. Ambos candidatos siguen un pensamiento ideológico diferentes tanto en discursos como en propaganda política, por lo que consideramos que son ejemplos ideales para poder determinar sus similitudes entre forma de trabajo y propuestas.

El objetivo de este trabajo es determinar con la ayuda de modelos de procesamiento natural de lenguaje, si podemos comparar los planes de gobierno y propuestas de campaña de ambos candidatos usando como cuerpo lingüístico los artículos en español de Wikipedia que abarca múltiples artículos en diferentes áreas (Mikolov, Chen, Corrado, & Dean, 2013).

El resto de este documento se encuentra organizado de la siguiente manera: la sección de materiales y métodos que describe de donde se obtuvieron los manifiestos, el cuerpo lingüístico utilizado, el modelo, tamaño de los vectores del algoritmo de Doc2Vec y describe las métricas de similitud que usamos para poder comparar nuestros resultados. Los resultados y la discusión cubren todos los valores que se obtuvieron del modelo y discutimos porqué ciertos dominios son o no similares. Finalmente, se presenta la conclusión alcanzada y se explica el trabajo que se puede realizar a futuro.

MATERIALES Y MÉTODOS

Corpus

El cuerpo lingüístico usado para entrenar los modelos de Doc2Vec son todos los artículos escritos en español en Wikipedia escritos hasta el febrero 21 del 2021. La compañía encargada de Wikipedia, Wikimedia, crea una copia de seguridad de todos sus artículos en todos los lenguajes dos veces al mes y es posible de obtener a través de sus volcados de datos en la página de Wikimedia[1]

Propuestas políticas

Las propuestas políticas fueron obtenidas en las páginas web oficiales de ambos candidatos, con Andrés Arauz[2] y con Guillermo Lasso[3]. Ambos planes de gobierno contienen todas las propuestas entregadas al Consejo Nacional Electoral del Ecuador que es el organismo encargado de organizar. Ambos planes de gobierno fueron tratados de tal manera de que se eliminen imágenes y tablas excluyéndolos del análisis semántico de los documentos. En el caso de Guillermo Lasso tenemos 28, 443 palabras para procesar y en el caso de Andrés Arauz 16, 961.

Algoritmo Doc2vec

El algoritmo de Word2Vec es una manera eficiente de generar representaciones numéricas de palabras, el objetivo del algoritmo es de convertir una gran cantidad de texto en un valor numérico sin perder la relación entre las palabras. (Jatnika, Bijaksana, & Suryani, 2019). En este estudio se utiliza de Gensim de Doc2Vec que es similar a Word2Vec. Con Word2Vec si usamos una palabra esta nos devuelve un vector que nos permite ver palabras que tengan un contexto similar que no significa que sea un sinónimo, sino que da palabras que

[1] https://dumps.wikimedia.org
[2] https://andresarauz.ec
[3] https://www.creo.com.ec

fueron usadas en contextos similares. Con Doc2Vec si se le alimenta con un documento, este divide los párrafos y lo convierte en un vector y permite ver documentos parecidos con el mismo texto. Igual que Word2Vec, Doc2Vec tiene diferentes modelos, memoria distribuida y lista de palabras distribuidas. (Mikolov, Chen, Corrado, & Dean, 2013)

DM (Distributed Memory – Memoria Distribuida)

El modelo de memoria distribuida entrena una red neuronal que intenta predecir palabras en medio de un párrafo sabiendo un promedio de todas las palabras que hay en el documento. Este modelo es parecido al modelo CBOW de Word2Vec que agrupa todas las palabras y predice palabras del medio, pero con la diferencia que los vectores se encuentran identificados por párrafo para identificar cada documento de manera única. (Le & Mikolov, 2014). La representación de este modelo se muestra en figura 1. Varias palabras son introducidas junto con la identificación del documento y el modelo predice una palabra

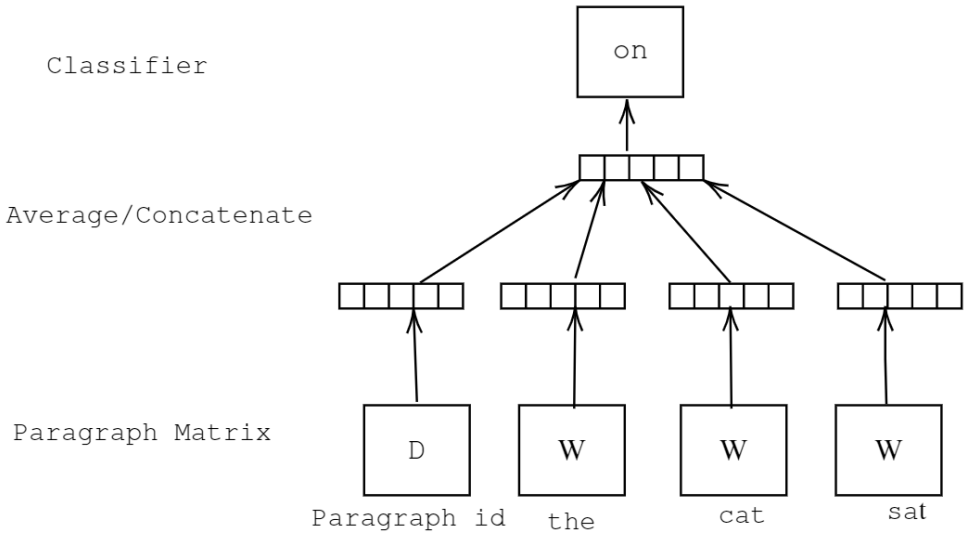

Figura 1. DM (Distributed Memory Model)

DBOW (Distributed Bag of Words – Lista de palabras distribuidas)

El modelo de lista de palabras distribuidas en cambio transforma todos los documentos en un vector del tamaño de todas las palabras únicas y cuenta las veces que cada palabra se repite. Con esta información intenta predecir las palabras del centro de los párrafos. Este modelo es más ligero y es recomendado para sistemas pequeños porque no guarda vectores de palabras sino solo pesos softmax (Le & Mikolov, 2014). La representación de este modelo se muestra en figura 2. Dada la identificación de un documento el modelo intenta determinar las palabras que se encuentren en el mismo.

