



Universitat d'Alacant
Universidad de Alicante

El análisis de opiniones en la
predicción política basados en
Twitter. Un nuevo enfoque

Jorge Arroba Rimassa



Tesis **Doctorales**

UNIVERSIDAD de ALICANTE

Unitat de Digitalització UA

Unidad de Digitalización UA



Instituto Universitario de Investigación en Informática

Escuela Politécnica Superior

El análisis de opiniones en la predicción política basados en Twitter. Un nuevo enfoque.

Jorge Arroba Rimassa

Tesis presentada para aspirar al grado de
DOCTOR POR LA UNIVERSIDAD DE ALICANTE

DOCTORADO EN INFORMÁTICA

Dirigida por:

Dr. Rafael Muñoz Guillena

Dr. Fernando Llopis Pascual

Octubre 2019

Agradecimientos

En primer lugar, quiero agradecer a mi Universidad Central del Ecuador, que me ha permitido servirle durante estos cuarenta años, como profesor, una actividad que me ha permitido estar en contacto con mentes lúcidas y ávidas de conocimiento, mis estudiantes.

Un reconocimiento imperecedero a Rafael y Fernando, mis directores de tesis, por su apoyo incondicional que desde el principio fueron encaminándome en esta investigación.

A mi familia, por su comprensión por los momentos que no pude compartir con ustedes y a ti Cynthia, gran parte de este trabajo te lo debo a ti.

Universitat d'Alacant
Universidad de Alicante

Resumen

La democracia representa la única forma “justa” de gobierno en los países, pues todos los ciudadanos pueden ejercer libremente su derecho a elegir a sus mandantes o escoger sobre temas de importancia, dados en los referéndum.

En tal sentido, el preservar los resultados electorales es una tarea de todos y una de las diversas formas de conseguirlo es tener un mecanismo por el cual pudiéramos predecir los resultados finales, para que no se manipulen o alteren.

Una forma es la realización de encuestas, estadísticamente confiables, pero que al momento de su ejecución se enfrentan a dificultades como: escasos presupuestos, marcos muestrales imperfectos, tiempos reducidos, trabajos de campo complejos, regulaciones electorales limitantes, etc.

Otra forma, es el aprovechar las nuevas formas de comunicación que tenemos las personas, las redes sociales, con las cuales expresamos libremente lo que sentimos y cuáles son nuestras preferencias, electorales también.

El utilizar la Inteligencia Artificial (IA), que en la actualidad está presente en diversos campos cotidianos del quehacer de los hombres, aplicado al análisis de la información que fluye a través de las redes sociales y específicamente al campo electoral es casi reciente.

Sin embargo, la forma en que este análisis se lo realiza tiene sus dificultades, dadas por la presencia de agentes exógenos que provocan distorsión y “ruido”; el presente trabajo incorpora una nueva metodología de predictibilidad electoral. Y es nueva porque considera que la unidad de análisis no son los mensajes sino más bien los electores, estos tienen su relevancia y ejercen la persuasión política a otros, también ellos están afincados en localidades geográficas diversas, por lo que se introducen los factores de ponderación geográficos, dada la asimetría existente en la utilización del Internet y el uso del Twitter, que es la red social, que se utilizó, dada la facilidad que presenta para la descarga de los mensajes.

La metodología se puso a prueba en uno de los campos más sensibles, que es el de predecir un resultado que tiene verificación, como son las elecciones políticas.

Esta metodología, será útil en los distintos estudios que se realicen en sociología, marketing y aquellos que requieren del conocer los comportamientos de las personas a través de las redes sociales.

Abstract

Democracy represents the only “fair” form of government in countries, as all citizens can freely exercise their right to elect their constituents or choose on important issues, given in the Referendum.

In this sense, preserving electoral results is everyone's task and one of the various ways to achieve this is to have a mechanism by which we could predict the final results, so that they are not manipulated or altered.

One way is to conduct statistically reliable surveys, but at the time of its execution it faces difficulties such as: reduced budgets, imperfect sampling frames, short times, complex field work, limiting electoral regulations, etc.

Another way is to take advantage of the new forms of communication that we have people, Social Networks, with which we freely express what we feel and what are our preferences, electoral too.

The use of Artificial Intelligence (AI), which is currently present in various daily fields of men's work, applied to the analysis of information that flows through Social Networks and specifically to the electoral field is almost recent.

However, the way in which this analysis is carried out has its difficulties, given by the presence of exogenous agents that cause distortion and "noise"; The present work incorporates a new methodology of electoral predictability. And it is new because it considers that the unit of analysis is not the messages but rather the electors, these have their relevance and exercise political persuasion to others, they are also based in different geographical locations, so weighting factors are introduced geographical, given the asymmetry in the use of the Internet and the use of Twitter, which is the Social Network, which was used, given the ease it presents for downloading messages.

The methodology was tested in one of the most sensitive fields, which is to predict a result that has verification, such as political elections.

This methodology will be useful in the different studies carried out in sociology, marketing and those that require knowing people's behaviors through Social Networks.

Contenido

Capítulo 1	11
Introducción	11
Definición del problema	14
Objetivo del trabajo	15
El estado del arte	16
Si predicen y no realizan análisis de polaridad	17
Si predicen y si realizan análisis de polaridad	18
No predicen	20
Software disponible	25
Las encuestas tradicionales. Métodos de predicción política	27
Algoritmos de aprendizaje de máquina	29
Preprocesamiento	32
Algoritmos	33
Capítulo 2	43
Propuesta general de la metodología	43
Contextualización de las redes sociales	43
Funciones de preprocesamiento, depuración y preparación	49
Funciones de polaridad de los mensajes	53
Mecanismos de imputación de polaridad al usuario/elector	54
La relevancia de los electores	56
Número de seguidores por mensaje	57
RT/Número de mensajes	57
Ratio ponderado de seguidores y RT	57
Asignación de votos heredados por la relevancia de cada usuario	58
Los factores de ponderación geográficos	61
El conteo de votos	63
El modelo final de predicción electoral	64
Capítulo 3	69
Casos de estudio. Implementación y evaluación	69
Elecciones Presidenciales en el Ecuador. Primera vuelta 2017	76
Elecciones Presidenciales en Chile. Segunda vuelta 2017	82
Consulta Popular en el Ecuador 2018	84

Utilizando el clasificador naïve Bayes	87
Utilizando el clasificador Random Forest	87
Utilizando el clasificador Support Vector Machine	88
Capítulo 4	91
Conclusiones	91
Publicaciones	93
Referencias Bibliográficas	95



Universitat d'Alacant
Universidad de Alicante

Figuras

FIGURA 1. CRONOLOGÍA HASTA EL 2006 DE LAS PRINCIPALES REDES SOCIALES.	16
FIGURA 2. PAÍSES SELECCIONADOS EN LA MUESTRA DE PROCESOS ELECTORALES QUE HAN UTILIZADO LAS MENCIONES O EL ANÁLISIS DE POLARIDAD.	25
FIGURA 3. CUADRANTE MÁGICO DE GARTNER'S 2019, PARA ANÁLISIS DE DATOS Y PLATAFORMAS DE APRENDIZAJE DE MÁQUINA.	26
FIGURA 4. DETERMINACIÓN DEL UMBRAL DE DECISIÓN EN UN CLASIFICADOR LINEAL.	34
FIGURA 5. EL FUNDAMENTO GEOMÉTRICO DEL SVM	34
FIGURA 6. TAXONOMÍA DE LOS MODELOS DE CLASIFICACIÓN.	36
FIGURA 7. OTRA TAXONOMÍA DE LOS MODELOS DE CLASIFICACIÓN CONSIDERANDO SI LOS ALGORITMOS SON SUPERVISADOS O NO.	36
FIGURA 8. LA FUNCIÓN DE POLARIDAD.	37
FIGURA 9. TIPOLOGÍA DE LOS ACTORES EN UNA RED SOCIAL.	45
FIGURA 10. MENSAJES CON SIMILAR POLARIDAD.	55
FIGURA 11. VARIOS MENSAJES CON DIFERENTE POLARIDAD ANTES QUE T.	55
FIGURA 12. PROCEDIMIENTO DE ASIGNACIÓN DE VOTOS A CADA USUARIO EN FUNCIÓN A SU RELEVANCIA.	60
FIGURA 13. METODOLOGÍA UTILIZADA EN LA PREDICCIÓN ELECTORAL.	67
FIGURA 14. PRINCIPALES REDES SOCIALES EN EL MUNDO	69
FIGURA 15. PERFIL DE LA AUDIENCIA DE LAS REDES SOCIALES	70
FIGURA 16. ASIMETRÍAS EN EL PERFIL DE LA AUDIENCIA DE TWITTER EN EL ECUADOR	78
FIGURA 17. ASIMETRÍAS EN EL PERFIL DE LA AUDIENCIA DE TWITTER EN CHILE	83
FIGURA 18. COMPARATIVO DE ELECTORES REALES Y ELECTORES TWITTER.	84

Universitat d'Alicante
Universidad de Alicante

Tablas

TABLA 1: POSICIONES SOBRE EL USO DE LOS TWEETS PARA EL ANÁLISIS PREDICTIVO	17
TABLA 2: PROCESOS ELECTORALES DE ACUERDO AL TIPO DE ANÁLISIS REALIZADO	23
TABLA 3: PROCESOS DE APRENDIZAJE DE MÁQUINA.	30
TABLA 4: ASIMETRÍA ETARIA DE LOS USUARIOS DE TWITTER.	71
TABLA 5: ASIMETRÍA POR REGIONES Y NIVEL DE DESARROLLO EN LA UTILIZACIÓN DEL INTERNET.	71
TABLA 6: RURALIDAD DE ALGUNOS PAÍSES DEL MUNDO.	72
TABLA 7: COMPOSICIÓN PROPORCIONAL DE LAS PROVINCIAS	73
TABLA 8: REPRESENTACIÓN MATRICIAL DE $RS_{A,T}$.	73
TABLA 9: DETERMINACIÓN DE LOS FACTORES DE PONDERACIÓN GEOGRÁFICOS	75
TABLA 10: COMPARATIVO DE LA VALORACIÓN POR TRES MÉTODOS DE CÁLCULO.	75
TABLA 11: ORDEN DE UBICACIÓN POR CADA UNO DE LOS MÉTODOS DE CÁLCULO	76
TABLA 12: PORCENTAJE DE LA POBLACIÓN DE 12 AÑOS Y MÁS QUE POSEE UNA CUENTA EN TWITTER (FUENTE: INEC, ECV. SEXTA RONDA 2013 – 2014)	77
TABLA 13: RESULTADO DE DIVERSAS ENCUESTAS REALIZADAS EN EL ECUADOR. 2016	79
TABLA 14: TWITTER DESCARGADOS, DUPLICADOS Y FACTOR DE REPITENCIA POR CADA CANDIDATO	79
TABLA 15: POLARIDAD IMPUTADA A LOS USUARIOS	80
TABLA 16. PARÁMETROS DE LA DISTRIBUCIÓN JOHNSON SL DISTRIBUCIÓN DE LA VARIABLE RELEVANCIA.	80
TABLA 17: COMPARATIVO DE RESULTADOS OFICIALES, DADOS POR LAS EMPRESAS ENCUESTADORAS Y LOS DATOS PREDICHOS POR LA METODOLOGÍA	82
TABLA 18: PORCENTAJE DE USO DE LAS REDES SOCIALES EN CHILE	82
TABLA 19: COMPARATIVO DE LOS RESULTADOS OFICIALES, RESULTADO DE LA EMPRESA ENCUESTADORA Y LOS PREDICHOS POR LA METODOLOGÍA PROPUESTA CON EL FACTOR DE PONDERACIÓN GEOGRÁFICO Y SIN EL FACTOR DE PONDERACIÓN EN LAS ELECCIONES PRESIDENCIALES DE SEGUNDA VUELTA DE CHILE	84
TABLA 20. CUENTAS PRINCIPALES Y SECUNDARIAS UTILIZADAS	86
TABLA 21. COMPARACIÓN DE RESULTADOS ENTRE EMPRESAS ENCUESTADORAS Y EL MODELO UTILIZADO	89

Capítulo 1

Introducción

Toda la huella que un hombre deja en medios digitales define el idiotipo de cada uno; es decir las diversas características propias de las personas, sus actitudes, sus creencias y en general su idiosincrasia.

Y la huella en el mundo digital de las personas, ya sea como usuario o como consumidor es lo que aporta o no a la web. Mensajería: mediante WhatsApp, correos electrónicos; visitas y mensajería en redes sociales: los likes, los comentarios; consultas Google: las búsquedas de cierto tipo de información, que va de académica, política, turismo, música, hágalo usted y demás interacción en el mundo virtual; van definiendo a las personas. El tiempo que dedica un hombre y está expuesto vía ordenador, televisión, escuchando mensajes radiales, ingresando en cajeros, y demás interacciones con el mundo digital van también formando el perfil de las personas.

Si se pudiera realizar un análisis detallado de toda la huella en el mundo digital de una persona, uno podría determinar sus preferencias, sus gustos, sus amistades, sus inquietudes y en general podríamos predecir sus comportamientos ante cualquier situación a futuro.

Este imaginario, esta suerte de Gran Hermano, todavía, no es posible implementar. No obstante, se puede realizar un seguimiento parcial del transitar de una persona en el mundo digital. Por ejemplo, cuáles son las consultas de una persona en Google; que hace una persona en las redes sociales; y otros seguimientos, que muchas veces están constreñidos con la ley, por el asunto de la privacidad de las personas.

Sin embargo, en esta privacidad, uno si puede analizar una de las *herramientas* que hoy por hoy han permitido poner a las personas de toda condición a “conversar” libremente, son las redes sociales. Hoy en día, a través de las redes sociales podemos opinar, sobre todo; en este contexto los líderes de opinión, las marcas de productos, las posiciones temáticas, entes orgánicos oficiales o privados, y los políticos, etc. interactúan con los diversos usuarios que a su vez opinan y toman alguna posición con respecto de éstos.

El realizar un análisis de esta información que fluye de manera bidireccional, entre actores principales y los usuarios, es lo

que da más potencia a las redes sociales; por cuanto estas se están convirtiendo en una herramienta para poder “influenciar” al público usuario. Entonces el análisis de una temática en una Red Social se convierte en un mecanismo que permite a los actores principales el saber su desempeño; si lo que expresan se posiciona bien, su grado de aceptación; es decir se tiene la medida para definir o mantener estrategias y cursos de acción con el fin de optimizar su comunicación.

Es que nada es más cierto, cuando Umberto Eco¹ manifiesta que: “Las redes sociales le dan el derecho de hablar a legiones de idiotas que primero hablaban sólo en el bar después de un vaso de vino, sin dañar a la comunidad. Ellos eran silenciados rápidamente y ahora tienen el mismo derecho a hablar que un premio Nobel. Es la invasión de los idiotas”.

Todas las personas, que tengan acceso a una red social, opinan, critican, advierten y emiten juicios de valor sobre cualquier tema, más aún si el tema es de corte político. Y en estos medios las personas se desnudan, se muestran tal como son. Claro está que hay miles que simplemente son actores pasivos.

Todas las personas, que tengan acceso a una red social, opinan, critican, advierten y emiten juicios de valor sobre cualquier tema y de entre las redes sociales, una de las que tiene más seguidores² y que permite su descarga para posterior análisis es el Twitter y por cuanto se desea utilizar los mensajes como herramienta para evaluar el desempeño de ciertos actores principales se deberá realizar un proceso de conversión para que la opinión dada en los mensajes equivalgan a los de la población en general, ya que los usuarios del Twitter son un subconjunto de estos. Este proceso de conversión es uno de los aportes de la presente investigación. La opinión se sustenta en datos factuales, como lo son el dar una aceptación del mensaje, el mencionar hechos que sustentan dicho mensaje, y también se fundamenta en datos de carácter subjetivos, toda vez que los usuarios las emiten con una carga emotiva.

Específicamente lo que se aborda en el presente trabajo es una propuesta metodológica sobre el procedimiento a seguir en la realización de pronósticos en la Red Social Twitter para la temática de las elecciones políticas, sin embargo, la metodología es aplicable en temáticas de otros campos como la sociología, marketing, etc. El tema es muy sensible, por cuanto, no es lo mismo evaluar una medida de rendimiento cuando no existe contraste frente al proceso de

¹ Periódico “La Stampa” el 10/06/2015, Turín, Italia.

² Según Hootsuite (<https://hootsuite.com/es/>) Twitter tiene 326 millones de cuentas registradas a nivel mundial, a enero del 2019.

evaluación de pronóstico con verificación, el caso que nos compete, que es el emitir el resultado de una elección antes que ésta ocurra.

El conocer los posibles resultados de cualquier proceso electoral es una necesidad que tiene la sociedad fundamentalmente para transparentar la democracia de los pueblos y un método para conocer los resultados previamente, es el uso de las encuestas tradicionales, que sean estadísticamente confiables.

Cuando (KATZ, LAZARSFELD, Y ROPER, 2005) formulan la noción de influencia y manifiestan que lo que importa es el conocer tres elementos: La audiencia: saber cuántas y cómo son las personas que atienden un mensaje; el Análisis del Contenido: que comprende el concepto de los mensajes emitidos y el Análisis de Efecto: el impacto de los mass media utilizados; para entender el contexto y las condiciones en que se realizaban las “campañas” en los medios de comunicación para modificar las opiniones y comportamientos, nunca supondrían que los electores, hoy en día, estarían totalmente comunicados con sus candidatos, sus mensajes son ahora personales.

Como se consiguió esto, por el apareamiento de las redes sociales; los políticos empiezan a interactuar con los electores directamente, recibiendo muestras de aceptación incondicional y también rechazo y repudio de otros. Por la facilidad que da el Twitter utilizaremos este microblogging para realizar predicciones electorales.

El lenguaje en la política es muy diferente al que usualmente se utiliza en el resto de actividades, generalmente cuando un conjunto de candidatos interactúa en redes sociales con sus potenciales electores, unos se limitan a presentar sus planes y programas de gobierno y otros en cambio se dedican a criticar al resto de sus oponentes. Los seguidores de éstos se encargan de dar argumentos en contra o a favor de las diversas posturas que los candidatos toman.

Esta forma de hacer política, es común, unos aducen que se debe dignificar la democracia y otros en cambio utilizando herramientas de persuasión se encargan de denostar a sus oponentes.

¿Qué cuál es mejor que otra.?. Usted debe tener su propia opinión. Pero lo que cuenta en último caso es la que moviliza más adherentes y por ende más votos a una u otra postura y su forma de hacer política o entenderla.

¿Pero de qué lenguaje hablamos? Del lenguaje que los cibernautas que utilizan en sus textos cotidianos para emitir sus criterios y este en la mayoría de los casos no siguen necesariamente

las reglas gramaticales y más bien utilizan expresiones idiomáticas propias de cada región.

Y además es fundamental caracterizar a las elecciones en los contextos geográficos, políticos, sociales y acceso a la tecnología en las que se desenvuelven.

Frente a estos problemas la propuesta que se presenta es una metodología que, utilizando las redes sociales, no para reemplazar las encuestas sino para considerarlas para la comparación de resultados y conseguir un MAE^3 muy reducido.

El documento ha sido organizado de la siguiente manera: en el capítulo 1, se aborda la definición del problema, el objetivo que se desea alcanzar, el estado del arte sobre la temática, las encuestas tradicionales como método de predicción política y los algoritmos de aprendizaje de máquina. En el capítulo 2 se presenta la propuesta general de la metodología que comprende la contextualización de las redes sociales, las funciones de preprocesamiento, depuración y preparación, las funciones de polaridad de los mensajes y luego los mecanismos de imputación de esta polaridad al elector, luego se considera la relevancia de los electores, los factores de ponderación geográficos; con estos elementos se aborda el conteo de votos y se presenta el modelo de la predicción electoral. El capítulo 3, trata sobre los casos de estudio, su implementación y evaluación; específicamente se consideran las elecciones presidenciales de Ecuador y Chile y una consulta popular realizada en el Ecuador. Finalmente, en el capítulo 4 se concluye con lo desarrollado en el presente trabajo.

Definición del problema

El problema central que aborda este trabajo es el de mejorar la predictibilidad en la Red Social Twitter mediante una nueva metodología de análisis dado que el conjunto de investigación (los tuiteros) es una muestra sesgada del universo de estudio (la población electoral) esto es: debemos determinar cómo se pueden optimizar los pronósticos en un proceso electoral.

El principal problema a resolver es determinar las características socio-demográficas de los tuiteros con el fin de ir ajustando la muestra obtenida hacia una muestra de los electores. En este problema se debe considerar la asimetría que existe en la tenencia y uso de la Red Social Twitter en la población en general.

También se debe establecer un mecanismo para ir determinando la polaridad a través del horizonte de tiempo en que dura la descarga de los datos. La información generada es dinámica y

³ MAE representa el Error Medio Absoluto.

evoluciona en el tiempo, puesto que los sujetos de estudio pueden ir emitiendo sus opiniones e ir adoptando posiciones políticas definidas en este lapso de tiempo, se deben considerar posibles cambios en lo que respecta a su decisión electoral final.

Finalmente, el procesamiento de esta información, que se va incrementando y se acumula en el tiempo, se lo deberá realizar en determinados hitos temporales y debe facilitar la implementación de diversas estrategias en los diversos ámbitos de aplicación.

Objetivo del trabajo

El objetivo del presente trabajo es el de aportar con una propuesta metodológica para mejorar las predicciones electorales que se convierten en un termómetro de los métodos que se utilizan, en la medida que sus resultados pueden ser contrastados con los datos reales

La metodología propuesta deberá permitir la evaluación de un pronóstico electoral, que lo notaremos por PE_{PM} (pronóstico electoral dado por la Propuesta Metodológica):

$$PE_{PM} = \{VP_{a,PM} | a \in A\}$$

En donde $VP_{a,PM}$ es el Valor Porcentual que obtuvo el candidato a del conjunto de Candidatos A , utilizando la propuesta metodológica.

Si definimos y notamos el resultado del proceso electoral real, emitido por un organismo oficial electoral como:

$$RE_{OOE} = \{VP_{a,OOE} | a \in A\}$$

En donde $VP_{a,OOE}$ es el Valor Porcentual del candidato a de acuerdo al Organismo Oficial Electoral⁴.

Utilizaremos como medida de eficiencia en la predicción electoral, el MAE , que se lo define como el promedio de las diferencias absolutas entre el valor real y el valor obtenido en el pronóstico de la encuesta e , $PE_e = \{VP_{a,e} | a \in A\}$ con $e=1, 2, \dots |E|$; donde E es el conjunto de los pronósticos presentados:

$$MAE_e = \frac{\sum_{a \in A} |VP_{a,OOE} - VP_{a,e}|}{|A|}$$

Entonces lo que se desea es:

⁴ Los Organismos Oficiales Electorales son los encargados de los procesos electorales en cada país, por ejemplo: en el Ecuador es el Consejo Nacional Electoral (CNE), en Chile es el Servicio Electoral (SERVEL) y en España es la Junta Electoral Central (JEC).

$$MAE_{PM} \leq \min_{e \in E} MAE_e$$

El estado del arte

El utilizar la información dada por las redes sociales para la determinación de cual candidato va a ganar un proceso electoral empezó cuando el uso de éstas se hizo masivo.

En la Figura 1, tomada de (BOYD, Y ELLISON, 2008) las redes sociales⁵ son muy recientes. Según las autoras, SixDegrees es la primera que permite a los usuarios de la Web: crear un perfil y manejar una lista de usuarios recorriendo su lista de conexiones y las realizadas por otros dentro de la Red.

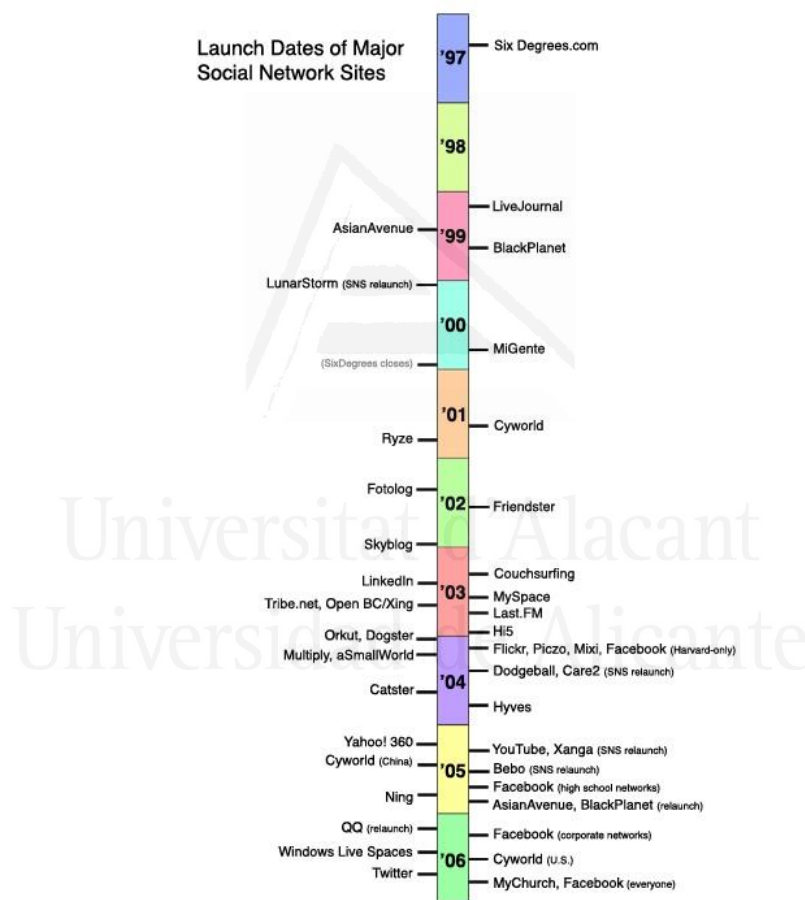


Figura 1. Cronología hasta el 2006 de las principales Redes Sociales.

⁵ Dicho esquema se sustenta en la definición siguiente de Red Social: Las redes sociales son los servicios en la web que permiten a las personas (1) construir un perfil público o semipúblico, (2) implementar una lista de otros usuarios con los que comparten una conexión, y (3) Navegar su lista de conexiones y las de otros dentro del sistema.

Y la Red que nos interesa empieza a partir del 2006, la Red Social Twitter.

Los estudios de Twitter y su uso en la política se originan una vez que esta tuvo usuarios masivos, así en el año 2007, (YEON-OK, Y WOO PARK, 2010) acuñan el lema 'Internet victory' en una campaña para elecciones primarias Presidenciales en Corea del Sur, este estudio utilizó a Twitter básicamente en los componentes de difusión de mensajes más no para realizar predicciones.

