Trabajo final: Modelado de tweets para campañas electorales

Jorge Guerra Laura Grandas Rafael Torres Ana María Patrón

22 de agosto de 2024

Introducción

En los últimos años, Twitter se ha convertido en una red social en la que más de 320 millones [1] de usuarios comparten, en máximo 140 caracteres, información de eventos a tiempo real, experiencias pasadas y opiniones sobre temas políticos, económicos, sociales, entre otros. Precisamente, previo a las elecciones de 2020 en Estados Unidos, esta herramienta fue usada por millones de personas para expresar sus preocupaciones y su posición política frente a las elecciones. En este sentido, el problema de este trabajo es 1) abordar si es posible clasificar temáticamente los tweets relacionados con la campaña electoral en EEUU; 2) de ser así, evaluar si esta clasificación temática presenta patrones de aglomeración geográfica y 3) evaluar si el tema predominante en cada distrito electoral se relaciona con que este elija a un candidato de un partido sobre el otro.

La relevancia de este trabajo consiste en su contribución al diseño de campañas electorales. Conocer de primera mano cuales son los principales temas que le preocupan a cada distrito electoral previo a la elecciones permite que los candidatos sepan que discurso dar en la competencia electoral en cada lugar. La herramienta que prove este trabajo es poderosa para el quehacer político en la medida en que permite realizar procesos de campaña más efectivos y de elección más conscientes. Cabe destacar que este trabajo es pionero en aplicar la metología LDA a campañas electorales y se espera que el método aquí creado para el caso de Estados Unidos pueda replicarse fácilmente a otro países, como Colombia[2].

El modelado de texto es un problema que ha sido abordado por diferentes autores. Blei, NG y Jordan (2003) en su paper *Latent Dirichlet Allocation* realizan una descripción teórica detallada de la Asignación Latente de Dirichlet (LDA). La presentan como un modelo probabilístico generativo que permite clasificar colecciones de datos discretos.En nuestro contexto de modelado de tweets para la elecciones de EEUU, este algoritmo nos permite evaluar si es posible hacer una clasificación temática de los tweets y de las palabras que contienen.

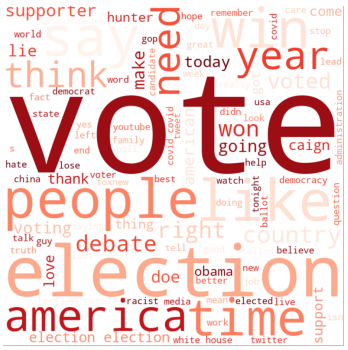
Por otro lado, Lansley y Longley (2016) en su paper *The geography of Twitter topics in London* estudian como la hora, de un día típico de 2013, en el que se publica un tweet se correlaciona con las actitudes expresadas en este. La contribución de este trabajo consiste en ser una aplicación del método LDA, en concreto al comportamiento social de los londinenses. La manera en que se manejaron los datos y en que se clasificaron en este paper enriquece nuestro arsenal para enfrentarnos a los retos empíricos con los datos nuestro trabajo.

[1] Twitter, 2021

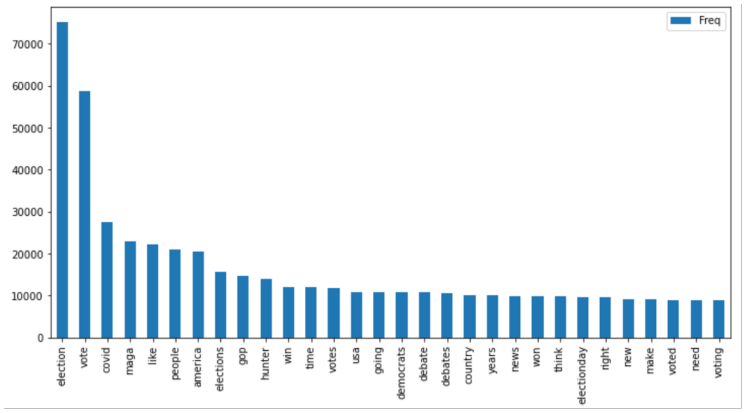
[2] El código usado se construyó autónomamente a lo largo del semestre y se encuentra disponible en https://github.com/jguerrae/Talleres-ML/tree/main/Proyecto%20final