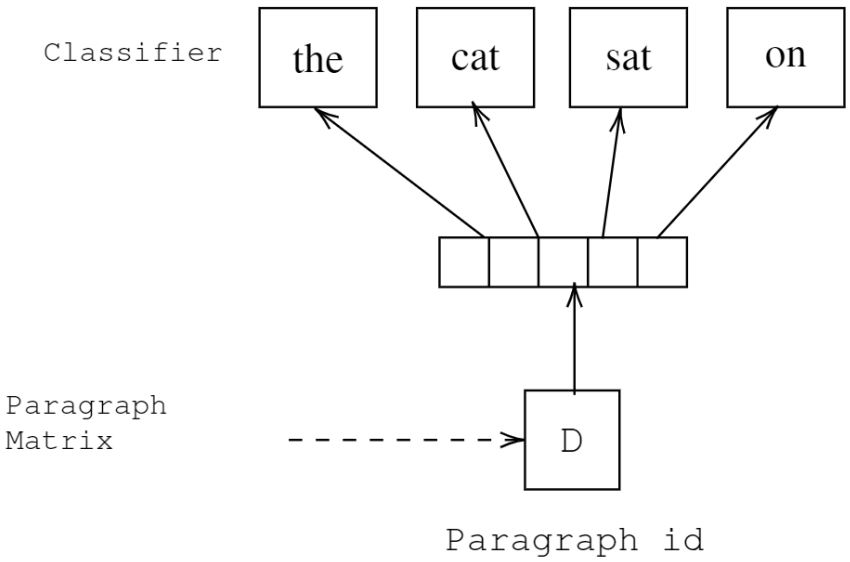

Figura 2. DBOW (Distributed Bag of Words)

ETIQUETADO DE PLANES DE GOBIERNO

El proceso de etiquetar los párrafos de los planes de gobierno se realiza manualmente siguiendo las instrucciones del Manifesto Coding Instructions 5th edition. Está dividido en 7 dominios que cubren múltiples puntos generales de las propuestas. (Krause, y otros, 2020). Se separan las propuestas de cada plan de gobierno en párrafos. Cada párrafo es analizado bajo

los 7 posibles dominios; y, a criterio del investigador principal se marca uno de esos dominios para ese párrafo. A continuación, se detalla cada dominio incluido en este estudio:

Dominio 1: Relaciones Internacionales

El dominio 1 representa el punto de vista del político en relaciones externas. Cubre ambas vistas positivas y negativas sobre relacionarse con naciones extranjeras, las referencias negativas sobre influencias extranjeras hacia otras naciones, sea estas políticas, militares o comerciales; y, cualquier mención contra el colonialismo. (Por ejemplo, estar en una posición en contra del FMI o el Banco Mundial). De igual manera, este dominio trata de propuestas sobre gasto militar relacionado con mantener la soberanía y paz con otros países; y, temas de cooperación internacional o políticas de aislamiento y apoyo a cortes internacionales, a la Organización de las Naciones Unidas y al reconocimiento de organizaciones internacionales (Volkens, y otros, 2020).

Dominio 2: Democracia Y Libertad

El dominio 2 representa el punto de vista del político en temas relacionados con el estado de libertad y derechos humanos. Como son el derecho a la libre expresión, libertad de prensa, derecho a la reunión y derecho a protesta. En democracia en cambio se encuentran temas como mantener, abolir o modificar la constitución de la nación (Volkens, y otros, 2020).

Dominio 3: Autoridad Política

El dominio 3 cubre en la manera de cómo el político maneja el gobierno. Así, este dominio trata de temas como la lucha contra la corrupción, apoyar la descentralización o la centralización del gobierno y encontrar maneras más eficientes o baratas para operaciones gubernamentales. Sin embargo, el tema más importante del dominio 3 es la manera en la que el político o el partido político se representa. Cualquier mención sobre cómo el político, el

partido político o su ideología son más competentes que la de su contrincante o atacar su ideología o sus propuestas son tomadas en cuenta en este dominio. (Volkens, y otros, 2020)

Dominio 4: Economía

El dominio 4 abarca los temas económicos como inversiones, impuestos o subsidios, es uno de los dominios más importantes de una nación con problemas económicos como Ecuador que los candidatos difieren ideológicamente. Abarcan temas que apoyen el libre mercado, el derecho a la propiedad privada e incentivo a empresas personales. El aumento o reducción de subsidios (no incluyen los subsidios a educación o salud) y cualquier manera de planeación económica, protección al consumidor, prevención de monopolios o favoritismos al corporativismo. Así mismo incluyen temas opuestos como proteccionismo a mercados internos, manejo keynesiano incrementando la demanda o gastos públicos. Toma en cuenta cualquier referencia al control de la economía a través de regulación de precios o cambio de salario mínimo. Incluye el favoritismo en el nacionalismo de la industria y propuestas para modernizar la tecnología y la infraestructura. Y también incluye cualquier apoyo a un enfoque marxista de la economía (Volkens, y otros, 2020).

