PREDICCIÓN ELECTORAL USANDO UN MODELO HÍBRIDO BASADO EN ANÁLISIS SENTIMENTAL: ELECCIONES PRESIDENCIALES DE COLOMBIA.

Mauro Callejas-Cuervo [1], Manuel A. Vélez-Guerrero [2]

[1]PhD en Energía y Control de Procesos. Docente Universidad Pedagógica y Tecnológica de Colombia. Grupo de Investigación en Software (GIS). Tunja – Colombia.
[2]Magister en Ingeniería. Investigador Universidad Pedagógica y Tecnológica de Colombia. Grupo de Investigación en Software (GIS). Tunja – Colombia.

Email: mauro.callejas@uptc.edu.co, manuel.velez@uptc.edu.co

RESUMEN

En Colombia, las redes sociales se han convertido en una poderosa herramienta para expresar opiniones políticas, especialmente durante el período de campaña en las elecciones presidenciales. Este trabajo propone un modelo híbrido para predecir el desenlace de la primera vuelta en las elecciones presidenciales de Colombia en 2018 cuyo objetivo es minimizar el error absoluto y mejorar la calidad de la predicción final. Para ello, las actividades de los usuarios en Twitter y Facebook fueron registradas y analizadas mediante algoritmos de inteligencia artificial, obteniendo como resultado una predicción precisa y coherente con la realidad. Como resultados principales se destaca que el RMSE del modelo híbrido propuesto ronda el 2,47%, superando en promedio el RMSE de las firmas encuestadoras tradicionales más prominentes del país. Adicionalmente también se predijo el valor del abstencionismo electoral con un error diferencial de 1,72% con respecto al valor real, demostrando la confiabilidad de la metodología propuesta.

Palabras clave: PLN; análisis sentimental; resultados electorales; inteligencia artificial; predicción.

Recibido: 19 de Septiembre de 2019. Aceptado: 04 de Diciembre de 2019.
Received: September 19, 2019. Accepted: December 04, 2019.

ELECTORAL PREDICTION USING A HYBRID MODEL BASED ON SENTIMENTAL ANALYSIS: PRESIDENTIAL ELECTIONS IN COLOMBIA

ABSTRACT

In Colombia, social networks have become a powerful tool to disseminate political opinions, especially during the campaign period in the national presidential elections. This paper proposes a hybrid model to predict the outcome of the first round of presidential elections in Colombia in 2018, which aims to minimize absolute error and improve the quality of the final prediction. User activities on Twitter and Facebook were recorded and analyzed with artificial intelligence algorithms, resulting in an accurate prediction consistent with reality. As a core result is highlighted that the RMSE of the hybrid model is around 2.47%, surpassing on average the RMSE of the country's most prominent traditional polling firms. Additionally, the value of electoral abstentionism was also predicted with a differential error of 1.72% in relation to the real value, demonstrating the reliability of the proposed methodology.

Keywords: *NLP; sentimental analysis; election results; artificial intelligence; prediction.*

Cómo citar este artículo: M. Callejas, M. Vélez. "Predicción electoral usando un modelo híbrido basado en análisis sentimental: elecciones presidenciales de Colombia." Revista Politécnica, vol. 15 no.30 pp.94-104, 2019. DOI:10.33571/rpolitec.v15n30a9

1. INTRODUCCIÓN

Los medios sociales, destacándose Twitter y Facebook, son cada vez más usados debido a su alto índice de penetración y popularidad, albergando información creciente de opiniones e interacciones, incluyendo una descripción de los comportamientos sociales. Debido al alto volumen de información, ha cobrado interés la predicción de fenómenos relacionados con los usuarios, entre los que destacan las predicciones de los eventos poblacionales. Desde que las publicaciones políticas han ganado importancia, la estimación de los resultados en las elecciones políticas ha sido uno de los objetivos más destacados, debido a la complejidad de la predicción de variables altamente fluctuantes como las involucradas en este proceso.

Alrededor del mundo y demostrado por varios estudios de caso, los investigadores han destacado el hecho de que las interacciones sociales reflejan fielmente los escenarios políticos futuros [1], [2]. Por ejemplo, en los primeros avances publicados en el tema [3] se ha demostrado una alta correlación entre el número de Tweets y los resultados electorales reales; y posteriormente se ha conseguido predecir con éxito elecciones políticas [4]. Aunque el desarrollo de este tema de estudio no es temprano, los avances recientes en técnicas específicas para el análisis de datos mediante inteligencia artificial han refinado notoriamente los procesos predictivos.

Las herramientas digitales de opinión son también una valiosa mina de información sobre el sentimiento colectivo [5], [6]. Estudios como [7] señalan que, mediante el uso de técnicas como el análisis de sentimientos, minería de datos y la lingüística computacional, es posible identificar y extraer información importante reflejada en hechos sociales palpables y no únicamente derivado del panorama político de una región.

Aunque la literatura actual ha demostrado que los datos de los medios sociales pueden representar en algunos casos la emoción del público y la intención de los votantes utilizando un modelo apropiado de extracción de datos [8], hay menor cantidad de estudios de caso en países en desarrollo que muestren una fusión de técnicas específicas para lograr el objetivo de predecir el resultado político en las elecciones presidenciales [9], [10].