En el año 2008, (HARIDAKIS, Y HANSON, 2009) implementan y estudian las e-campañas políticas, específicamente de Obama y McCain en el 2008.

Se puede decir que, (TUMASJAN ET AL., 2010), (LARSSON, Y MOE, 2012) y (GAYO-AVELLO, 2011) son los pioneros utilizando un modelo predictivo en Twitter para pronosticar los resultados de elecciones en Alemania, Suecia y EEUU respectivamente.

El realizar análisis de polaridad con el fin de predecir resultados electorales y otros determinados comportamientos utilizando las redes sociales tiene posiciones. Desde las más extremas que manifiestan que definitivamente no tiene sentido el hacer cualquier intento de análisis de la información de las redes sociales hasta aquellas que propenden su uso y abuso para tratar de predecir los diversos comportamientos de los actores electorales.

Se muestra en la Tabla 1 una clasificación de las posiciones en cuanto si las redes sociales se pueden utilizar o no para predecir resultados y el uso o no de métodos para determinar la polaridad de los mensajes.

SI PREDICEN		NO PREDICEN
No realizan análisis de polaridad		
Si realizan análisis de polaridad		

Tabla 1: Posiciones sobre el uso de los tweets para el análisis predictivo

Procedemos a describir los aportes de los investigadores que se los puede ubicar en una u otra posición y que han utilizado sus resultados en distintas actividades y más aún en los procesos electorales, algunos han utilizado diversas técnicas de análisis de polaridad a más del simple conteo de las menciones. Podemos, sin embargo, decir que han existido posiciones contradictorias y métodos de implementación muy diversos.

Si predicen y no realizan análisis de polaridad

En esta posición están aquellos que manifiestan que el análisis de los tweets sirve para predecir resultados, pero sin utilizar

métodos para detectar la polaridad, es decir trabajando con el número total de menciones que puede tener uno u otro candidato.

En los resultados entregados por (TUMASJAN ET AL., 2010) en las elecciones parlamentarias en Alemania del 2009; se analiza la elección federal alemana para determinar si Twitter es usado como un foro político y si los mensajes en línea son acordes al sentimiento político fuera de línea, usando el software de análisis de texto "Linguistic Inquiry and Word Count" (LIWC)⁶; demostrando que Twitter se usa ampliamente para la deliberación política. El hallazgo es que: la cantidad de mensajes que mencionan un partido refleja el resultado de la elección.

Una vez que los diversos investigadores se dieron cuenta de que los electores empezaban a emitir sus opiniones y comentarios en las redes sociales, se empezó a descargar la información de ellos para realizar el análisis, sin embargo, el referente seguía siendo las encuestas tradicionales. Entonces se empezó a comparar en principio, los resultados de las encuestas y los análisis básicos de las redes sociales; (ZARRELLA, 2010) "se eligió al azar treinta disputas electorales. El 71% de los casos, el candidato con mayor número de seguidores era también el candidato que ocupaba la primera posición en las encuestas" en las elecciones legislativas del 2009 en EEUU.

Esta aparente correlación es la que motiva a la comunidad académica y política a tomar en serio el análisis de las redes sociales.

(BERMINGHAM, Y SMEATON, 2011) utilizando la Elección General de Irlanda del 2011, como un estudio de caso y mediante las medidas basadas en el volumen como el análisis del sentimiento concluyen que son predictivos.

En el trabajo de (FERNÁNDEZ CRESPO, 2013) en las elecciones generales en España del 2011, en las elecciones autonómicas a la Comunidad de Madrid en 2011, en las elecciones autonómicas a la Región de Murcia en 2011 y en las elecciones al parlamento de Cataluña en 2010; utilizando simplemente el conteo del número de menciones por una opción electoral en el Twitter logra acertar los resultados.

Todos estos autores coinciden, que para tener buenos resultados en las predicciones se deberá contar con un gran número de casos.

Si predicen y si realizan análisis de polaridad

En esta posición, se tienen aquellos que utilizan y realizan el análisis de mensajes de Twitter para predecir, pero incorporan los procedimientos de determinación de la polaridad de los mensajes.

⁶ LIWC es un programa que analiza textos diseñado por James W. Pennebaker, Roger J. Booth y Martha E. Francis en el 2007.

Diversos autores, en distintos procesos electorales en el mundo, han utilizado diversas técnicas de análisis de polaridad a más del simple conteo de las menciones, muchos de estos resultados son icónicos por la difusión que se realizó de ellos, un ejemplo son las elecciones generales de 2010 en el Reino Unido, reportadas por una agrupación no afiliada a ningún partido político, *tweetminster*⁷ cuyo objetivo es: “hacer que la política sea más social, desarrollar mejores formas de evaluar la opinión pública y aumentar el acceso de todos para hacer una contribución a la vida política en el Reino Unido” utilizando el Twitter.

Los trabajos de predicción electoral en redes sociales, empiezan a desarrollarse también en lengua española, así (COTELO, CRUZ, Y TROYANO, 2012) presentan una metodología para la obtención de un corpus de mensajes de Twitter de las elecciones generales españolas del 2011 y proponen un método basado en grafos que permite la descarga de mensajes de una temática determinada, que adapta dinámicamente las consultas de los temas relacionados que aparecen.

De manera similar, analizando las redes sociales en lengua española, para las elecciones presidenciales españolas de 2011 (BORONDO ET AL., 2013), presentan una “nueva medida para estudiar el apoyo político en Twitter, que llamamos Soporte Relativo” y caracterizan al usuario utilizando “análisis de los patrones estructurales y dinámicos de las redes complejas que surgen de las redes de mención y retweet. Nuestros resultados sugieren que la atención colectiva está dirigida por una fracción muy pequeña de usuarios.”

Analizando a los políticos y a los principales partidos en las elecciones autonómicas andaluzas de 2012, (DELTELL, CLAES, Y OSTESO, 2013) han demostrado que por medio de Twitter “se pueden predecir los sentimientos y las tendencias políticas de una comunidad determinada”.

Las investigaciones empiezan a proliferar en todas las regiones así el autor (MONTESINOS GARCÍA, 2014) en las elecciones primarias en Chile el 2013; las utiliza para una elección Primaria de un Partido Político.

(MAHMOOD ET AL., 2013) analizan el impacto de los tweets para predecir el ganador de las elecciones de 2013 en Pakistán y dan como ganador a un partido que ganó por mayoría en una provincia pero que sin embargo perdió y concluyen que este análisis en Twitter si tuvo influencia en una determinada provincia.

⁷ <https://www.theguardian.com/media/2010/apr/30/social-media-election-2010>. (Febrero, 2019).

En su documento, (CERON, CURINI, Y IACUS, 2015), describen “las ventajas de usar el Análisis de Sentimiento Agregado Supervisado (SASA) de las redes sociales para pronosticar los resultados electorales” y mediante una variante del método ReadMe, que utiliza funciones de polaridad de varias categorías analizan en las redes sociales varias elecciones celebradas entre 2011 y 2013 en Francia, Italia y los Estados Unidos. Y concluyen que: “Los pronósticos electorales también son más precisos en los países con mayor penetración de Internet y dada la presencia de sistemas electorales basados en la representación proporcional.”

En el documento, (SINGH, Y SAWHNEY, 2018), se analizan trabajos anteriores sobre el resultado político de las elecciones realizadas en distintas regiones tratando de determinar “la técnica más estable y apropiada para predecir los resultados de las elecciones.”

En su artículo, (CHOY ET AL., 2012), con los datos de Twitter y analizando el sentimiento y las discusiones en línea pronostican el próximo presidente de los Estados Unidos en el 2012.

En su artículo (BEAUCHAMP, 2017), combina 1200 encuestas a nivel estatal durante la campaña presidencial de 2012 en los Estados Unidos con más de 100 millones de Tweets políticos, las medidas basadas en Twitter mediante un examen de las características textuales revelan que los temas y eventos asociados con los cambios de opinión, pueden ser útiles para la estrategia de campaña en tiempo real.

Concluyen, (TJONG KIM SANG, Y BOS, 2012), que no es suficiente contar los mensajes de Twitter que mencionan nombres de partidos políticos, sino que se deben realizar un análisis del sentimiento para que las predicciones sean tan buenas como las encuestas de opinión obtenidas tradicionalmente.

(TSAKALIDIS ET AL., 2015), proceden a predecir los resultados de las elecciones en Alemania, los Países Bajos y Grecia en 2014. Sin embargo, manifiestan que se han realizado “varios trabajos en este dominio y que muchos se basaron estrictamente en los datos de Twitter y han demostrado ser ineficaces cuando se probaron en diferentes elecciones”. También resaltan el hecho de que la mayoría de las investigaciones se publican después de haberse efectuado las elecciones.

No predicen

En esta posición en cambio mencionamos a (GAYO-AVELLO, METAXAS, Y MUSTAFARAJ, 2011) como defensores de los que manifiestan que el análisis de Twitter no puede predecir ningún resultado por los valores que se obtienen; manifiestan que podría ir

mejor si se utilizaran otras técnicas. Resultados erráticos y valores del MAE altos, hacen que los autores no recomienden el uso de análisis en el Twitter; ellos utilizan las técnicas que se había utilizado para predecir los resultados del Congreso de los EE.UU. en el 2010 y manifiestan que: “no encontramos una correlación entre los resultados del análisis y los resultados electorales, lo que contradice los informes anteriores”.

En la Tabla 2 se mencionan diversos procesos electorales en las que destaca el tipo de análisis que se utilizó, en lo que respecta a si contabilizan las menciones básicamente y las que si utilizan herramientas de análisis de polaridad u otra técnica.

No.	PAÍS/AÑO/PROCESO	TIPO	MAE	FUENTE
1	Alemania. 2009. Parlamento Nacional	M.	1,65	(TUMASJAN ET AL., 2010)
2	Alemania. 2009. Parlamento Nacional	M.	9,36	(JUNGHERR, JÜRGENS, Y SCHOEN, 2011)
3	Alemania. 2009. Parlamento Nacional	M.	1,57	(JUNGHERR, JÜRGENS, Y SCHOEN, 2011)
4	Alemania. 2009. Parlamento Nacional	M.	2,17	(JUNGHERR, JÜRGENS, Y SCHOEN, 2011)
5	Alemania. 2009. Parlamento Nacional	M.	4,00	(JUNGHERR, JÜRGENS, Y SCHOEN, 2011)
6	Alemania. 2013. Parlamento Nacional	M.	5,29	(*)
7	Alemania. 2014. Parlamento a la Unión Europea	AP.		(TSAKALIDIS ET AL., 2015)
8	Canadá. 2012. Asamblea Legislativa en Alberta	AP.		(MAKAZHANOV, Y RAFIEI, 2013)
9	Chile. 2013. Primarias presidenciales	AP.		(MONTESINOS GARCÍA, 2014)
10	Ecuador. 2013. Presidencial	AP.		(GAURAV ET AL., 2013)
11	España. 2010. Parlamento de Cataluña	M.		(FERNÁNDEZ CRESPO, 2013)
12	España. 2011. Autonómicas de Madrid	M.		(FERNÁNDEZ CRESPO, 2013)
13	España. 2011. Autonómicas de Murcia	M.		(FERNÁNDEZ CRESPO, 2013)
14	España. 2011. Generales de España	M.		(FERNÁNDEZ CRESPO, 2013)
15	España. 2012. Autonómicas de Andalucía	M.		(*)
16	Francia. 2011. Primarias partido Socialista (primera vuelta)	M.	5,90	(*)
17	Francia. 2011. Primarias partido Socialista (segunda vuelta)	M.	2,60	(*)
18	Francia. 2012. Legislativas (primera vuelta)	AP.	2,38	(CERON, CURINI, Y IACUS, 2015)
19	Francia. 2012. Presidencial (primera vuelta)	M.	3,38	(NOORALAHZADEH, ARUNACHALAM, Y CHIRU, 2013)
20	Francia. 2012. Presidencial (primera vuelta)	M.	6,70	(*)
21	Francia. 2012. Presidencial (primera vuelta)	AP.	4,65	(*)
22	Francia. 2012. Presidencial (primera vuelta)	AP.		(*)
23	Francia. 2012. Presidencial (segunda vuelta)	AP.	3,30	(CERON, CURINI, Y IACUS, 2015)
24	Grecia. 2014. Parlamento a la Unión Europea	AP.		(TSAKALIDIS ET AL., 2015)
25	Holanda. 2011. Senado	M.	1,33	(TJONG KIM SANG, Y BOS, 2012)
26	Holanda. 2011. Senado	AP.	2,00	(TJONG KIM SANG, Y BOS, 2012)
27	Holanda. 2012. Casa de Representantes	M.	2,20	(SANDERS, Y VAN DEN BOSCH, 2012)
28	Holanda. 2012. Casa de Representantes	M.	1,90	(SANDERS, Y VAN DEN BOSCH, 2012)
29	Holanda. 2014. Parlamento a la Unión Europea	AP.		(TSAKALIDIS ET AL., 2015)
30	India. 2014. Parlamento Nacional	AP.		(ALMATRAFI, PARACK, Y CHAVAN, 2015)
31	India. 2015. Asamblea de Nueva Delhi	AP.		(SRIVASTAVA ET AL., 2015)

32	Indonesia. 2014. Legislativas de Jakarta	AP.	(RAMADHAN, NURHADRYANI, Y HERMADI, 2014)
33	Indonesia. 2014. Presidencial	AP.	(DWI PRASETYO, Y HAUFF, 2015)
34	Irlanda. 2011. Parlamento General	M. 5,58	(BERMINGHAM, Y SMEATON, 2011)
35	Irlanda. 2011. Parlamento General	AP. 3,67	(BERMINGHAM, Y SMEATON, 2011)
36	Italia. 2011. Administrativa Italiana	M. 16,1	(GIGLIETTO, 2012)
37	Italia. 2011. Alcaldía de Milán	AP. 2,60	(*)
38	Italia. 2012. Primaria de la Coalición Centro-Izquierda (primera vuelta)	M. 6,36	(*)
39	Italia. 2012. Primaria de la Coalición Centro-Izquierda (primera vuelta)	M. 9,72	(*)
40	Italia. 2012. Primaria de la Coalición Centro-Izquierda (primera vuelta)	AP. 9,27	(*)
41	Italia. 2012. Primaria de la Coalición Centro-Izquierda (primera vuelta)	AP. 8,65	(*)
42	Italia. 2012. Primaria de la Coalición Centro-Izquierda (primera vuelta)	AP. 1,96	(CERON, CURINI, Y IACUS, 2015)
43	Italia. 2012. Primaria de la Coalición Centro-Izquierda (segunda vuelta)	AP. 1,50	(CERON, CURINI, Y IACUS, 2015)
44	Italia. 2012. Regional de Sicilia	M. 2,66	(*)
45	Italia. 2013. Parlamento General	M. 2,81	(*)
46	Italia. 2013. Parlamento General	M. 12,5	(*)
47	Italia. 2013. Parlamento General	AP. 1,62	(CERON, CURINI, Y IACUS, 2015)
48	Italia. 2013. Primaria de la Liga Norte	AP. 0,50	(*)
49	Italia. 2013. Primaria del partido Demócrata	M. 8,51	(*)
50	Italia. 2013. Primaria del partido Demócrata	AP. 9,17	(*)
51	Italia. 2013. Regional de Lombardía	AP. 1,59	(*)
52	Japón. 2010. Casa de Consejeros	M. 3,38	(*)
53	Japón. 2010. Casa de Consejeros	M. 3,39	(*)
54	Paquistán. 2013. Presidencial	AP.	(MAHMOOD ET AL., 2013)
55	Paraguay. 2013. Presidencial	AP.	(GAURAV ET AL., 2013)
56	Portugal. 2011. Presidencial	M. 3,54	(FONSECA, Y LOUCA, 2011)
57	Portugal. 2011. Presidencial	M. 10,1	(FONSECA, Y LOUCA, 2011)
58	Singapur. 2011. Parlamento	M. 5,23	(SKORIC ET AL., 2012)
59	Singapur. 2011. Presidencial	AP. 6,07	(CHOY ET AL., 2012)
60	UK. 2010. Parlamento	M. 3,00	(TWEETMINSTER, 2010)
61	UK. 2010. Parlamento	AP. 9,71	(LAMPOS, 2012)
62	UK. 2010. Parlamento	AP. 15,8	(*)
63	UK. 2010. Parlamento	AP. 3,63	(LAMPOS, 2012)
64	UK. 2010. Parlamento	AP.	(LAMPOS, 2012)
65	UK. 2012. Alcaldía de Londres	AP. 2,47	(*)
66	US. 2008. Presidencial en el estado de California	AP. 0,42	(GAYO-AVELLO, 2011)
67	US. 2008. Presidencial en el estado de Carolina del Norte	AP. 16,4	(GAYO-AVELLO, 2011)
68	US. 2008. Presidencial en el estado de Florida	AP. 14,8	(GAYO-AVELLO, 2011)
69	US. 2008. Presidencial en el estado de Indiana	AP. 14,2	(GAYO-AVELLO, 2011)
70	US. 2008. Presidencial en el estado de Misuri	AP. 18,0	(GAYO-AVELLO, 2011)
71	US. 2008. Presidencial en el estado de Ohio	AP. 7,49	(GAYO-AVELLO, 2011)
72	US. 2008. Presidencial en el estado de Texas	AP. 20,3	(GAYO-AVELLO, 2011)
73	US. 2010. Congreso	M. 21,8	(DIGRAZIA ET AL., 2013)
74	US. 2010. Legislativas	M.	(ZARRELLA, 2010)
75	US. 2010. Senado de California	M. 3,80	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
76	US. 2010. Senado de California	AP. 6,30	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)

77	US. 2010. Senado de California	AP.	4,60	(*)
78	US. 2010. Senado de Colorado	M.	24,6	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
79	US. 2010. Senado de Colorado	AP.	12,4	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
80	US. 2010. Senado de Delaware	M.	26,5	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
81	US. 2010. Senado de Delaware	AP.	19,8	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
82	US. 2010. Senado de Kentucky	M.	39,6	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
83	US. 2010. Senado de Kentucky	AP.	1,20	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
84	US. 2010. Senado de Massachusetts	M.	6,30	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
85	US. 2010. Senado de Massachusetts	AP.	1,20	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
86	US. 2010. Senado de Nevada	M.	2,10	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
87	US. 2010. Senado de Nevada	AP.	4,70	(METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011)
88	US. 2010. Senado de Nevada	AP.	1,90	(*)
89	US. 2012. Presidencial	M.	17,9	(NOORALAHZADEH, ARUNACHALAM, Y CHIRU, 2013)
90	US. 2012. Presidencial	AP.	1,80	(WASHINGTON ET AL., 2013)
91	US. 2012. Presidencial	AP.	16,0	(WASHINGTON ET AL., 2013)
92	US. 2012. Presidencial	AP.	3,63	(WASHINGTON ET AL., 2013)
93	US. 2012. Presidencial	AP.	1,29	(*)
94	US. 2012. Presidencial	AP.	0,47	(CHOY ET AL., 2012)
95	US. 2012. Presidencial	AP.	0,02	(CHOY ET AL., 2012)
96	US. 2012. Presidencial	AP.		(CERON, CURINI, Y IACUS, 2015)
97	US. 2012. Primaria del partido Republicano de Carolina del Sur	M.	2,76	(SHI ET AL., 2012.)
98	US. 2012. Primaria del partido Republicano de Florida	M.	3,70	(SHI ET AL., 2012.)
99	US. 2012. Primaria del partido Republicano de Iowa	M.	3,10	(JENSEN, Y ANSTEAD, 2013)
100	US. 2012. Primaria del partido Republicano de New Hampshire	M.	4,50	(SHI ET AL., 2012)
101	US. 2013. Primaria del partido Republicano de Massachusetts	AP.	16,6	(*)
102	Venezuela. 2013. Presidencial	AP.		(GAURAV ET AL., 2013)

Tipo: M.= Menciones, AP.= Análisis de Polaridad

(*) Referidos de (CERON, CURINI, Y IACUS, 2015)

Tabla 2: Procesos electorales de acuerdo al tipo de análisis realizado

Hemos seleccionado una muestra de diversos procesos electorales del mundo en los que se utilizó menciones y análisis de sentimientos y de los resultados obtenidos en procesos similares hemos procedido a comparar los valores del *MAE*, la comparación siguiente la realizamos únicamente en aquellos procesos eleccionarios en los que se puedan contrastar uno y otro método, así por ejemplo en los procesos de: elección Presidencial en Primera Vuelta en Francia el año del 2012, el pronóstico de (NOORALAHZADEH, ARUNACHALAM, Y CHIRU, 2013) fue el que menor *MAE* tuvo, con un 3,38 ellos utilizaron las menciones.

Para las elecciones del Senado en Holanda el año del 2011, los pronósticos de (TJONG KIM SANG, Y BOS, 2012) utilizando las

menciones obtuvieron un $MAE=1,33$ inferior a los valores obtenidos por los investigadores que utilizaron análisis de polaridad.

En las primarias del Partido Demócrata en Italia, en el año del 2013 los pronósticos presentados en (CERON, CURINI, Y IACUS, 2015) se realizó utilizando el conteo de menciones, el MAE de 8,51 aunque elevado estuvieron más próximos a los resultados reales obtenidos.

Para las elecciones del Parlamento del Reino Unido, el 2010, se presentaron algunos pronósticos y los que mejores resultados obtuvieron fueron los presentados por (TWEETMINSTER, 2010) con un MAE de 3,00; utilizando menciones.

En el año del 2010, en la elección del Senado del estado de California, el pronóstico de (METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011) con un $MAE= 3,80$ utilizando las menciones fue el más pequeño.

Para las elecciones del Parlamento General en Irlanda en el 2011, los investigadores (BERMINGHAM, Y SMEATON, 2011) realizaron el análisis utilizando análisis de polaridad y menciones, el pronóstico que realizaron utilizando el análisis de polaridad obtuvo un MAE de 3,67, inferior al obtenido utilizando únicamente las menciones.

Para las elecciones de las Primaria de la Coalición Centro-Izquierda (primera vuelta) realizada en Italia en el año del 2012, (CERON, CURINI, Y IACUS, 2015) con un $MAE = 1,96$ realizaron el pronóstico más acertado.

Y para las elecciones del Parlamento General en Italia, los mismos investigadores (CERON, CURINI, Y IACUS, 2015) volvieron a utilizar el análisis de polaridad para conseguir el MAE de 1,62 que fue el más bajo de todos los otros pronósticos que realizaron.

En Estados Unidos, en el año 2010, se realizaron elecciones para Senadores en diversos estados; en Colorado, Delaware, Kentucky y Massachusetts los investigadores (METAXAS, MUSTAFARAJ, Y GAYO-AVELLO, 2011) obtuvieron los mejores pronósticos utilizando análisis de polaridad. En estas mismas elecciones para Senadores en el estado de Nevada, (CERON, CURINI, Y IACUS, 2015) utilizando también análisis de polaridad obtuvieron un $MAE = 1,90$ siendo el más bajo de los otros pronósticos que se realizaron.

(CHOY ET AL., 2012) en la elección presidencial de los Estados Unidos en el año del 2012, utilizaron para realizar su pronóstico el análisis de polaridad obteniendo un MAE de 0,02 que fue realmente notable.

En la Figura 2 se presentan los países en los que se han desarrollado las elecciones consideradas y destaca que su uso es muy extendido en las diversas regiones del mundo.

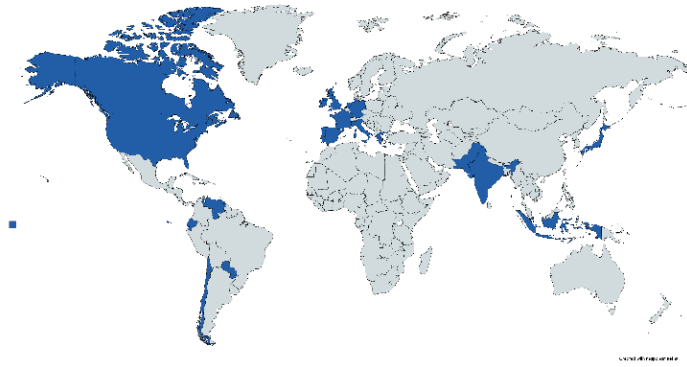


Figura 2. Países seleccionados en la muestra de procesos electorales que han utilizado las menciones o el análisis de polaridad.

De acuerdo a los resultados obtenidos podríamos dar una muy ligera ventaja a los pronósticos que utilizan el análisis de polaridad frente a los que simplemente realizan el cálculo en base a las menciones de los candidatos o partidos políticos.

A continuación, nos referiremos al uso de herramientas informáticas referentes a la predicción electoral en las redes sociales utilizando diversas técnicas de análisis de polaridad e incorporando mecanismos de predicción electoral.

Software disponible

Hoy en día existen algunas herramientas finales que de una manera u otra son capaces de incorporar tecnologías de lenguaje humano para proporcionar infraestructuras analíticas.

Específicamente para el campo que nos compete, que es la predictibilidad de elecciones utilizando la Red Social Twitter existen muy pocas opciones, la mayoría de las existentes manejan como caja cerrada el método de asignación de polaridad y permiten utilizar ciertos algoritmos únicamente. En la Figura 3 se presentan algunas herramientas que se destacan y utilizan diversos algoritmos de Machine Learning.



Figura 3. Cuadrante Mágico de Gartner's 2019, para Análisis de Datos y Plataformas de Aprendizaje de Máquina.

Necesariamente se debe recurrir a software especializado, si se desean implementar ciertos cambios metodológico conceptual.

En la medida de que el análisis de sentimientos es una de las áreas de investigación más importantes en la informática, (FELDMAN, 2013), manifiesta que hay más de 7.000 artículos han sido escritos sobre el tema. Cientos de nuevas empresas están desarrollando soluciones de análisis de sentimientos y los principales paquetes estadísticos, como SAS y SPSS, incluyen módulos de análisis de sentimientos.

En este sentido, hay innumerables aplicaciones para el análisis de sentimientos, según (GUTIÉRREZ ET AL., 2018) mencionan a: Atribus que en tiempo real es capaz realizar un seguimiento a un determinado usuario filtrando la búsqueda a través de palabras clave; Natural Opinions, de manera similar realiza esta búsqueda y determina automáticamente lo más relevante; Textalytics es una herramienta de análisis de texto que entrega ciertos elementos que tengan significado sobre cualquier temática para posteriormente se los pueda procesar; Sentimentviz en cambio trabaja con textos reducidos e incompletos y los visualiza; TweetReach a partir de textos de twitter presenta diferentes estadísticas y el alcance de los tweets; SocialBro en cambio permite gestionar y analizar las diferentes comunidades de twitter; SumAll, permite establecer estadísticas sobre los seguidores de las redes Facebook, Instagram, Twitter, LinkedIn y YouTube

(AGULLÓ ET AL., 2015), presentan ElectionMap, una aplicación web que analiza las opiniones en Twitter de partidos políticos. “Las opiniones de los usuarios sobre estas entidades son clasificadas según su valoración y posteriormente representadas en un mapa geográfico para conocer la aceptación social sobre agrupaciones políticas en las distintas regiones de la geografía española.”