Dominio 5: Estado De Bienestar Y Calidad De Vida

El dominio 5 son los temas que abarcan el bienestar y la calidad de vida de la nación. Esto incluye cualquier política para proteger el medio ambiente (siempre y cuando no referencie impuestos ya que estos son del dominio 4), derechos de los animales y preservación de la naturaleza en general. Incluye las políticas que promuevan el financiamiento en artes, deportes y cultura como campos deportivos, museos y librerías. La implementación de políticas que garanticen la equidad entre clases y grupos sociales no privilegiados, propuestas sobre distribución equitativa de recursos, la eliminación de discriminaciones raciales y sexuales.

Finalmente incluyen temas económicos que no cubre el Dominio 4 que es el incremento o la limitación de gasto estatal en salud y educación (Volkens, y otros, 2020).

Dominio 6: Fábrica De La Sociedad

El dominio 6 son las propuestas que tienen un impacto con las interacciones sociales de la población. Apoyar o desaprobar ideales nacionalistas o de patriotismo, seguir con moralidad tradicional o religiosa como propuestas por una familia tradicional o apoyar la composición actual de familia moderna, recurrir a la restricción de libertades para proteger al estado en caso de subversión. Promover o rechazar la inmigración y políticas de cambio en la ley y el orden como cambio de penas, descriminalización de sustancias psicotrópicas y cambios en políticas de prostitución y apuestas. En Ecuador es muy importante recalcar dos tópicos de este dominio, uno de ellos siendo propuestas sobre multiculturalismo y diversidad migratoria y siendo el otro tópico derechos indígenas, incluyendo motivar o disuadir la participación de este grupo social en el gobierno. (Volkens, y otros, 2020)

Dominio 7: Grupos Sociales

El dominio 7 abarca menciones y propuestas enfocadas a todos los grupos sociales, pueden ser clases sociales como clase alta, media o baja, grupos de trabajadores profesionales y no profesionales incluyendo uniones de trabajadores que demanden políticas de mejores condiciones de trabajo. Cabe recalcar que, aunque se incluyan uniones de trabajadores, se aplica tanto para apoyarlos como por rechazarlos. Otros grupos sociales como desempleados, grupos agricultores, grupos minoritarios y finalmente, individuos que no se encuentren en ningún grupo económico independiente de la edad. Cualquier referencia directa hacia estos grupos se lo clasifica dominio 7. (Volkens, y otros, 2020)

Métricas de similitud

Similitud De Coseno

La similitud de coseno es una función matemática que cuando se calcula con dos vectores nos da un número entre 0 y 1. Este número representa la diferencia del ángulo entre dos vectores. Si el valor tiende hacia 1, significa que los dos vectores son similares, si el valor tiende hacia 0 significa que los vectores son diferentes. Esta métrica a resultado útil para calcular similitudes en modelos realizados con Word2Vec y Doc2Vec (Alonso, Volkens, Cabeza, & Gómez, 2012) (Bilbao-Jayo & Almeida, 2018).

$$similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}} = \cos(\theta)$$

Figura 3. Ecuación de similitud de coseno

Norma L2

La norma L2 calcula la distancia entre dos vectores en un espacio euclidiano, a diferencia de la similitud de coseno la norma L2 puede dar cualquier valor a partir de 0, mientras más grande sea la medida significa que es mayor la diferencia entre vectores. (Knapp, 2005)

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Figura 4. Ecuación de norma L2

Configuración del Experimento

Etiquetado de planes de gobierno

Los planes de gobierno contienen una portada, diseños gráficos, imágenes y tablas que no contribuyen con la semántica del texto por lo que solo dejamos el texto sin formato. El proceso de clasificar cada párrafo de los planes de gobierno es manual bajo la observación de un experto para poder determinar si un párrafo corresponde a uno o a otro dominio, en el caso de que un párrafo abarque múltiples dominios se escoge el que tenga mayor influencia.

Entrenamiento del modelo Doc2vec

El cuerpo lingüístico que utilizamos para entrenar los modelos de Doc2Vec son todos los artículos escritos de Wikipedia, se utiliza este cuerpo lingüístico por ser sumamente amplio y abarcar múltiples temas.

Configuración del modelo

Creamos tanto modelo DBOW como modelo DM en configuraciones similares con excepción del tamaño del vector que genera el modelo donde se usa un vector de dimensión de 100, 200 y 500. Palabras que no superan una frecuencia absoluta de 19 son ignoradas. El valor para revisar palabras que rodean cada palabra para analizar contextos (windows size) está limitado en 8 y los modelos están entrenados en 10 épocas. El valor de *windows size* y épocas depende del tamaño y la composición del cuerpo lingüístico. Cuando se tienen muchos documentos las épocas necesarias se reducen debido a la variedad lingüística, si se tienen pocos documentos como cuerpo lingüístico se requiere de un mayor número de épocas, de otra manera el modelo no podrá distinguir diferencias. El *windows size* en cambio depende del tamaño de los documentos, si los documentos son cortos, se recomienda un *windows size* pequeño para que asocie palabras entre documentos. Si el *windows size* es grande, esto nos permite distinguir palabras usadas entre documentos.

