

Análisis poblacional del panorama político y perspectivas para las elecciones presidenciales de 2022 en Colombia.

1. Introducción

Hoy en día y desde siempre la política ha representado uno de los ámbitos más marcados en la sociedad, la postura política de un individuo usualmente recae en quien considera representa sus ideales (aunque esto no siempre ocurra); La política se encuentra intrínsecamente relacionada con poder y este resulta ser codiciado por un grupo puntual de individuos de la sociedad, específicamente los partidos políticos y a su vez de todos sus filiales y contribuyentes pues aunque la figura de Presidente represente un sustantivo singular de este dependen muchos de los altos cargos y toma de decisiones futuras para todos los habitantes.

Aunque la política en Colombia se ha distinguido por el bipartidismo desde la constitución del 1991 se da vía libre al multipartidismo con lo cual los partidos políticos cobran una figura de gran influencia y a su vez los encargados de representarlos y competir con sus homólogos. En la actualidad compañías destacadas como GUARUMO, INVAMER, MOE y el CNC, Centro Nacional de Consultoría son los encargados de realizar encuestas, sondeos y datos respecto al panorama político en Colombia proporcionando una visión general a la población respecto a las preferencias y favoritismos políticos de la mayoría, pero además proporcionando la llave maestra a los partidos y sus candidatos respecto a cual grupo social de sufragantes podría ser crucial ganar en la siguiente contienda electoral.

El presente proyecto de Ciencia de datos empleando el proceso estándar entre industrias para la minería de datos (CRISP-DM) pretende estudiar las estadísticas políticas registradas en los últimos periodos electorales dando un panorama de necesidades y preferencias de los ciudadanos el cual resulta de gran interés y utilidad para partidos políticos, agencias de publicidad, gremios sociales-económicos y otras diferentes esferas sociales.

Estructuración del sistema electoral colombiano

Sistema Electoral Colombiano: Sistema de elección popular donde los integrantes de los partidos políticos son escogidos para los diferentes cargos gubernamentales. Se encuentra a cargo del CNE (Consejo Nacional Electoral) y Registraduría Nacional del Estado Civil, su propósito es velar por el buen cumplimiento de las normas electorales del Código electoral colombiano

CNE (Consejo Nacional Electoral): -Organismo autónomo e independiente parte de la organización electoral establecido mediante el artículo 120 de la constitución política colombiana -Se encuentra conformado por 9 magistrados elegidos en Congreso

Registraduría Nacional del Estado Civil: Igualmente, que el CNE es una institución descentralizada de carácter público encargada de la organización electoral Según establece el artículo 266 de la Constitución Política de Colombia, el Registrador Nacional del Estado Civil será escogido por los presidentes de la Corte Constitucional, la Corte Suprema de Justicia y el Consejo de Estado, mediante concurso de méritos organizado según la ley.

Código electoral colombiano: Actualmente rige el código electoral establecido en el decreto 2241 de 1986 en el cual se estipulan autoridades electorales, delegados, partes, funciones, disposiciones y capacidades de las distintas partes de los miembros electorales Actualmente se encuentra en trámite como ley estatutaria un código electoral en Congreso (Cámara y Senado deciden)

Objetivo general:

Analizar los comportamientos electorales de la elección de la figura Presidencial en Colombia en los últimos años mediante los datos registrados de sufragios en las anteriores contiendas analizando variables como la dominancia política, la representatividad y no representatividad, voter turnout, ergonomía electoral entre otras variables contrastando con las proyecciones actuales de los candidatos presidenciales para el periodo electoral 2022-2026.

Objetivos específicos:

- Dar un panorama político por departamentos con base en sus comportamientos electorales previos
- Adquirir bases de datos referentes a las elecciones de los periodos 2014-2018 y 2018-2022 de los datos públicos recopilados por gobierno y registraduría
- Establecer las distintas correlaciones entre las variables presentadas para sacar mayor provecho de las bases proporcionadas combinando datos por su ubicación, participación y preferencia política

Criterios de éxito del Proyecto

- Tener un rango de proximidad con base en el análisis hecho frente a las encuestas actuales para los distintos departamentos y regiones.
- Concluir en tiempo oportuno el análisis electoral propuesto.
- Obtener posibles escenarios de los cuales se puedan derivar estrategias de utilidad para los distintos actores interesados en la próxima contienda electoral.