En su documento, (FERNÁNDEZ ET AL., 2015), mencionan a Social Rankings que es una aplicación web que analiza en tiempo real a diversas entidades en las redes sociales. “Detecta y analiza las opiniones sobre estas entidades utilizando técnicas de análisis de sentimientos para generar un informe visual de su valoración y su evolución en el tiempo.”

El paquete de software ReadMe para R, desarrollado por (HOPKINS ET AL., 2012) utilizando como entrada un esquema de categorización elegido por el usuario (por ejemplo, sentimiento positivo a negativo) o cualquier otro conjunto de categorías mutuamente excluyentes y exhaustivas y un subconjunto de documentos de texto clasificados manualmente en las categorías dadas. ReadMe, utiliza el aprendizaje supervisado para determinar la categoría de los documentos que no hayan sido codificados.

Dentro del software existente, es destacable la siguiente herramienta que incorpora características socio-demográficas; así (FERNÁNDEZ ET AL., 2017), describen Social Analytics y su visión sobre cómo deberían funcionar este tipo de sistemas, con una interfaz simple y optimizada que permita responder las necesidades de los usuarios.

También, (UREÑA LÓPEZ ET AL., 2014) manifiestan que el proyecto ATTOS centra su actividad en el estudio y desarrollo de técnicas de análisis de opiniones, Para ello se estudian parámetros tales como la intensidad de la opinión, ubicación geográfica y perfil de usuario, entre otros factores, para facilitar la toma de decisiones.

Las encuestas tradicionales. Métodos de predicción política

Si el objetivo es encontrar un estimador de una cierta variable de interés, en nuestro caso la medida de rendimiento es la votación de los candidatos en el marco de un proceso electoral, referente a una población sin tener que realizar un censo, se utiliza la encuesta, que es una técnica metodológica de investigación, en la cual se consulta a un segmento de la población, llamada muestra, sobre el valor de la variable de interés para luego inferir los resultados en la población en general.

Para la selección de la muestra se debe realizar el diseño muestral, que consiste en establecer el método con el cual se ubicará a las unidades a ser investigadas. En el diseño se establecen los

parámetros de nivel de confianza y el del error estadístico que se relacionan como lo establece la fórmula:

$$Prob(|Y_M - Y_P| \leq e) = 1 - \alpha$$

En donde Y_M representa el valor estimado de la variable de interés calculado con los datos de la muestra, Y_P es el verdadero valor de la variable de estudio en la población, $1-\alpha$ es el nivel de confianza y e representa el error. De esta manera una investigación realizada utilizando encuestas nos permite establecer el intervalo de confianza en el que el verdadero valor de la variable puede variar, realizando observaciones en la muestra seleccionada con métodos estadísticos.

La utilización de las encuestas permite obtener estos intervalos de confianza para el estimador Y_M , mientras que en estudios no probabilísticos o de muestras sesgadas estos intervalos no pueden ser estimados.

El diseño muestral se convierte en el eje central y consiste en elegir una muestra que sea estadísticamente representativa de la población que se va a investigar, dicha determinación de los individuos de las unidades primarias de investigación se la puede realizar de diversas formas y todas éstas dependen de los parámetros previamente definidos de la confiabilidad y el nivel de error máximo que se permite. Entre los métodos más utilizados podemos mencionar a: el Muestreo Aleatorio Simple MAS, el Muestreo Estratificado, el Muestreo por Conglomerados, el Muestreo Multietápico y otros métodos de muestreo particulares.

Los Métodos de Predicción Política son muy diversos y tienen diversos usos, los hay para definición de estrategias de gobernabilidad, campañas políticas, transparencia de la democracia o simplemente para información de los electores.

Específicamente refiriéndonos a las encuestas, que generalmente es uno de los métodos de predicción política más utilizado se realizan Base Line, Tracking, Exit Poll, Conteo Rápido (ARROBA, 2000) y cada una de ellas miden la intencionalidad de voto a una muestra estadísticamente representativa de la población utilizando diversos métodos de recolección de la información, pero todos extrapolan los resultados al del universo utilizando los respectivos intervalos de confianza.

El año 2016 puede ser visto como el año en que las encuestas fracasaron. El caso de PODEMOS en España, el caso de las elecciones presidenciales del Perú, el caso de la elección presidencial en los Estados Unidos por mencionar los más sonados. En estos procesos, la mayoría de los sondeos de opinión que se realizaron, utilizando métodos de encuestas tradicionales; dieron resultados totalmente opuestos a los que sucedieron.

Qué las muestras fueron mal diseñadas, que las preguntas estaban sesgadas, que el ausentismo superó las expectativas, que el “voto de castigo” y demás explicaciones se han esgrimido, para intentar justificar las causas de estos errores.

Como el periodista Juan Cuví⁸ mencionaba en su artículo: “Quizás por recelo, por instinto de conservación o por una falsa complacencia, muchos electores prefieren obviar en las encuestas una posición que se pondrá de manifiesto en medio del secreto de las urnas”.

Para indicar a veces el temor o la indiferencia de las personas frente a las encuestas.

También, el periodista, Erasmo Quintana⁹, menciona en su artículo “El voto vergonzante”: “La psicología de este pueblo es la que es” para indicar que los encuestados prefieren decir cualquier cosa en vez de manifestar su opinión.

Y seguramente tienen razón. Es que el ciudadano, el elector, el hombre común que es parte de una muestra; tiene su propia idiosincrasia y motivaciones; que no se reflejan en las preguntas que se les realiza.

En este punto, surge con más fuerza, el uso de las diferentes redes sociales, para conocer el sentir y la idiosincrasia del elector común.

Este es el aporte de la presente investigación, el de realizar una analogía entre las encuestas y la extracción de información mediante el Twitter.

Algoritmos de aprendizaje de máquina

Cuando se dispone de un corpus, que en nuestro caso lo constituye las descargas de Twitter, se desea determinar, en base al lenguaje utilizado en el texto, el grado de adhesión positiva o negativa y de esta forma poder determinar la valoración de cada uno de los actores principales o candidatos o posiciones en caso de que se traten de Referéndums, Consultas Populares u otra forma de elección popular.

Esta actividad debe ser permanente, mientras dure la campaña política, que es cuando los usuarios de Twitter expresan sus opiniones, y como los mensajes van incrementándose se los deberá procesar automáticamente.

⁸ Periódico “El Comercio”, Quito, Ecuador, 5/11/2016.

⁹ Publicación “Norte Gran Canaria”, España, 14/07/2016.

El analizar textos es una actividad del análisis de contenido, (KRIPPENDORFF, 1990) lo define como “una técnica de investigación destinada a formular a partir de ciertos datos, inferencias reproducibles y válidas que puedan aplicarse a su contexto” qué en la actualidad, se la aborda a través de Inteligencia Artificial, IA.

Dentro de la rama de Inteligencia Artificial, hay diversos métodos que se encargan de analizar los documentos mediante la minería de textos que clasifican las actitudes de los usuarios que son sus juicios de valor o su estado emocional mediante diversos algoritmos de Aprendizaje de Máquina. Estos algoritmos buscan determinar la polaridad de cada uno de los mensajes u otro tipo de clasificación.

Los mecanismos de clasificación deberán primero determinar el conjunto de las posibles clases con que se desean analizar los textos; dichas clases deben representar las actitudes, juicios de valor, estados afectivos o emocionales y para el análisis político se acostumbra a trabajar con las clases: “negativo”, “neutro” o “positivo”; también se utiliza una clasificación de: “enfado”, “tristeza”, “divierte” o “felicidad”.

Existen diversas formas de abordar el proceso de Aprendizaje de Máquina, pero todos tienen pasos y procesos básicos que se deberán realizar como se indican en la Tabla 3¹⁰.

Formulación del problema	¿Cuál es el problema? ¿Cómo se puede resolver el problema?
Preparación de los datos	Selección de datos Preprocesamiento de los datos Transformación de los datos
Elección de Algoritmos	Escoger uno o varios algoritmos
Optimización de los resultados	Adecuación de los algoritmos Ingeniería de características especiales
Presentación de los resultados	Solución Hallazgos Limitaciones Conclusiones

Tabla 3: Procesos de Aprendizaje de Máquina.

La definición del problema planteado tiene que ver con la conceptualización de lo que se desea medir o evaluar, se deberán definir los objetivos y determinar los mecanismos de resolución del problema planteado. Como se abordará en el capítulo 2, se deberán establecer la Temática del problema, los Actores involucrados, así

¹⁰ Adaptada de Jason Brownlee on February 12, 2014 in Machine Learning Process. Tweet Share.

como la temporalidad u horizonte de tiempo en el cual se debe resolver el problema.

La preparación de los datos tiene que ver con la selección y creación del conjunto de datos con los cuales se trabajará, esto implica la búsqueda y descarga de los datos disponibles para luego consolidarlos en un solo conjunto de datos.

El preprocesamiento de los datos no sólo tiene que ver con la limpieza y depuración de los mismos en lo que concierne a los datos atípicos (outliers) y los datos perdidos (missing), sino que implica incluir métodos de reducción de dimensionalidad, transformación de características y atributos de los datos. También se deberán escoger los mecanismos de transformar los datos con el fin de que los algoritmos, con los que se trabajará, sean más eficientes.

La elección del o los algoritmos, consiste en determinar el o los métodos específicos para el reconocimiento de patrones o detección de la polaridad. También se deberá determinar el mecanismo de implementación hasta tener un resultado que esté acorde a parámetros de precisión satisfactorios.

La optimización de los resultados se consigue muchas veces realizando una adecuación de los algoritmos, o aplicando Ingeniería de características especiales que puedan tener los datos, es decir utilizando ciertos procesos como combinar algoritmos o métodos.

La presentación de los resultados es el proceso en el cual se presentan las soluciones al problema planteado originalmente, destacándose los hallazgos, pero dejando constancia de las limitaciones existentes, que muchas veces tienen que ver con el uso de determinados algoritmos de polaridad utilizados.

Para finalmente presentar las conclusiones que en última instancia es el aporte en la generación de conocimiento sobre el problema que se resuelve.

Uno de los mayores problemas que existe en el análisis de polaridad es la subjetividad que puede estar presente en los documentos o mensajes. La subjetividad entonces se la deberá entender cuando el documento proviene de la opinión del autor o la expresión de un sentimiento o como los "estados privados" que en el trabajo de (MONTOTO, MARTÍNEZ-BARCO, Y BALAHUR, 2012) manifiestan: "el análisis de subjetividad se ocupa de la detección de "estados privados" (opiniones, emociones, sentimientos, creencias, especulaciones)"

Los temas de privacidad y manipulación también están inherentes en los estudios, así (PANG, Y LEE, 2008) expresan: "que hay temas relacionados con la privacidad y la manipulación" al referirse

al impacto económico que puede existir en el desarrollo de servicios de acceso a la información orientados a la opinión.

En el trabajo de (ZHANG ET AL., 2010) los temas de privacidad son considerados y expresan que: “en los problemas de diseño para la seguridad y privacidad de las redes sociales, descubrimos que existen conflictos de diseño inherentes entre estos y los objetivos de diseño tradicionales de las redes sociales,”.

Preprocesamiento

La fase previa para realizar cualquier tipo de análisis es el preprocesamiento, pues este permite preparar el conjunto de datos, que en nuestro caso son las descargas de Twitter mientras dure la campaña política.

Uno de los fines que persigue el preprocesamiento es facilitar la implementación de los diversos algoritmos existentes; (ZUBAIR ASGHAR ET AL., 2014) manifiestan que: “la reducción del espacio de la característica, la eliminación de la redundancia y la evaluación del rendimiento de los métodos híbridos de selección de la característica pueden ser la dirección futura del trabajo de investigación”

Para el análisis de polaridad las operaciones en el preprocesamiento son básicas y como se menciona en la investigación de (HEMALATHA, SARADHI VARMA, Y GOVARDHAN, 2013), “las técnicas de preprocesamiento implementadas en algoritmos especialmente diseñados para realizar un análisis de sentimientos, son fundamentales.”

Como mencionan (EFFROSYNIDIS, SYMEONIDIS, Y ARAMPATZIS, 2017) el preprocesamiento es el primer paso que se debe realizar, además comparan diferentes técnicas de preprocesamiento y aplican tres algoritmos de polaridad en textos de Twitter y concluyen; “las técnicas como eliminar números y reemplazar palabras alargadas mejoran la precisión, mientras que otras como eliminar la puntuación no lo hacen.”

Esta fase de preprocesamiento no hay como soslayarla, pues permite mejorar la precisión de los clasificadores, así lo mencionan (ANGIANI ET AL., 2016), “nuestro objetivo es resaltar la importancia de las técnicas de preprocesamiento y mostrar cómo pueden mejorar la precisión del sistema.”

Un aspecto muy importante en esta fase del preprocesamiento lo constituye la negación, pues dependiendo del contexto ésta invierte la polaridad del texto y lo clasifica en otra categoría, un mecanismo para su detección y afectación es el dar seguimiento del árbol sintáctico del texto, en su artículo (JIMÉNEZ ZAFRA, MARTÍNEZ CÁMARA, Y UREÑA LÓPEZ, 2014) trabajan con el árbol

de dependencias, “Para abordar el fenómeno de la negación se han definido una serie de reglas que permiten determinar el ámbito de una partícula negativa en una frase a partir de su árbol de dependencias.”. Otra forma menos precisa es determinar la paridad de la negación e imputar la categoría respectiva.

Una característica que se debe tomar en cuenta en la aplicación de los algoritmos de polaridad es la dimensión de la matriz asociada de las palabras, que lo trataremos en el capítulo 2 y que tiene que ver con la Ley de Zipf, así (MONTEMURRO, 2001) menciona “la ley de Zipf-Mandelbrot se revisa en el contexto de la lingüística” y un aspecto derivado tiene que ver con el procesamiento de los hashtags y su relación con la mencionada Ley de Zipf, (PÉREZ, CONEJERO, Y FERRI, 2017) en su investigación concluyen: “la similitud de la distribución de frecuencia de la popularidad del hashtag con respecto a la ley de Zipf” que tiene implicaciones si se desea investigar las redes de comunalidades, útil para el análisis político de tendencias y sirve como elemento para la definición de estrategias en campañas políticas.

Algoritmos

Dados los conjuntos X (espacio de input o de características) e Y (espacio de output o de etiquetas o de clases) la función de polaridad se la define como:

$$f : X \rightarrow Y$$

También se le llama un clasificador,

Dependiendo de las propiedades de f se puede realizar una taxonomía¹¹ de los algoritmos existentes, pueden ser:

Modelos Geométricos

Estos utilizan características geométricas del espacio de input, tales como planos de separación, transformaciones lineales, distancias y métricas asociadas.

Por ejemplo, un clasificador lineal se lo puede construir de la siguiente manera: sea p el centro de masa positivo y n el centro de masa negativo, entonces un límite de decisión se lo obtiene en el punto medio de la línea entre los centros de masa positivo y negativo. En la Figura 4 se presenta el umbral de decisión, sea la ecuación $w \cdot x = t$ con $w=p-n$; el punto medio $(p+n)/2$ está en el límite de decisión,

¹¹ Peter A. Flach. Intelligent Systems Laboratory, University of Bristol, United Kingdom.

luego $t = (p-n) \cdot (p+n)/2 = (||p||^2 - ||n||^2)/2$ representa el umbral de decisión.

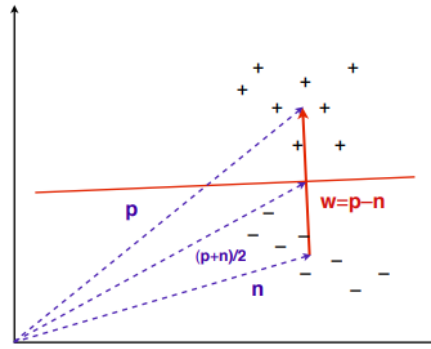


Figura 4. Determinación del umbral de decisión en un clasificador lineal.

Claro está que el umbral de decisión debería estar condicionado a otro tipo de indicador, como lo es el centro de masa considerado anteriormente; un indicador que mejoraría al clasificador anterior es uno que permita maximizar la distancia entre las líneas punteadas, ver la Figura 5.

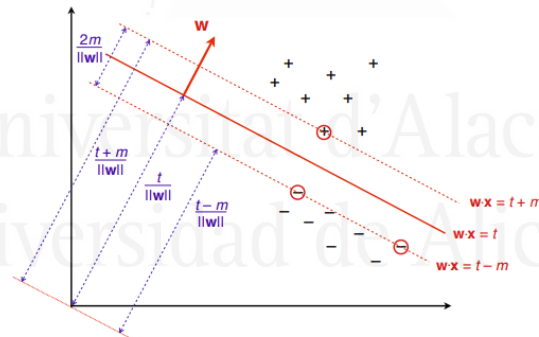


Figura 5. El fundamento geométrico del SVM

La decisión de permitir un acotamiento en la decisión es lo que se conoce como las Máquinas de Soporte Vectorial¹², los círculos en los datos de entrenamiento son los vectores de soporte que son los más cercanos en la decisión de pertenecer a una categoría u otra. El algoritmo SVM busca *maximizar* el valor de $m/||w||$. Se pueden reescalar los valores de t , $||w||$ y m ; Si escogemos entonces $m=1$, el

¹² VAPNIK, V., and A. CHERVONENKIS, 1964. A note on one class of perceptrons. *Automation and Remote Control*

problema entonces es equivalente a resolver el problema de *minimizar* $\|w\|$.

El problema anterior es también equivalente a *maximizar* $1/2 \|w\|^2$ y la solución de este es el SVM.

Formalmente el problema es:

$$\{(w^*, t)\} = \underset{w, t}{\operatorname{argmin}} \frac{1}{2} \|w\|^2 \text{ s.a.: } y_i(w \cdot x_i - t) \geq 1, 1 \leq i \leq n$$

Modelos Probabilísticos

Estos utilizan en cambio las distribuciones de probabilidad en las variables o columnas del espacio de input X. Se trata de reducir la incertidumbre modelándose a través de las probabilidades de pertenencia a una categoría o clase Y.

Por ejemplo, un clasificador probabilístico es el conocido como naive Bayes; que utiliza las funciones de distribución de probabilidades de X y asume un supuesto (ingenuo) de independencia de X. Suponemos que $Y = \{Y_1, Y_2, \dots, Y_K\}$

$$Prob(Y_i|X) = \frac{Prob(X|Y_i)Prob(Y_i)}{\sum_{k=1}^K Prob(X|Y_k)Prob(Y_k)}$$

Y la regla de asignación entonces es:

$$\text{Asignar } Y_i \text{ si } Prob(Y_i|X) = \max_k (Prob(Y_k|X))$$

Modelos Lógicos

Este tipo de modelos se basan más bien en relaciones lógicas que se puede derivar de la estructura del espacio de input X.

Se pueden construir expresiones lógicas y árboles de derivación con el fin de ir caracterizando cada una de las clases de Y.

Una taxonomía alternativa plantea en cambio una clasificación de acuerdo al cómo se resuelven los modelos en la práctica. **Modelos de Grupos**, que va separando el espacio X y en cada segmento aplica un determinado modelo. Y los **Modelos de Calificación**, en los que se aplica un solo modelo a todo el espacio X. En la Figura 6 se presentan esta clasificación de los modelos y la posición que toman los algoritmos básicos de polaridad.

Combinándose ambas taxonomías se pueden mencionar algunos algoritmos que pueden pertenecer a una u otra clasificación si se aceptan ciertas relajaciones en las aplicaciones de los modelos.

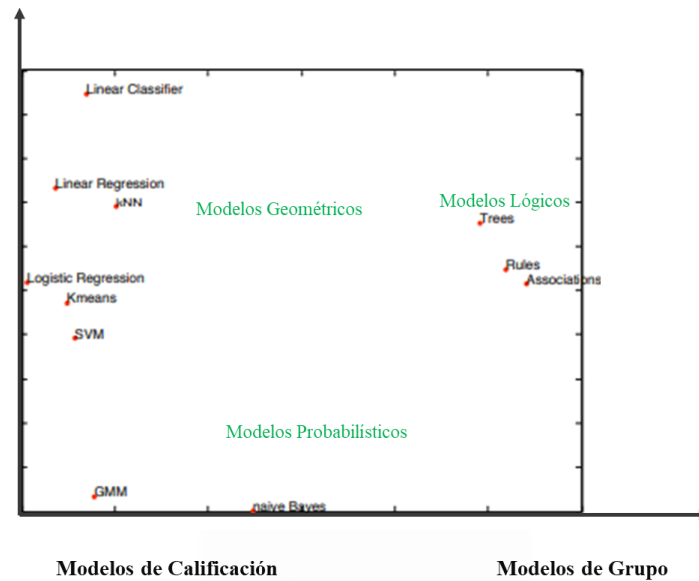


Figura 6. Taxonomía de los Modelos de Clasificación.

Si a esta clasificación propuesta se añade el hecho de que en la implementación de los algoritmos estos se hacen de acuerdo al hecho de que son supervisados o no podemos representarlos de acuerdo al siguiente esquema, de la Figura 7.

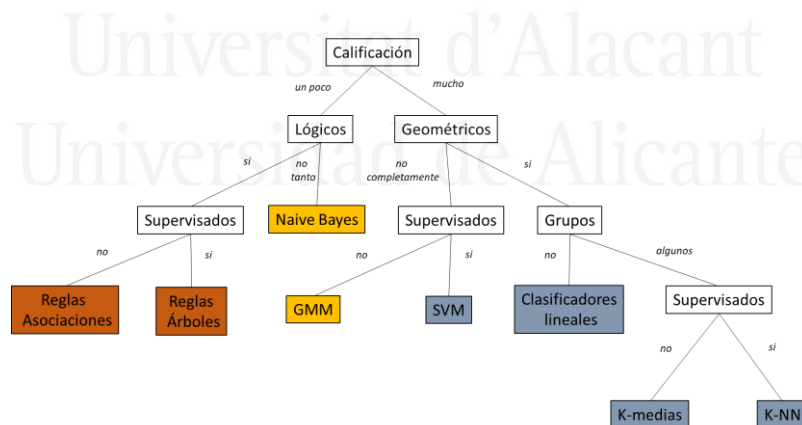


Figura 7. Otra taxonomía de los Modelos de Clasificación considerando si los algoritmos son supervisados o no.

Los colores indican el tipo de modelo, de izquierda a derecha lógicos (café), probabilístico (dorado) y geométrico (azul).

Cuando a la función de polaridad se le incorporan diferentes procesos que se les puede aplicar empiezan a aparecer distintas clasificaciones que relacionan los algoritmos propiamente dichos con su forma de ir mejorando su rendimiento, o cuando se está

analizando espacios que contienen opiniones diversas sobre una temática definida.

Si el objetivo es poder determinar una función de polaridad hay que contemplar otros elementos al momento de medianamente intentar una taxonomía sobre los algoritmos. Consideraciones sobre lo que se conoce como Métodos de Conjunto, Aprendizaje Reforzado y dado que se desea analizar redes sociales, lo que implica analizar textos, se deben considerarse también los Enfoques basado en Lexicones.

En la actualidad existen innumerables algoritmos, muchos de ellos son variantes cosméticas de otros, pero lo cierto es que uno puede combinar algoritmos de Aprendizaje de Máquina, aquellos que son Supervisados o no. Esta combinación es la aplicación de los Métodos de Conjunto con diversos procedimientos que recomiendan fundamentalmente el overfitting y el lowfitting. También se puede utilizar lo que tiene que ver con simulaciones de tipo Monte Carlo en los resultados obtenidos, el Aprendizaje de Refuerzo permite ir obteniendo resultados más precisos si se trabaja con datos que van incrementándose, como son los provenientes de redes sociales.

En la Figura 8 se presentan los algoritmos y métodos, más comunes, que se utilizan para determinar la Función de Polaridad.

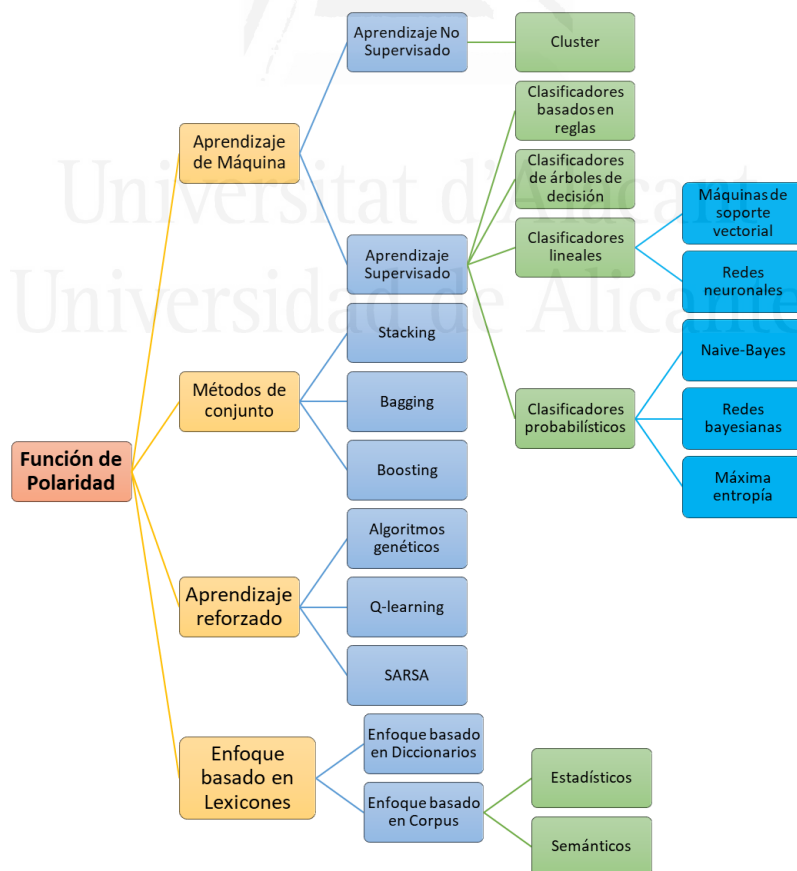


Figura 8. La Función de Polaridad.

Los algoritmos de polaridad han sido extensamente desarrollados a partir de los años 90 y diversos investigadores han hecho importantes aportes, así (SEBASTIANI, 2002) manifiesta que la eficacia de los clasificadores de texto ha mejorado a partir de los años 90 y se debe al apareamiento de los métodos de Machine Learning permitiendo mejoras frente a otros métodos poco sofisticados y débiles, como Rocchio. Las ventajas de este enfoque, que “se basa en técnicas de aprendizaje automático sobre el enfoque de ingeniería del conocimiento, que utiliza la definición manual de un clasificador por expertos son: una buena efectividad, ahorros considerables en términos de fuerza de trabajo experta y una fácil portabilidad a diferentes dominios.”