RESULTADOS Y DISCUSIÓN

Con todos los planes de gobierno etiquetados en diferentes dominios podemos hacer la comparación de todo el documento. Y, en cada uno de los dominios podemos crear nuevos documentos enfocados en las partes del plan de trabajo que correspondan a los temas de cada dominio. Los resultados nos brindan una similitud en ambos documentos si los comparamos por completo, pero por dominios separados, tenemos diferencias notables de similitud. Como podemos en las figuras 3-8 con diferentes modelos y dimensiones de vectores, existe un patrón entre los dominios con una mayor similitud y menor similitud.

Dominios con mayor similitud

Los dominios 1,2 y 6 tienen una alta similitud. En el dominio 1, que trata de relaciones internacionales, ambos candidatos expresan un fuerte deseo con relacionarse con otros países y ambos mencionan los mismos países con los que desean trabajar. Las únicas diferencias notables son que el candidato Andrés Arauz expresa su deseo con establecer relaciones UNASUR y con el candidato Guillermo Lasso de establecer relaciones con la Alianza del Pacífico. Ambos candidatos se comprometieron a respetar relaciones internacionales actuales con la excepción del candidato Arauz mencionando una única vez que no seguiría con los desembolsos del FMI.

La similitud con el dominio 2 es la mas alta en todos los modelos considerando los distintos sets de características (dimensiones de vectores). Este dominio cubre la democracia y la libertad donde ambos candidatos se comprometen a mantener la democracia del país, respetar la separación de poderes, proteger los derechos humanos y respetar la constitución. No hay diferencias importantes ya que ambos candidatos se comprometen a mantener el estado democrático y de derecho.

El dominio 6 se refiere al tejido de la sociedad, donde ambos candidatos proponen mejorar la ley y el orden, conservar al Ecuador como un estado laico y apelan al patriotismo de los ciudadanos, hacen un importante hincapié en que la sociedad ecuatoriana se vea unida y capaz de ayudarse entre ella. Resaltan la importancia de como Ecuador es una nación pluricultural y de lo favorable que es contener diferentes regiones indígenas con diferentes fondos culturales. La diferencia notoria que se encuentra entre candidatos empieza con el candidato Lasso en proponer la eliminación de la tabla de consumo de estupefacientes mientras que el candidato Arauz propone una rehabilitación social con uso de fondos públicos y apela a la solidaridad del vecindario de reintroducir estos grupos a la sociedad.

Dominios con menor similitud

Los dominios 3, 4, 5 y 7 tienen la menor similitud. Esto tiene sentido porque es precisamente los dominios de autoridad política, economía, estado de bienestar y grupos sociales. Con el dominio 3 hay una clara diferencia de ideas sobre como manejar el poder del estado. Ambos concuerdan en la lucha contra la corrupción política pero sus propuestas para combatirla difieren, así el candidato Arauz propone sistemas estatales para regulación y control que contribuye al centralismo de poder político a diferencia del candidato Lasso que propone lo contrario, es decir, una descentralización progresiva del estado con observadores tanto de funcionarios públicos, privados como observadores internacionales. También en la manera en como ejercen su autoridad política para resaltar su competencia en gobernar con el candidato Arauz empleando términos como “neoliberalismo”, “morenismo” o “trujillato” para socavar al otro candidato, y con el candidato, Lasso, empleando palabras como “correísmo” y “socialismo del siglo 21”, para describir la mala administración que realizaría el otro candidato.

El dominio referente a la economía, el dominio 4, también tiene fuertes diferencias entre los candidatos. Ambos candidatos están de acuerdo en mantener o elevar ciertos subsidios

como el subsidio a la electricidad y aumentar la infraestructura tecnológica del país, pero difieren ampliamente en cómo mejorar la economía del país. El candidato Lasso propone una economía de libre mercado con incentivos a pequeños empresarios, reducción de impuestos y aumento de salario mínimo mientras que el candidato Arauz propone una planeación económica con regulación de mercado, proteccionismo al mercado local frente al extranjero y un manejo keynesiano de economía que propone aumentar la demanda pública mejorando los gastos a servicios sociales.

El quinto dominio es un tema importante porque involucra el gasto del estado en educación, salud, inversión en cultura y políticas medioambientales. Ambos candidatos se comprometen a cuidar el medio ambiente sin embargo difieren en la manera en qué políticas implementar. El candidato Lasso propone protección a la naturaleza bajo la figura de economía naranja que tiene como principio aprovechar solo recursos naturales necesarios y minimizar el impacto de contaminación ambiental, completamente opuesto a lo que propone el candidato Arauz a una política de proteger regiones acuíferas y proteger las zonas amazónicas del Ecuador. En educación los dos candidatos están de acuerdo en una educación pública gratuita y de calidad, pero el candidato Arauz prioriza otorgar becas y reabrir las escuelas que se crearon en el gobierno anterior mientras que Lasso prefiere reabrir escuelas rurales que habían sido cerradas para dar prioridad a las escuelas que el candidato Arauz proponía reabrir y que han cerrado debido a la pandemia del Covid-19 o incluso fueron cerradas antes por baja ocupación. Ambos proponen facilidad de inversiones para arte, deporte y cultura con la diferencia de que el candidato Arauz prefiere programas gubernamentales de apoyo mientras que el candidato Lasso prefiere facilitar apoyo a través de la empresa privada.