Este trabajo, por tanto, pretende contribuir al desarrollo de modelos de predicción híbridos, que mezclan el Procesamiento del Lenguaje Natural (PLN), el Análisis de Sentimientos (AS) y el análisis de datos numéricos producto de encuestas tradicionales, mostrando el caso de aplicación directa en las elecciones presidenciales colombianas de 2018.

El contenido de este artículo está organizado del siguiente modo: en la sección II se presenta una revisión de la literatura que muestra las principales técnicas utilizadas en la predicción del entorno político; en la sección III se presentan los antecedentes regionales; la sección IV describe la metodología propuesta, seguido de la sección V que profundiza en el enfoque propuesto, la descripción del escenario de prueba y el desarrollo de las técnicas utilizadas en el modelo híbrido. Por último, la sección VI muestra las conclusiones de los resultados y la discusión de los hallazgos.

2. REVISIÓN DE LITERATURA

Se presenta la revisión de la literatura que resume de manera concisa las diferentes metodologías para predecir el comportamiento electoral. También se presenta una revisión de las técnicas que se han utilizado para el análisis de los sentimientos, la extracción de patrones y la fusión de información de diferentes fuentes.

Análisis de sentimientos y otras técnicas de predicción.

Uno de los primeros casos de estudio particulares en usar el Análisis de Sentimientos (AS) como herramienta directa de predicción estuvo a cargo de la Universidad de Essex [11], donde se realizo la predicción de los resultados electorales en Francia, contrastado con la metodología propuesta en [3]. Este es uno de los primeros estudios en resaltar que el volumen de tweets es menos significativo que el sentimiento o polaridad expresado en los mismos. El método propuesto también considera tweets neutros relacionados con candidatos específicos, lo que ha demostrado aumentar la precisión de las predicciones realizadas.

Otros trabajos que muestran las capacidades del AS para la predicción de resultados electorales se presentan en [12], [13], donde se toman como estudio de caso las elecciones generales de cuatro países. Los resultados sugieren que las predicciones basadas en Twitter y AS pueden producir resultados electorales precisos basado en la actividad digital de un país y la recepción pública de estos medios.

Dos estudios independientes realizados en India [14], [15], apuntan directamente al uso de técnicas de AS soportadas por Máquinas de Soporte Vectorial (SVM) y Naive Bayes para la predicción de resultados electorales en el país. Como un resultado en común, se destaca que la precisión de las técnicas de procesamiento empleadas depende tanto de la cantidad como de la calidad de los datos etiquetados usados para el entrenamiento del algoritmo. Estos desarrollos comparten algunas limitaciones similares, donde los datos marcados manualmente no fueron suficientes para proporcionar resultados más precisos, un aspecto relevante a la hora de definir la polaridad sentimental de un *tweet*. Los autores sugieren que el modelo desarrollado, aunque crucial para definir los resultados de las predicciones electorales, debería combinarse con otros tipos de información, abriendo así la posibilidad de utilizar algoritmos de análisis mixtos o híbridos.

Trabajos publicados en relación a algoritmos de aprendizaje de máquina, análisis de sentimientos y redes sociales fueron publicados en [16] y [17], donde se clasifican los datos de *Twitter* en polaridad de sentimientos utilizando diferentes clasificadores de aprendizaje supervisado. Se concluye que los resultados son buenos, aunque aún hay problemas no triviales como el sesgo de la muestra y la comprensión automática del contenido textual. Un análisis a fondo que explora este contexto es presentado en [18], en donde resalta que las características de las redes sociales brindan elementos importantes en las encuestas electorales, y que su exclusión puede inhibir nuestra comprensión del entorno y la toma de decisiones.

Tomando como estudio de caso las elecciones presidenciales de Estados Unidos en 2016, tres estudios independientes [19]-[21] muestran una caracterización completa de los usuarios de redes sociales como Twitter y sus preferencias políticas.

El objetivo de la predicción de los resultados definitivos es encontrar una correlación entre una polaridad sentimental y una tendencia en el comportamiento políticos. Se destaca que los datos obtenidos en la etapa de recolección de la información son tan importantes como el procesamiento en sí mismo. Se concluye que es importante obtener métricas adecuadas de comparación a la hora de probar los modelos de predicción, ya que la elección de métricas incompletas y erróneas llevan a una dificultad en la obtención de resultados cercanos a la realidad.

Para finalizar, trabajos desarrollados en estudios de caso geográficamente distanciados como España, Inglaterra y Estados Unidos [22]–[24] brindan resultados interesantes en sus propios contextos. Se establece que la localización geográfica juega un papel importante cuando se realizan las predicciones basadas en comportamientos sociales ya que los usuarios de internet hacen uso de las herramientas de una forma distinta. Se concluye que el uso de las técnicas elegidas para el procesamiento de la información en redes sociales es acertado y se encuentra validada en sinnúmero casos de estudio a lo largo del tiempo.

Algoritmos mixtos o híbridos para el mejoramiento de predicciones.

Los algoritmos híbridos buscan mejorar los resultados de las predicciones realizadas por sistemas convencionales, así como resolver problemas conocidos, algunos de ellos expuestos en [25]. Otros trabajos muestran la relación que hay entre hechos sociales aparentemente aislados y las elecciones políticas de determinada región. Tal es el caso de [26] que brinda la posibilidad de analizar otras fuentes de información para establecer conclusiones definitivas acerca de las predicciones electorales basadas en AS e inteligencia colectiva.