Para mejorar la precisión de los algoritmos, los autores (VIMALKUMAR, Y AHMEDABAD, 2016) presentan una serie de enfoques, “La principal contribución de este documento es dar una idea de que la selección cuidadosa de características y los enfoques de clasificación existentes, pueden brindar una mayor precisión.”

(TRIPATHY, AGRAWAL, Y RATH, 2016) realizan una evaluación de los siguientes algoritmos de aprendizaje: Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD) y Support Vector Machine (SVM), comparándolos con las medidas de precisión de cada uno recomiendan utilizar los n- gram.

Un aporte diferente lo consignan (XU ET AL., 2017) que manifiestan que los algoritmos Time-User Sentiment/Topic Latent Dirichlet Allocation (TUS-LDA) modela sentimientos y al realizar sus experimentos en China, en Sina Weibos superan a otros modelos en las tareas de clasificación de sentimientos.

Un enfoque alternativo es utilizar el análisis de componentes principales y los árboles de decisión, según (JOTHEESWARAN, Y KOTHEESWARAN, 2016) precisan que para mejorar la precisión de los clasificadores se puede utilizar los Árboles de Decisión y el Análisis de Componentes Principales (PCA), este último utilizado como reductor de la dimensionalidad de dimensiones en la medida que genera los componentes principales que son combinaciones lineales de diversas características que pueden estar altamente correlacionadas.

En su trabajo (WU ET AL., 2007) determinan cuales son los algoritmos más influyentes en la comunidad de investigación, que son: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes y CART.

En el presente trabajo una parte fundamental es el poder identificar el lugar de residencia u otra característica socio-demográfica de los usuarios, los autores (GARCÍA, LANCHAS SAMPABLO, Y PRIETO RUIZ, 2013) lo confirman, estas características son básicas cuando se realizan estudios de opinión.

También los autores (MARTÍNEZ CÁMARA ET AL., 2011) comparan los algoritmos de análisis de polaridad SVM y Naïve Bayes y concluyen que SVM se desempeña mejor; lo importante es que utilizan corpus de películas en idioma español.

Otra forma de determinar la polaridad es trabajar con los n-gram incorporándolo en los algoritmos de Naive Bayes, SVM, Red Neuronal y Cluster, que de acuerdo a la investigación realizada por (KUMAR SINGH, Y SHAHID HUSAIN, 2014), “la precisión de SVM es mejor que los otros tres métodos cuando se utilizó la característica n-gram”.

Dado que las aplicaciones en las redes sociales son más numerosas, el trabajar con diccionarios de palabras y sobre léxicos es otro tema de investigación, así (ESULIY, Y SEBASTIANI, 2007) escriben SentiWordNet, que es un utilitario léxico para que cada clasificador automático asocie uno de los conjuntos de WordNet un triplete de puntajes que son el resultado de la combinación de ocho clasificadores ternarios, “todos caracterizados por niveles de precisión similares pero con un comportamiento de clasificación extremadamente diferente”.

(TABOADA ET AL., 2010) en su investigación sobre léxicos para determinar el sentimiento de un texto desarrollan un calculador de la orientación semántica SO-CAL y lo aplican como clasificador de polaridad utilizando diccionarios de palabras etiquetadas con su orientación semántica, polaridad y fuerza, e incorporan la negación.

Generalmente se utilizan léxicos etiquetados para entrenar a un clasificador de sentimientos, sin embargo, estos enfoques no se adaptan bien en diferentes dominios, los enfoques basados en corpus requieren un esfuerzo de anotación manual. En este documento (HE, Y ZHOU, 2011) proponen un clasificador inicial utilizando información previa de un léxico de sentimiento dado que contenga las etiquetas de sentimiento.

(VELIKOVICH ET AL., 2010), incorporan una nueva óptica con léxicos que no requieren del clásico WordNet, para construir un léxico en inglés especial que contiene jerga, faltas de ortografía, expresiones de varias palabras, y manifiestan que se obtiene un rendimiento superior a los léxicos clásicos incluido uno derivado de WordNet.

Cuando se trata de determinar la polaridad de un tweet, que es un caso especial de los textos, en el sentido de que se dispone de un espacio reducido, en número de palabras, para expresar un discurso y si este está en lenguaje figurado el problema es mucho más complejo pues se trata de textos cortos en los que en la mayoría de los casos no se respetan las reglas gramaticales; (ESCORTELL PÉREZ, GIMÉNEZ FAYOS, Y ROSSO, 2017) presentan diversos léxicos de

emociones, sobre datos descargados de Twitter como la ironía y el sarcasmo.

En esta misma línea de analizar los textos de los microblogs (KESHAVARZ, Y SANIEE ABADEH, 2017) proponen un nuevo algoritmo genético para resolver el problema de optimización y encontrar léxicos para clasificar el texto.

También cuando se utilizan léxicos con orientación semántica de palabras positivas y negativas (MOHAMMAD,DUNNE, Y DORR, 2009) proponen un mecanismo para generar un léxico de orientación semántica utilizando un tesoro similar a Roget y un conjunto de afijos.

Se han desarrollado esquemas de tipo híbrido en que se combinan léxicos etiquetados y aprendizaje automático, en la investigación de (FERNÁNDEZ, GÓMEZ, Y MARTÍNEZ-BARCO, 2014) “el léxico se genera de manera automática a partir de un corpus etiquetado, y se asigna a cada término del texto una puntuación para cada polaridad. El aprendizaje automático se encarga de combinar las puntuaciones de cada término del texto para decidir la polaridad de ese texto.”

Para determinar la clasificación de la polaridad de tweets escritos en español, (VILARES, ALONSO, Y GÓMEZ-RODRÍGUEZ, 2013) utilizan también un método híbrido combinando las técnicas de aprendizaje supervisado con el conocimiento lingüístico obtenido extrayendo información morfológica, sintáctica y semántica para realizar tareas de clasificación con cuatro y seis categorías.

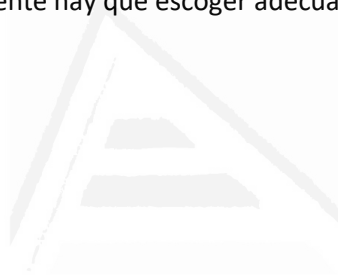
(COTELO ET AL. , 2015) reconocen que los textos de Twitter son muy especiales, toda vez que existe una limitación en la longitud de los textos, la calidad de los mismos es baja pues muchas veces no se respetan las reglas gramaticales y además existen oraciones que son propias de la localidad geográfica en las que se genera; sin embargo, utilizan la información de los Follow para mejorar la clasificación.

El uso de las redes sociales, específicamente el Twitter, para predecir distintos comportamientos de los usuarios es una de las grandes líneas de investigación que se han desarrollado, así (HEMALATHA, SARADHI VARMA, Y GOVARDHAN, 2013) presentan un mecanismo de predicción en el área de la inteligencia de negocios, “Los resultados del análisis de tendencias se mostrarán como tweets con diferentes secciones con resultados positivos, negativos y neutrales.”

Cuando (KATZ, LAZARSELD, Y ROPER, 2005) plantearon el concepto de “flujo en dos pasos” en la que se prioriza la relación entre el público y los líderes de opinión que son “los receptores de los medios de comunicación, sintetizando, analizando y procesando la información que reciben. Posteriormente toman el papel de emisor y

transmiten la información ya procesada a su público, adecuándola a sus preocupaciones, eliminando datos irrelevantes, añadiendo interpretaciones, sintetizando puntos complicados y coloquializando o refinando el lenguaje de los medios a un lenguaje que se adecúe mejor al nivel de cada subpúblico específico”, (WU ET AL., 2011) redefinen a estos líderes de opinión con las “listas” introducidas en Twitter, para distinguir entre usuarios de élite o celebridades, bloggers y representantes de medios de comunicación y otras organizaciones formales, y entre los usuarios comunes y sobre esta clasificación, manifiestan que el 50% de las URL son generadas por usuarios de élite de 20 K, donde los medios producen más información, pero las celebridades son las más seguidas. Detectan una homofilia en los grupos y los bloggers retransmiten más información

Como afirma (FELDMAN, 2013) hay más de 7.000 artículos sobre el análisis de sentimientos, en el año 2013, lo que demuestra que el problema de utilizar uno u otro modelo para la función de polaridad, simplemente hay que escoger adecuadamente.



Universitat d'Alacant
Universidad de Alicante

Capítulo 2

Propuesta general de la metodología

Antes de emprender en los análisis de las redes sociales, hay que empezar a contextualizar su concepto, debemos definir formalmente sus principios con el fin de explotar de mejor manera sus características.

También detallaremos el proceso de conversión para que su uso sea más extendido y que los resultados a posteriori sean más cercanos a la realidad y reflejen el verdadero sentir de los usuarios.

Contextualización de las redes sociales

La forma de comunicarse entre las personas se establece de muy diversas formas, hay los que se comunican con otra persona y otros que simplemente dan su alocución a un gran público. Existen diferentes características en la comunicación de personas: el modo (por ejemplo: con palabras, con fotos, con canciones, etc.), el idioma (por ejemplo: inglés, español, binario, etc.), el medio (por ejemplo: a través del aire, mediante el ordenador, mediante un teléfono, etc.), etc.

En estos escenarios de comunicación que se pueden plantear existen las *directivas*, entendiéndose como las políticas para establecer la comunicación. Hay políticas de contenidos (por ejemplo: no comunicar temas políticos, no hablar de temas de religión, no publicar fotos de violencia o de sexo explícito, etc.), de la longitud de la comunicación (por ejemplo: la longitud puede ser en tiempo, hablar solo cinco minutos; escribir únicamente 280 caracteres, etc.), el medio de difusión (por ejemplo: libros, revistas, conferencias, papers, en ordenador, etc.), los grupos (se pueden establecer grupos abiertos, restringidos, seleccionados, etc.) y así se establecen diversas características que permiten o prohíben la forma de comunicar, el qué, el cuándo y el cómo.

Por ejemplo, el presente documento es una comunicación entre nosotros los autores del mismo con todos aquellos que lo lean; las *directivas*, son los que establecen las políticas: formatos, número máximo de hojas, temáticas a comunicar, medio impreso o digital a presentar, etc.

A continuación, definiremos formalmente el concepto de Temática en una Red Social, que nos permitirá introducir la metodología del proceso de predicción electoral usando el Twitter.

Empezamos definiendo los mensajes que se transmiten entre las personas, que es una información de tipo subjetiva, por cuanto es la apreciación o interpretación de una o varias personas sobre un hecho en particular y representa su perspectiva personal.

Podemos considerar que un mensaje es un vector $p+1$ dimensional:

$$m\$ = (v_1, v_2, \dots, v_p, t\$)$$

dónde las v_i son variables que identifican al usuario que emite un mensaje y además determinan el contexto en el que se realiza el mensaje, estas están determinadas por las *directivas* y $t\$$ contiene el contenido que puede ser: texto, foto, imagen, video, audio, etc. Nosotros consideraremos que $t\$$ contiene un texto.

Las variables asociadas v_i son por ejemplo: la fecha y hora cuando se realizó el envío del mensaje, el lenguaje que se utilizó, la localidad desde que se lo emitió, el identificativo del emisor del mensaje, el número de seguidores que tiene el que emite el mensaje, etc.

Si definimos el conjunto:

$$\mathcal{M} = \{m\$ | m\$ es un mensaje\}$$

como el conjunto de los mensajes.

Y como éstos se realizan siguiendo unas políticas y formatos propios dados por las *directivas*, la biupla:

$$RS = (\mathcal{D}, \mathcal{M})$$

se llama una Red Social donde \mathcal{D} representa las *directivas* y \mathcal{M} todos los mensajes que han transitado utilizando obviamente las directivas.

El conjunto A de todos los usuarios que envían o reciben mensajes en una RS se llama el conjunto de *Actores* y están determinados unívocamente por su identificativo, una de las variables v_i .

Hay que distinguir dos tipos de actores: los actores activos que emiten cualquier mensaje, ya sea para dar a conocer un hecho en particular, una opinión o simplemente desean retransmitir el contenido $t\$$ de un mensaje $m\$$ y otros que simplemente reciben los mensajes, son usuarios que simplemente observan la transmisión de

t \$, ellos no participan en la emisión de ningún mensaje ni opinión sobre nada, ellos son los actores pasivos.

Podemos seleccionar dentro de los actores activos, un grupo al cual se le desea realizar su seguimiento y análisis, los llamaremos los actores activos principales o simplemente actores principales. En la Figura 9, se muestran los tipos de actores que se pueden clasificar en una Red Social.



Figura 9. Tipología de los actores en una Red Social.

Como los actores pasivos no interactúan con el resto, tampoco se puede determinar sus características ni saber qué y cómo opinan, estos quedan fuera de cualquier análisis. Es decir, restringiremos el conjunto A sólo a los actores activos y a los actores principales.

Los actores activos del actor principal $a \in A$ serán todos los u que “interactúan con” a .

Podemos definir en la Red Social, la relación “interactúan con” como:

$$R: A \times A \rightarrow \{0,1\}$$

$$(u, a) \rightarrow R(u, a)$$

Donde:

$$R(u, a) = \begin{cases} 1 & \text{si } u \text{ sigue al actor principal } a \\ 0 & \text{si } u \text{ no sigue al actor principal } a \end{cases}$$

El hecho de seguir a un actor significa que usted emite una opinión t \$ sobre el actor en una Red Social.

Sin pérdida de generalidad, en la medida que uno conoce y ha definido perfectamente cuales son los actores principales a los que

se desea analizar, mientras que los actores activos, más bien son desconocidos, el conjunto de actores A , será definido en base a los actores principales.

Es decir $A = \{a_1, a_2, \dots, a_r\}$ será el conjunto de r actores principales.

El conjunto de todos los actores activos del actor principal a_k se notará por:

$$U_{a_k} = \{u_{1,k}, u_{2,k}, \dots, u_{q_k,k}\}$$

donde $k=1,2, \dots, r$ y q_k representa el número de usuarios del actor principal a_k

Los mensajes del actor activo $u_{j,k}$ para $j=1,2, \dots, q_k$ serán:

$$m_{j,1,k}, m_{j,2,k}, \dots, m_{j,n_{j,k},k}$$

donde $n_{j,k}$ es el número total de mensajes que envía el usuario j -ésimo al actor principal k -ésimo.

Y la representación matricial de los mensajes para el actor principal a_k será:

$$Mm_{a_k} = \begin{pmatrix} m_{1,1,k} \\ m_{1,2,k} \\ \vdots \\ m_{1,n_{1,k},k} \\ m_{2,1,k} \\ m_{2,2,k} \\ \vdots \\ m_{2,n_{2,k},k} \\ \vdots \\ m_{q_k,1,k} \\ m_{q_k,2,k} \\ \vdots \\ m_{q_k,n_{q_k,k},k} \end{pmatrix}$$

$$= \begin{pmatrix} v_{1,1,k,1} & v_{1,1,k,2} & \dots & v_{1,1,k,p} & t_{1,1,k} \\ v_{1,2,k,1} & v_{1,2,k,2} & \dots & v_{1,2,k,p} & t_{1,2,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{1,n_{1,k},k,1} & v_{1,n_{1,k},k,2} & \dots & v_{1,n_{1,k},k,p} & t_{1,n_{1,k},k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{q_k,1,k,1} & v_{q_k,1,k,2} & \dots & v_{q_k,1,k,p} & t_{q_k,1,k} \\ v_{q_k,2,k,1} & v_{q_k,2,k,2} & \dots & v_{q_k,2,k,p} & t_{q_k,2,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{q_k,n_{q_k,k},k,1} & v_{q_k,n_{q_k,k},k,2} & \dots & v_{q_k,n_{q_k,k},k,p} & t_{q_k,n_{q_k,k},k} \end{pmatrix}$$

el orden de esta matriz es: $(\sum_{j=1}^{q_k} n_{j,k}) \times (p + 1)$.

Sea un actor principal a_k con matriz asociada $Mm\$_{a_k}$ y sea a_h otro actor principal con matriz asociada $Mm\$_{a_h}$ definimos la operación matricial unión entre estas dos matrices: $Mm\$_{a_k} \vee Mm\$_{a_h}$ de la siguiente manera:

$$\begin{aligned} \text{Si } Mm\$_{a_k} &= \begin{pmatrix} v_{1,1,k,1} & \dots & t\$_{1,1,k} \\ \vdots & \ddots & \vdots \\ v_{q_k,n_{q_k,k},k,1} & \dots & t\$_{q_k,n_{q_k,k},k,k} \end{pmatrix} \text{ y } Mm\$_{a_h} \\ &= \begin{pmatrix} v_{1,1,h,1} & \dots & t\$_{1,1,h} \\ \vdots & \ddots & \vdots \\ v_{q_h,n_{q_h,h},h,1} & \dots & t\$_{q_h,n_{q_h,h},h,h} \end{pmatrix} \end{aligned}$$

entonces:

$$Mm\$_{a_k} \vee Mm\$_{a_h} = \begin{pmatrix} v_{1,1,k,1} & \dots & t\$_{1,1,k} \\ \vdots & \ddots & \vdots \\ v_{q_k,n_{q_k,k},k,1} & \dots & t\$_{q_k,n_{q_k,k},k,k} \\ v_{1,1,h,1} & \dots & t\$_{1,1,h} \\ \vdots & \ddots & \vdots \\ v_{q_h,n_{q_h,h},h,1} & \dots & t\$_{q_h,n_{q_h,h},h,h} \end{pmatrix}$$

Y tendremos la siguiente representación matricial uniendo las matrices asociadas para todos los actores de A de acuerdo a la ecuación 1:

$$Mm\$ = \bigvee_{k=1}^r Mm\$_{a_k} \quad (1)$$

El conjunto de todos los mensajes del conjunto de actores A en la temporalidad $T=[t_i, t_f]$ donde t_i y t_f representan la fecha de inicio y fin de recepción de los mensajes; se notará por: $RS_{A,T}$ y lo denominaremos la **Temática de los actores A en la temporalidad T en la red social RS**, siendo su representación matricial $Mm\$$.

Para una Red Social, una temática es por ejemplo “una campaña publicitaria” lanzada hace un mes; en donde uno o varios actores principales son los que desean “vendernos” un producto y también otros actores principales, los que quieren que no lo “compremos”, seguramente sus competidores; los actores activos vendrían a ser todos los usuarios que reciben los mensajes de estos actores principales que nos muestran las virtudes y defectos del producto. La temporalidad será el período que va desde hace un mes hasta la actualidad.

Si vamos a realizar el análisis de una temática en una red social y como ésta queda claramente caracterizada por las directivas \mathcal{D} , por el conjunto de mensajes \mathcal{M} , por el conjunto de actores principales A y por la temporalidad T y puesto que trabajaremos con la Red Social Twitter, las directivas están claramente identificadas y

conceptualizadas; sabemos que cada contenido de un mensaje tiene una extensión máxima de 280 caracteres, las variables v_i que se utilizarán son: **id**: código del actor activo, su DNI en Twitter, **created_at**: la fecha en que se publicó el tuit; **location**: la localidad donde se ubica el actor activo y **text**: el contenido del tuit que nosotros lo notamos con $t\$$.

Y entonces ¿qué significa el realizar el análisis de $RS_{A,T}$?.

Es dar respuesta a una serie de interrogantes que se pueden plantear, básicamente es conocer el valor de un estimador de una variable o variables de interés conocidas como medidas de rendimiento o desempeño. En el ejemplo precedente puede ser: el porcentaje de usuarios que aceptarán el “producto”.

Para el caso de procesos eleccionarios estas medidas de rendimiento son el porcentaje de aceptación de cada uno de los candidatos del proceso electoral que estamos analizando.

Para proceder con el análisis de $RS_{A,T}$, consideremos un instante τ dentro de la temporalidad T:

$$\tau \in T = [t_i, t_f]$$

Consideremos la última columna de la matriz asociada $Mm\$$:

$$t\$_{\tau} = \begin{pmatrix} t\$_{1,1,1} \\ t\$_{1,2,1} \\ \vdots \\ t\$_{1,n_1,1} \\ t\$_{2,1,1} \\ t\$_{2,2,1} \\ \vdots \\ t\$_{2,n_2,1} \\ \vdots \\ \vdots \\ t\$_{q_1,1,1} \\ t\$_{q_1,2,1} \\ \vdots \\ t\$_{q_1,n_{q_1},1} \\ \vdots \\ \vdots \\ t\$_{1,1,r} \\ t\$_{1,1,r} \\ \vdots \\ \vdots \\ t\$_{q_r,n_{q_r},r} \end{pmatrix}$$

Este vector o matriz tiene las siguientes dimensiones:

$$\left(\sum_{k=1}^r \sum_{j=1}^{q_k} n_{j,k} \right) \times 1$$

El documento asociado a $RS_{A,T}$ lo definimos como el conjunto \mathcal{D} que tiene todos los contenidos:

$$\dot{D} = \{t\$_{1,1,1}, t\$_{1,2,1}, \dots, t\$_{q_r, n_{q_r, r}, r}\}$$

Se deben definir funciones de depuración y/o preparación sobre los mensajes de una *RS* con el fin de que la siguiente fase, la determinación de la polaridad sea más eficiente; a continuación, se enuncian algunas de dichas funciones que afectan al vector $t\$_r$.

Funciones de preprocesamiento, depuración y preparación

Hay diversas formas de realizar el preprocesamiento, pero todas están encaminadas a realizar una depuración de los datos, a reducir la dimensionalidad de las matrices asociadas para que sean utilizadas por los diferentes algoritmos o el algoritmo de polaridad que se utilice.

En este aspecto deben darse prioridad a aquellos algoritmos que permitan manejar texto, opiniones políticas en concreto y teniendo en cuenta el lenguaje que se utilice, definitivamente no es lo mismo preprocesar en inglés o en español o en mandarín.

Sin embargo, de las funciones más utilizadas en el preprocesamiento, depuración y preparación son:

- Furl's

Que consiste en eliminar todas las direcciones web presentes en un mensaje se retiren. Generalmente las notificaciones o menciones a una dirección web son para hacer adhesivo o contrastatoria una opinión de otro actor.

El análisis de estas direcciones es fundamental al momento de realizar una estrategia política, se utilizan por ejemplo para analizar comunalidades y grupos políticos diversos.

- F@-#

Esta es solo aplicable a la Red Social Twitter. Es la función que retira todas las menciones a otros usuarios o a otros grupos que se identifican a través de estos símbolos. El uso de ellos sirve para la realización de estrategias.

La eliminación o no del # es más controversial, generalmente se elimina solo el símbolo como tal y no el texto que va después de él. #Alicante se convierte en Alicante.

- Frt

Esta función también es solo aplicable a la Red Social Twitter. Es la función que retira todos los RT de los mensajes.

Existen diversas posiciones en el utilizar o no dicha función; lo más aconsejable es trabajar con dos dominios, el uno en que se los

incluye, es decir se acepta que la cita del mensaje de otro actor ha sido aceptada por él y lo comparte con sus contactos o se retira el mensaje toda vez que no se la puede considerar como una opinión propia de él.

- Fsw

Se la utiliza más como función de preparación para la reducción de dimensionalidad de las matrices asociadas que utilizan las máquinas de aprendizaje para poder determinar la función de polaridad, retira los stop words de los mensajes.

Entendiéndose como stop words a aquellas palabras que no aportan a la información final del mensaje.

También el retiro o no depende de la temática de análisis, según¹³, “Esto hace que la indexación cambie mucho respecto a lo que podría ser si nuestro idioma no tuviera estas palabras. Pero como estas son indispensables en el lenguaje y es algo que demuestra que nuestro contenido es de y para humanos, tenemos que utilizarlos.”

Generalmente los stop words son los artículos, pronombres, preposiciones y otras expresiones que no inciden en mayor medida y se las utiliza para dar estructura al lenguaje, pero el contenido no se afecta en el carácter del mensaje.

- Fzh

Función que utiliza mecanismos especiales como el de Zipf-Heaps.

En lingüística se ha definido una heurística conocida como ley de Heaps, que determina el número de palabras no repetidas que tiene un documento, esta es de tipo potencial y viene dada por la ecuación 2.

$$W(n) = Cn^{\beta} \quad (2)$$

En donde $W(n)$ representa el número de palabras no repetidas que tiene un documento, de cualquier lengua, n es el número de palabras totales que tiene el documento y C y β son constantes propias para cada lengua y documento.

La dimensionalidad de la matriz asociada de palabras viene a ser:

$$\left(\sum_{k=1}^r \sum_{j=1}^{q_k} n_{j,k} \right) \times W(n)$$

¹³ <http://googleseo.marketing/lista-de-stop-words-o-palabras-vacias-en-espanol/> (Marzo 2019)

El lingüista Zipf formula su ley empírica, que da cuenta de la distribución de palabras en cualquier lengua y viene dada por la ecuación 3:

$$F(m) = \frac{K}{m^\alpha} \quad (3)$$

Donde $F(m)$ representa la frecuencia de la palabra m -ésima más frecuente y K y α son parámetros propios de cada lengua y documento. De la ecuación 3 tenemos que:

$$\log F(m) = \log K - \alpha \log m \quad (4)$$

Luego las palabras que tengan frecuencia 1 serán aquellas que $\log(F(m)) = 0$, es decir si no consideramos estas palabras que no aportan ninguna información adicional desde el punto de vista del algoritmo que se utilice. De acuerdo a la ecuación 4, m tiene el siguiente valor:

$$m = 10^{\log K / \alpha}$$

Determina la cota inferior para la cual se tiene que $\text{Var } w_i = 0$.

Entonces una reducción significativa en dimensionalidad de la matriz asociada de palabras, se puede obtener eliminando en una primera aproximación estas palabras con frecuencia 1. Y la nueva dimensionalidad de esta matriz será:

$$\left(\sum_{k=1}^r \sum_{j=1}^{q_k} n_{j,k} \right) \times (Cn^\beta - 10^{\log K / \alpha})$$

Todo depende del volumen dimensionalidad de la matriz asociada, muchas veces se utilizan parámetros de hasta el 5% como cota de eliminación de palabras.

- Fdetección-de-bots

En la actualidad la presencia de agentes que asumen la identidad de un elector común es muy usual. Si se utiliza la metodología que se propone un conjunto de mensajes idénticos se convierten en una unidad de análisis, sin embargo, se debe eliminarlos teniendo presente algunas consideraciones: idéntico id del usuario y muchos mensajes repetidos, o gran presencia en la cuenta de un determinado actor principal.