Recursos

- Computadora con lenguaje de programación Python 3.8.5 y librerías: Pandas, Numpy, Openpyxl, Twint, Re, Time, TextBlob
- Datos proporcionados por entidades gubernamentales sobre resultados y sondeos de elecciones, preferencias partidarias y candidaturas anteriores.
- Equipo de expertos y analistas en contiendas políticas y asesores de campañas

inconformidad, con efectos políticos” y agrega que “el voto en blanco constituye una valiosa expresión del disenso a través del cual se promueve la protección de la libertad del elector

Votos nulos: Cuando la marcación efectuada por el elector no permite determinar de manera clara la intención del sufragante, es decir que no es posible identificar a favor de qué candidato votó.

Tarjetones no marcados: La tarjeta no marcada es aquella en la cual a pesar de que el ciudadano la recibe, no la marca y la deposita en la urna sin haber seleccionado alguna opción.

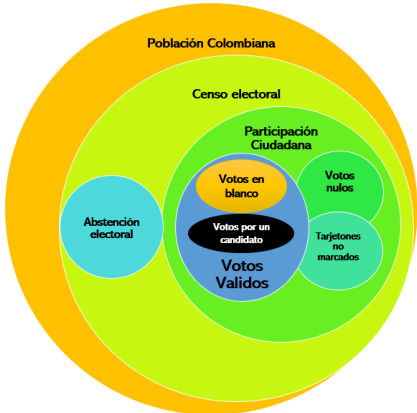


Figura 1: Jerarquía de variables presentes en los datos

Datos.

Para entender los datos obtenidos primero debemos entender la definición de las variables presentadas en los datos, así como su jerarquía y correlaciones para un mejor entendimiento de estos y sus implicaciones.

Code: La Codificación de la División Político-Administrativa de Colombia es una nomenclatura estandarizada, diseñada por el DANE para la identificación de entidades territoriales (departamentos, distritos y municipios), áreas no municipalizadas y centros poblados, mediante la asignación de un código numérico único.

Municipio

Censo Poblacional: El censo de población y vivienda realizado consiste en contar y caracterizar las personas residentes en Colombia, así como las viviendas y los hogares del territorio nacional

Censo Electoral: El censo electoral determina el número de electores que se requiere para la validez de los actos y votaciones. Es también el instrumento técnico, elaborado por la Registraduría Nacional del Estado Civil, que le permite a la Organización Electoral planear, organizar, ejecutar y controlar los certámenes electorales y los mecanismos de participación ciudadana”.

Abstención: Abstención, en ciencia política, es el acto por el cual un potencial votante en unas elecciones decide no ejercer su derecho al voto, ya sea en unas elecciones generales o si, en un procedimiento parlamentario, el representante está presente en la votación, pero no vota. En ambos casos, el sujeto se atiene al resultado del voto de los electores que sí votaron, al igual que ocurre con el voto en blanco.

Votos en blanco: De acuerdo con la sentencia C-490 de 2011 de la Corte Constitucional, que declaró la exequibilidad de la Ley 1475 (Reforma Política), el voto en blanco es “una expresión política de disenso, abstención o

Al igual que las variables una caracterización descriptiva de la construcción y obtención de las bases de datos es presentada a continuación:

ARCHIVO	2014PresidentePrimeraVuelta	2014PresidenteSegundaVuelta	2018PresidentePrimeraVuelta	2018PresidenteSegundoVuelta
DESCRIPCIÓN	Tabla de datos registrados en las elecciones presidenciales del 25 de mayo de 2014 informando código de DANE, municipio, departamento, candidatos, votos, participaci	Tabla de datos registrados en las elecciones presidenciales del 15 de junio de 2014 informando código de DANE, municipio, departamento, candidatos, votos, participaci	Tabla de datos registrados en las elecciones presidenciales del 27 de mayo de 2018 informando código de DANE, municipio, departamento, candidatos, votos, participaci	Tabla de datos registrados en las elecciones presidenciales del 17 de junio de 2018 informando código de DANE, municipio, departamento, candidatos, votos, participaci
CONSTRUCCIÓN	Misión de Observación Electoral MOE	Misión de Observación Electoral MOE	Misión de Observación Electoral MOE	Misión de Observación Electoral MOE
FUENTE DE DATOS	Registraduría Nacional del Estado Civil	Registraduría Nacional del Estado Civil	Registraduría Nacional del Estado Civil	Registraduría Nacional del Estado Civil
FECHA	Noviembre de 2014	Noviembre de 2014	Enero de 2019	Enero de 2019

Figura 2: Metada de los datos utilizados

El primer aspecto para analizar es la demografía electoral presentada en torno a la participación ciudadana, analizar el crecimiento o decrecimiento de esta implica factores tanto sociales como inherentes. Nuestro periodo para analizar es el inmediatamente más cercano en donde se evidencia el siguiente incremento de la participación y decrecimiento del abstencionismo.

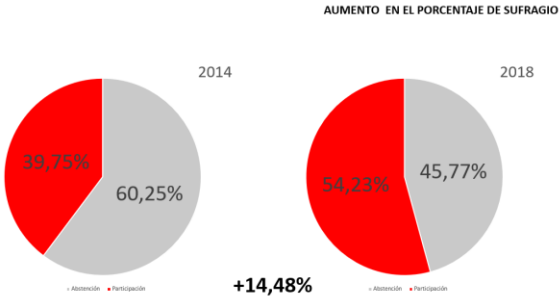


Figura 3: Porcentajes de participación y abstención en 2014 y 2018

Comprender las variables de participación y abstención siempre ha sido un tema de gran atención, implica cuestiones como cercanía y control de puestos de votación, observación y

transparencia en el sufragio, inscripción de cédulas etc. además de fundamentalmente la concentración de población y sus condiciones ya sea bien ruralidad o urbanidad, por lo anterior resulta fructífero realizar un análisis específico por territorios pues algunos presentan problemas puntuales que afectan directamente la partición. La división de censo y participación electorales de la última contienda presidencial por departamentos correspondiente a 2018-II se presenta a continuación.

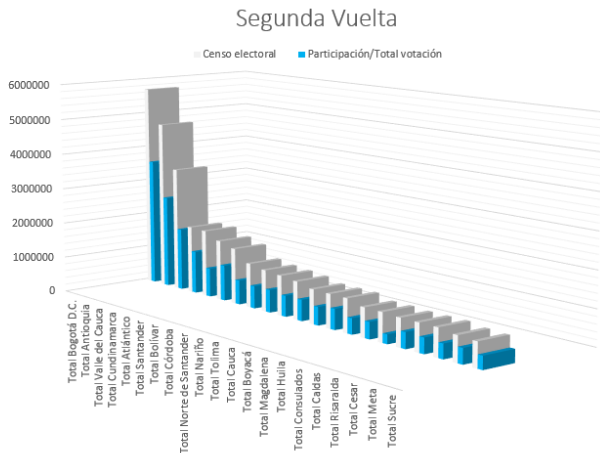


Figura 4: Censo frente a participación por departamentos

Algo importante a destacar del análisis fragmentando por ubicación geográfica es la concentración de poder de los territorios lo cual es claramente razonable que se ubique en los departamentos que poseen las principales capitales con gran presencia económico-política. La distribución de esta en el territorio colombiano es consistente y sólida durante los últimos 30 años y se divide de la siguiente manera