El séptimo dominio es complicado de determinar, ya que ambos candidatos mencionan en general todos los grupos y ofrecen su apoyo a todos los grupos sociales del país, sin

embargo, la razón por la que podemos ver que no hay similitud es por la forma en que ambos candidatos se refieren a estos grupos. Así, el candidato Arauz menciona especialmente grupos de clases sociales pobres, grupos indígenas, grupos LGBT y otras minorías mientras que, el candidato Lasso toma un enfoque más abierto mencionando un Ecuador con todos los grupos sociales unidos, pero resalta a los grupos campesinos, grupos indígenas, a grupos de jóvenes que deseen emprender su propio negocio y a las mujeres. Si bien los dos candidatos intentan tener un discurso que abarque todos los grupos sociales en el Ecuador, se ve que sus discursos están focalizados a ciertos grupos.

Comparación entre planes de gobiernos

En el caso de la comparación de documento completo tenemos fluctuaciones entre los modelos en las tres dimensiones inclusive en las figuras 5,7 y 9 podemos ver que el modelo DM genera una distancia euclidiana tan grande que supera la distancia generada por el manifiesto dividido en dominios, tiene sentido que tengan una similitud de coseno que puede superar la similitud que tienen otros dominios ya que ambos manifiestos hablan del Ecuador y abarca temas relevantes para el país. El modelo DM es un modelo que puede ser muy específico con el contexto en el que analiza los documentos, es por eso que el resultado de similitud de coseno con este modelo marca mayormente que hay una similitud más alta del promedio comparado con la similitud de coseno del modelo DBOW. Sin embargo, el modelo DBOW generaliza más los documentos y los resultados no son tan diferentes si tomamos en cuenta la distancia euclidiana, el defecto de DBOW es que puede terminar con una similitud alta a pesar de que tengamos en cuenta que existen realmente diferencias en los documentos si ambos documentos son extensos.