Datos: descripción y tratamiento

Los datos se obtuvieron de la base de datos US Election 2020 Tweets disponible en Kaggle. Esta base se compone de una base con tweets que apoyaban a Trump y otra con tweets que apoyaban a Biden. Los tweets se publicaron entre el 15 de octubre y el 8 de noviembre de 2020, en total son 1.7 M de tweets y 23 variables, entre estas están ubicación (reportada por wifi), latitud, longitud,likes, retweets y una dummy que toma el valor de 1 si el tweet si apoya a Biden y 0 a Trump. Esta base se filtró para los tweets que se publicaron en EEUU. Además, se realizó una limpieza teniendo en cuenta aspectos como la separación de textos en los bloques básicos de análisis o tokenización; la gramática, la eliminación de pronombres, conectores, preposiciones, artículos y demás palabras irrelevantes para la clasificación o stopwords; así como la eliminación de palabras comunes del país que no aportaran nada a la clasificación. Finalmente se obtuvo una base con 394.390 observaciones. La figura 1 muestra que las palabras más comunes en la base resultante son election, vote y covid, las cuales aparecieron más de 70.000, 55.000 y 25.000 veces, respectivamente.



(a) Nube de palabras más comunes



(b) Frecuencia de palabras más comunes

Figura 1: Descriptivas

Algoritmos

A fin de abordar el problema planteado se emplearon dos métodos. Para la clasificación temática de los tweets se emplea el método de aprendizaje no supervisado de Asociación Latente de Dirichlet (LDA). Para evaluar si esta clasificación presenta patrones de aglomeración geográfica, dado que contabamos con la ubicación de los tweets y los distritos, se usará un sistema de georreferenciación (GIS) para unir características electorales a tweets. Este sistema también permitirá hacer el cotejo del partido ganador con el tema que predominaba en cada distrito electoral.

I.*Asociación Latente de Dirichlet (LDA):*

Este modelo tiene un proceso probabilístico generativo de tres niveles, el cual asume que cada tema se compone de un conjunto de palabras y a su vez cada tweet es una mezcla de varios conjuntos de probabilidad de un tema (Blei et al. 2003). En palabras sencillas, el modelo asume que cada documento es una colección de temas con una distribución ponderada determinada por los términos que la componen.

El paso inicial fue definir el número de temas. El algoritmo LDA se aplicó a la base de datos y se calculó para el caso con 1 tópico hasta el caso con 20 tópicos. Para cada caso, se obtuvo una distribución visual donde el tamaño del círculo representa la proporción de tweets que tienen

una probabilidad más elevada de pertenecer al tema y la distancia entre los círculos refleja las diferencias (lejanía) o similitudes (cercanía) entre los temas. Con base en esta distribución visual, se consideró que el caso con 7 tópicos era el mejor escenario (figura 2).

En el siguiente paso, cada uno de los 7 tópicos se nombró teniendo en cuenta las palabras que lo

Figura 2: Escenario con 7 tópicos

caracterizaban. En concreto, los 7 temas se pueden entender así:

- **Tópico 1: Estímulos a la inversión**
  Algunos de sus términos más importantes son "impuestos", "negocios", "dinero" y "familia", aspectos clave de la campaña de Trump. El tema parece atraer a republicanos moderados y demócratas e independientes de derecha. Podría considerarse que se inclina a favor de Trump. Geográficamente, en el norte, centro y oeste del país, es dominante en los distritos electorales clasificados como "seguros"[3] para cualquiera de los partidos. Sin embargo, en los estados del sur, como Nuevo México, Georgia y Florida, es dominante en los distritos en los que perdería el partido incumbente, como NM-2 y GA-7.

- **Tópico 2: Campaña directa a Trump**
  Los términos más relevantes son los lemas de la campaña de Trump "MAGA"(Make America Great Again) y "KAG"(Keep America Great), además de términos como "victoria aplastante[4] MAGA", "victoria" y "patriota", términos que se refieren a los partidarios de Trump y perspectiva de una gran victoria republicana en la noche de las elecciones. Así, estos términos apuntan hacia "Vote por Trump".