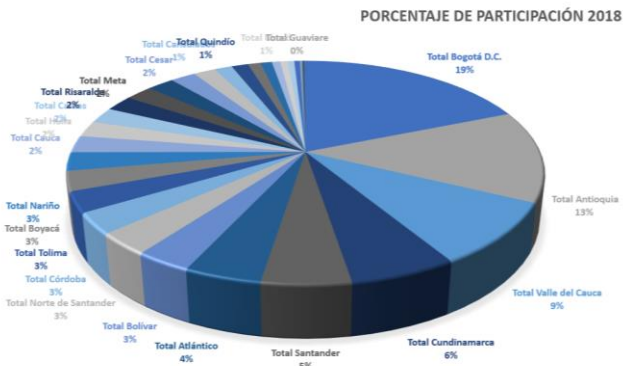


Figura 5: Porcentaje de poder de voto en los departamentos de Colombia

Esta concentración permite derivar una observación primordial, las cuantiosas sumas de los rubros de campañas electorales que son destinadas a las campañas en las capitales y en general esto no depende solo del candidato sino además de las posibles alianzas del partido político en general con los distintos gremios económico-sociales pues si bien el partido que gane la figura presidencial tiene una gran connotación este además está detrás de segundas figuras como gobernadores, alcaldes, ediles, representantes a la cámara, senadores, ediles ..etc.

En 2018 en primera vuelta el partido político Alianza verde mediante el candidato Sergio Fajardo obtuvo la ciudad de Bogotá, única mayoría en un territorio del país frente a sus otros contrincantes y aunque mantenía un relativo tercer puesto en el resto de los departamentos el centro de su campaña siempre fue la capital y esto se vio reflejado en las urnas rozando ligeramente al candidato Gustavo Petro del partido Colombia Humana en un poco menos de dos puntos porcentuales.

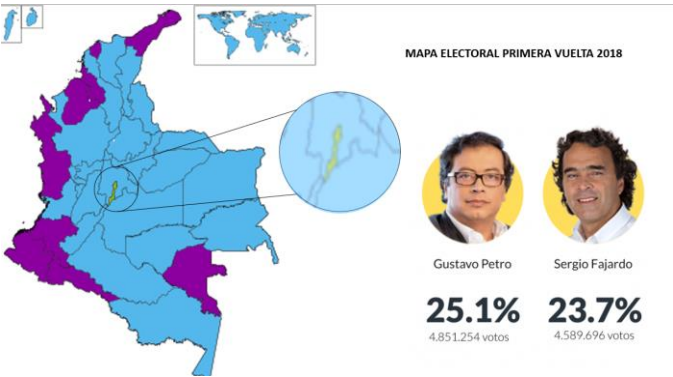
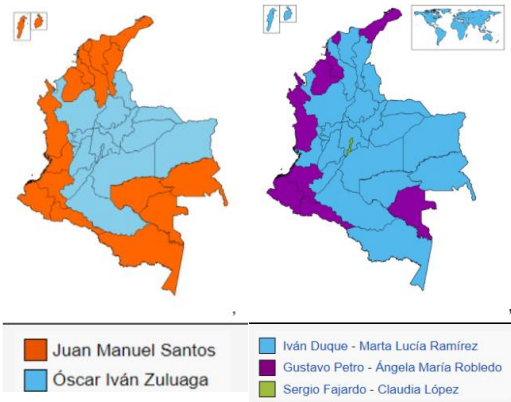


Figura 6: Mapa electoral 2018 Primera vuelta

Transiciones electorales asociadas a Partidos Políticos.

Desde el bipartidismo planteado en la introducción se han visto ligaduras políticas de ciertas regiones del país con respecto a algunos partidos políticos ya sea porque por sus mismos ciudadanos son militantes de estos o de alguna forma los ciudadanos y población en general se ha visto beneficiada por sus políticas. Colombia en los últimos años ha mostrado batallas no solo entre partidos políticos de una misma corriente política, es el caso en 2014 del Partido la U en conjunto con la denominada 'Unidad nacional' vs el recién creado 'Centro Democrático' partido del expresidente Álvaro Uribe el cual tenía como ministro de defensa a su ahora nuevo contrincante Juan Manuel Santos.