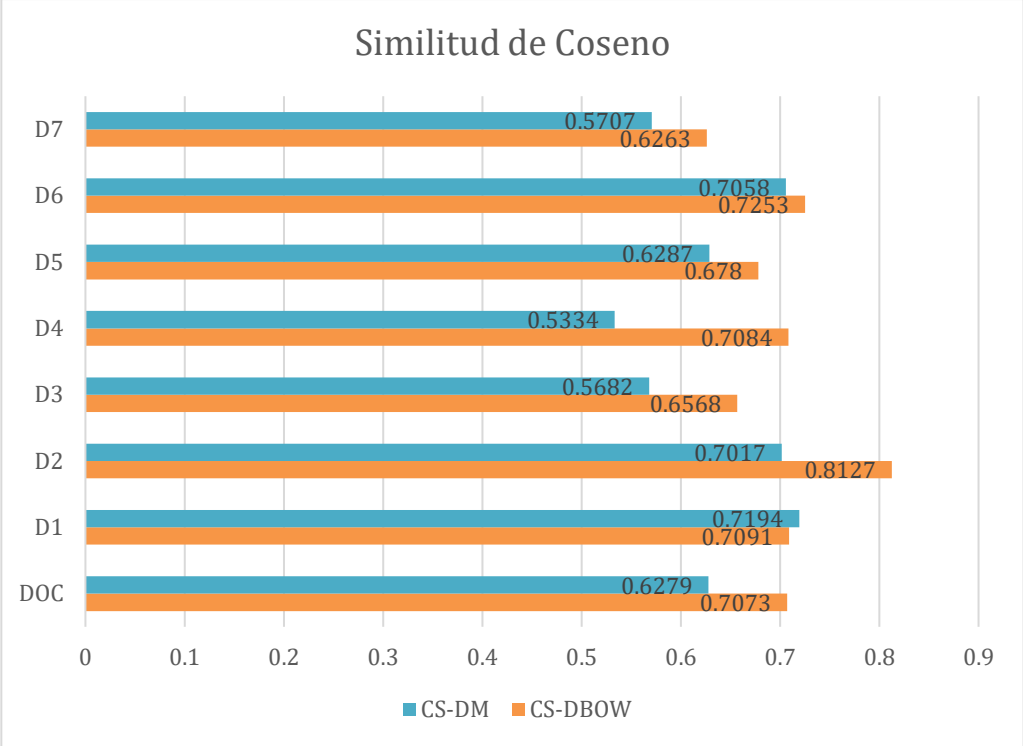

Figura 5. Similitud de coseno con vector de 100 dimensiones

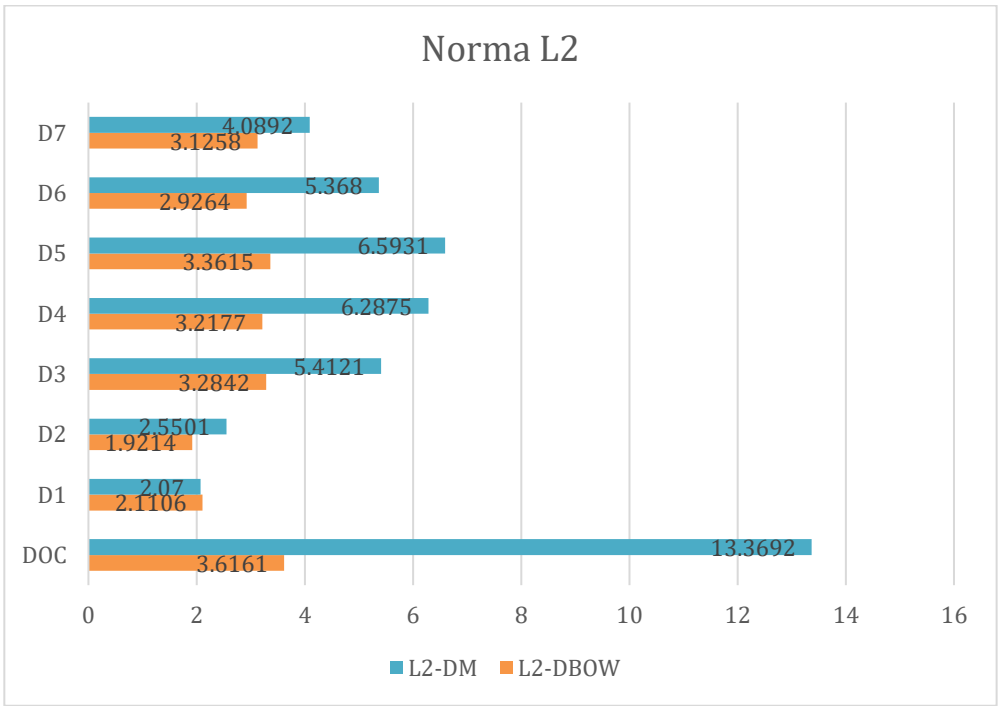

Figura 6. Norma L2 con vector de 100 dimensiones

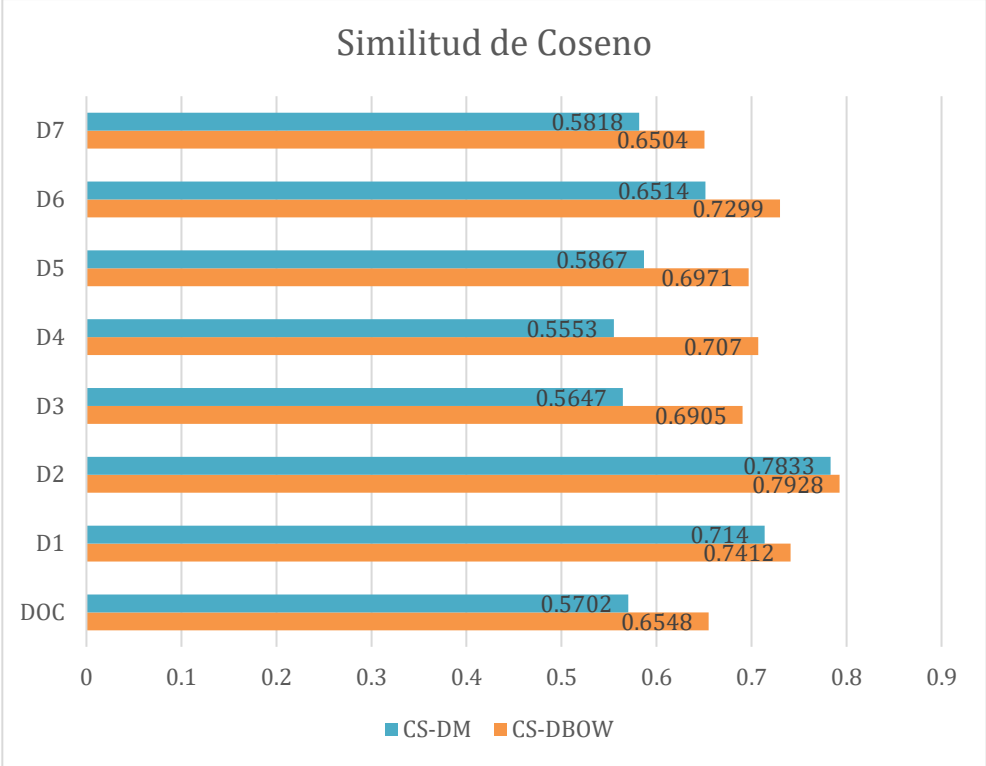

Figura 7. Similitud de coseno con vector de 200 dimensiones

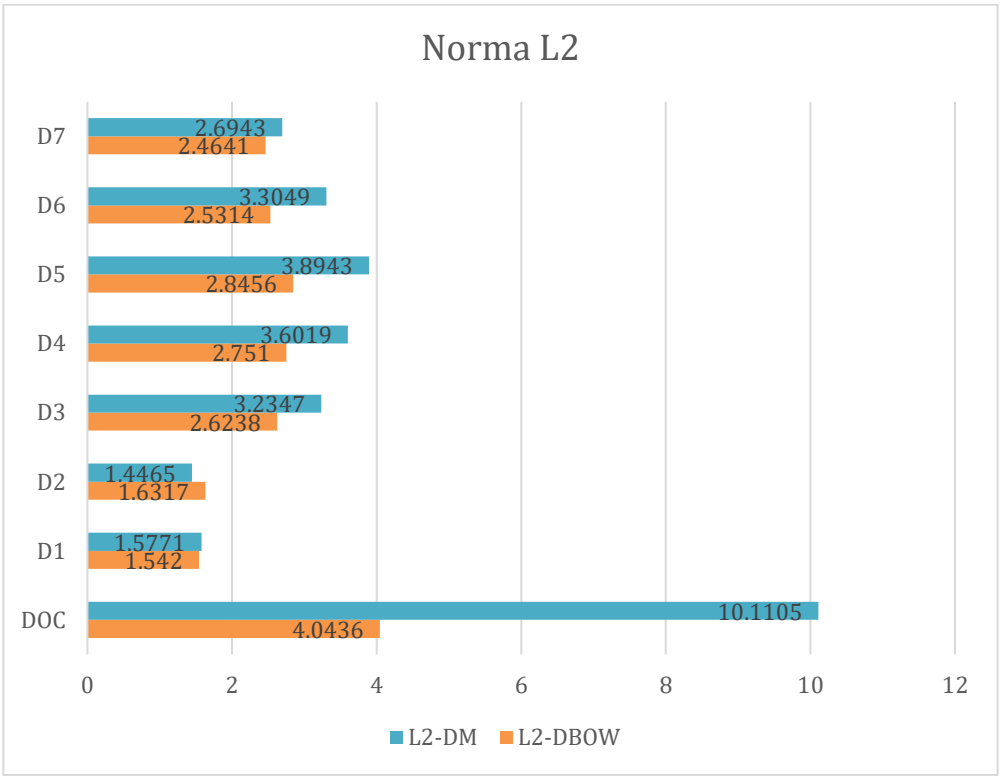

Figura 8. Norma L2 con vector de 200 dimensiones

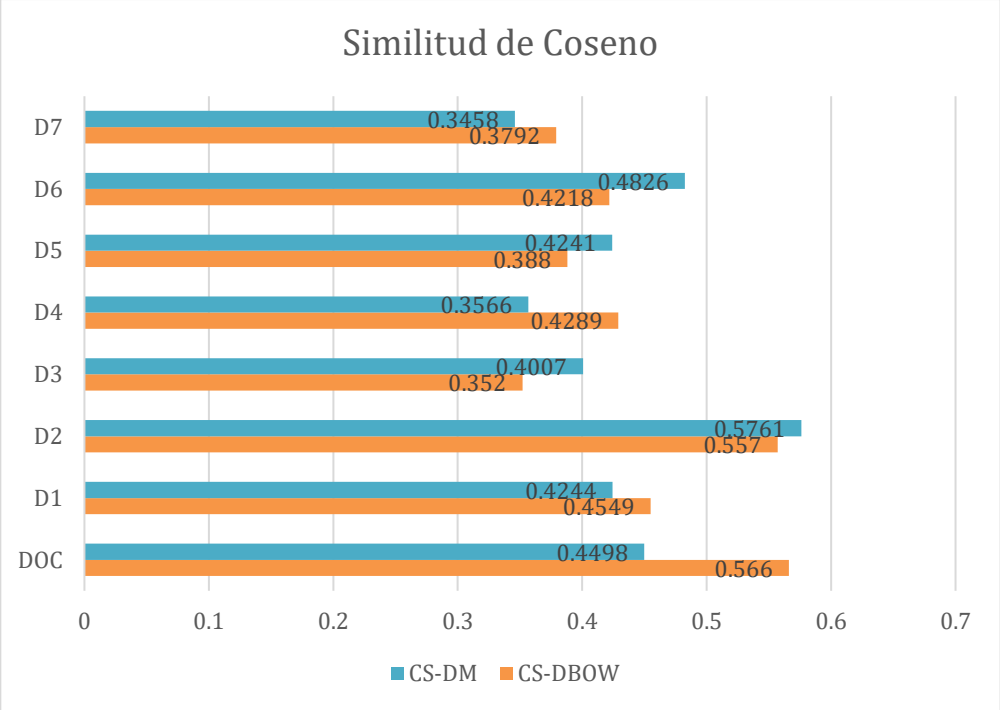

Figura 9. Similitud de coseno con vector de 500 dimensiones

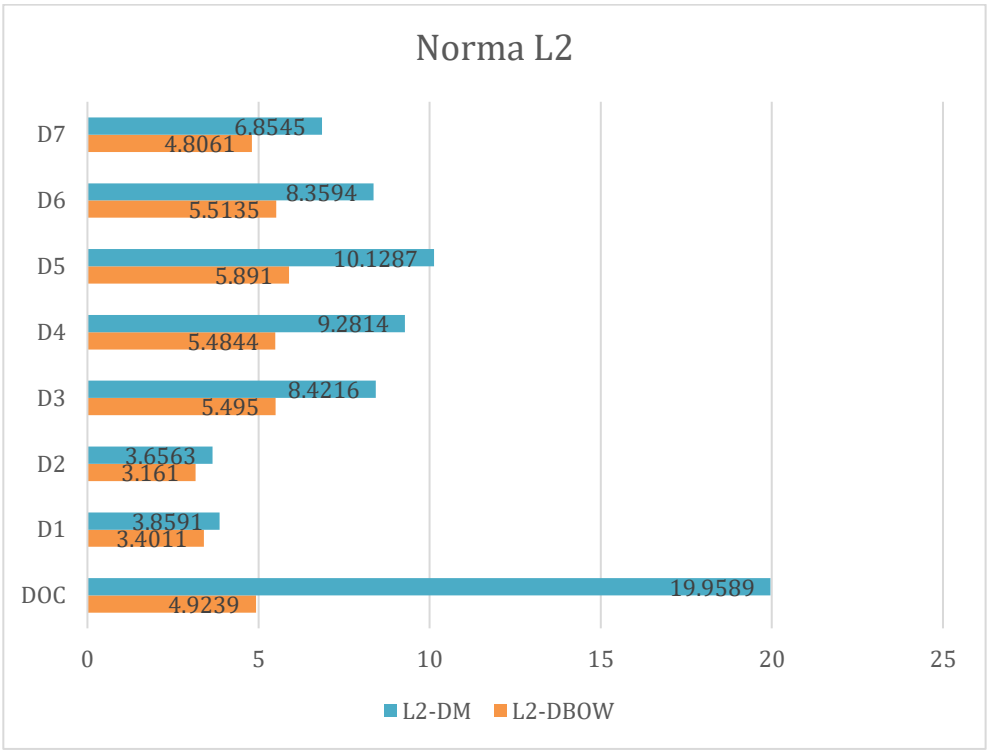

Figura 10. Norma L2 con vector de 500 dimensiones

CONCLUSIONES

Hemos llegado a la conclusión que la comparación de vectores puede ofrecer información sobre cuán similares pueden ser los planes de gobierno, esto es comprensible ya que ambos planes de gobierno intentan llegar a la mayor cantidad de votantes posible y abordar los temas más comunes del momento que se necesitan para convencer a la población. Dependiendo del modelo elegido podemos obtener diferentes interpretaciones entre similitud del tema (sobre todo útil usando el modelo DBOW). Pero, con el modelo DM podemos ver que distingue mucho mejor el contexto. Los resultados entre DM y DBOW en la comparación de los manifiestos completos nos indican que el modelo DBOW nos puede dar un resultado más fiable en documentos extensos ya que la similitud de coseno está acorde a la distancia euclidiana; sin embargo, comparando con el modelo DM la distancia euclidiana es notoria. No podemos decir lo mismo de las comparaciones entre dominios donde la similitud de coseno está acorde a la distancia euclidiana por lo que podemos concluir que para comparaciones generales es ideal usar el modelo DBOW, pero para comparaciones de temas específicos el modelo DM resulta mejor para comparar los planes de gobierno por partes.