- **Tópico 3: Fraude electoral**
  Sus términos más relevantes apuntan hacia la perspectiva de un posible fraude electoral, debido al uso de boletas por correo. Términos dominantes como "correo", "fraude", "amañado", "margen" y "recuentos", sugieren que la gente estaba preocupada por la transparencia de la elección. Cabe resaltar que esta idea del fraude electoral fue fuertemente promovida por Trump y un vasto sector del Partido Republicano, meses antes de las elecciones.

[3] Esta clasificación la hacen los medios y analistas, y varía entre cadenas. Un distrito se considera seguro para un partido si la probabilidad de perder la elección es muy baja.

[4] "Landslide"

- **Tópico 4: Día de elecciones**
  Este tópico parece ser concerniente al momento puntual del día de las votaciones. Los términos más comunes son elections, election day y sus combinaciones. Es posible concluir también, que este tópico es neutral debido a la presencia de palabras para ambos partidos. Por un lado, ser observan acrónimos como maga (Make Great America Again) y GOP (Grand Old Party) por parte del partido Republicano y palabras como voteblue para el partido Demócrata. Cabe aclarar que aparecen mencionados estados como North Carolina, Florida, Ohio, Wisconsin y Pennsylvania. Estos estados fueron clave para ganar las elecciones, ya que sus resultados fueron sorpresivos y previo a las elecciones, presentaban indecisión. La discusión parece girar en torno a la importancia de votar por un candidato, en particular en los principales estados indecisos.

- **Tópico 5: COVID 19**
  Evidentemente, el tópico 5 al contener palabras como virus, pandemic, disease, covid, cases, deaths, es relacionado a la pandemia por el COVID 19 y la crisis económica y sanitaria derivada. Las preocupaciones por las muertes, la crisis económica y las medidas políticas indican que la respuesta al COVID-19 es un tema de debate importante para los votantes.Cabe resaltar que la mayoría de la propaganda demócrata se centró en responsabilizar a Trump de las muertes de COVID en los EE. UU.

- **Tópico 6: Medios de comunicación en torno a las elecciones**
  Sus términos más relevantes apuntan hacia las redes sociales y la prensa. Las menciones a "Twitter" y "Facebook" vienen junto con "NBC", "New York Post" y "ABC". Este tema está relacionado con el intento de los demócratas de convencer a otras personas de que voten en azul citando las redes sociales y los medios demócratas. El "bloque secundario"de bluewave, wtpblue, wtpsenate, voteblue ayuda a ubicar este tema como a favor de los democratas. Este tema también invita implícitamente a los demócratas a votar por las elecciones de la Cámara y el Senado, no solo en las presidenciales.

- **Tópico 7: Campaña directa a Biden**
  Profundamente a favor de Biden y anti Trump. Se refiere a los escándalos de violación a derechos humanos de la administración Trump. Incluye términos como racista, fracking, padres, hijos, blm (las vidas de los afroamericanos importan), morir, antidumping, asesinato en la cárcel, veterinario para la ciencia, niños, negligencia antomicida. Es decididamente anti trump e incluye otros términos como votehimout, dump, anti, voteblue, blm, votebluetosaveamerica, byedon, dump. Hay algunos términos relacionados con la propaganda republicana (s.a. MAGA), pero esos términos también se encontraron en los tweets demócratas que critican a Trump, por lo que su inclusión no es una razón suficiente para considerar el tema como Neutral.