En el campo de la fusión de la información con técnicas como algoritmos de análisis mixtos, el trabajo de Wang [27] resulta fundamental al ser pionero en el incremento de la fiabilidad de los resultados electorales usando información de análisis de sentimientos con inteligencia colectiva. La comprobación del nuevo modelo se realizó por medio del análisis de las elecciones taiwanesas de 2014.

Otro estudio relevante en el área de la fusión de información para extraer datos predictivos es realizado en [28], quienes utilizaron técnicas de Big Data para predecir los resultados de las elecciones en referendos, donde los algoritmos de procesamiento híbrido o conjunto también han sido propuestos para predecir las intenciones de votos de los usuarios en Internet. En [29] se propone el clasificador conjunto resultado de combinar dos clasificadores de aprendizaje base. El rendimiento del clasificador propuesto se ha comparado con varios métodos tradicionales de análisis de sentimientos, como el SVM y el Naive Bayes, lo que ha dado como resultado que el clasificador conjunto se desempeñe mejor que los clasificadores independientes con la mayoría de las predicciones realizadas.

Para concluir, el equipo de Bansal [30], propone la fusión de técnicas de AS convencionales con la técnica denominada *Hybrid Topic Based Sentiment Analysis* (HTBSA), un modelo propuesto con el objetivo de capturar relaciones de palabras y coocurrencias en *tweets*. Se concluye que el sistema es capaz de realizar predicciones más precisas en comparación con el análisis de sentimiento explícito basado en frases, sin la necesidad de un pre-entrenamiento ni anotaciones humanas.

Predicción de intenciones políticas en Latinoamérica.

Uno de los trabajos más destacados del contexto latinoamericano fue desarrollado por Cerón-Guzmán [31], cuyo estudio de caso son las elecciones presidenciales de Colombia en 2014. Este documento es relevante ya que muestra el camino a seguir para predecir las intenciones de voto en los países en desarrollo. Sin embargo, concluyen que "los resultados experimentales muestran que los métodos de inferencia basados en datos de Twitter no son consistentes", indicando un punto de partida para futuras investigaciones.

En el caso de [32] el método propuesto está destinado a predecir las elecciones parlamentarias, donde se resalta una precisión del 98,72% en la categorización de *tweets*. Sin embargo, el diccionario político construido únicamente clasifica los resultados en dos conjuntos de alineación (oposición y gobierno), con lo que se distorsiona el resultado de la información sin predecir directamente los porcentajes de votos.

Algunas técnicas interesantes basadas en análisis de sentimientos son presentadas en [33]., donde no se realizó un pronóstico sobre los porcentajes electorales, pero sí sobre la popularidad del ex presidente ecuatoriano Rafael Correa. Se concluye que, aunque los sentimientos hacia el máximo líder político ecuatoriano no influyeron en los resultados electorales de su partido, los medios sociales como Twitter son igualmente determinantes en la imagen que una personalidad refleja.

Por último, Rodríguez y su equipo presenta en [34] un método para predecir los resultados de las elecciones presidenciales en Chile, donde compara diferentes algoritmos para clasificar y extraer características lingüísticas. Destaca especialmente la precisión del método propuesto, en el que los resultados se predijeron con un MAE del 0,51%.

Conclusiones de la revisión.

Gran parte del trabajo muestra enfoques relacionados con el AS en las redes sociales, donde predomina *Twitter* por la facilidad de acceso al contenido de los usuarios. Este enfoque es validado por investigadores de todo el mundo, que han hecho predicciones fiables basadas en la inteligencia colectiva, pero deja parcialmente inexplorado el protagonismo de otras redes como *Facebook*.

En el caso de las metodologías para la inclusión de datos de otras fuentes, se logra un consenso común mediante el logro de mejoras significativas en los resultados finales de las predicciones. Se destaca la contribución de estos trabajos al estado del arte, ya que perfeccionan las metodologías desarrolladas por otros investigadores desde el auge de las técnicas de inteligencia computacional.

3. ANTECEDENTES

El sistema de elecciones presidenciales en Colombia permite elección directa del presidente cada cuatro años [35], sin posibilidad de reelección desde el año 2019. A partir de 1991, el presidente de Colombia es elegido a través de un sistema de dos vueltas: si ningún candidato recibe la mayoría de los votos en la primera vuelta (es decir, el 51% del total de los votos), se celebra una segunda vuelta entre los dos primeros candidatos con el mayor número de votos [36].

Para el presente caso, la investigación tiene como ventana de estudio únicamente la primera vuelta electoral de la contienda presidencial, cuyo proceso de campaña se inició oficialmente el 27 de febrero y culminó con la celebración electoral el 27 de mayo de 2018.

Según el estudio de la Misión de Observación Electoral en Colombia (MOE) [37], las elecciones presidenciales de 2018 contó con características que propiciaron un mejor ambiente para la investigación. Se presentaron opiniones radicalmente diferentes, opuestos y en competencia entre los candidatos a la presidencia nacional, lo que permitió que el debate público fuera más intenso.

Finalmente, se destaca que fueron seis los candidatos presidenciales que compitieron en la primera vuelta. Algunos de los candidatos inscritos inicialmente retiraron posteriormente su candidatura, los cuales no fueron tenidos en cuenta dentro del desarrollo de este estudio.