- Fematización

Es la función que convierte cada palabra del texto del mensaje a su forma no flexionada, es decir la convierte en el lema o raíz de la palabra.

Es también una función que es útil considerarla cuando las dimensiones de la matriz asociada para el algoritmo a utilizarse.

- Femoticones

Existen diversos mecanismos de depuración de los mensajes que tienen que ver con la eliminación de caracteres especiales como los emoticones, que son diversos símbolos utilizados para expresar algo. Muchos de ellos son de uso universal y otros son locales.

Para todos es sabido que una expresión como “*!!!!&***” implica que un elector está en desacuerdo sobre cierta temática u actor.

El prescindir de ellos o no es también discutible. Existen en la actualidad diversos data set de emoticones con su valoración.

- Fabreviaturas

Esta función, que se utiliza para la preparación de la matriz asociada de palabras, sirve para normalizar ciertas expresiones no gramaticales. Por ejemplo “ud”, “q” se reemplazarán a su forma reconocida.

- Fminúsculas

También es una función de utilidad en la preparación de la matriz asociada, y consiste en convertir todas las letras a su forma en minúscula. De esta forma se logra estandarizar las palabras lo que permite reducir la dimensionalidad de la matriz asociada, puesto que palabras con grafías aparentemente distintas son una misma.

- Fsegmentación

Esta es una función especial si se va a utilizar un método de detección de polaridad del tipo de Enfoque Basado en Lexicones. La función separa los párrafos en oraciones para posteriormente realizar un análisis por diccionario u corpus.

Su implementación tiene sus dificultades toda vez que el punto, que es el separador de oraciones se lo utiliza en la identificación numérica o como terminación de abreviatura u otros usos gramaticales que se le pueda dar.

- Ftokenización

Esta función se la utiliza para separar el texto del mensaje en palabras. Es la que permite empezar a dimensionar la matriz asociada.

La descripción precedente de las funciones de preprocesamiento, depuración y preparación son de las más utilizadas y se deberán implementar todas aquellas que permitan mejorar el rendimiento de la función de polaridad, y también depende del algoritmo a utilizar, por ejemplo, si se va a utilizar el Enfoque Basado en Lexicones se puede utilizar el procedimiento de la

Ftokenización y luego el etiquetado de palabras. Que puede ser la valoración de la palabra o la etiqueta según su sintaxis como artículo, sustantivo, verbo, etc.

Entonces tendremos la función $f_{LIMPIEZA}$ que puede ser la composición de una o varias de las funciones particulares que hemos considerado, es decir se pueden aplicar sucesivamente éstas funciones y tendremos la siguiente expresión:

$$x'_{\tau} = f_{LIMPIEZA}(t_{\tau})$$

El vector x'_{τ} es el que contiene los mensajes ya depurados y su dimensión es la misma que t_{τ} ; y es sobre este vector resultante que deberemos determinar su polaridad; para lo cual existen numerosas “herramientas” que lo realizan.

Funciones de polaridad de los mensajes

La polaridad o aceptabilidad de un mensaje está definido como la medida de “imputabilidad” a una determinada clase predefinida. La imputabilidad es el asociar a cada mensaje con un concepto definido de antemano sobre objetos o tópicos distintos; así se requiere disponer de modelos, procedimientos o “herramientas” que obtengan y estructuren la información, dada en los mensajes que son de tipo subjetivo.

De manera general se debe considerar el conjunto $C = \{c_1, c_2, \dots, c_v\}$ de clases.

Lo que se busca es entonces una función que asigne a cada mensaje una clase de C es decir:

$$\begin{array}{ccc} f: & D & \rightarrow C \\ & t_{\tau_i} & \rightarrow f(t_{\tau_i}) = c_j \end{array} \quad (5)$$

ó que le asigne varias clases de C , en cuyo caso buscamos una función multiclase del tipo:

$$\begin{array}{ccc} f: & D & \rightarrow C^{*v} \\ & t_{\tau_i} & \rightarrow f(t_{\tau_i}) = (c_{j_1}, c_{j_2}, \dots, c_{j_v}) \end{array}$$

Donde $C^* = C \cup \{\cdot\}$ puesto que no necesariamente a cada mensaje t_{τ_i} se le asignan v clases de C .

El caso de una función multiclase se utiliza cuando un mensaje puede tomar varios valores del conjunto C , se utiliza cuando queremos analizar las diversas clases que un mensaje puede tener.

Para el caso de decisión política, generalmente la función que se busca es del tipo dado por la expresión 5 y el conjunto de clases

tomará tres valores: a favor, neutro o en contra y le asignaremos los valores de clase 1, 0 o -1 respectivamente.

$$f(t\$_i) = \begin{cases} 1 & \text{si } t\$_i \text{ es un mensaje a favor} \\ 0 & \text{si } t\$_i \text{ es un mensaje neutral} \\ -1 & \text{si } t\$_i \text{ es un mensaje en contra} \end{cases}$$

Esta función f está dada por la mayoría de los métodos de Minería de Textos mencionados en el capítulo 1, Algoritmos de aprendizaje de máquina.

Para nuestro caso, los mensajes que sean neutrales o sean en contra significan que el usuario o el elector no manifiestan una decisión política favorable para un determinado actor principal, tomaremos únicamente dos valores 1 o 0; a favor o no a favor. Para lo cual construiremos nuestra función que tome estos dos valores y la denominaremos función de polaridad y la notaremos por $f_{POLARIDAD}$.

$$f_{POLARIDAD}(t\$_i) = \frac{f(t\$_i) + |f(t\$_i)|}{2} \quad (6)$$

La polaridad (función de polaridad) la aplicaremos al vector depurado $x'\$_\tau$ y para el caso de decisión política tendremos:

$$x\$_\tau = f_{POLARIDAD}(x'\$_\tau) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{|D|} \end{pmatrix}$$

En donde x_i representa la polaridad del mensaje i –ésimo mensaje, el vector $x\$_\tau$ tiene como la misma dimensión que $x'\$_\tau$.

Para el caso de estudios de predicción electoral, en la medida de que la unidad de análisis se convierte en los usuarios distintos que existan y que se hayan descargado sus mensajes hasta un tiempo τ es que la polaridad a través del tiempo puede variar. Lo que se requiere es que cada usuario adopte una única polaridad a favor o no a favor, por lo que se debe imputar la polaridad al usuario en base a la polaridad de los mensajes.

Mecanismos de imputación de polaridad al usuario/elector

El principio básico que hemos considerado es que la unidad de análisis lo constituye el elector, nuestro objetivo no es el realizar estrategias de campaña, sino que es el de predecir un resultado electoral utilizando las descargas de la Red Social Twitter, en tal sentido deberemos construir una función de imputación de las polaridades de los mensajes de un usuario a una sola polaridad, que será la asignada a cada usuario.

Pueden darse dos casos. El primer caso se presenta cuando solo existe un mensaje como en la Figura 10. a o varios mensajes, pero

con similar polaridad Figura 10.b. La polaridad del mensaje se imputará al usuario.

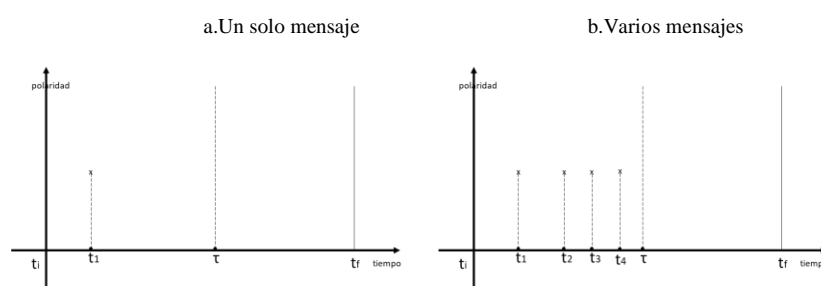


Figura 10. Mensajes con similar polaridad.

El otro caso, en el que existan varios mensajes con polaridades diversas en diferentes tiempos, todos antes del tiempo de procesamiento τ . Estamos considerando el caso de a favor o no a favor. En la Figura 11 se observa, por ejemplo, la polaridad de un usuario a través del tiempo y sus valores en tiempos antes que τ ; los tres mensajes enviados por un elector, tienen una misma polaridad en los tiempos t_1 y t_3 mientras que en el tiempo t_2 esta cambia de valor.

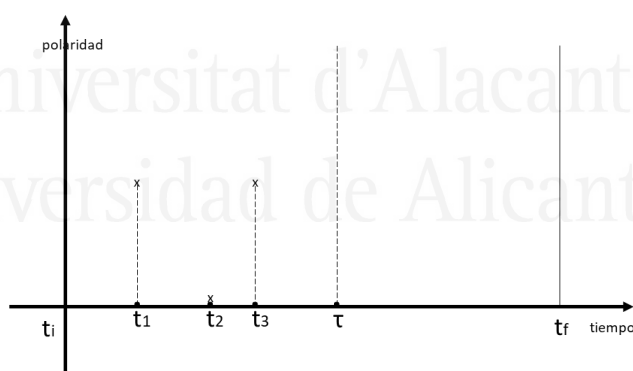


Figura 11. Varios mensajes con diferente polaridad antes que τ .

Se pueden utilizar algunos tipos de imputación. Una, por ejemplo, de las mayorías. Mayor presencia de una determinada polaridad sobre la otra debería imputarse esta polaridad. Para el caso del ejemplo, deberemos entonces asociar el valor de polaridad positiva al elector a cuando procesemos en τ .

Otro mecanismo de imputación sería, dado que la temática en nuestro caso es la política y un elector, puede haber cambiado de

opción política; lo que debería imputarse es la polaridad del último mensaje antes que τ . En nuestro ejemplo la polaridad positiva.

Cambios muy recurrentes en la polaridad de los mensajes de un mismo usuario son síntoma de que la función de polaridad no discrimina bien o simplemente proviene de la subjetividad o volubilidad del elector.

Con este procedimiento de imputación se consigue en cierta medida eliminar la presencia de bots o de actores cercanos a un determinado actor principal que se encargan de este tipo de actividad, inflar los mensajes hacia un determinado actor.

Sea $f_{\text{IMPUTACIÓN-POLARIDAD}}$ la función de imputación de polaridad que hayamos considerado, tendremos entonces:

$$z\$_{\tau} = f_{\text{IMPUTACIÓN-POLARIDAD}}(x\$_{\tau}) = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ \vdots \\ z_N \end{pmatrix}$$

En donde z_i es la polaridad del i -ésimo usuario o elector y N es el número de electores distintos que existen en $RS_{A,T}$ antes del tiempo τ .

La relevancia de los electores

La relevancia es un indicador que determina el grado de importancia que pueda tener un objeto, y en términos de comunicación, se puede decir que un mensaje es más relevante que otro por el nivel e influencia que pueda tener en los electores. Dentro de los líderes de opinión esta influencia se da debido a dos factores fundamentalmente, su popularidad y su prestigio.

Nuestra propuesta para medir la relevancia está dada por una función que depende de los seguidores que tenga, por el grado de aceptación y por la difusión que tenga dicho mensaje. ¿Pero en qué puede incidir la relevancia de un mensaje en el quehacer político? Básicamente en los adeptos y en los adeptos potenciales que este pueda generar.

Para Twitter se han definido diversas métricas, las más generales son los *engagement*, sobre la relevancia de los mensajes y que se utilizan para medir la importancia de una cuenta de Twitter, en nuestro caso nos centraremos en la relevancia de un mensaje o tweet. En las descargas de los mensajes existen diversas variables asociadas, las v_i de $m\$$, las más utilizadas son: el Número de seguidores que está dado por **favoriteCount** (Número de veces que los usuarios han marcado este mensaje como favorito) y el número de retweets o RT dado por **retweetCount** (Número de veces que el

mensaje ha sido retweeteado). Existen diversas métricas para poder medir la relevancia; algunas métricas usuales son:

Número de seguidores por mensaje

Es muy simple la métrica y en cierta medida da la importancia de la cuenta pues determina el número de seguidores que tiene. Esta métrica no toma en cuenta el tiempo en Twitter que lleve el usuario. Pues un nuevo usuario que haya conseguido similares seguidores que otro, pero en mitad de tiempo, es que éste es más relevante y potencialmente más seguido para el instante τ .

RT/Número de mensajes

Una de las fuentes con que se nutren de mensajes el Twitter son justamente los retweets, pues son una forma que tienen los otros usuarios de manifestar que lo que dice otro actor es importante o interesante y que además lo comparte. Cuantos más retweets se tenga entonces será más relevante. Esta medida se lo normaliza dividiendo para el número de mensajes que se hayan descargados antes de τ .

Ratio ponderado de seguidores y RT

Otra métrica que se utiliza es una ponderación entre el número de seguidores y la difusión de los mensajes que a través de los reenvíos se realiza, es decir los RT. Esta combinación de estos indicadores en cierta medida mide la difusión de los mensajes y además la expansión potencial que ellos tienen.

Se puede entonces definir la relevancia mediante la siguiente ecuación 7.

$$f_{RELEVANCIA} = \frac{\alpha favoriteCount + \beta retweetCount}{N_{mensajes}} \quad (7)$$

Los parámetros α y β son arbitrarios y dependen de la Temática de investigación para dar mayor o menor importancia a uno de estos indicadores. En análisis político se pretende dar una mayor puntuación a aquellos mensajes que tienen mayor difusión, pero fundamentalmente que tenga una gran relevancia, en comparación a otros mensajes, por lo que se lo debe normalizar dividiendo para el número de mensajes.

Hay que recordar que nuestra unidad de análisis es el usuario y no los mensajes, entonces se debe seguir un proceso de imputabilidad de la relevancia de mensajes a una relevancia de usuario. Se deberá tomar en cuenta los Mecanismos de imputación de polaridad al usuario/elector tratados.

Sea el proceso que se siga tendremos una función que valore la relevancia de los usuarios dados por una función que la denotaremos como $f_{RELEVANCIA}$ expresada por la ecuación 8.

$$r_{\tau}^{\$} = f_{RELEVANCIA}(V) = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix} \quad (8)$$

En donde r_i es la relevancia del i -ésimo usuario o elector y N es el número de electores distintos que existen en $RS_{A,\tau}$ antes del tiempo τ . V representa la matriz formada por las primeras p columnas de $Mm\$$. La mayoría de las funciones de relevancia en base a los diversos *engagement* para redes sociales utilizan a lo sumo tres o cuatro variables v_i .

A continuación, estableceremos un supuesto de que esta relevancia que tiene cada usuario debe permitirle tener *cierta influencia política*, obtener una recompensa en lo que tiene que ver con su decisión política. Claro está que deberán ser muy pocos los que la obtengan.

En política, la única recompensa que se pueda obtener, son los votos. Electores con alta relevancia deberán poder persuadir a otros electores, pero la persuasión solo la tendrán ciertos electores, la gran mayoría asumiremos que no podrá incidir en otro elector más.

A esta persuasión política la podemos describir como una función que dada una determinada relevancia de una persona le asigna votos adicionales, con la misma polaridad; Debe notarse que la relevancia es una variable aleatoria, en vista de que provienen de una muestra de electores, conformada por aquellos ciudadanos que envían sus mensajes por Twitter. A continuación, trataremos una función planteada por los autores, pero que se podrá modificar de cualquier manera, manteniéndose la hipótesis de que muy pocos tienen *cierta influencia política*.

Asignación de votos heredados por la relevancia de cada usuario

Deberemos definir ciertos parámetros para la asignación de votos a aquellos electores que poseen gran influencia. Proponemos una función de asignación de votos a dos niveles, lo que significa que asignaremos k_2 votos a aquellos electores que tengan una relevancia superior o igual que el cuantil p_2 , les asignaremos k_1 votos ($k_1 < k_2$) si en cambio la relevancia que poseen estos electores es mayor o igual que el cuantil p_1 pero menor que el cuantil p_2 ($p_1 < p_2$). Para aquellos electores en cambio que su relevancia es menor que p_1 les asignaremos 1 voto, para expresar que ellos no pueden incidir en más electores y su votación será la de él mismo.

Para mantener la hipótesis planteada sobre la persuasión política, se deberán asignar valores reducidos de k_i así como que los cuantiles p_i sean lo más cercanos al 100%.

Lo que deseamos determinar es entonces una función $f_{\text{ASIGNACION-VOTOS}}$ que dada una relevancia r_i le asigne una votación, como la establecida en la ecuación 9.

$$f_{\text{ASIGNACION-VOTOS}}(r_i) = \begin{cases} k_2 & \text{si } r_i \geq F^{-1}(\rho_2) \\ k_1 & \text{si } F^{-1}(\rho_1) \leq r_i < F^{-1}(\rho_2) \\ 1 & \text{si } r_i < F^{-1}(\rho_1) \end{cases} \quad (9)$$

Donde F^{-1} representa la inversa de la función de distribución acumulada de la variable aleatoria relevancia.

La determinación de dicha función tiene sus dificultades, no siempre existen inversas de funciones de distribución acumulada o no siempre se puede calcular estas inversas (para la familia de las distribuciones normales, por ejemplo). Sin embargo, como estamos trabajando con datos muestrales en el tiempo τ y disponemos con el vector r_{τ} determinamos su función de distribución acumulada de probabilidad mediante el establecimiento de la siguiente hipótesis nula, H_0 : El conjunto de datos se ajusta a una función ϕ . Esta la aceptaremos o no con un nivel $1 - \alpha$ de significación. Toda vez que utilizaremos cualquier analizador estadístico estos ajustan a distribuciones conocidas, por tal motivo notamos con ϕ y que aceptamos de aquí en adelante que $\phi \sim F$. De manera similar cualquier lenguaje puede implementar la opción **distrinv(.)**¹⁴, la función propuesta se la puede observar en la Figura 12.

En esta asignación de votos debe considerarse la cota superior en el total de votos, que viene dada por:

$$(1 - \rho_2) k_2 + (\rho_2 - \rho_1) k_1 + \rho_1$$

¹⁴ Se trata de la distribución inversa.

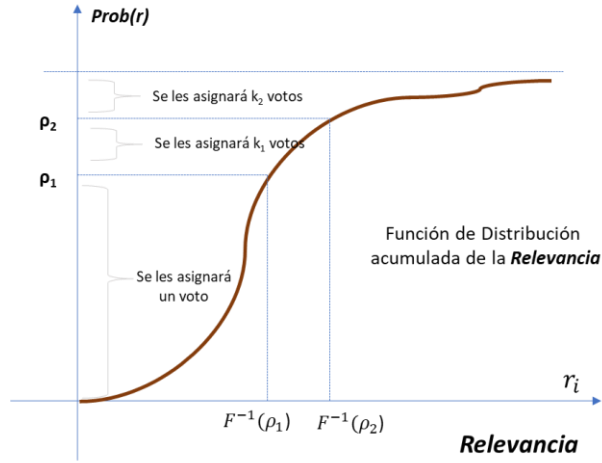


Figura 12. Procedimiento de asignación de votos a cada usuario en función a su relevancia.

Independientemente del mecanismo de implementación que se siga o no se tendrá una función que asigne votos en base a la relevancia que tenga cada usuario (sino se considera se utilizará la función identidad) que la denotaremos como $f_{ASIGNACIÓN-VOTOS}$ expresada por la ecuación 10.

$$v_{\tau}^{\$} = f_{ASIGNACION-VOTOS}(r_i) = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix} \quad (10)$$

En donde v_i es la asignación de votos del i-ésimo usuario o elector y N es el número de electores distintos que existen en $RS_{A,T}$ antes del tiempo τ .

Existe otra consideración en lo que tiene que ver con los votos que posee cada usuario o elector, con su polaridad definida, está a favor o no está a favor, y tiene que ver con el carácter estadístico muestral con el que se debe proceder.

Una vez que estamos consignando datos y opiniones de una Red Social, Twitter en nuestro caso y que de acuerdo a nuestro objetivo, acerca de mejorar los pronósticos electorales, hay que compararse con las encuestas que mencionaba (ARROBA, 2000) en el sentido de que toda encuesta realizada por métodos estadísticos poseen un diseño muestral y en este considerando existe una componente exógena que tiene que ver con la localidad geográfica que tiene el elector, y las elecciones que deseamos evaluar son también de ámbitos geográficos y dependiendo del tipo de elección las hay nacionales, locales, etc.

En este considerando hay que mencionar el poder de penetración que tienen nuestras observaciones y datos con los

universos reales de estudio en donde nos evaluamos. ¿Cuántos, dónde y quiénes son los electores que mensajean twitter?

El siguiente tema que consideramos trata estos interrogantes.

Los factores de ponderación geográficos

El factor de ponderación geográfico es el mecanismo por el cual se deberá realizar un proceso de conversión para que la opinión dada en los tuits equivalga a los de la población en general, dado que los usuarios del Twitter son un subconjunto de estos.

Como los procesos de análisis de una temática se realizan en una localidad geográfica determinada, habrá que determinar las divisiones administrativas de menor nivel. Todo dependerá de a que nivel se tienen los datos desagregados de electores y usuarios de Twitter en las divisiones administrativas inferiores. En El Ecuador la división es por: provincias, cantones y parroquias, en España es por: Comunidades Autonómicas, provincias y municipios y dependiendo de la localidad geográfica en que se desee predecir un resultado se deberá disponer de esta información que es de carácter exógena.

En vista de que el único dato “demográfico” que se dispone es el de la localidad desde la que se realizan los tuits y puesto que los datos de localidad no son concretos en algunas situaciones, trataremos de identificar la localidad y parametrizarla con la provincia de origen donde fue emitido el mensaje y para aquellos que no tengan ubicación en ninguna de estas categorías serán definidas como otros.

Nosotros proponemos el siguiente procedimiento para determinar el valor del factor de ponderación geográfico.

Definamos el conjunto de *provincias*, escogeremos esta división simplemente para ejemplificar la aplicación del método, y de sus localidades de la siguiente manera:

$$P = \{P_i | P_i \text{ es una provincia}, i = 1, 2, \dots, NP\}$$

Donde NP es el número de provincias que tiene la localidad geográfica de estudio. Y consideremos un nivel inferior en su división política administrativa. Para cada provincia i , formamos su conjunto P_i de localidades $L_{i,k}$ con $k=1, 2, \dots, n_i$, sean:

$$P_i = \{L_{i,1}, L_{i,2}, \dots, L_{i,n_i}\} \quad \forall i, i = 1, 2, \dots, NP$$

Primero hay que determinar, para cada elector j , $j=1, 2, \dots, N$ el número de menciones (con polaridad positiva, $z_j=0$) en total que existen pertenecientes a la provincia del elector j . El único dato que se tiene es el de la localidad loc_j que es una de las v_i variables de la

matriz V^{15} y puede darse el caso que esta no esté definida o esté simplemente expresada mediante algún coloquialismo. En este caso esta localidad pasará al grupo de *otros* o los *no ubicados*. Según (PEREGRINO, TOMÁS, Y LLOPIS, 2013) “obtener un lugar real a partir de estos datos es un problema muy complejo”.

Es por eso que en el conjunto P_i de localidades de la provincia i , se deberá ser lo más extensivo, en el sentido de poner con el mayor nivel de detalle las localidades o zonas administrativas de menor nivel.

El número de menciones de provincias en la matriz se lo determinará mediante la ecuación 11:

$$N_j = \left(\left(z_j \sum_{i=1}^{NP} |\{loc_j\} \cap P_i| \left[\sum_{k=1}^N |\{loc_k\} \cap P_i| z_k \right] - 1 \right) \right) \left| \{loc_j\} \cap \bigcup_{i=1}^{NP} P_i \right| z_j + 1 \quad (11)$$

Para $j=1,2, \dots, N$

(El símbolo $|\cdot|$ representa el cardinal del conjunto).

En donde la variable N el número de usuarios o electores, NP el número de provincias existentes en la localidad geográfica de análisis.

A continuación, determinamos el total de localidades que pertenezcan a alguna provincia de las consideradas. Este valor se lo calcula mediante la ecuación 12:

$$D = \sum_{k=1}^N z_k \left| \{loc_k\} \cap \bigcup_{i=1}^{NP} P_i \right| \quad (12)$$

Luego la proporción muestral para el usuario j está dada por la ecuación 13:

$$M_j = \frac{N_j}{D} \quad (13)$$

para $j=1,2, \dots, N$

Proceso similar se deberá realizar para obtener la proporción del universo de electores, para cada provincia i deberemos disponer de la información sobre la proporción Poblacional (electoral) pP_i , pero esta proporción no es válida para aquellas localidades de la categoría *otros* o los *no ubicados* por lo que se deberá realizar una reconversión, mediante la siguiente ecuación 14:

¹⁵ En Twitter, las variables para identificar la localidad o loc_j son las siguientes: **locate** que es la definición que el dueño de la cuenta registró y **UTM** que proporciona las coordenadas geográficas dadas por el usuario. Esta última variable por lo general no es muy utilizada, por el porcentaje reducido de aquellos que la tienen implementada.

$$U_j = z_j \sum_{i=1}^{NP} |\{loc_j\} \cap P_i| p_i \quad (14)$$

para $j=1,2, \dots, N$

Luego el factor de ponderación geográfico que se aplicará a cada elector j , con polaridad z_j ubicado en la localidad loc_j tanto para aquellas que tengan localidades ubicadas, como localidades de tipo otros o las no ubicados se estimará mediante la siguiente ecuación 15:

$$fpg_j = \left(\left(\frac{U_j}{M_j} - 1 \right) \left| \{loc_j\} \cap \bigcup_{i=1}^{NP} P_i \right| + 1 \right) z_j \quad (15)$$

para $j=1,2, \dots, N$

La función que se utilice la denotaremos por $f_{\text{FACTOR-DE-PONDERACIÓN-GEOGRÁFICO}}$ la misma que permitirá reconvertir la muestra de electores de Twitter hacia una muestra de electores de la localidad geográfica de análisis.

$$fpg_{\tau} = f_{\text{FACTOR-DE-PONDERACIÓN-GEOGRÁFICO}}(P, \{P_i\}_{i=1}^{NP}, \{p_i\}_{i=1}^{NP}, z_{\tau}) = \begin{pmatrix} fpg_1 \\ fpg_2 \\ \vdots \\ fpg_N \end{pmatrix} \quad (16)$$

Donde fpg_i es el factor de ponderación geográfico del usuario i -ésimo y N es el número de electores distintos que existen en $RS_{A,T}$ antes del tiempo τ .