El tamaño del vector generado por el modelo también es importante, los vectores de 100, 200 y 500 tienen resultados similares para determinar qué dominios son más similares entre ambos manifiestos. Podemos determinar, por los resultados el vector de tamaño 500 es demasiado grande y empieza a sufrir de *overfitting*, donde el modelo es demasiado específico y empieza a determinar que todos los documentos son diferentes debido a la cantidad baja de similitud de coseno. El tamaño de vector 100 o menor en cambio generaliza demasiado los documentos y empieza a sufrir de *underfitting*, que es cuando el modelo empieza a decir que todos los documentos se parecen y eso explica porque los resultados de la similitud de coseno son altas. Ambos casos nos sirven para determinar que el modelo no debe ni ser demasiado

estricto ni generalizar demasiado ya que entre los extremos se empiezan a perder las diferencias entre los dominios, por lo que concluimos que es el tamaño de vector ideal es 200 ya que los valores de similitud de coseno y distancia euclidiana son acordes.

En este trabajo se ha usado como cuerpo lingüístico los artículos escritos en español de Wikipedia, puede ser un cuerpo lingüístico que usa el español neutro y que funciona para comparaciones muy generales. Como trabajo a futuro se podrían crear nuevos modelos con cuerpos lingüísticos especializados en los mismos temas que abarcan los dominios, aunque si se realiza esto hay que tomar en cuenta el cambio de hiper parámetros, los parámetros que se usaron en este trabajo son recomendados para cuerpos lingüísticos muy grandes y generalizados.

REFERENCIAS BIBLIOGRÁFICAS

Abercrombie, G., Nanni, F., Batista-Navarro, R., & Ponzetto, S. P. (11 de 2019). Policy Preference Detection in Parliamentary Debate Motions. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (págs. 249–259). Hong: Association for Computational Linguistics. doi:10.18653/v1/K19-1024

Alonso, S., Volkens, A., Cabeza, L., & Gómez, B. (2012). *The content analysis of manifestos in multilevel settings: Exemplified for Spanish regional manifestos.* Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). Retrieved from http://hdl.handle.net/10419/57768

Ares, C., & Volkens, A. (2017). ¿Por qué y cómo se está extendiendo el Manifesto Project a América Latina? *Revista Espa textasciitildenola de Ciencia Política*, 115-135. doi:10.21308/recp.43.05