Las transiciones políticas reflejan la poca representación que los ciudadanos ven en los candidatos presentes, esto lleva a la creación de pequeños grupos y movimientos que van ganando fuerza mediática, es el caso del posicionamiento del partido Colombia Humana representación de la corriente de izquierda política, la cual por cierto divide el último mapa político obligando no solo a nuevamente formar alianzas entre partidos de una misma corriente para ganar mayorías sino además confrontar relativamente estas nuevas corrientes entre sí.



La consistencia de una decisión de voto en un departamento en las diferentes contiendas es un punto clave a analizar pues esta fidelización es lo que permite anticipar que es muy probable que un candidato del partido favorito gane las elecciones en esta región. Este periodo por venir 2022 las cartas son claras en el juego, nuevamente se los dos extremos políticos ideales en su gran mayoría contrarios y la corriente de centro de la cual toma algunas políticas de cada uno.

Centro Democrático



Antioquia	1375965
Bogotá D.C.	989744
Valle del Cauca	527622
Cundinamarca	479936
Santander	442219

Figura 8: Cantidad de votos obtenidos por el partido CD en 2018-I

Es natural que los brazos políticos de votos de los partidos estén en las concentraciones de este, en el presente estudio no solo se refleja la preferencia de los territorios a un candidato en primera estancia sino además el porcentaje de transferencia de votos que podría ser adquirido frente a un escenario en el que el candidato escogido ya no se encuentra en el tarjetón, como lo son las segundas contiendas electorales.

Antioquia

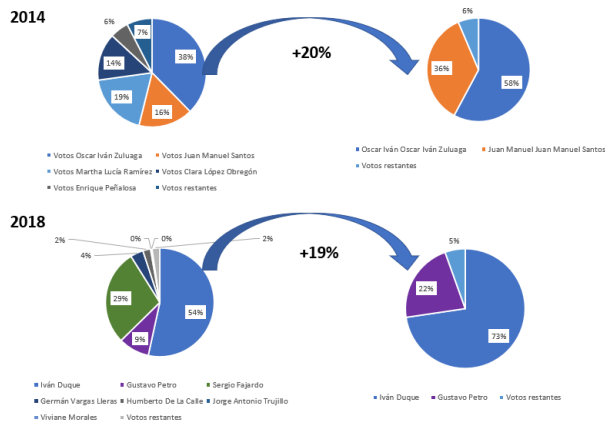


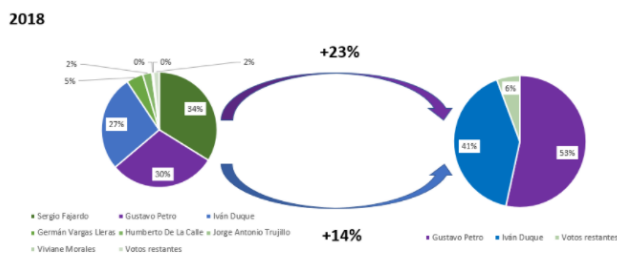
Figura 9: Transferencia de votos primer segunda vuelta del CD

Colombia Humana



Bogotá D.C.	1099955
Valle del Cauca	485867
Atlántico	331979
Córdoba	250437
Bolívar	242471

Bogotá



Machine Learning

Como lo vimos en los datos del proyecto, estos no solo ofrecen la cantidad de votos obtenidos por un candidato, sino que además generan todo un entorno y caracterización demográfica respecto al entorno político-social y preferencias ciudadanas, el cual resulta de gran importancia poder observar y comprender. Así mismo, dentro del desarrollo del proyecto se planteó el contraste cuantitativo como medida de desempeño y éxito de las cifras obtenidas de las distintas variables a analizar frente a los sondeos actuales realizados por las firmas encuestadoras donde si bien se posee un alto nivel de congruencia respecto a las proyecciones y los resultados en las urnas también se presenta un margen de error considerable pues aunque las firmas encuestadoras intentan tomar un grupo muestral característico este aún puede verse aún reducido en comparación con el tamaño real de la población. Hoy en día el Machine Learning es fuertemente aplicado en el entorno social, desde la publicidad mostrada a cada usuario hasta el mismo algoritmo que los clasifica entre sí y por lo tanto el ámbito político no se queda atrás pues hoy en día las redes sociales son el primer lugar para divulgar las posturas respecto a algún tema y en particular políticas. La aplicación del presente proyecto al Machine Learning que se pretende abordar es como las redes sociales pueden establecer métricas más exactas de las posturas políticas de los individuos tomando no solo un grupo más cuantioso sino además con mayor activismo y presencia en divulgación e influencia clave en la sociedad de hoy en día.