4. METODOLOGÍA

De manera general, se establece que la investigación se divide en tres etapas fundamentales: I) Monitoreo de la actividad en torno a las campañas presidenciales en las redes sociales, incluyendo opiniones y polaridad sentimental hacia los candidatos o sus propuestas. II) Seguimiento y recopilación de datos numéricos de encuestas, que son realizadas por empresas privadas para los medios de comunicación en Colombia y difundidas al público durante el período de la campaña electoral. III) Fusión de la información y obtención de los resultados finales, mostrados con mayor detalle en la Figura 1.

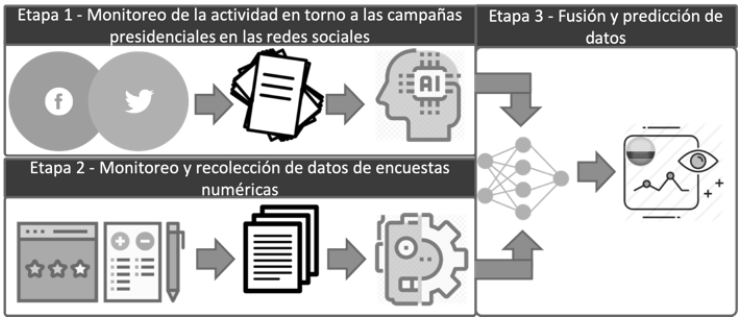

Fig. 1. Metodología propuesta para el desarrollo de la investigación.

Cabe señalar que las etapas I y II se llevan a cabo en paralelo, ya que en el transcurso de la campaña electoral los datos de los usuarios se recogen tanto en las redes sociales como en los medios de comunicación tradicionales, a través de encuestas realizadas y divulgadas públicamente. El seguimiento y obtención de resultados finales son realizados *pre-hoc*, es decir antes de la celebración de las elecciones presidenciales en primera vuelta.

5. DESARROLLO

Esta sección presenta los resultados obtenidos por cada etapa desarrollada, el desarrollo del modelo híbrido y los resultados finales obtenidos en la culminación de todo el proceso.

Conjunto de datos: redes sociales y sondeos de opinión.

La construcción de los conjuntos de datos es un paso fundamental para el desarrollo de esta investigación. La base de datos consta de tres conjuntos de datos independientes, cuya información se extrajo de las redes sociales, específicamente *Twitter* ("TWCo18") y *Facebook* ("FBCo18"), añadiendo datos de encuestas y sondeos de opinión ("OPCo18"). Estos datos fueron recolectados en un período de 120 días, antes y durante el período de la campaña electoral, desde enero 15, 2018 hasta mayo 15, 2018. Adicionalmente se incluyó la información del histórico sobre el abstencionismo electoral en Colombia. Esta última información constituye el dataset ("EACo18"), y es empleado para la predicción del abstencionismo electoral.

En cada red social se recogieron datos de usuarios anónimos y se rastrearon las páginas de opinión más destacadas del país, así como las cuentas de los candidatos oficiales, personalidades y políticos prominentes. Se han tenido en cuenta las cuentas de algunos medios de comunicación nacionales y de opinión ciudadana en relación con la mención de uno de los *hashtags* o etiquetas propietarias de la campaña de un candidato, ya que esto puede afectar a la polaridad sentimental de otros usuarios.

Durante el período de campaña electoral, más de veinte encuestas y sondeos de opinión pública fueron publicados, constituyendo un cuerpo de información prominente que puede alimentar confiablemente el modelo híbrido de este desarrollo. Se hizo seguimiento a las principales encuestas publicadas en los medios de comunicación más importantes de Colombia. Como es mostrado en la Tabla 1, un total de 14 encuestas fueron analizadas.

Tabla 1. Información de encuestas y sondeos de opinión que componen el dataset OPCo18.

Encuestador	Fecha	Muestras
Datexco	Ene. 20,	1220
Guarumo	Feb. 2,	2187
YanHass	Feb. 6,	1251
Invamer	Feb. 10,	1200
Cifras y Conceptos	Feb. 12,	2813
Centro Nacional de Consultoría	Feb. 18, 2018	1187
Centro Nacional de Consultoría	Feb. 22, 2018	1175
Centro Estratégico Latinoamericano de Geopolítica	Feb. 28, 2018	1200
Cifras y Conceptos	Mar. 1,	2960
Guarumo	Mar. 4,	3425
Centro Nacional de	Mar. 8,	1192
Invamer	Mar. 27,	1200
Datexco	Abr.14,	1049
Invamer	Abr. 27,	1200

Un resumen consolidado de estas encuestas puede consultarse en [38]. Para este caso en particular, el conjunto de datos OPCo18 se realizó y consolidó manualmente, ya que los datos de estas encuestas se publican en medios tradicionales como prensa, radio y televisión.

Por otro lado, el abstencionismo electoral en Colombia es a menudo alto. Como lo registra [39], las causas de este fenómeno son variadas, influyendo desde el pensamiento político de la ciudadanía, la escasa polarización de los partidos políticos tradicionales, entre otros. Por esta razón, es de gran importancia predecir esta variable determinante en los resultados de las elecciones presidenciales, ya que proporciona información adicional sobre el comportamiento político en el país.

Se registraron manualmente los datos de abstencionismo electoral en Colombia teniendo como fuente de información principal el centro de datos históricos de la Registraduría Nacional de Colombia [40], de la cual se extraen 17 datos históricos comprendidos desde 1978 hasta 2016.

Etapa I: análisis de datos de redes sociales.