Bilbao-Jayo, A., & Almeida, A. (2018). Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks, 14*, 1550147718811827. doi:10.1177/1550147718811827

Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. *Procedia Computer Science, 157*, 160-167. doi:https://doi.org/10.1016/j.procs.2019.08.153

Knapp, A. (2005). *Basic real analysis: along with a companion volume Advanced real analysis.* Birkhäuser.

Krause, W., Lehmann, P., Theres, M., Merz, N., Regel, S., & WeÃŸels, B. (2020). The Manifesto Data Collection: South America. Version 2020b. *The Manifesto Data*

Collection: South America. Version 2020b. Berlin: Wissenschaftszentrum Berlin fÃ¼r Sozialforschung. doi:10.25522/manifesto.mpdssa.2020b

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *CoRR, abs/1405.4053*. Obtenido de http://arxiv.org/abs/1405.4053

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Efficient Estimation of Word Representations in Vector Space*.

Ramadhanti, N. R., & Mariyah, S. (2019). Document Similarity Detection Using Indonesian Language Word2vec Model. *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, (págs. 1-6). doi:10.1109/ICICoS48119.2019.8982432

Volkens, A., Burst, T., Krause, W., Lehmann, P., MatthieÃŸ, T., Merz, N., . . . Zehnter, L. (2020). The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2020b. *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2020b*. Berlin: Wissenschaftszentrum Berlin fÃ¼r Sozialforschung. doi:10.25522/manifesto.mpds.2020b

Zapata, P., & Peignier, S. (2017). *Análisis del Discurso Socialista Latinoamericano Basado en Inteligencia Artificial.* Instituto Internacional de Integración Convenio Andrés Bello. Obtenido de https://hal.archives-ouvertes.fr/hal-01766739

Zhu, L., Wang, G., & Zou, X. (2016). A Study of Chinese Document Representation and Classification with Word2vec. *2016 9th International Symposium on Computational Intelligence and Design (ISCID), 1*, págs. 298-302. doi:10.1109/ISCID.2016.1075