Definición del problema

Salida: La buena ejecución del problema de ML debería proporcionar si una persona es afín o no a un candidato o hay una posible intención de voto de los distintos ciudadanos en los diferentes territorios y departamentos a un candidato proporcionando una polaridad positiva negativa o neutra, esto de forma análoga a lo obtenido en la analítica de datos, pero esta vez mediante el análisis de las opiniones y comentarios publicados en tiempo real en las distintas redes sociales de gran flujo de texto y fácil accesibilidad como por ejemplo Twitter.

Entrada: Para la recopilación de la entrada de datos a trabajar en el proceso de analítica de datos en este caso puntual los tweets, se deben tener en cuenta distintos factores de selección y clasificación, esto con el fin de obtener una mayor calidad en la información en materia de magnitud y contenido, algunos de los más relevantes para tener en cuenta podrían ser:

- **Selección reciente:**

Si bien una parte de la población es fiel a sus posturas unas las cuestiones que más se analiza en una contienda electoral es el 'sentimiento popular' y la variación de este por lo cual un escaneo a la red de a lo más 3-4 meses previos a las elecciones contienen información más sólida.

- **Mayor cantidad de tráfico social**

Existen horas del día donde el tráfico de redes sociales es mucho mayor comparado con otras franjas horarias y esto en

política fuertemente conectado cuando los medios de comunican masiva se pronuncian respecto a algún tema que de cierta forma directa o indirecta atañe a algún candidato o partido político y lo anterior cruzado cuando las personas tienen más presencia en sus redes como lo son las horas sociales y no laborales.

- **Anclaje a eventos políticos**

Este se deriva del factor inmediatamente anterior y es que cuando algún personaje político se pronuncia o actúa de determinada forma respecto a algún tema esto genera una ola de comentarios y reacciones de con alto flujo y expansión en la red.

- **Geolocalización**

Resulta más que necesario seccionar los datos por territorios pues de esto depende el dominio de un candidato en un territorio pues esto se logra en la obtención de un sufragante y la expansión en sus círculos sociales más próximos como los son familiares y amigos ,aunque si bien cada voto cuenta por el análisis de datos visto en el proyecto la concentración del poder de voto se encuentra en las principales capitales; De igual forma el histórico de afinidad de partidos políticos a departamentos

Una vez recopilados los tweets producidos en determinado tiempo debe ser posible filtrarlos para solo seleccionar los que son afines a nuestro estudio político y esto se logra mediante el previo entrenamiento del modelo el cual claramente debe ser supervisado pues la conexión léxica con las distintas expresiones y la biblioteca de posibles nexos para deducir un sentimiento de intención de voto es bastante amplia

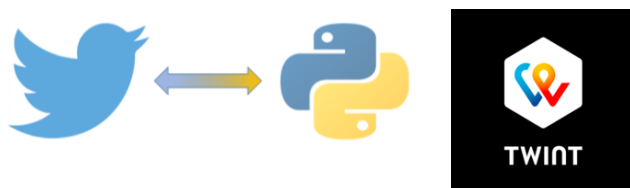


Figura 12: Librerías utilizadas para el Scraping Social Network

Hoy en día es muy común la extracción de Data de las redes sociales(“Scraping”) para diversos tipos de estudios de públicos de alcance, análisis de preferencias etc. Por lo cual en la industria se han venido desarrollando aplicaciones y algoritmos que faciliten la búsqueda y filtración de datos. Para la extracción de los tweets se tomó la librería Twint de Python la cual proporciona herramientas de gran ayuda para la filtración de información por localización, contenido de texto, fecha entre otras funciones.