Para el primer proceso, se utiliza a nivel general técnicas de Procesamiento del Lenguaje Natural (PLN) y Análisis Sentimental (AS) para determinar el nivel de favorabilidad e imagen de los candidatos. Los datos recolectados en los conjuntos de datos han sido clasificados según la polaridad del sentimiento, pudiendo ser Positivos, Negativos o Neutrales, según la metodología propuesta por [34].

En este desarrollo se utilizó específicamente un modelo de LSVM (*Linear Support Vector Machine*) como algoritmo de clasificación. En primer lugar, los datasets TWCo18 y FBCo18 son separados en tres grupos: conjunto de entrenamiento (30%), conjunto de validación (10%) y conjunto de predicción (60%).

Como se detalla en la metodología de [34], el clasificador es entrenado para todos los candidatos que participaron en la primera vuelta electoral. Los resultados de esta etapa se presentan en la Fig. 2.

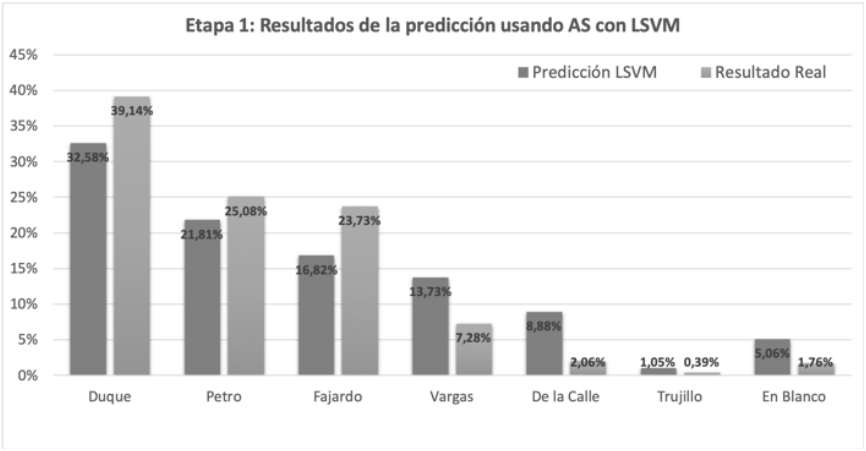

Fig. 2. Resultados de la predicción porcentual utilizando el clasificador LSVM en los datasets TWCo18 y FBCo18 versus los resultados reales obtenidos en la contienda.

Para medir la cantidad de error en la predicción, se hace uso del Root Mean Squared Error (RMSE), el cual brinda un mayor nivel de penalización que el MSE a medida que el error crece. EL RMSE se calcula como se muestra en la Ec. 1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

De los resultados observados en esta etapa del análisis, se puede intuir que los candidatos menos polarizantes tuvieron mejor aceptación entre los usuarios de Internet, lo que aumentó su popularidad.

Además, debido a un número relativamente alto de *tweets* clasificados como Neutros, o aquellos que contienen referencias cruzadas entre candidatos, el número de votos en blanco también fue particularmente alto.

A pesar de que los resultados son cercanos, el RMSE sigue siendo relativamente alto en comparación a otros trabajos destacados en las referencias documentales.

Etapa II: análisis de los datos de las encuestas y sondeos de opinión.

Para el segundo proceso se utiliza Facebook Prophet [41], el cual puede producir predicciones numéricas precisas mediante la inferencia de datos de entrada. En esta etapa, los datos presentes en OPCo18 son analizados como las observaciones en series de tiempo del modelo propuesto. Los resultados de esta etapa se presentan en la Fig. 3.

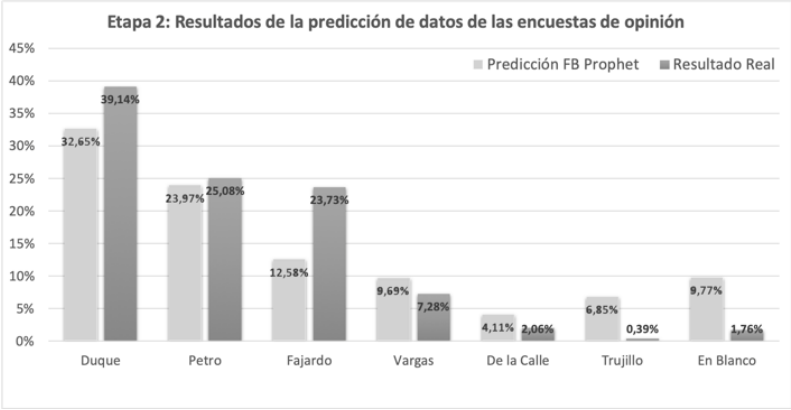

Fig. 3. Resultados de la predicción porcentual utilizando Facebook Prophet en el datasets OPCo18 versus los resultados reales obtenidos en la contienda.

Teniendo en cuenta los resultados derivados de las predicciones en la Etapa 1 y en la Etapa 2, es evidente una correlación. Esto puede deberse a que el nivel de favoritismo de un candidato o de sus propuestas está vinculado al sentimiento popular causado, que se transfiere a otros medios no digitales como la prensa, la radio y la televisión.

Etapa III: modelo híbrido, fusión de información y resultados finales.