El conteo de votos

Dentro de la forma de contabilizar los votos hay que tomar una decisión y tiene que ver con la no inclusión de los votos asignados en cuyo caso cada elector solo dispondrá de un voto, el suyo. Si consideramos en cambio, esta ganancia de votos, o conversión dada por la relevancia en votos potenciales que se pueden generar; aquellos mensajes que son más relevantes incidirán en mayor magnitud en el electorado que aquellos que no poseen tanta relevancia. Esta forma de cuantificar o de convertir los votos se asume que se aplicarán en aquellos electores que todavía no han decidido por quién votar, los indecisos, que generalmente en el inicio de una campaña electoral son una gran mayoría y a medida que se acercan las elecciones van disminuyendo puesto que ellos van alineándose con una u otra postura. Esta forma de ir adscribiéndose a una u otra postura se da por muchas razones, la que nos interesa es la que se da en base a información relevante que estos indecisos

tienen a su disposición y que les incide de tal manera que los hace adoptar una posición definida.

Para obtener el conteo de votos o la valoración final para cada uno de los actores principales $A = \{a_1, a_2, \dots, a_r\}$ evaluaremos su votación sumando para todos los electores su votación asignada v_k su polaridad z_k afectándole con el factor de ponderación geográfico fp_g_k , entonces sea $a \in A$ un actor principal de A , su valoración estará dada por la ecuación 17:

$$V_a = \sum_{k=1}^N z_k v_k fp_g_k |\{a\} \cap \{a_k\}| \quad (17)$$

Donde N es el número de electores distintos, con el fin de normalizar las valoraciones de cada actor principal a , utilizaremos la valoración porcentual de cada actor a , en el que estos pueden variar en el intervalo $[0,100]$ y se calcula mediante la ecuación 18:

$$VP_a = \frac{V_a}{\sum_{a \in A} V_a} \times 100 \quad (18)$$

El modelo final de predicción electoral

La metodología utilizada sigue una serie de pasos que están determinadas e inician desde la definición de la Temática de la Red Social Twitter, que implica la definición de los actores principales A , o candidatos que están enfrentados en una elección en una determinada localidad geográfica; así como la determinación de la temporalidad T que se empleará, es decir t_i será el instante en que se empezará la descarga de los mensajes y t_f será la fecha en que se realizará la elección.

Para la extracción de datos en Twitter se deberán considerar a los actores principales a_k y determinar las cuentas $@ a_k$ para realizar las descargas de los mensajes durante la temporalidad T . De esta forma se van construyendo las matrices $Mm\$_{a_k}$, en la que se deberá incluir en esta el código del actor principal a_k que será considerado como un campo más de dentro de las variables v_i asociadas a cada mensaje $m\$$. Utilizando entonces la ecuación 1, se dispondrá de la matriz de mensajes $Mm\$$ en el instante $\tau < t_f$.

A continuación, aplicaremos el procedimiento de preprocesamiento, depuración y preparación aplicándose las funciones: *Furl's*, *F@-#*, *Frt*, *Fsw*, *Fzh*, *Fdetección-de-bots*, *Flematización*, *Femoticones*, *Fabreviaturas*, *Fminúsculas*, *Fsegmentación*, *Ftokenización* u otras que permitan que la determinación de la función polaridad sea más eficiente. Estas funciones aplicadas definen la función $f_{LIMPIEZA}$ que afecta al vector de los textos $t\$_\tau$ de $Mm\$$, para obtener el vector depurado $x' \$_\tau$ a través de: $x' \$_\tau = f_{LIMPIEZA}(t\$_\tau)$.

Para la determinación de la polaridad de los mensajes existen diversas estrategias que se pueden implementar, sin embargo hay que tener en cuenta las siguientes consideraciones: que los tiempos de análisis son limitados pues se debe contar con los resultados antes de t_f , además el número de los mensajes¹⁶ van incrementándose a medida que va acercándose la fecha de la elección y las características de los mensajes, que son de máximo 280 caracteres que contienen muchos casos diversas expresiones idiomáticas, propias de cada localidad geográfica, expresiones “irónicas”, palabras con errores ortográficos, abreviaturas, emoticones, etc.

Una vez definido el mecanismo de determinación de la polaridad de los mensajes, contaremos con el vector $x\$_\tau$ a través de: $x\$_\tau = f_{POLARIDAD}(x' \$_\tau)$. Definiremos dos categorías de polaridad, está a favor codificado como uno o no está a favor codificado como cero.

Como se había establecido, la unidad de análisis, no son los mensajes, sino los usuarios electores, se deberá definir un mecanismo de imputación de la polaridad de los mensajes de un elector a la polaridad del elector, a través de: $z\$_\tau = f_{IMPUTACIÓN-POLARIDAD}(x\$_\tau)$. Disponemos entonces del vector $z\$_\tau$ que contiene las *polaridades* de los electores.

Nuestra propuesta incluye un indicador que determina el grado de importancia que pueda tener un mensaje en base a su grado de aceptación, la difusión que tenga u otras características para medir el nivel de influencia que pueda tener en los electores. A este indicador lo hemos definido como la relevancia y se lo calcula mediante un *engagement*; utilizando un procedimiento de imputación de mensajes a elector, similar al utilizado para la polaridad, se contará con una función: $r\$_\tau = f_{RELEVANCIA}(V)$ donde $r\$_\tau$ es la relevancia de cada elector dado por su popularidad y prestigio. Pero esta relevancia sólo sirve para permitir que los electores tengan una real influencia en otros electores, influencia que es medida a través de una *asignación* de votos, la hipótesis básica es que: “Electores con alta relevancia deberán poder persuadir a otros electores, pero la persuasión solo la tendrán ciertos electores, la gran mayoría asumiremos que no podrá incidir en otro elector más.”. Esto lo determinamos a través de: $v\$_\tau = f_{ASIGNACION-VOTOS}(r_i)$ en la que el vector $v\$_\tau$ contiene los votos asignados a cada elector dados por su relevancia.

¹⁶ En el icónico trabajo (TUMASJAN ET AL., 2010), mencionan que se descargaron 104.003 mensajes. (BEAUCHAMP, 2017) analizó aproximadamente 120 millones de tweets para la elección presidencial del 2012 en Estados Unidos.

Los autores, también consideramos el factor de ponderación geográfico que es el procedimiento de conversión, para que la opinión dada en los tuits equivalga a los de la población electoral, puesto que los usuarios del Twitter son una muestra de ellos. Para el cálculo de estos factores son necesarios la determinación de los siguientes datos: $P, \{P_i\}_{i=1}^{NP}, \{pP_i\}_{i=1}^{NP}$ los dos primeros conjuntos son la determinación de las divisiones político-administrativas que tiene la localidad geográfica en la que se realizan las elecciones que se analizan, como primer nivel las *provincias* y al interior de ellas las localidades y lugares geográficos que la conforman y el tercer conjunto de datos representa la fracción de electores que representan las circunscripciones geográficas de primer nivel.

Mediante la ecuación 12 obtenemos el vector $fpg\$_r$ que contiene el factor de ponderación geográfico de los electores.

Finalmente se procede al conteo de votos que es el procedimiento con el cual obtenemos las valoraciones porcentuales de cada candidato, o actor principal α , dado por $\{VP_\alpha\}_{\alpha \in A}$.

En la Figura 13, se presenta la Metodología para la predicción electoral.

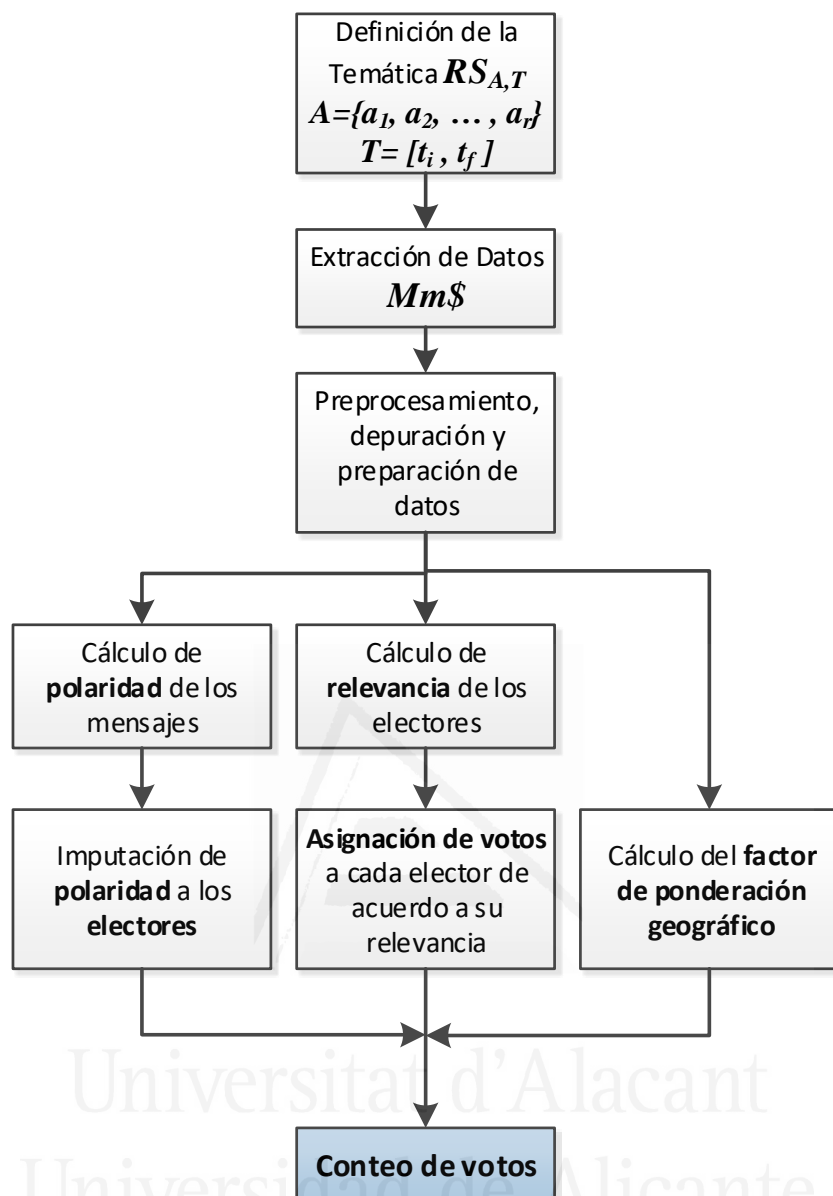


Figura 13. Metodología utilizada en la predicción Electoral.

Capítulo 3

Casos de estudio. Implementación y evaluación

Previa a la etapa de implementación y evaluación de la Metodología utilizada en la predicción Electoral, hay que considerar el carácter de la Red Social Twitter, creada el 2006 como un servicio de microblogging, que permite enviar y recibir mensajes de hasta 280 caracteres, originalmente eran 140 caracteres, que se ha convertido en una de las más utilizadas en el presente. En la Figura 14¹⁷, se presentan los números de los usuarios de las distintas redes sociales, destacándose la presencia de China, Qzone, Doyin/TikTok, Sina Weibo, Douban y Baidu Tieba.



Figura 14. Principales redes sociales en el mundo

Cada una de ellas tiene especificidades muy particulares y sirven para distintos públicos objetivos con funcionalidades específicas cada una de ellas. Sin embargo, Twitter se destaca de las otras por cuanto es la que concentra temáticas variadas y no se concentra en “*historias de vida*”, LinkedIn por ejemplo es una red profesional para puestos de trabajo, Facebook en cambio es la que permite la publicación de casi todo tipo de contenidos, destacándose los asuntos muy personales, Instagram en cambio permite compartir imágenes y videos personales de los usuarios.

Estas diversas funcionalidades han hecho que cada Red Social tenga más adeptos de públicos específicos y en lo que respecta al género de las personas, las mujeres en general participan menos de la interacción en las diversas redes, así por ejemplo de los usuarios de

¹⁷ Readaptado de Hootsuite, <https://hootsuite.com/es/>

Facebook, el 43% son mujeres; en Snapchat en cambio el 62% de sus usuarios son mujeres. En la Figura 15 representamos las audiencias de las principales redes segmentada por el género y particularmente en Twitter, esta asimetría en el uso de acuerdo al género de las personas es más marcada, el 65% de los usuarios son hombres; también se presenta la composición etaria de los usuarios de Twitter.

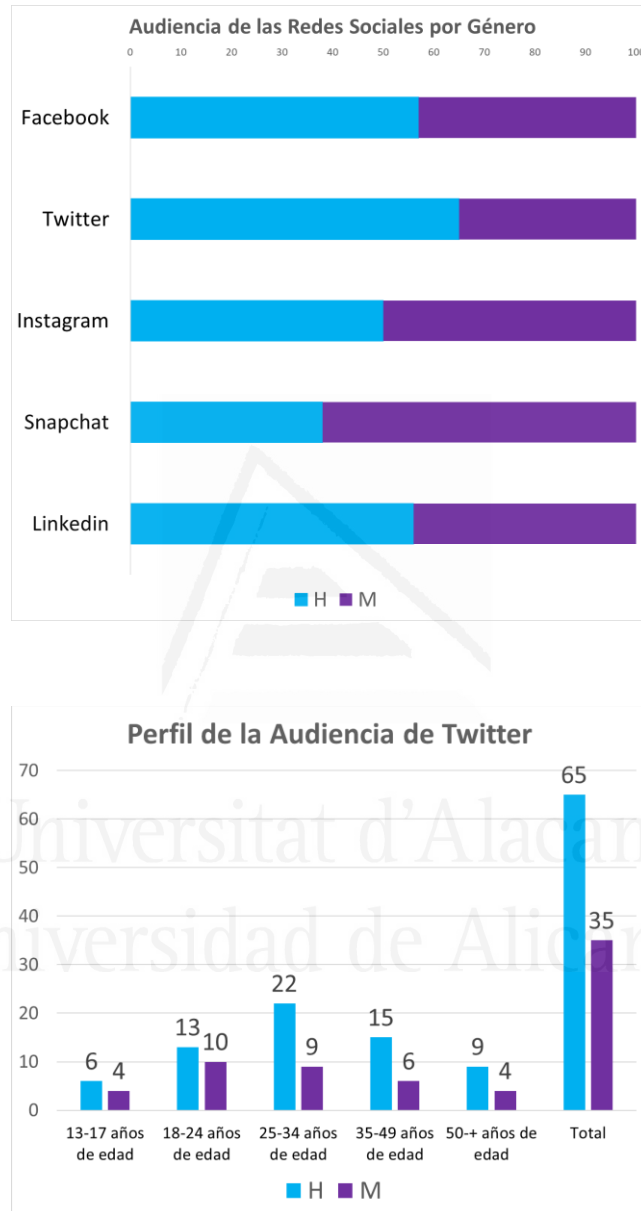


Figura 15. Perfil de la Audiencia de las redes sociales

Hemos agrupado los rangos de edad para los usuarios de Twitter y los hemos comparado con la población mundial¹⁸ y se puede observar en la Tabla 4 una sobrerrepresentación de los jóvenes y una subrepresentación de las personas de mayor edad.

¹⁸ https://www.indexmundi.com/es/mundo/distribucion_por_edad.html

Composición etaria de la población mundial vs usuarios de Twitter (%)		
Rango de edad	Mundo	Twitter
13-34	38	64
35-49	23	22
50+	39	14

Tabla 4: Asimetría etaria de los usuarios de Twitter.

Claro está que estas desigualdades en el uso del Twitter y en general de todas las redes sociales, se ahondan más si consideramos a las personas que habitan en las diversas regiones del mundo y por el tipo de nivel de desarrollo que éstos poseen, como se indica en la Tabla 5, el porcentaje de personas que tenían acceso al Internet en el mundo era del 47,1% de acuerdo a la Unión Internacional de Telecomunicaciones (UIT).

Porcentaje de personas que utilizan Internet			
Por regiones	%	Por nivel de desarrollo	%
Europa	79,1	Desarrollados	81,0
América	65,0	En desarrollo	40,1
CEI	66,6	Subdesarrollados	15,2
Estados Árabes	41,6		
Asia/Pacífico	41,9		
África	25,1		

Tabla 5: Asimetría por regiones y nivel de desarrollo en la utilización del Internet.

Finalmente mencionamos que estas desigualdades son aún mayores si se considera las zonas en las que habitan las personas, las zonas rurales en el mundo, que concentran al 44,7 %¹⁹ de la población, han estado en su mayoría excluidas de las tecnologías de la información, ellas difícilmente interactúan y opinan en las redes sociales.

En la Tabla 6 se presenta el porcentaje de ciertos países referenciales del mundo²⁰, para destacar la diferencia que existe entre la zona de residencia.

¹⁹ Estimaciones de personal del Banco Mundial sobre la base de las Perspectivas de la urbanización mundial de las Naciones Unidas.

²⁰ https://datos.bancomundial.org/indicador/SP.RUR.TOTL.ZS?name_desc=false

**% de Ruralidad de diferentes países
del mundo en el 2018**

Valor de referencia: MUNDO = 45 %	
Nombre del país	%
Bermudas	0
Mónaco	0
Chile	12
Estados Unidos	18
España	20
Ecuador	36
Papua Nueva Guinea	87

Tabla 6: Ruralidad de algunos países del mundo.

Una de las características más importantes, para los analistas de datos, es que Twitter permite que se descarguen²¹ los contenidos de los mensajes en distintos formatos. Este hecho ha permitido que la comunidad académica haya desarrollado diversos avances en el tratamiento de la información.

Pero definitivamente las muestras son sesgadas. En efecto, cada país, cada localidad geográfica tiene su propia realidad. En la práctica se deben reconocer las limitaciones de la metodología propuesta. No *per se*. La limitación básica está dada por las características de asimetría y accesibilidad al uso de Twitter. A la ruralidad de las localidades, al acceso a las tecnologías de la información, al género del usuario, a la edad, a la disposición de interactuar de manera específica en esta Red y finalmente que le interese la política en cualquiera de sus facetas.

Con el fin de ir corrigiendo el hecho de que los datos con que se trabajan provienen definitivamente de muestras sesgadas, la Metodología que se plantea introduce la utilización de los factores de ponderación geográficos, en este particular se debe ser lo más exhaustivo que se pueda e ir descomponiendo las ubicaciones de los mensajes de Twitter en localidades ubicables geográficamente.

Mediante el siguiente ejemplo se demuestra el por qué se debe utilizar los factores de ponderación geográficos.

Sea $P = \{P_1, P_2, P_3\}$ una determinada entidad geográfica con $NP=3$ provincias, donde cada una de estas tienen las siguientes localidades (comarcas, municipios, cantones, etc.):

²¹ Instagram también permite que se descargue la información, pero su contenido básicamente de “imágenes” ha hecho que su uso para fines predictivos no se haya desarrollado lo suficiente.

$$P_1 = \{ L_1, L_2 \}, P_2 = \{ L_3, L_4, L_5 \} \text{ y } P_3 = \{ L_6 \}.$$

Y sean los actores principales son: $A = \{aA, aB, aC, aD\}$

Las proporciones poblacionales de cada provincia se muestran en la Tabla 7, siendo la provincia P_3 la que concentra la mayor cantidad de población electoral.

Provincias	Proporción poblacional pP_i
P_1	0,35
P_2	0,25
P_3	0,40
Total	1,00

Tabla 7: Composición proporcional de las Provincias

En la Tabla 8 se muestra una representación matricial de $RS_{A,T}$ (asumimos que la temporalidad ha sido recabada en las fechas previstas). Con el fin de facilitar asumimos que existen 16 actores activos diferentes y que cada uno ha emitido un mensaje en la cuenta de cada uno de los candidatos que están en contienda y los denominaremos como u_i .

Observación	localidad	usuario	Candidato	contenido
obs	loc	u_i	a	t\$
1	L1	1	aA	es un amigo bueno y colaborador
2	L1	2	aA	el es amigo colaborador
3	L1	3	aA	es un amigo malo mentiroso y corrupto
4	L3	4	aA	se supone que viajó por corrupto
5	L4	5	aA	el es malo mentiroso y corrupto
6	L4	6	aA	el es colaborador con todos
7	L5	7	aB	el es eficaz y amigo
8	L2	8	aB	el viajó por ineficaz
9	L2	9	aB	siempre ha sido corrupto
10	L6	10	aC	el es eficaz y colaborador pero a veces es mentiroso
11	L1	11	aD	el es un amigo bueno y mentiroso
12	L1	12	aD	el es un amigo mentiroso y corrupto
13	L4	13	aD	el se ha comportado como corrupto mentiroso e ineficaz
14	L7	14	aD	era un amigo eficaz colaborador y a veces malo
15	L7	15	aD	ha sido un corrupto ineficaz
16	L7	16	aD	a él le dijeron ineficaz malo y corrupto

Tabla 8: Representación matricial de $RS_{A,T}$.

Supongamos que dichos mensajes ya están depurados y para la determinación de la polaridad utilizaremos el método del Enfoque Basado en Lexicones, ver la figura 8, utilizando los Enfoques de Diccionarios con palabras sin etiquetar y sin asignarles factores de pesaje. Se utiliza la métrica del conteo de palabras y se decide por el método de las mayorías. Uno de los métodos más simples utilizado como función de polaridad. Los diccionarios que utilizaremos son los siguientes:

$$DP = \{amigo, bueno, colaborador, eficaz\}$$

$$DN = \{malo, mentiroso, corrupto, ineficaz\}$$

Para efectos del análisis que se requiere realizar, utilizaremos la ecuación 6 para que la polaridad tome únicamente dos valores: 1 si es positivo y 0 en caso contrario.

El conteo de palabras positivas (cp) en cada contenido, así como el de palabras negativas (cn) ha sido evaluada para los contenidos $t\$$ de cada una de los mensajes consideradas en la matriz $Mm\$$, para de esta manera calcular la polaridad (pol_j). El siguiente paso es el de evaluar los factores de ponderación geográficos para cada uno de los electores que en nuestro caso corresponde a los mensajes.

Utilizando las ecuaciones 11, 12 y 13 para ir evaluando los términos de N_j y D , con estos valores calculamos la proporción muestral dada por M_j . Se evalúa la proporción de los términos del universo poblacional empleando la ecuación 14 y finalmente tenemos que el factor de ponderación geográfico fpg_j dado por la ecuación 15, para cada elector que se presenta en la Tabla 9.

Observación	Candidato	contenido	Conteo de positivos	Conteo de negativos	polaridad	factor de ponderación geográfico
obs	a	t\$	cp	cn	polj	fpg
1	aA	es un amigo bueno y colaborador	3	0	1	0,70
2	aA	el es amigo colaborador	2	0	1	0,70
3	aA	es un amigo malo mentiroso y corrupto	1	3	0	0,00
4	aA	se supone que viajo por corrupto	0	1	0	0,00
5	aA	el es malo mentiroso y corrupto	0	3	0	0,00
6	aA	el es colaborador con todos	1	0	1	0,75
7	aB	el es eficaz y amigo	2	0	1	0,75
8	aB	el viajo por ineficaz	0	1	0	0,00
9	aB	siempre ha sido corrupto	0	1	0	0,00
10	aC	el es eficaz y colaborador pero a veces es mentiroso	2	1	1	2,40
11	aD	el es un amigo bueno y mentiroso	2	1	1	0,70
12	aD	el es un amigo mentiroso y corrupto	1	2	0	0,00

13	aD	el se ha comportado como corrupto mentiroso e ineficaz	0	3	0	0,00
14	aD	era un amigo eficaz colaborador y a veces malo	3	1	1	1,00
15	aD	ha sido un corrupto ineficaz	0	2	0	0,00
16	aD	a el le dijeron ineficaz malo y corrupto	0	3	0	0,00

Tabla 9: Determinación de los factores de ponderación geográficos

Finalmente corresponde realizar el cálculo de la valoración para cada uno de los candidatos o actores principales utilizando la ecuación 17 y también la valoración porcentual mediante la ecuación 18.

En la Tabla 10, en la primera columna se muestra la valoración porcentual realizada para cada uno de los actores principales o candidatos utilizando el procedimiento descrito en el capítulo 2, es decir determinando la polaridad de los mensajes y calculando el factor de ponderación geográfico de cada elector, en la siguiente columna tenemos la valoración, como tradicionalmente se realiza, utilizando únicamente la polaridad y sin utilizar factores de ponderación geográficos y finalmente la valoración calculada únicamente con las menciones de cada usuario/elector, en la que no se utilizan ni el cálculo de las polaridades ni el cálculo de los factores de ponderación geográficos.

Valoración porcentual de cada Actor por tres métodos			
actor	polaridad y factor de ponderación geográfico	polaridad	menciones
aA	30,7	42,9	37,5
aB	10,7	14,3	18,8
aC	34,3	14,3	6,3
aD	24,3	28,6	37,5
Total	100,0	100,0	100,0

Tabla 10: Comparativo de la valoración por tres métodos de cálculo.

Los resultados son contradictorios. El mejor valorado utilizando polaridad y factores de ponderación geográficos, combinados es el actor principal aC con un 34,3% que resulta ser el peor valorado con los otros dos métodos. Hay que observar que los mensajes emitidos de la provincia 3, la más grande deben reescalarsse por cuanto hay que darle a esta provincia su verdadera proporción que tiene en la población de estudio.

En la Tabla 11 se muestra la ubicación de cada actor principal o candidato de acuerdo a cada uno de los métodos que se emplean. En esta se observa que los métodos de únicamente polaridad y el de menciones casi arrojan resultados similares.

Orden de valoración por cada método			
polaridad y factor de ponderación	polaridad	menciones	Ranking
aC	aA	$aA-aD$	1
aA	aD	aB	2
aD	$aB-aC$	aC	3
aB			4

Tabla 11: Orden de ubicación por cada uno de los métodos de cálculo

Teniendo en cuenta lo mencionado anteriormente acerca de las muestras sesgadas que se obtienen al trabajar con mensajes de Twitter, pero incorporando lo expuesto en la Metodología de predicción Electoral, hemos realizado tres casos de estudio.

Elecciones Presidenciales en el Ecuador. Primera vuelta 2017

La metodología fue aplicada en las elecciones presidenciales en el Ecuador que se realizaron el 19 de febrero del 2017, de acuerdo a la legislación ecuatoriana esta elección sería la de primera vuelta. Ecuador es un país que está compuesto por 24 provincias y en donde todos los ecuatorianos residentes en el exterior pueden votar, convirtiéndose así en otra “provincia” más. El nivel en el uso de Twitter no llega de manera uniforme a todos, existen regiones y zonas en donde su alcance y penetración está muy por debajo de la media nacional. Está claro entonces que “la voz” de ciertos electores no estará presente en igual proporción que los habitantes de otras zonas.

En la Tabla 12 se muestran: el total de la población, el total de los que manifiestan tener una cuenta en una Red Social (en el Ecuador el nivel de posesión de cuentas en Facebook es casi mayoritario, del total que manifiestan tener una cuenta el 97,9% manifestaron tenerla) y el total de la población que tienen cuenta en Twitter, a estos totales se les asocia sus respectivos porcentajes. Los datos hacen referencia a la población de más de doce años de edad. Segmentada de acuerdo a ciertas variables básicas como son: la zona de residencia, la región geográfica en que se habita (El Ecuador está

constituida por cuatro regiones geográficas, Costa, Sierra, Amazonía y la región insular Galápagos) y el quintil de consumo²².