Datos no balanceados y sesgos

Un posible sesgo sistemático puede provenir de la cantidad de interpretaciones que pueden tener hoy en día las expresiones incluso por más cortas que estas sean, en el trabajo de toma y filtración se deben tener varias capas de procesamiento que consideren o descarten situaciones como:

- Monosílabos sin conexión, cadenas de caracteres ilegibles
- Tweets que no presentan conexión con algún tema político
- Imágenes

- Links, sitios de redirección, tweets automatizados de página clasificadas como Spam
- Expresiones coloquiales o determinadas formas de referirse a una cosa, tema o sujeto de una forma no tan literal
- Expresiones en tono irónico o subjetivo donde no se refleje con inmediata claridad una verdadera favorabilidad o desfavorabilidad

Si bien estos filtros pueden descartar el posible ruido de la cantidad de datos también están las formas directas de anclar los tweets de importancia directa como los son nombrar de forma explícita al candidato, así como las tendencias y hashtags.

Una vez refinados y filtrados los tweets se procede a la clasificación por candidato esto con el fin de obtener de forma independiente una posición esto para evitar caer en juicios imparciales o conjuntos pues esto afectaría la efectividad del método.

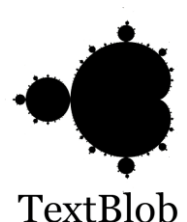
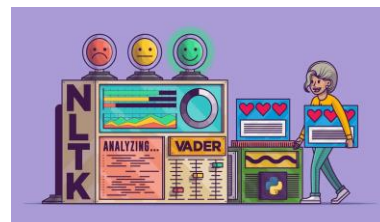
Finalmente, el último paso para retornar a un análisis de datos social con la nueva data encontrada se define en establecer la polaridad del dato en este caso tweet, esto es más conocido y usado fuertemente en la industria aplicando la ciencia de datos como ‘análisis de sentimientos’, de forma análoga al scrapping también se han desarrollado librerías y herramientas para este fin.



Figura 13: Análisis de sentimientos en la industria

Vader y TextBlob son dos de los analizadores de sentimientos más usados, al ser ambos de entrenamiento nativo en idioma inglés cualquier análisis requiere el previo de un traductor de lenguaje, el entrenamiento de un analizador resulta ser una tarea nada sencilla pues se traduce en enriquecer un diccionario computacional tanto como el lenguaje humano sin contar los posibles sesgos e inconsistencias mostradas anteriormente.

La salida de una polaridad permite deducir sentimientos con respecto a un sujeto, objeto o tema, y aunque existe una numerable conexión de palabras con sentimientos, como por ejemplo ‘Amor’ y ‘Odio’ también existen charlas intermedias que dan lugar a la subjetividad y el error, esto justifica la necesidad de un muestreo considerable para obtener mayor apreciación de las posiciones de los ciudadanos



Aplicación y deducción del panorama.

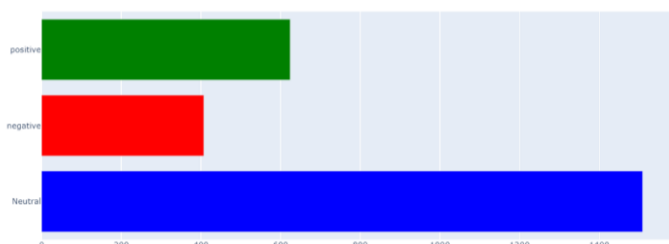
Hoy septiembre de 2021 Colombia posee 30 precandidatos presidenciales, algunos de ellos con el respaldo de su partido político y otros buscan la inscripción mediante firmas, lo cierto es que de estos 30 personajes llegaran a los más cinco o seis al tarjetón final, es decir aún restan muchas posibles estrategias, alianzas y 'coaliciones' para ganar fuerza entre corrientes políticas de donde es claro que siempre intentaran mantener congruencia entre sus ideales.



Figura 13: Candidatos presidenciales por partido 2022

Un scrapping en la red pretende tomar una muestra de la población con mayor amplitud para obtener resultados más exactos sin embargo este a su vez depende fuertemente del entrenamiento previo.