Finalmente, la predicción de los resultados electorales se realiza en la fusión de datos, para lo cual se utiliza un modelo de aprendizaje basado en celdas *Long Short-Term Memory* (LSTMs).
Usando como función de pérdida el RMSE y como optimizador el descenso estocástico por gradiente, se busca encontrar un óptimo local adecuado después de un número suficiente de iteraciones. En concreto, el modelo recurrente utilizado es un modelo secuencial de una capa, donde hay 10 nodos LSTM con una forma de entrada (2,7), que corresponde a dos entradas dadas a la red (Predicción de la Etapa 1 y Predicción de la Etapa 2) con siete valores cada una (6 Candidatos presidenciales y voto en blanco).

Los resultados de las predicciones realizadas independientemente en la Etapa 1 y Etapa 2 del desarrollo de este trabajo se tratan como un nuevo conjunto de datos, el cual es nuevamente sub-dividido en conjunto de entrenamiento y conjunto de prueba.

El modelo de esta etapa se considera de carácter híbrido, ya que integran los datos de predicción realizados en las dos etapas anteriores. La Fig. 4 muestra los resultados obtenidos usando LSTMs, y la Tabla 2 muestra la medición del error de las predicciones contra los resultados de la contienda electoral.

Tabla 2. Medición del error de las predicciones por etapa.

Etapa	1: AS con LSVM	2: Pronóstico con FB Prophet	3: Modelo hibrido con LSTMs
RMSE	5,36%	6,37%	2,47%

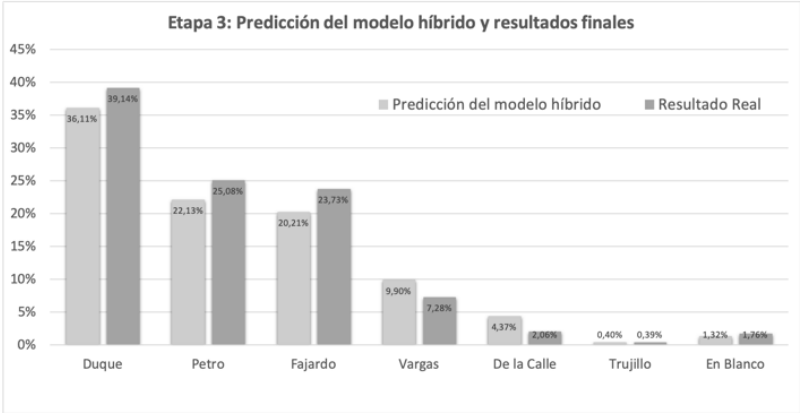

Fig. 4. Resultados de la predicción porcentual utilizando LSTMs fusionando los resultados de la Etapa 1 y 2

.

Como se ha evidenciado, el modelo que utiliza la predicción basada en LSTMs y el dataset los resultados de las predicciones independientes de la Etapa 1 y Etapa 2, proporciona resultados finales más cercanos al resultado real de la primera vuelta de las elecciones presidenciales en Colombia 2018. De manera comparativa, los resultados de las etapas anteriores y el resultado final se muestran en la Fig. 5.

Fig. 5. Comparación de resultados por etapas

.

Al tomar los resultados de la Etapa 3 y ubicarlos en términos de otras predicciones hechas por algunas de las firmas encuestadoras más prominentes del país, es evidente que el modelo híbrido propuesto proporciona los resultados más cercanos a la realidad. En la Tabla 3 se hace un resumen que muestra los resultados de las predicciones de las firmas encuestadoras mencionadas anteriormente.

Candidato	Resultados Reales	Nuestra Predicción		as y Conceptos (Predicc		Invamer (Predicción)		Guarumo (Predicción)	
		Predicción	Error	Predicción	Error	Predicción	Error	Predicción	Error
I. Duque	39,14%	36,11%	3,04%	34,90%	4,24%	41,30%	2,16%	23,60%	15,54%
G. Petro	25,08%	22,13%	2,96%	18,35%	6,73%	31,00%	5,92%	23,10%	1,98%
S. Fajardo	23,73%	20,21%	3,52%	14,05%	9,68%	13,30%	10,43%	8,10%	15,63%
G. Vargas	7,28%	9,90%	2,62%	21,90%	14,62%	7,90%	0,62%	6,30%	0,98%
H. de la Calle	2,06%	4,37%	2,31%	4,20%	2,14%	2,50%	0,44%	4,10%	2,04%
J. Trujillo	0,39%	0,40%	0,01%	0,00%	0,39%	2,10%	1,71%	1,50%	1,11%
Voto en Blanco	1,76%	1,32%	0,44%	5,40%	3,64%	1,90%	0,14%	13,20%	11,44%
		RMSE:	2,47%	RMSE:	7,45%	RMSE:	4,66%	RMSE:	9,46%

Fig. 6. Comparación de nuestra propuesta versus las predicciones de firmas encuestadoras en Colombia

.

Si se compara el desempeño general de nuestra propuesta con el de las empresas de encuestas, la metodología propuesta en esta investigación proporciona mejores resultados con al menos un 2,19% de confiabilidad con respecto a los sistemas de encuestas tradicionales. Se resalta que los todos los resultados presentados en la Tabla 3 fueron obtenidos con anterioridad (*pre hoc*) al desarrollo de la jornada electoral.

Abstencionismo electoral.

El modelo propuesto para la Etapa 3, basado en nodos LSTMs también se utiliza para predecir el abstencionismo electoral en Colombia. Se presenta en la Tabla 3 un resumen acerca de la predicción de esta variable.

Tabla 3. Resumen de resultados para el Abstencionismo Electoral.