Nivel de desagregación por zona, región y quintil de consumo	Total (Población de 12 años y más)	% Población	Total (Población de 12 años y más que posee una cuenta en alguna red social)	Total (Población de 12 años y más que utiliza Twitter)	% utilización Twitter
Área					
Nacional	12.052.548	100,0	4.995.474	1.019.607	100,0
Urbano	8.263.136	68,6	4.092.522	924.908	90,7
Rural	3.789.412	31,4	902.952	94.699	9,3
Región Natural					
Sierra	5.484.261	45,5	2.380.433	444.339	43,6
Costa	5.974.453	49,6	2.399.306	550.026	53,9
Amazonía	571.784	4,7	201.919	22.200	2,2
Galápagos	22.049	0,2	13.816	3.043	0,3
Quintiles de Consumo					
Total	11.970.446	100,0	4.963.688	1.008.924	100,0
Quintil 1	2.052.456	17,1	449.428	38.236	3,8
Quintil 2	2.262.343	18,9	725.434	83.614	8,3
Quintil 3	2.396.254	20,0	1.000.982	148.934	14,8
Quintil 4	2.552.687	21,3	1.223.247	259.659	25,7
Quintil 5	2.706.706	22,6	1.564.597	478.482	47,4

Tabla 12: Porcentaje de la población de 12 años y más que posee una cuenta en Twitter (Fuente: INEC, ECV. Sexta Ronda 2013 – 2014)

Los que poseen una cuenta de Twitter a nivel nacional son 1.019.607 de personas, que representa el 8,5% de la población de más de 12 años de edad y el 20,4% de la población de más de doce años que poseen una cuenta en alguna Red Social. Con referencia a las variables de segmentación consideradas tenemos que: de acuerdo a la zona de residencia del usuario, los pobladores de las zonas rurales son los que menos lo poseen, tan solo 94.699 tienen cuenta en Twitter que representa el 10,5% de los que tienen una cuenta en red social. El 11% de los pobladores de la región geográfica del oriente ecuatoriano tienen cuenta de Twitter con respecto a los que si poseen una cuenta en alguna Red Social; muy por debajo de los otros pobladores de las otras regiones del país; se tiene que los habitantes que se pertenecen al quintil 3 e inferior de consumo están por debajo

²² Una forma de medir el nivel socio-económico es a través del consumo de las personas. Si de este indicador se obtienen los quintiles, en donde el primer quintil, estarán las personas de menor consumo y en el quintil cinco, aquellos que tienen mayor cantidad económica de consumir

de la media nacional en lo referente a poseer una cuenta en Twitter (con respecto a aquellos que poseen cuenta en alguna Red Social).

Esta asimetría en la tenencia de Twitter se la representa en la Figura 16. Se representan los porcentajes de la población de más de doce años del Ecuador de acuerdo a algún criterio de segmentación acompañado del porcentaje de la población de los que utilizan Twitter con respecto a su total.

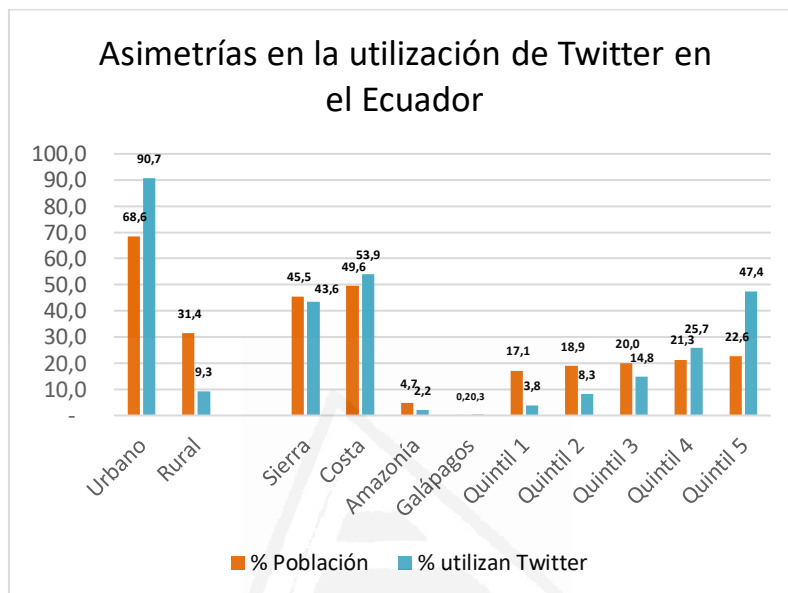


Figura 16. Asimetrías en el perfil de la Audiencia de Twitter en el Ecuador

Teniendo presente de que existen asimetrías y por ende se trabajarán con muestras sesgadas, pues el acceso y tenencia de una cuenta en Twitter es desigual, es que es más prioritario el utilizar los factores de ponderación geográficos.

El primer paso entonces consistió en definir la temática de investigación, para lo cual se necesitaban los nombres de los potenciales candidatos. De acuerdo a los resultados de diferentes encuestas realizadas por diferentes empresas en octubre del 2016 y publicada en la revista Vistazo, citadas en el artículo “¿Quién es quién? en el duelo de las encuestas” los candidatos Moreno, Lasso, Viteri y Moncayo se disputaban con alguna opción los primeros lugares para la presidencia, y el 10,3% del electorado estaría por un grupo de cuatro candidatos: Bucaram, Espinel, Zuquilanda y Pesántez. A pesar de la anticipación con el día de las elecciones se estaban ya perfilando los potenciales finalistas, y se notaba una aparente definición en los resultados, se presenta en la Tabla 13, los resultados de las encuestas y se observa que una de ellas alcanza un **MAE = 2,2**.

A estos cuatro últimos candidatos, para efectos prácticos se les denominó como “otros”. En tal sentido se definieron las cuentas

oficiales en Twitter de los candidatos y del partido político que los apoyaba para la extracción de los tuits de los usuarios que los siguen.

Encuestas realizadas en el Ecuador y presentadas el 20/10/2016						
PreCandidatos	Consejo Nacional Electoral	CIE ES	CEDATOS	Perfiles de Opinión	Opinión Pública Ecuador	VISOR
Lenin Moreno	39,4	43,0	35,0	49,0	44,0	34,6
Guillermo Lasso	28,2	16,0	21,0	15,5	14,0	27,7
Cynthia Viteri	16,3	8,0	7,0	10,6	4,0	16,0
Paco Moncayo	6,7	3,0	5,0	2,3	1,0	11,4
Otros	9,4	30,0	32,0	22,6	37,0	10,3
Total	100,0	100,	100,0	100,0	100,0	100,0
MAE		9,7	9,0	9,1	12,9	2,2

Tabla 13: Resultado de diversas encuestas realizadas en el Ecuador. 2016

Como temporalidad T se definió como fecha de inicio t_i el 20 de diciembre del 2016 y como t_f el 14 de febrero del 2017.

Existen muchas aplicaciones gratuitas y de pago que permiten descargar la información del Twitter. Utilizando la herramienta: Tags de Google se empezó la descarga siguiendo las observaciones de (TUMASJAN ET AL., 2010) sobre la periodicidad de la toma de datos.

Fueron descargados un total de 823.135 mensajes correspondientes a los usuarios que seguían estas cuentas oficiales de todos los candidatos, en la Tabla 14, se presentan el número de los mensajes descargados por cada actor principal o candidato, los mensajes no duplicados es decir los que se deben imputar a cada elector y el factor de repitencia o mensajes promedios por cada elector. De la totalidad de mensajes se redujeron a 132.379 electores.

Candidato	No de tuits descargados	No de tuits no duplicados	Factor de repitencia
Lenin Moreno	427.190	40.656	10,5
Guillermo Lasso	128.976	27.544	4,7
Cynthia Viteri	84.183	15.005	5,6
Paco Moncayo	71.197	15.700	4,5
Otros	111.589	33.473	3,3
Total	823.135	132.378	6,2

Tabla 14: Twitter descargados, duplicados y factor de repitencia por cada candidato

Depuramos ciertos valores del campo de la localidad donde fue emitido el mensaje de Twitter, por ejemplo existen casos en

donde encontramos las palabras: “quito”, “uio”, “Quito!!!” u otra expresión asociada a la ciudad de Quito, (capital del Ecuador y que se pertenece a la Provincia de Pichincha) debemos encontrar esas expresiones y asociarlas a un solo nombre que será solo “Pichincha”, realizamos una depuración para las 24 provincias del Ecuador, para localidades extranjeras y aquellas que no son localidades o que no entran en la clasificación antes mencionada o no poseen el campo Localidad se les asignó la localidad *otros*.

Para la determinación de la polaridad utilizamos el método del Enfoque Basado en Lexicones, para lo cual se construyó un lexicon de opinión y se lo adaptó a las características y modalidades del español de Ecuador, este lexicon es un mecanismo lingüístico que contiene la polaridad de cada palabra, ya sea positiva o negativa. Se construyó entonces un diccionario de palabras positivas (1.625) y otro de palabras negativas (2.942) asignándoseles un valor equiprobable a cada uno de los términos. Como métrica o función de polaridad se procedió al conteo de palabras y si existía un saldo positivo se le asignó un valor de 1 o a favor, si el saldo era negativo en cambio se le asignó la polaridad de -1 o en contra y finalmente si este saldo era cero, se decía que el mensaje tenía polaridad 0 o neutro.

En la Tabla 15 se presentan el número de electores que fueron asignados de acuerdo a la función de polaridad que se utilizó y su imputación, el 20,7% tuvieron polaridad positiva y el 79,3% son no a favor.

Candidato	En contra	Neutro	A favor	Total
Lenin Moreno	2.520	27.611	10.525	40.656
Guillermo Lasso	2.370	17.031	8.143	27.544
Cynthia Viteri	1.095	9.926	3.984	15.005
Paco Moncayo	1.134	12.917	1.649	15.700
Otros	3.368	26.948	3.157	33.473
Total	10.487	94.433	27.458	132.378

Tabla 15: Polaridad imputada a los usuarios

Se realizó el cálculo de la variable Relevancia utilizando la función *Ratio ponderado de seguidores y RT* descritos en el capítulo 2, se asignó los valores $\alpha = 1$ y $\beta = 3$ y determinamos su función de densidad, la que más se ajusta a los datos de la variable Relevancia es una de tipo Johnson SI, cuyos parámetros se describen en la Tabla 16:

Tipo	Parámetro	Estimación
Forma	γ	0,973
Forma	δ	0,012
Localización	θ	0,000
Escala	σ	1

Tabla 16. Parámetros de la Distribución Johnson SI Distribución de la variable Relevancia.

Se utilizó la prueba de Kolmogorov-Smirnov-Lilliefors, K-S-L, (PEDROSA, 2015) para la bondad de ajuste y se aceptó la hipótesis nula $H_0 = \text{los datos provienen de una distribución Johnson SI}$. Con esta función de densidad se determinó su respectiva función de distribución acumulada CDF, para definir los umbrales en los cuales daríamos la conversión de los votos dados por el valor de la relevancia. Utilizamos como umbral el cuantil 97,5%, $Q_{97,5\%}$, por sobre el cual si la Relevancia es mayor se le daría a ese elector un total de tres votos, asumiendo una hipótesis conservadora de que su mensaje podría incidir en otras dos personas más, si la Relevancia es mayor que el cuantil 95%, $Q_{95\%}$, pero menor al $Q_{97,5\%}$ se le asignará dos votos, el podrá incidir en un elector más. Para valores menores que el $Q_{95\%}$ asumimos que no tendría incidencia en otros electores.

Como se muestra en la Tabla 17 se presenta el comparativo entre los resultados oficiales dados por el Consejo Nacional Electoral del Ecuador, CNE, los resultados de las principales empresas dedicadas a la realización de encuestas políticas en el Ecuador presentados el 8 de febrero del 2017²³; las empresas MARKET y CEDATOS ligadas a los partidos de oposición y CIESS, Perfiles de Opinión y Opinión Pública Ecuador cercanas al oficialismo y los datos procesados siguiendo la metodología propuesta; utilizando el factor de ponderación geográfico y sin utilizarlo.

La variabilidad de los resultados entre las empresas encuestadoras siempre pone en tela de duda a las empresas en primer lugar y luego, lo que es más complejo, sobre la técnica de las encuestas electorales. En el caso del candidato Moreno, existe una diferencia máxima de 10,9% entre los resultados de las empresas encuestadoras.

Candidato	CNE	Resultado de las empresas encuestadoras					PROPUESTA	
		MARKET	CIESS	Perfiles de Opinión	Opinión Pública Ecuador	CEDATOS	Sin utilizar el factor de ponderación geográfico	Utilizando el factor de ponderación geográfico
Lenin Moreno	39,4	32,4	43,3	41,7	42,6	38,6	38,3	41,7
Guillermo Lasso	28,1	20,8	21,3	19,0	22,7	25,7	29,7	27,7
Cynthia Viteri	16,3	23,0	12,6	16,7	17,1	16,7	14,5	14,0
Paco Paco	6,7	13,1	10,8	8,3	9,2	9,2	6,0	6,6

23

https://es.wikipedia.org/wiki/Anexo:Sondeos_de_intenci%C3%B3n_de_voto_para_las_elecciones_generales_de_Ecuador_de_2017

Otros	9,5	10,7	12,0	14,3	8,5	9,7	11,5	10,1
Total	100	100	100	100	100	100	100	100
MAE		5,7	4,2	3,6	2,6	1,3	1,4	1,1

Tabla 17: Comparativo de resultados oficiales, dados por las empresas encuestadoras y los datos predichos por la metodología

Como se muestra en la Tabla 17, el error más pequeño que cometió una de las empresas encuestadoras fue CEDATOS, con un **MAE = 1,3** frente a la predicción realizada analizando los mensajes de Twitter con la metodología propuesta usando los factores de ponderación geográficos con un **MAE=1,1** hacen que esta metodología sea una alternativa en la predicción electoral. También la predicción sin utilizar los factores de ponderación tiene un buen desempeño.

Elecciones Presidenciales en Chile. Segunda vuelta 2017

La elección de la primera vuelta presidencial de Chile se realizó el 19 de noviembre del 2017 y los candidatos finalistas para la segunda vuelta presidencial fueron: Piñera y Guillier. En tal sentido SERVEL, que es el organismo oficial encargado de los procesos electorales, convocó para elección de segunda vuelta en Chile, el 17 de diciembre del 2017.

Chile está compuesto por 15 regiones y los chilenos residentes en el exterior también pueden votar, convirtiéndose esta en una *región* adicional. El nivel de ruralidad de Chile es del 12%, ver en la Tabla 6, lo que incide que el uso de las redes sociales sea mayor en comparación con otros países de la región. Así el uso de las redes sociales es muy difundido y Facebook se convierte en la más utilizada, el nivel en el uso de Twitter alcanza al 47% de la población, en la Tabla 18²⁴ se presentan los porcentajes de uso de las principales redes sociales en Chile.

Redes Sociales	%
Facebook	93
Instagram	74
Twitter	47
Linkedin	32

Tabla 18: Porcentaje de uso de las redes sociales en Chile

Según un estudio de la empresa AnalITIC²⁵, el 59% de los tuiteros son hombres frente al 41% que son mujeres, lo que indica

²⁴ Según el estudio de las empresas Cadem y Jelly

²⁵ <https://www2.deloitte.com/uy/es/pages/home/articulos/Tendencias-en-Analytics-2016.html>. Descargado en marzo, 2019

que su uso no llega de manera uniforme a todos, las asimetrías también se presentan en relación a la variable rango etario del usuario; en la Figura 17 las personas que están comprendidas en el rango de 22 a 51 años de edad son los que más utilizan el Twitter en comparación a la población en general, siendo las personas de más de 51 años las que menos utilizan el Twitter en relación a la población en general. Entonces “la voz” de ciertos electores no estará presente cuando se descarguen los mensajes de Twitter y por ende la muestra será sesgada.

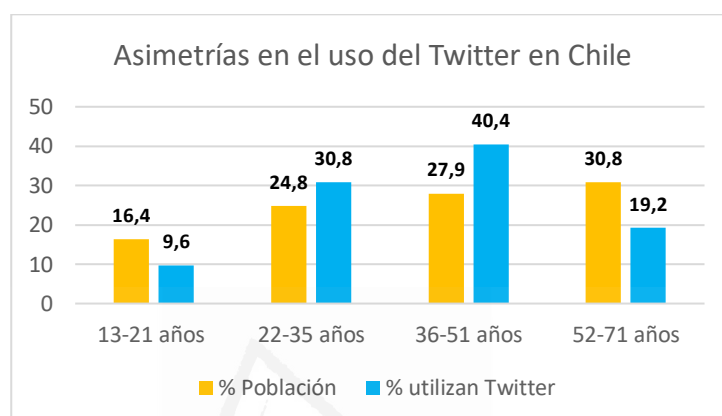


Figura 17. Asimetrías en el perfil de la Audiencia de Twitter en Chile

Se aplicó la metodología descrita en el capítulo 2 y para la definición de la Temática se consideraron entonces como actores principales a Piñera y Guillier para lo cual se utilizaron sus cuentas y las de sus partidos políticos que los respaldaban para la extracción de los tuits de los actores activos que los seguían en Twitter.

Para la temporalidad T se definió como fecha de inicio t_i el 20 de noviembre del 2017 y como t_f el 15 de diciembre del 2017. Como mencionamos la elección de la primera vuelta presidencial de Chile fué el 19 de noviembre del 2017; es por esto que no se pudo seguir la recomendación de (TUMASJAN ET AL., 2010) sobre el tiempo de descarga de los datos.

Utilizando la herramienta Social Analytics de la Universidad de Alicante, España, se realizó la descarga, dicho software además permite el cálculo de la polaridad utilizando aproximación basada en skipgrams. Se descargaron un total de 89.162 mensajes de usuarios de estas cuentas.

Para el cálculo de los factores de ponderación geográficos se tuvo que establecer la proporción Poblacional (electoral) p_{Pi} de cada Región de Chile, así por ejemplo la Región Metropolitana de Santiago concentra al 39,4% de los electores nacionales y al finalizar la descarga se tuvieron un 65,7% de electores de twitter de esta Región, que interactuaron en la red a través de sus mensajes, esta asimetría se presenta en la Figura 18.

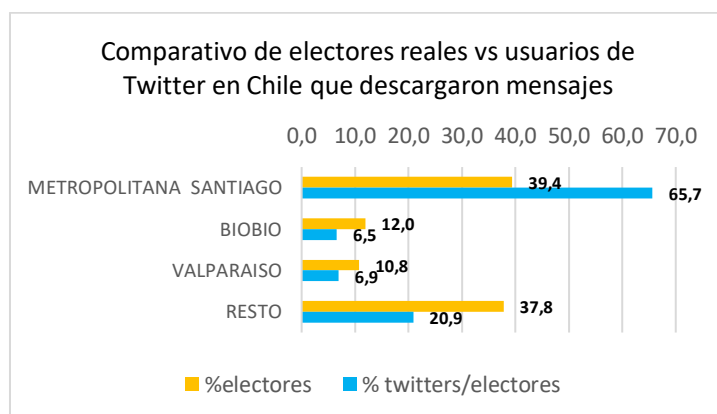


Figura 18. Comparativo de electores reales y electores Twitter.

En la Tabla 19 se muestra el comparativo de los resultados oficiales del Servicio Electoral de Chile, SERVEL, los resultados de la principal empresa encuestadora de Chile, CADEM; que predijo empate técnico y los datos procesados siguiendo la metodología propuesta; utilizando el factor de ponderación y sin utilizarlo.

Candidatos	Resultado oficial SERVEL	Encuestadora CADEM	Con factor de ponderación	Sin factor de ponderación
S. Piñera	54,6	50,9	53,8	53,1
A. Guillier	45,4	49,1	46,2	46,9
MAE		3,7	0,8	1,5

Tabla 19: Comparativo de los resultados oficiales, resultado de la empresa encuestadora y los predichos por la metodología propuesta con el factor de ponderación geográfico y sin el factor de ponderación en las elecciones presidenciales de segunda vuelta de Chile

El $MAE=0,8$ demostraron que esta metodología sirve como una alternativa válida en la predicción de los resultados electorales.

Consulta Popular en el Ecuador 2018

En la mayoría de los referéndums, se le consulta a la población sobre una pregunta en particular que equivalen a una elección de tipo usual entre dos candidatos, el candidato A y el candidato B. Un caso particular es cuando los gobiernos nacionales o locales convocan a un referéndum en la que se consulta sobre diversos temas y la población debe elegir para cada tema entre dos posiciones; si o no, a favor o en contra, de acuerdo o en desacuerdo; posiciones antinómicas que generalmente polarizan la opinión del

país o de la localidad; estos procesos democráticos equivalen a un conjunto de elecciones, cada una de dos candidatos.

Si un referéndum en particular posee n preguntas, existirán 2^n diversas formas de votar, pero la mayoría de las personas tienen ya una posición definida, o se está a favor en todas o se está en contra en todas las preguntas. Salvo, claro está, si existe una pregunta con un marcado sesgo para que la población se incline mayoritariamente a favor o en contra.

Con el objetivo de predecir los resultados de un referéndum utilizando el análisis en las redes sociales se deberá evaluar el parecer de los usuarios en la Red sobre una y otra posición. Si se utiliza la red social Twitter, como mecanismo para analizar la opinión de los electores, las diferentes formas de votar no se las puede detectar debido al espacio reducido, máximo 280 caracteres, que se dispone.

Con el fin de formalizar consideraremos un referéndum de tipo binario, aquel en que los ciudadanos tienen únicamente dos opciones para decidir y podemos definir la posición de un ciudadano j , con $j=1,2,\dots,m$; en donde m representa el número de opiniones descargadas en Twitter con respecto a la pregunta i , con $i=1,2,\dots,npr$; en donde npr representa el número de preguntas de un referéndum; de la siguiente manera:

$$x_{i,j} = \begin{cases} 1, & \text{si el individuo } j \text{ está a favor de la pregunta } i \\ 0, & \text{caso contrario} \end{cases}$$

Obviamente $x_{i,j}$ sigue una distribución Bernoulli y lo que deseamos estimar es la proporción de la población que está a favor en la pregunta i , es decir la variable X_i :

$$X_i = \frac{\sum_{j=1}^m x_{i,j}}{m}$$

Como se mencionó, frente a la dificultad de estimar este valor individual de X_i para todas las n preguntas del referéndum utilizando los tweets, es que nosotros determinaremos un estimador de tipo general, que dé cuenta del grado de apoyo general mediante la siguiente expresión:

$$X = \frac{\sum_{i=1}^{npr} X_i}{npr}$$

Este estimador X representa el promedio de las proporciones de cada pregunta del Referéndum y al provenir de una muestra de

individuos que han dado su opinión mediante un tweet es una variable aleatoria.

En el Ecuador se desarrolló el Referéndum y Consulta Popular el 4 de febrero del 2018, proceso que contenía un total de siete preguntas de tipo binario, pues había que responder por el SI o por el NO a cada una de ellas, en tal sentido aplicamos la metodología expuesta en el capítulo 2 con las consideraciones que hemos anotado acerca de la medición de la variable X , que representa el promedio simple de los valores a cada una de las siete preguntas que se efectuaron.

Para la definición de la temática del Referéndum y Consulta Popular del Ecuador había dos actores políticos que las lideraban, el presidente actual Lenin Moreno y el expresidente Rafael Correa, el primero lideraba la opción por el SI y el segundo por la opción del NO. Las cuentas oficiales de estos actores políticos se las denominó las principales y adicionalmente se consideraron un conjunto de cuentas secundarias y hashtags pertenecientes a diversos colectivos políticos que apoyaban una u otra opción, estas cuentas fueron consideradas para que la base de documentos fuera diferente cuando se iba a realizar el aprendizaje por cuanto se deseaba incorporar factores diversos para mejorar la predictibilidad del modelo propuesto y además el hecho de que no se deseaba determinar cuál método de clasificación era mejor que el otro.

La asignación a cuentas a descargar para cada una de las posiciones y por cada método se describen en la Tabla 20:

Posición	Cuentas principales	Cuentas secundarias por cada método		
		NB	RF	SVM
SI	@Lenin	@7vecessi	@35PAIS	#EcuadoSí
			@sietevecesSI	#EcuadorSi
NO	@MashiRafael	@SomosmasEc	#SiporEcuador	
			@EcuadorRC	#NoALaConsulta
			#DilesNo	

Tabla 20. Cuentas principales y secundarias utilizadas

La temporalidad T o período de descarga de la información se definió como inicio t_i el 18 de enero del 2018 y t_f el 2 de febrero del 2018, es decir hasta dos días antes del proceso electoral.

En la determinación de la función de polaridad, tratado en el capítulo 1 existen diversos procedimientos, detallados en la Figura 8, uno de ellos son los basados en la mezcla de modelos, para mejorar la precisión de los clasificadores, estos métodos se denominan Métodos de Conjunto, puesto que combinan los resultados de más de un modelo.

Para configurar el método de conjunto se deben determinar entonces los métodos que se van a combinar y para realizar el análisis se escogieron los de aprendizaje supervisado que es una técnica para determinar una función de polaridad a partir de datos de entrenamiento y que sea capaz de predecir el valor de cualquier dato de entrada. Los datos de entrenamiento son vectores depurados, formados por las palabras de un tweet, a los cuales se los ha entrenado de manera manual con un valor de polaridad; y se procede a validar dicha función con un conjunto de datos de prueba. Una vez que el clasificador tenga la precisión deseada se utiliza dicha función para clasificar el resto de datos.

Se procedió a seleccionar de entre los métodos de aprendizaje supervisado al Clasificador Probabilístico naïve Bayes, al Clasificador de árboles de decisión Random Forest y al Clasificador con kernel lineal Support Vector Machine, los dos primeros métodos fundamentados en las probabilidades subyacentes de los mensajes y el tercer método fundamentado en la geometría de su representación.

Una vez que cada conjunto de datos fue debidamente entrenado y evaluado por cada uno de los métodos el siguiente paso fue el de realizar la implementación, es decir obtener el valor del estimador de la proporción de aceptación de la opción SI (que lo notaremos por X), obteniéndose los siguientes resultados:

Utilizando el clasificador naïve Bayes

El clasificador de naïve Bayes considera que un texto es positivo o negativo si están presentes o ausentes ciertos términos y que éstos contribuyen independientemente a la probabilidad final de pertenecer a una categoría u otra, para determinar estas probabilidades condicionadas se debe enseñar y clasificar que textos son positivos o negativos.

Dentro de las medidas de evaluación para determinar si el entrenamiento fue correcto o si el método si aplica se utilizó la *exactitud* o Accuracy = 78,3%.

Calculando el error en la estimación con este método se obtuvo un $MAE_{NB}=1,31$.