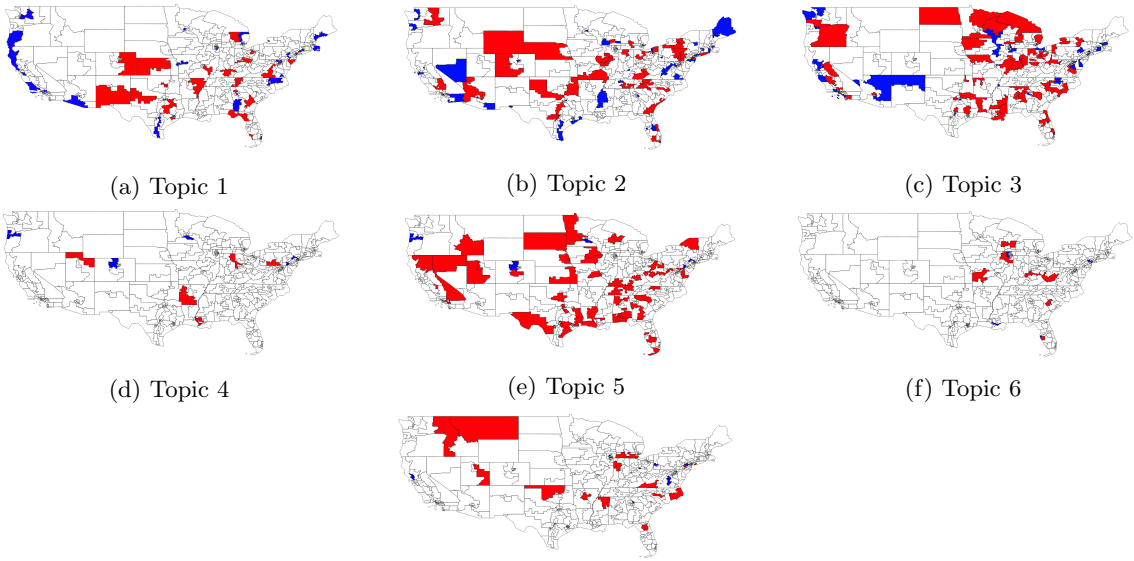
El resultado de aplicar este algoritmo fue clasificar temáticamente los 394.390 tweets sobre las elecciones de 2020 en EE.UU

II. Sistema de georreferenciación (GIS):

Para construir este sistema, el paso inicial fue asignar a cada tweet un peso dado por W = 2R + L, donde L representa el número de me gusta y R el de retweets, esta metodología se extrajo de Perdana et al. Seguidamente; usando las coordenadas de cada tweet, se recortó cada punto (tweet) al mapa de distritos del Congreso de 2020, el cual cuenta con 435 distritos[5]. Por lo tanto, cada tweet se asignó a un solo distrito del Congreso. Luego, teniendo el peso tweet W y su tema dominante, se definió que un tema era dominante en un distrito si y solo si la suma del peso sobre los tweets dentro de él era el máximo entre los temas. El resultado de este algoritmo fue obtener un mapa para cada tema, donde se resaltaban los distritos electores en donde dicho

[5]El directorio de distritos usado se encuentra disponible en https://www.270towin.com/2020-election-results-live/house/

tema era predominate. Después de esto; se identificó el partido ganador por cada distrito con los datos recopilados por Associated Press. Posteriormente, para cada mapa obtenido anteriormente, se asignó un color rojo a los distritos en los que dominaba el tema de interés y en los que el partido ganador fue el Republicano y azul en los que dominaba el tema de interés y en los que el partido ganador fue el Demócrata. La figura 3 muestra el resultado de este algoritmo.



(a) Topic 1 (b) Topic 2 (c) Topic 3

(d) Topic 4 (e) Topic 5 (f) Topic 6

(g) Topic 7

Figura 3: Tema dominante y asignación de escaños en la Cámara. En cada panel, el distrito del Congreso está coloreado si el tema en el título es dominante allí. En rojo, distritos electorales ganados por republicanos. En azul, distritos electorales ganados por demócratas.

Resultados

Como muestra el gráfico a, el tema 1 es discutido tanto en la costa este como en la costa oeste y es un tema que preocupa casi que parejamente a los seguidores del Partido Democráta como al Republicano. Por otro lado, no es posible distinguir con claridad que el tema de los estímulos a la inversion preocupe más a los seguidores de Biden o de Trump.El tema 2 tiende a preocupar en mayor medida a los seguidores del Partido Republicano como se evidencia en el gráfico b. Esto es un resultado alentador en cuanto según la clasificación que realizamos con LDA, este era el tema que directamente apoyaba a Trump.No obstante, no es posible evidenciar patrones de aglomeración geográfica.