Item	Abstencionismo electoral (Previsto)	Abstencionismo electoral (Resultado)	Error absoluto
Valor	44,92%	46,64%	1,72%

El modelo basado en LSTMs es óptimo para la predicción del abstencionismo, donde el margen de error con respecto a los resultados reales evidenciados es mínimo. Se toma este resultado

como un punto agregado y diferenciador al análisis anteriormente propuesto, ya que complementa la información predicha de manera fundamental.

6. DISCUSIÓN Y CONCLUSIÓN

El desarrollo de esta investigación encontró que las predicciones basadas en los sentimientos de la gente expresados a través de plataformas digitales pueden producir resultados precisos para predecir los resultados de las elecciones.
El principal hallazgo sugiere que el nivel de fiabilidad aumenta si la información de las interacciones sociales se mezcla con datos de encuestas numéricas (televisión, radio y prensa), según lo evidenciado en la Etapa 3, el cual obtuvo el valor de RMSE más bajo en comparación a las anteriores etapas. (Error Cuadrático Medio RMSE inferior al 2,5%).

Considerando que la tarea de predecir los resultados es difícil por varias razones, incluyendo la variabilidad del clima político en el tiempo y las formas de campaña, el resultado de la predicción es exitoso. Al contrastar los resultados obtenidos con nuestra metodología versus el margen de error de algunas de las firmas encuestadoras en la tarea de predicción de los resultados finales, se evidencia una clara ventaja al usar el modelo hibrido (2,19% menos error que Invamer, 4,98% menos error que Cifras y Conceptos y 6,99% menos error que Guarumo).

Con relación a otros estudios como los presentados por Cerón-Guzmán [31] y Rodríguez [34], la confiabilidad de las predicciones realizadas por este estudio de caso es inferior, esto debido a que no se incluyó dentro de la metodología un sistema de eliminación de ruido, lo cual introduce errores significativos a la hora de realizar las predicciones finales. Sin embargo, los resultados siguen siendo cercanos a la realidad política desarrollada en el país. Con los resultados obtenidos se concluye que la elección correcta de los algoritmos inteligentes funciona de manera apropiada cuando se trata de datos de la vida real y fuera de un contexto académico.

7. REFERENCIAS

[1] K.-B. Shibu Kumar, V.-S. Devi, K.-K. Rajeev, y A. Bhatia. Probabilistic algorithms for election result prediction. Int. Conf. Soft Comput. Mach. Intell. ISCMI 2014, pp. 79–82, 2014.

[2] C. Ganser y P. Riordan. Vote expectations at the next level. Trying to predict vote shares in the 2013 German federal election by polling expectations. Elect. Stud., vol. 40, pp. 115–126, 2015.

[3] A. Tumasjan, T.-O. Sprenger, P.-G. Sandner, y I.-M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media Predict., vol. 30, no. 2, pp. 178–185, 2010.

[4] E. Sang y J. Bos. Predicting the 2011 Dutch Senate Election Results with Twitter. Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguist., no. 53, pp. 53–60, 2012.

[5] J. Lee y Y. Choi. Expanding affective intelligence theory through social viewing: Focusing on the South Korea's 2017 presidential election. Comput. Human Behav., vol. 83, pp. 119–128, 2018.

[6] M. Korakakis, E. Spyrou, y P. Mylonas. A survey on political event analysis in Twitter. Proc. - 12th Int. Work. Semant. Soc. Media Adapt. Pers. SMAP 2017, pp. 14–19, 2017.

[7] Y. Arslan, A. Birturk, B. Djumabaev, y D. Küçük. Real-time Lexicon-based sentiment analysis experiments on Twitter with a mild (more information, less data) approach. IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018, pp. 1892–1897, 2018.

[8] B. Heredia, J. Prusa, y T. Khoshgoftaar. Exploring the Effectiveness of Twitter at Polling the United States 2016 Presidential Election. IEEE 3rd Int. Conf. Collab. Internet Comput. CIC 2017, vol. 2017, pp. 283–290, 2017.

[9] A.-M. Koli, M. Ahmed, y J. Manhas. An Empirical Study on Potential and Risks of Twitter Data for Predicting Election Outcomes. in Emerging Trends in Expert Applications and Security, vol. 841, no. January, Springer Singapore, 2019, pp. 725–731.

[10] Y. Zhao y E. Santos. A Failure of Collective Intelligence. IEEE/WIC/ACM Int. Conf. Web Intell., pp. 361–366, 2018.

[11] L. Wang y J. Q. Gan. Prediction of the 2017 French election based on Twitter data analysis. 9th Comput. Sci. Electron. Eng. Conf. CEEC, pp. 89–93, 2017.

[12] A. Saifuddin, J. Kokil, y M. S. Marko. Tweets & Votes - A 4 Country Comparison of Volumetric and Sentiment Analysis Approcahes. Proc. 10th Int. Conf. Web Soc. Media, no. ICWSM, pp. 507–510, 2016.

[13] M. Ramzan, S. Mehta, y E. Annapoorna. Are tweets the real estimators of election results?. 10th Int. Conf. Contemp. Comput. IC3, pp. 1–4, 2018.

[14] P. Sharma y T. S. Moh. Prediction of Indian election using sentiment analysis on Hindi Twitter. IEEE Int. Conf. Big Data, pp. 1966–1971, 2016.