El valor de X evaluado por este método lo notamos por X_{NB} tuvo un valor de 0,6896 para la opción SI.

Utilizando el clasificador Random Forest

El método de Random Forest optimiza el poder de clasificación de los árboles de decisión; por cuanto se construye un árbol de decisión normal y cada vez que se realiza una división se

utiliza un subconjunto reducido de características aleatorias; así se construyen diversos árboles aleatorios y finalmente se los promedia para obtener el modelo final. Estos pasos ayudan a reducir la correlación entre estos árboles para minimizar la varianza del árbol final.

Como medida de evaluación se obtuvo el Accuracy = 82.4%. Y de manera individual este método arroja un $MAE_{RF}=2,93$.

Para el estimador X o promedio, que lo notamos por X_{RF} tuvo un valor de 0,7058 para la opción del **SI**.

Utilizando el clasificador Support Vector Machine

El SVM es un método analítico, que no se basa en la probabilidad de los términos y observaciones y consiste en un problema de programación cuadrática que se resuelve mediante su dual por el método de multiplicadores de Lagrange. Lo que busca el método es un hiperplano óptimo en el espacio de características utilizando una función de kernel. La solución consiste en la combinación lineal de los vectores de soporte de este hiperplano.

Como indicador se obtuvo Accuracy = 86,4%. En este método se tiene un $MAE_{SVM}=3,13$.

Y el valor de X que lo notaremos por X_{SVM} tuvo un valor de 0,6452 para la opción **SI**.

De acuerdo a los métodos de conjunto que combina los resultados de más de un modelo se procedió a realizar la estimación del valor X mediante la siguiente expresión:

$$X = \frac{X_{NB} + X_{RF} + X_{SVM}}{3}$$

Obteniéndose el valor $X= 68,02 \%$ para la opción del **SI** y de **31,98%** para la opción del **NO**.

Nótese que cada uno de los métodos aplicados de manera individual hubieran producido un mayor MAE , esta es una de las ventajas de los Métodos de Conjunto, pues al combinar diferentes métodos se consigue una reducción en la varianza total. Si se tratara de evaluar cual método fue más eficiente, el NB alcanza menor MAE pero su Accuracy es la más baja. Claro está que uno no conoce los resultados y no se puede saber a que método se le debería dar mayor peso, en tal sentido una medida conservadora es el de dar igual peso a cada resultado parcial.

En la Tabla 21 se contrastan los resultados oficiales dados por el Consejo Nacional Electoral, CNE, con los resultados dados por firmas encuestadoras del Ecuador, los resultados de un medio de prensa y los resultados del MODELO considerado.

Resultados oficiales, de encuestadoras y del modelo						
Opciones	Oficiales	Encuestadoras y prensa				Modelo
	CNE	CEDATOS	CLIC K	DIAGNÓSTICO	EL UNIVERSO	
SI	67,65	76,24	73,11	75,94	78,8	68,02
NO	32,35	23,76	26,89	24,06	21,2	31,98
MAE		8,59	5,46	8,29	11,15	0,37

Tabla 21. Comparación de resultados entre empresas encuestadoras y el modelo utilizado

Un $MAE=0,37$ demuestra que la metodología propuesta es eficiente y se convierte en una buena alternativa para predecir resultados electorales.

Capítulo 4

Conclusiones

La investigación que se presenta en este trabajo, es una propuesta metodológica sobre el procedimiento a seguir en la realización de pronósticos electorales en la Red Social Twitter, el tema es muy sensible, no es lo mismo evaluar una variable cuando no existe contraste frente al proceso de pronosticar existiendo verificación. Las elecciones políticas se han convertido en un indicador sobre los procedimientos que se utilizan, en la medida que sus resultados pueden ser contrastados con los datos reales. En tal sentido la metodología propuesta debe conseguir un *MAE* muy reducido, inclusive menor que los resultados proporcionados por las empresas u organismos que realizan encuestas tradicionales (estadísticamente diseñadas).

Esta propuesta pretende mejorar la predictibilidad en la Red Social Twitter mediante una nueva metodología de análisis y dado que los tuiteros son una muestra sesgada de la población electoral hay que determinar las características socio-demográficas de ellos con el propósito de ir adaptando la muestra observada hacia una muestra estadísticamente significativa de los electores. Pero el principal problema y limitación que existe es la asimetría que existe en la tenencia y uso de Twitter en la población electoral en general, sus usuarios son mayoritariamente hombres afincados en zonas urbanas y además son jóvenes.

La metodología se fundamenta en que la unidad de análisis lo constituye el elector y NO los mensajes descargados del Twitter, nuestro objetivo no es realizar estrategias de campaña, sino que es predecir un resultado electoral y en tal sentido se deben construir funciones de imputación de ciertas características de los mensajes de un usuario a una sola característica, que será asignada a cada usuario.

Una característica básica es la polaridad de un mensaje, que es la medida de asignación o clasificación a una clase predefinida, generalmente en política utilizamos la clase a favor o no a favor, y en la actualidad el estado del arte ha permitido que se dispongan de varias funciones de clasificación y al combinarse diferentes características de ellas las clasificaciones se han hecho más precisas; por lo que esta búsqueda de la polaridad ya no es una limitación. La información que se descarga es dinámica y evoluciona en el tiempo ya que los usuarios emiten sus opiniones y adoptan posiciones políticas definidas en el tiempo en que dura la campaña política y hay que considerar posibles cambios respecto a su decisión electoral final. Cambios frecuentes de polaridad de los mensajes de un mismo

usuario se debe a que la función de polaridad no discrimina bien o simplemente proviene de la subjetividad o volubilidad del elector.

Estas funciones de imputación de polaridad que se definen transfieren la polaridad de los mensajes de un usuario al usuario, con este procedimiento de imputación también se consigue disminuir radicalmente la presencia de bots que afectan artificialmente o manualmente los seguidores de un candidato.

Otra característica básica es la relevancia, que determina el grado de importancia que pueda tener un mensaje dado por el nivel e influencia que pueda tener en los electores. Para determinar la relevancia de los mensajes se utilizará una métrica de *engagement* que mida básicamente la difusión y popularidad.

Las funciones de imputación de la relevancia de los mensajes la transferirán a él usuario. Pero esta relevancia que tiene cada usuario debe permitirle tener *cierta influencia política*, que en política se traduce en votos, electores con alta relevancia deberán poder persuadir a otros electores, pero la persuasión solo la tendrán ciertos electores, la gran mayoría asumiremos que no podrá incidir en otro elector más.

Con el fin de ir corrigiendo el hecho de que las muestras de los usuarios de Twitter son sesgadas, la metodología propuesta incluye el factor de ponderación geográfico que es el mecanismo por el cual se realiza un proceso de conversión para que la opinión dada en los tuits equivalga a los de la población electoral en general. Como el único dato “demográfico” que se dispone es la localidad de donde se realizan los tuits se deberá identificar la localidad y parametrizarla con alguna división administrativa existente y que se conozcan su número de electores.

La metodología ha sido probada en temas políticos y es aplicable en diversas temáticas de otros campos como la sociología, la investigación de mercados, el marketing digital, etc. A pesar de las dificultades en la asimetría de utilizar la Red Social Twitter, la propuesta que se presenta es una alternativa válida para la predicción o simplemente la descripción de cualquier temática social y no para reemplazar las encuestas tradicionales sino para considerarlas para la comparación de resultados y conseguir un *MAE* muy reducido.

En este sentido, algunos retos de trabajos futuros son la incorporación de otras variables sociodemográficas que permitan reducir el sesgo al trabajar con las muestras de los tuiteros; debido a que las asimetrías de los usuarios de Twitter se presentan por: la variable género, se deberá a través de los metadatos ir reconociendo a cual pertenece cada usuario; la variable edad, mediante las expresiones coloquiales en los contenidos de los mensajes se puede en cierta forma ir determinando el rango de edad del usuario; la

variable zona de residencia, de manera similar a la ubicación geográfica se deberá determinar si el usuario vive en zonas urbanas o rurales.

La relevancia de ciertos usuarios, generalmente líderes de opinión y su verdadera influencia sobre los otros electores es una línea que deberá investigarse, toda vez que la persuasión política se manifiesta a través de las comunidades de Twitter.

Publicaciones

A partir de la investigación realizada en el presente trabajo se realizaron algunas publicaciones, las cuales son las siguientes:

“The Weighting Factors to Improve Predictability on Twitter”

Jorge Arroba Rimassa, Rafaél Muñoz Guillena y Fernando Llopis. Febrero 27, 2018. <https://doi.org/10.4236/ti.2018.91005>. Technology and Investment, 2018, 9, páginas 68-79, <http://www.scirp.org/journal/ti>, ISSN Online: 2150-4067.

“Using the Twitter social network as a predictor in the political decision”

Jorge Arroba Rimassa, Rafaél Muñoz Guillena, Fernando Llopis y Yoan Gutierrez CICLing 2018, 19th International Conference on Computational Linguistics and Intelligent Text Processing, March 18 to 24, 2018, Hanoi, Vietnam.

“Relevance as an enhancer of votes on Twitter”

Jorge Arroba Rimassa, Fernando Llopis y Rafaél Muñoz Guillena. 2nd International Conference on Advanced Research Methods and Analytics (CARMA2018). Universitat Politècnica de Valencia, Valencia, Spain 2018. DOI: <http://dx.doi.org/10.4995/CARMA2018.2018.8311>.

“Harvesting Opinions in Twitter for Sentiment Analysis”

Juan Guevara, Joana Costa, Jorge Arroba y Catarina Silva. 2018 13th Iberian Conference on Information Systems and Technologies (CISTI) . 13-16 June 2018. DOI: 10.23919/CISTI.2018.8399226. Publisher: IEEE. Caceres, Spain.

“GPLSI at TREC-2019, using twitter to detect emotions”

Javi Fernández, Fernando Llopis, Yoan Gutiérrez, Jorge Arroba y Patricio Martínez-Barco: 28th TREC 2019: Gaithersburg, MD, USA Pendiente de publicación (Noviembre 2019).

"Using the geographic weighting factors can improve the predictions based on Twitter"

Jorge Arroba Rimassa, Rafael Muñoz Guillena, Fernando Llopis y Yoan Gutierrez. 7TH INTERNATIONAL SYMPOSIUM ON LANGUAGE & KNOWLEDGE ENGINEERING LKE 2019. Technological University Dublin, Tallaght Campus. Pendiente de publicación (Noviembre 2019).



Universitat d'Alacant
Universidad de Alicante

Referencias Bibliográficas

- Agulló, F., A. Guillén, Y. Gutiérrez, y P. Martínez-Barco. 2015. ElectionMap: una representación geolocalizada de intenciones de voto hacia partidos políticos sobre la base de comentarios de usuarios de Twitter. *Procesamiento del Lenguaje Natural*, Revista nº 55, páginas 195-198.
- Almatrafi, O., S. Parack, y B. Chavan. 2015. Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. 9th International Conference on Ubiquitous Information Management and Communication, páginas 41-47.
- Angiani, G., L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, y S. Manicardi. 2016. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. Conference: Kdweb 2016, At Cagliari, páginas 1-11.
- Arroba, J. 2000. ¿Cuándo y cómo se hace un sondeo flash?. *Revista Latinoamericana de Comunicación, Chasqui*. Num. 70, DOI: <https://doi.org/10.16921/chasqui.v0i70.1355>, páginas 43-45.
- Beauchamp, N. 2017. Predicting and Interpolating State-level Polls using Twitter Textual Data. *American Journal of Political Science*, 2017 - Wiley Online Library. <http://dx.doi.org/10.7910/DVN/RJAUNW>, páginas 1-36
- Bermingham, A., y Smeaton, A. 2011. On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Work-shop on Sentiment Analysis, Psychology*, páginas 2-10.
- Borondo, J., A. J. Morales, J. C. Losada, y R. M. Benito. 2013. Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential Election as a case study. arXiv: 1309.5014v1. physics.ph-soc, páginas 1-8.
- Boyd, D. M., y N. B. Ellison. 2008. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13 (2008), páginas 210–230.
- Ceron, A., L. Curini, y S. Iacus. 2015. Using social media to forecast electoral. *Statistica Applicata - Italian Journal of Applied Statistics* Vol. 25(3), páginas 239-261.
- Choy, M., M. Cheong, M. Nang Laik, y K. Ping Shung. 2012. US Presidential Election 2012 Prediction using Census Corrected Twitter Model. arXiv: 1211.0938. *Computers and Society*, páginas 1-12.
- Cotelo, J. M., F. Cruz, F. J. Ortega, y J. A. Troyano. 2015. Explorando Twitter mediante la integración de información estructurada y no estructurada. *Procesamiento del Lenguaje Natural*, Revista nº 55, septiembre de 2015, páginas 75-82.
- Cotelo, J. M., F. L. Cruz, y J. A. Troyano. 2012. Generación adaptativa de consultas para la recuperación temática de tweets. *Procesamiento de Lenguaje Natural*, Revista 48, páginas 57-64.

Deltell, L., F. Claes, y J. M. Osteso. 2013. Predicción de tendencia política por Twitter: Elecciones Andaluzas 2012. *Ámbitos*, núm. 22, enero-junio, 2013. Universidad de Sevilla, páginas 1-15.

DiGrazia, J., K. McKelvey, J. Bollen, y F. Rojas. 2013. F. More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. Published: November 27, 2013. <https://doi.org/10.1371/journal.pone.0079449>.

Dwi Prasetyo, N., y C. Hauff. 2015. Twitter-based election prediction in the developing world. 26th ACM Conference on Hypertext and Social Media, páginas 149–158.

Effrosynidis, D., S. Symeonidis, y A. Arampatzis. 2017. A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis. J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, DOI: 10.1007/978-3-319-67008-9 31, páginas 394–406.

Escortell Pérez, M. A., M. Giménez Fayos, y P. Rosso. 2017. El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter. *Procesamiento del Lenguaje Natural*, Revista nº 58, marzo de 2017, páginas 85-92.

Esuliy, A., y F. Sebastiani. 2007. SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining. Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche Pisa, Italy, Technical Report 2007-TR-02, páginas 1-26.

Feldman, R. 2013. Techniques and Applications for Sentiment Analysis. *Communications of the acm*. vol. 56, no. 4, doi:10.1145/2436256.2436274, páginas 82-89.

Fernández Crespo M. 2013. Predicción electoral mediante análisis de redes sociales. Memoria para optar al grado de Doctor de la Universidad Complutense de Madrid.

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y Á. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, Revista nº 58, páginas 141-148.

Fernández, J., J. M. Gómez, y P. Martínez-Barco. 2014. Análisis de sentimientos multilingüe en la Web 2.0. *Actas V Jornadas TIMM*, Cazalla de la Sierra, España, 12-JUN-2014, páginas 19-21

Fernández, J., Y. Gutiérrez, J. M. Gómez, y P. Martínez-Barco. 2015. Social Rankings: análisis visual de sentimientos en redes sociales. *Procesamiento del Lenguaje Natural*, Revista nº 55, páginas 199-202.

Fonseca, A., y J. Louca. 2011. Political Opinion Dynamics in Social Networks: the Portuguese 2010-11 Case Study. *researchgate.net*, páginas 1-31.

García, O.M., J. Lanchas Sampablo, y D. Prieto Ruiz. 2013. Characterising social media users by gender and place of residence. *Procesamiento del Lenguaje Natural*, Revista nº 51, páginas 57-64.

Gaurav, M., A. Srivastava, A. Kumar, y S. Miller. 2013. Leveraging candidate popularity on twitter to predict election outcome. 7th Workshop on Social Network Mining and Analysis ACM, páginas 7-15.

- Gayo-Avello, D. 2011. Don't turn social media into another "Literary Digest" poll. In *Communications of the ACM*, 54(10), páginas 121–128.
- Giglietto, F. 2012. If Likes Were Votes: An Empirical Study on the 2011 Italian Administrative Elections. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, páginas 1-6.
- Gutiérrez, Y., J. M. Gómez, F. Llopis, L. Canales, y A. Guillén. 2018. Plataforma inteligente para la recuperación, análisis y representación de la información generada por usuarios en Internet. *Procesamiento del Lenguaje Natural*, Revista nº 61, páginas 127-130.
- Haridakis, P., y G. Hanson. 2009. Social Interaction and Co-Viewing With YouTube: Blending Mass Communication Reception and Social Connection. *Journal of Broadcasting & Electronic Media*. Volume 53, 2009 - Issue 2, páginas 317-335.
- He, Y., y D. Zhou. 2011. Self-training from labeled features for sentiment analysis. *Information Processing and Management*. doi:10.1016/j.ipm.2010.11.003, páginas 606–616.
- Hemalatha, I., G. P. Saradhi Varma, y A. Govardhan. 2013. Sentiment Analysis Tool using Machine Learning Algorithms. *International Journal of Emerging Trends & Technology in Computer Science*. Volume 2, March – April 2013. ISSN 2278-6856, páginas 105-109.
- Hopkins, D., G. King, M. Knowles, y S. Melendez. 2012. ReadMe: Software for Automated Content Analysis. Available from <http://GKing.Harvard.Edu/readme> under the Creative Commons Attribution, páginas 1-9
- Jensen, M., y N. Anstead. 2013. Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process. *Wiley Online Library*. <https://doi.org/10.1002/1944-2866.POI329>, páginas 161-182.
- Jiménez Zafra, S. M., E. Martínez Cámara, y L.A. Ureña López. 2014. Desafíos del Análisis de Sentimientos. *Actas V Jornadas TIMM*, Cazalla de la Sierra, páginas 15-18.
- Jotheeswaran, J., y S. Koteeswaran. 2016. Feature Selection using Random Forest method for Sentiment Analysis. *Indian Journal of Science and Technology*, Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, páginas 1-7.
- Jungherr, A., P. Jürgens, y H. Schoen. 2011. Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment". *Social Science Computer Review*. <https://doi.org/10.1177/0894439311404119>.
- Katz, E., P. F. Lazarsfeld, y E. Roper. 2005. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. ISBN-10: 1412805074. ISBN-13: 978-1412805070. Editor: Routledge; Edición: 2 (30 de septiembre de 2005).

- Keshavarz, H., y M. Saniee Abadeh. 2017. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs, Knowledge-Based Systems, KNOSYS 3806, DOI: 10.1016/j.knosys.2017.01.028, páginas 1-48.
- Krippendorff, K. 1990. Metodología de análisis de contenido: teoría y práctica. Edit. PAIDOS comunicación. ISBN 84-7509-627-1.
- Kumar Singh, P., y M. Shahid Husain. 2014. Methodological Study of Opinion Mining and Sentiment Analysis. International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014, DOI: 10.5121/ijsc.2014.5102, páginas 11-21.
- Lamos, V. 2012. On voting intentions inference from Twitter content: a case study on UK 2010 General Election. arXiv preprint arXiv:1204.0423, páginas 1-11.
- Larsson, A. O., y H. Moe. 2012. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society, 14 (5), páginas 729-747.
- Mahmood, T., I. Tasmiyah, A. Farnaz, L. Wajeeta, y M. Atika. 2013. Mining Twitter Big Data to Predict 2013 Pakistan Election Winner. Conference Location: Lahore, Pakistan. Publisher: IEEE. DOI: 10.1109/INMIC.2013.6731323, páginas 49-54.
- Makazhanov, A., y D. Rafiei. 2013. Predicting political preference of Twitter users. International Conference on Advances in Social Networks Analysis and Mining, páginas 298-305.
- Martínez Cámara, E., M. T. Valdivia Martín, J. M. Perea Ortega, y L. A. Ureña López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. Sociedad Española para el Procesamiento del Lenguaje. Volumen 47, páginas 163-170.
- Metaxas, P.T., E. Mustafaraj, y D. Gayo-Avello. 2011. How (not) to predict elections". In Proceedings of PASSAT/SocialCom 2011, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE Computer Society, Los Alamitos, CA, USA, páginas 165-171.
- Mohammad, S., C. Dunne, y B. Dorr. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, ACL and AFNLP, páginas 599-608.
- Montemurro, M., 2001. Beyond the Zipf-Mandelbrot law in quantitative linguistics. Universidad Nacional de Córdoba, Argentina. www.elsevier.com/locate/physa, páginas 567-578.
- Montesinos García, L. 2014. Análisis de sentimientos y predicción de eventos en Twitter. Memoria para optar al título de Ingeniero Civil Eléctrico de la Universidad de Chile.
- Montoyo, A., P. Martínez-Barco, y A. Balahur. 2012. Subjectivity and sentiment analysis: An overview of the current state of the area and

envisaged developments. *Decision Support Systems*, doi:10.1016/j.dss.2012.05.022, páginas 675–679.

Nooralahzadeh, F., V. Arunachalam, y C. Chiru. 2013. Presidential elections on twitter—analysis of how the US and French election were reflected in tweets. *19th International Conference on Control Systems and Computer Science (CSCS)*, páginas 240–246.

Pang, B., y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. Vol. 2, No 1-2, páginas 1–135.

Pedrosa, I. 2015. Pruebas de bondad de ajuste en distribuciones simétricas. *Universitas Psychologica*, páginas 245-254.

Peregrino, F. S., D. Tomás, y F. Llopis. 2013. Every move you make i'll be watching you: geographical focus detection on Twitter. *7th Workshop on Geographic Information Retrieval, ACM*, páginas 1-8.

Pérez, J., A. Conejero, y C. Ferri. 2017. Zipf's and Benford's laws in Twitter hashtags Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, España, páginas 84–93.

Ramadhan, D., Y. Nurhadryani, y I. Hermadi. 2014. Campaign 2.0: Analysis of Social Media Utilization in 2014 Jakarta Legislative Election. In *ICACIS 2014*, páginas 102 – 107.

Sanders, E., y A. Van den Bosch. 2012. Relating Political Party Mentions on Twitter with Polls and Election Results, *Radboud University Nijmegen*, páginas 1-4.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *Consiglio Nazionale delle Ricerche, Pisa, Italy. ACM Computing Surveys*, Vol. 34, No. 1, páginas 1-47.

Shi, L., N. Agarwal, A. Agrawal, R. Garg, y J. Spolstra. 2012). Predicting US Primary Elections with Twitter. <http://snap.stanford.edu/social2012/papers/shi.pdf>.

Singh, P., y R. S. Sawhney. 2018. Influence of Twitter on Prediction of Election Results. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-10-6875-1_65, páginas 665-673.

Skoric, M., N. Poor, P. Achananuparp, E. Lim, y J. Jiang. 2012. Tweets and Votes: A Study of the 2011 Singapore General Election. *Hawaii International Conference on System Sciences*, páginas 1-10.

Srivastava, R., H. Kumar, M.P. Bhatia, y S. Jain. 2015. Analyzing Delhi assembly election 2015 using textual content of social network. *Sixth International Conference on Computer and Communication*, páginas 78-85.

Taboada, M., J. Brooke, M. Tofiloski, K. Vol, y M. Stede. 2010. Lexicon-Based Methods for Sentiment Analysis. *Association for Computational Linguistics, Computational Linguistics Volume 37, Number 2*, páginas 1-42.

Tjong Kim Sang, E., y J. Bos. 2012. Predicting the 2011 Dutch Senate Election Results with Twitter. *Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, Avignon, France*, páginas 1-8.

- Tripathy, A., A. Agrawal, y S. K. Rath. 2016. Classification of Sentiment Reviews using N-gram Machine Learning Approach. ESWA 10597. Expert Systems With Applications. DOI: 10.1016/j.eswa.2016.03.028, páginas 1-30
- Tsakalidis, A., S. Papadopoulos, A. Cristea, y Y. Kompatsiaris. 2015. , Predicting Elections for Multiple Countries Using Twitter and Polls. . In Predictive Analytics. IEEE INTELLIGENT SYSTEMS, páginas 10-17.
- Tumasjan, A., T. Sprenger, P. Sandner, y I. Welp. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, páginas 178-185.
- Tweetminster. 2010. Can Word-of-Mouth Predict the General Election Result? A Tweetminster Experiment in Predictive Modeling, <http://www.scribd.com/doc/29154537/Tweetminster-Predicts>. Last access: 05/10/2015.
- Ureña López, L. A., R. Muñoz Guillena, J. A. Troyano Jiménez, y M. T. Martín Valdivia. 2014. ATTOS: Análisis de Tendencias y Temáticas a través de Opiniones y Sentimientos. Procesamiento del Lenguaje Natural, Revista nº 53, páginas 151-154.
- Velikovich, L., S. Blair-Goldensohn, K. Hannan, y R. McDonald. 2010. The viability of web-derived polarity lexicons. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, páginas 777–785.
- Vilares, D., M. A. Alonso y C. Gómez-Rodríguez. 2013. Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico. Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, páginas 127-134.
- Vimalkumar, B. V., y B. M. Ahmedabad. 2016. Analysis of Various Sentiment Classification Techniques. International Journal of Computer Applications. Volume 140 – No.3, India, páginas 22-27.
- Washington, A.L., F. Parra, J.B. Thatcher, K. LePrevost, y D. Morar. 2013. What is the correlation between Twitter, polls and the popular vote in the 2012 presidential election? APSA 2013 Annual Meeting Paper. American Political Science Association.
- Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Hua Zhou, M. Steinbach, D. J. Hand, y D. Steinberg. 2007. Top 10 algorithms in data mining. Knowl Inf Syst. DOI 10.1007/s10115-007-0114-2, páginas 1-37.
- Wu, S., J. M. Hofman, W. A. Mason, y D. J. Watts. 2011. Who Says What to Whom on Twitter. International World Wide Web Conference Committee (IW3C2). ACM 978-1-4503-0637-9/11/03, páginas 705-714.
- Xu, K., G. Qi, J. Huang, T. Wu, y X. Fu. 2017. Detecting bursts in sentiment-aware topics from social media. Knowledge-Based Systems, <https://doi.org/10.1016/j.knosys.2017.11.007>, páginas 1-11.
- Yeon-Ok, L., y H. Woo Park. 2010. The Reconfiguration of E-Campaign Practices in Korea: A Case Study of the Presidential Primaries of 2007. Article

first published online. <https://doi.org/10.1177/0268580909346705>.
Volumen: 25 issue: 1, páginas 29-53.

Zarrella, D. 2010. The Social Media Marketing Book. Published by O'Reilly Media, Inc. CA 95472.

Zhang, C., J. Sun, X. Zhu, y Y. Fang. 2010. Privacy and Security for Online Social Networks: Challenges and Opportunities. IEEE Network, July/August 2010, páginas 13-18.

Zubair Asghar, M., A. Khan, S. Ahmad, y F. Masud Kundi. 2014. A Review of Feature Extraction in Sentiment Analysis. Journal of Basic and Applied Scientific Research. TextRoad Publication. ISSN 2090-4304, páginas 181-186.



Universitat d'Alacant
Universidad de Alicante