Del método utilizado y analizado en diferentes contextos siempre se posee una gran cantidad de neutralidad en la clasificación de la polaridad y esto se deriva precisamente de la dificultad y amplitud del entrenamiento computación de las intenciones del lenguaje.



Conclusiones

- Un análisis territorial y regresivo del comportamiento de electoral presentado en los últimos periodos es efectivo porque si bien existen pendularidad en la intención de voto se mantiene la congruencia en la corriente política de la representatividad del ciudadano
- A primera estancia es difícil constatar naturalmente los porcentajes actuales porque a fecha de esta publicación recién se debaten las candidaturas finales sin embargo de la baja participación por candidato presentada en las últimas encuestas es equiparable a los datos registrados por departamentos de concentración de poder analizados anteriormente
- El análisis de variables de interés secundario para los partidos y la población como lo son la abstención y los votos en blanco crea un marco efectivo para correlacionar con los entornos sociales y factores que producen dichas cifras en cada periodo
- La incursión de la ciencia de datos y la inteligencia artificial es cada vez más prominente, este proyecto fue un prototipo prematuro de las grandes cantidades de datos, análisis y rubros que son destinados a esta causa política los cuales arrojan tasas de éxito muy altas en los países donde se ha implementado.
- El criterio de éxito de una aplicación más amplia de este proyecto depende en gran medida del entrenamiento de máquina para la interpretación de sentimientos lo cual es una basta labor.
- El nivel de éxito propuesto en el modelo de machine learning claramente es mucho más prominente y preciso frente a los sondeos pues logra recopilar localización espacial y temporal en tiempo real en unos segundos de procesamiento cosa que no logran las encuestas pues dependen en mayor medida del factor humano.

Juan Camilo García Malagón
Introducción a la ciencia de datos 2021-1
Universidad Nacional de Colombia
Departamento de Matemáticas
juacgarciamal@unal.edu.co
Septiembre de 2021

Referencias

- <https://www.moe.org.co/>
- <https://www.registraduria.gov.co/>
- <https://participacion.mininterior.gov.co/sites>
- <https://www.datoselectorales.org/>
- [predicción electoral usando un modelo híbrido basado en análisis sentimental: elecciones presidenciales de colombia .https://revistas.elpoli.edu.co/index.php/pol/article/view/1651/1359](https://revistas.elpoli.edu.co/index.php/pol/article/view/1651/1359)
- [método de predicción electoral a través de modelo basado en análisis de sentimientos en](#)
- [twitter](#)
- http://opac.pucv.cl/pucv_txt/txt-8000/UCC8094_01.pdf
- <https://www.semana.com/nacion/articulo/la-proliferacion-de-candidatos-presidenciables-de-cara-al-2022/202100/>
- https://elecciones1.registraduria.gov.co/pre_pre_s_2018/resultados/2html/resultados.html
- https://elecciones1.registraduria.gov.co/pre_pre_s_2014/resultados/2html/resultados.html
- [Análisis de Sentimientos con TextBlob y VADER en Python - ▷ Cursos de Programación de 0 a Experto © Garantizados \(unipython.com\)](#)
- <https://textblob.readthedocs.io/en/dev/>
- http://www.colombia.com/especiales/elecciones_2007/preguntas/preguntas3.asp
- http://www.colombia.com/especiales/elecciones_2007/preguntas/preguntas3.asp
- <http://www.eltiempo.com/politica/partidos-politicos/elecciones-2014-en-colombia-abstencion-llego-a-60-por-ciento/14035636>
- https://www.registraduria.gov.co/?page=Elecciones_2018
- https://www.registraduria.gov.co/?page=Elecciones_2014
- <https://blog.apify.com/how-to-scrape-tweets-and-more-on-twitter-59330e6fb522/>
- [Sentiment Analysis on Twitter data regarding 2020 US Elections | by Pritam Guha | Towards Data Science](#)
- <https://github.com/twintproject/twint>