[15] J. Ramteke, S. Shah, D. Godhia, y A. Shaikh. Election result prediction using Twitter sentiment analysis. Proc. Int. Conf. Inven. Comput. Technol. ICICT, vol. 1, 2017.

[16] P. Juneja. Casting Online Votes: To Predict Offline Results Using Sentiment Analysis by machine learning Classifiers. 8th ICCCNT, 2017.

[17] M. Coletto, C. Lucchese, S. Orlando, y R. Perego. Electoral Predictions with Twitter: a Machine-Learning approach Introduction and Related Work. Proc. 6th Ital. Inf. Retr. Work, 2017.

[18] D. Leiter, A. Murr, E. Rascón Ramírez, y M. Stegmaier. Social networks and citizen election forecasting: The more friends the better. Int. J. Forecast., vol. 34, no. 2, pp. 235–248, 2018.

[19] J.A. Caetano, J. Almeida, y H.T. Marques-Neto. Characterizing politically engaged users' behavior during the 2016 us presidential campaign. IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, pp. 523–530, 2018.

[20] B. Kostadinov. Predicting the Next US President by Simulating the Electoral College. J. Humanist. Math., vol. 8, no. 1, pp. 64–93, 2018.

[21] A. Hernandez-Suarez, *et al*. Predicting political mood tendencies based on Twitter data. 5th Int. Work. Biometrics Forensics, IWBF, pp. 1–6, 2017.

[22] M.S. Lewis-Beck y C. Tien. Candidates and campaigns: How they alter election forecasts. Elect. Stud., vol. 54, pp. 303–308, 2018.

[23] S. Martin-Gutierrez, J.-C. Losada, y R.-M. Benito. Recurrent Patterns of User Behavior in Different Electoral Campaigns: A Twitter Analysis of the Spanish General Elections of 2015 and 2016. Complexity, pp. 1–15, 2018.

[24] R. Johnston, T. Hartman, y C. Pattie. Predicting general election outcomes: campaigns and changing voter knowledge at the 2017 general election in England. Quality and Quantity, Springer Netherlands, pp. 1–21, 2018.

[25] D. Hussein. A survey on sentiment analysis challenges. J. King Saud Univ. Eng. Sci., vol. 30, no. 4, pp. 330–338, 2018.

[26] U. Khan y R.-P. Lieli. Information flow between prediction markets, polls and media: Evidence from the 2008 presidential primaries. Int. J. Forecast., vol. 34, no. 4, pp. 696–710, 2018.

[27] M.-H. Wang y C.-L. Lei. Boosting election prediction accuracy by crowd wisdom on social forums. 2016 13th IEEE Annu. Consum. Commun. Netw. Conf. CCNC, pp. 348–353, 2016.

[28] A. Mavragani y K.-P. Tsagarakis. Predicting referendum results in the Big Data Era. J. Big Data, vol. 6, no. 1, p. 3, 2019.

[29] M. Ankit y N. Saleena. An Ensemble Classification System for Twitter Sentiment Analysis. Procedia Comput. Sci., vol. 132, pp. 937–946, 2018.

[30] B. Bansal y S. Srivastava. On predicting elections with hybrid topic-based sentiment analysis of tweets. 3rd Int. Conf. Comput. Sci. Comput. Intell., vol. 135, no. 1, pp. 346–353, 2018.

[31] J.-A. Cerón-Guzmán y E. León-Guzmán. A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election. IEEE Int. Conf. Big Data Cloud Comput. BDCloud, pp. 250–257, 2016.

[32] R. Castro y C. Vaca. National leaders' twitter speech to infer political leaning and election results

in 2015 Venezuelan parliamentary elections. IEEE Int. Conf. Data Min. Work, pp. 866–871, 2017.

[33] O. Hidalgo, R. Jaimes, E. Gomez, y S. Lujan-Mora. Sentiment analysis applied to the popularity level of the ecuadorian political leader Rafael Correa. Int. Conf. Inf. Syst. Comput. Sci, pp. 340–346, 2018.

[34] S. Rodríguez *et al*. Forecasting the Chilean electoral year: Using twitter to predict the presidential elections of 2017. Lecture Notes in Computer Science LNCS, vol. 10914, pp. 298–314, 2018.

[35] G. Roland, *et al*. Colombia's electoral and party system: Proposals for reforms, 2000.

[36] D.-M. Hanratty, S.-W. Meditz, y R.-A. Hudson, Colombia: a country study, vol. 1, no. 1. 2010.

[37] Misión de Observación Electoral MOE. Political Context of the 2018 presidential election in Colombia. 2018.

[38] H.-K. Sonneland y Americas Society Council of the Americas. Poll Tracker: Colombia's 2018 Presidential Election. 2018. Disponible en: https://www.as-coa.org/articles/poll-tracker-colombias-2018-presidential-election. [Consultado el: 25-Feb-2019].

[39] A.-P. Torres-Espinosa y J. Ferri-Durá, Abstención electoral en Colombia. Desafección política, violencia política y conflicto armado, 1st ed. Bogotá: Universidad Complutense de Madrid, 2013.

[40] Registraduría Nacional del Estado Civil de Colombia. Resultados Elecciones Presidenciales 2018 Primera Vuelta. 2018. Disponible en: https://www.colombia.com/elecciones/2018/resultados/presidente.aspx?C=P1. [Consultado el: 27-Feb-2019].

[41] S.-J. Taylor and B. Letham. Forecasting at Scale. PeerJ Prepr., pp. 1–25, 2017.