El gráfico c muestra que el tema 3 presenta una agomeración geográfica hacia el este del país. Así, en los distritos como Dakota AL, Minessota 8, Wisconsin 7, entre otros, las personas están preocupadas por un posible fraude electoral.Además, el que la mayor parte de los distritos que aparezcan resaltados tenga color rojo, muestra que este es un tema más recurrente entre los seguidores del Partido Republicano y de nuevo, esto es consistente con los resultados obtenidos al aplicar el método de Asignación Latente de Dirichlet.

Asimismo, el tema 4 no parece preocupar mucho a los estadounidenses pues son pocos los distritos electorales en donde este es el tema predominante. Así, el día de las elecciones en sí mismo no genera mucha preocupación entre democrátas y republicanos. Este tema tampoco presenta patrones de aglomeración geográfica.Por otro lado, el gráfico e muestra que el tema 5 le preocupa en mayor medida a los distritos con mayoría republicana. Este es el tema del COVID- 19 y se

concentra geográficamente en la zona sureste así como en los distritos de Dakota del Sur AL y Minessota 7.El tema 6 es el tema de los medios de comunicación en torno a las elecciones. Como muestra el gráfico f, este tema definitivamnete se encuentra aglomerado en el este y preocupa en mayor medida a los distritos en donde hay mayoría republicana.

Finalmente, el gráfico g muestra que el tema 7 se concentra en el norte, en los distritos de Montana e Idaho 1 y 2, y se encuentra esparcido en el este y en el centro del país. Este es el tema que apoyaba directamente a Biden y estaba en contra de Trump. Sin embargo, el que en los distritos resaltados tengan en mayor medida color rojo es inquietante. No obstante, la explicación que se le da a esto es que si bien, twitter es un medio que puede reflejar la ideología politica de una parte de la población, no implica que todos los votante hayan expresado sus preocupaciones a través de esta red social. Así, pudo darse el caso que en estos distritos, quienes apoyaban a Biden tenían mayor acceso a twitter respecto a los que apoyaban a Trump, pero que los últimos eran mayoría. Las elecciones que ganó Trump mostraron cómo las redes sociales y los medios más populares no necesariamente reflejan la ideología de la mayoría de la población.

Conclusiones

Las campañas electorales son un tema álgido de los países. Los candidatos buscan mostrarle a los electores porque ellos son la mejor opción, para eso idean planes de campaña que recojan las necesidades, que ellos creen, tienen los diferentes votantes. Twitter es una herramineta que permite expresar posturas y actitudes políticas de una manera informal y sin rodeos. Para el caso de las elecciones de Estados Unidos de 2020, el método de Asignación Latente de Dirchlet permitió procesar estos tweets y clasificarlos temáticamente en 7 grupos. Esta clasificación es alentadora y permite identificar cuales eran los temas que le preocupaban a cada uno de los 435 distritos electorales.

Asimismo, estos resultados permitieron evaluar que temas presentaban patrones de aglomeración geográfica. Así, el tema del fraude electoral y el rol de los medios de comunicación era una preocupación frecuente en el este; los estímulos a la inversion eran un tema que preocupaba a los votante de las costas, tanto este como oeste, y que el tema del día de las elecciones no presentaban un patrón de aglomeración geográfica definido. Finalmente, fue posible identificar que los temas 2 y 5 eran temas de apoyo republicano, y que en la mayoría de distritos donde estos eran los temas predominantes los republicanos ganaron las elecciones. Asimismo, para el tema 7 no fue posible establecer esta relación. En este sentido, se plantean retos como el de como incluir las posturas políticas de los votantes que no cuentan con acceso a twitter y que pueden definir el rumbo de las elecciones en algunos distritos electorales.

En conclusión, los resultados son alentadores para el caso de Estados Unidos, muestran un comportamiento tendencial de los votantes en cada distrito y el ejercico puede replicarse para otro países de mundo. Este método puede ser una herramienta eficaz en los procesos de campaña, en tanto permiten al votante comunicarle al candidato cuales son sus necesidades y al candidato dar el discurso político que el votante quiere escuchar.

Referencias:

Sokolova et al (2016). Topic Modelling and Event Identification from Twitter Textual Data. Tomado de: https://arxiv.org/abs/1608.02519
Lansley, G. Longley, P (2016) The geography of Twitter topics in London. Computers, Environment and Urban Systems 58 (2016) 85–96
Blei, M. Ng, Y. Jordan, M. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022