ISSN 1870-4069

Análisis de sentimientos para el estudio del comportamiento en noticias de COVID-19 en redes sociales como series de tiempo

Jesús Hernádez Padilla, Hiram Calvo, Ildar Batyrshin

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Laboratorio de Ciencias Cognitivas Computacionales,
México

jhernandezp2019@alumno.ipn.mx, hcalvo@ipn.mx,
batyr1@cic.ipn.mx

Resumen. En la última década, los usuarios en redes sociales se han incrementado exponencialmente. Las redes sociales cada vez están más inmersas en la vida humana y en consecuencia cada vez tienen mayor influencia en la información de los usuarios: qué les gusta, qué les disgusta, que los motiva, etc. Twitter en la actualidad es una excelente fuente para generar un corpus robusto de las emociones de las personas y en consecuencia es una fuente muy importante para experimentar y comprobar si la información textual fortalece los modelos de series de tiempo en medida que una figura pública influye en las aplicaciones de redes sociales en mayor o menor medida. Un ejemplo reciente es el comportamiento del bitcoin a partir del interés que mostró Elon Musk como una divisa para adquirir productos tesla, otro que es el se analizara en este ejercicio es el comportamiento de las personas frente al primer bimestre cuando la Organización Mundial de la Salud declaró el inicio de pandemia el 11 de marzo de 2020. Se ha encontrado que si se representan las emociones o sentimientos reflejados en Twitter. Se puede analizar una tendencia en los sentimientos textuales en redes sociales y grupos o clusters de tipos de usuarios agrupados por emociones o sentimientos. En este trabajo se analizaron *tweets* recopilados durante el inicio de la pandemia, en el bimestre marzo y abril de año 2020. Se evaluaron los modelos de clasificación de regresión logística, *random forest*, *modelos bayesianos* y máquinas de soporte vectorial. Adicionalmente se buscó entender y observar la tendencia del comportamiento de la fluctuación de los tweets dentro de la red a partir de una serie temporal.

Palabras clave: Análisis de sentimientos, COVID-19, twitter, series de tiempo, redes sociales, análisis de comportamiento, agrupamiento por emociones.

Jesús Hernández Padilla, Hiram Calvo, Ildar Batyrshin

Sentiment Analysis for the Study of Behavior in COVID-19 News on Social Media as Time Series

Abstract. In the last decade, the number of social media users has grown exponentially. These platforms are increasingly embedded in daily human life and, as a result, have a significant influence on users' information consumption — what they like, dislike, what motivates them, and more. Today, Twitter is an excellent source for generating a robust corpus of human emotions, making it a valuable resource for experimenting with and validating whether textual information can enhance time series models — especially when public figures influence how content spreads across social networks. A recent example is the behavior of Bitcoin following Elon Musk's expressed interest in accepting it as payment for Tesla products. Another example, analyzed in this work, is the public's reaction during the first two months after the World Health Organization declared the COVID-19 pandemic on March 11, 2020. This study demonstrates that emotions or sentiments expressed on Twitter can be modeled to analyze trends over time and cluster user types by sentiment. Tweets collected during March and April 2020 were analyzed. Sentiment classification models — including logistic regression, random forest, Bayesian models, and support vector machines — were evaluated. Additionally, temporal analysis was conducted to understand the behavior and fluctuation of tweets over time.

Keywords: COVID-19, twitter, time series, social media, behavior analysis, emotion clustering.

1. Introducción

Algunos trabajos de los investigadores sobre el comportamiento de las finanzas, han arrojado que el mercado puede ser influenciado por las reacciones emocionales de los agentes económicos encargados elementalmente del consumo.

Esto brinda una oportunidad estratégica a los agentes destinados exclusivamente a la producción de bienes y servicios. Pues de algunos años a nuestra actualidad, en algunas investigaciones se ha encontrado que las aplicaciones de las redes sociales son bastante rentables para predecir las acciones o la demanda de mercado, específico o general como necesidades médicas, insumos de primera necesidad . Esto dependiendo de la temporalidad, estacionalidad y tendencia del comportamiento de los demandantes de servicios y de los bienes. A partir de la percepción de las personas (*inlfluencers*) y productos inmersos en el mercado.

Por otro lado, no solo un producto nuevo o innovador, o una inclinación preferencial de alguna figura influyente en los grupos sociales genera una tendencia en las aplicaciones de redes sociales y en el mundo real. Pues también los problemas políticos y de salud crean nuevas tendencias sociales solo que

con mayor correlación en el sector farmacéutico y de seguridad social. Es por ello que la situación por la que atravesamos frente a la pandemia declarada por la Organización Mundial de la Salud el 11 de marzo de 2020 a causa del virus SARS-Cov-2, es que en este trabajo se busca analizar algunos tipos de *tweets* recopilados durante tiempos de pandemia. Específicamente de marzo y abril, 2020.

La humanidad siempre se ha preocupado por intentar predecir el futuro. Predecir el clima ha sido muy importante una vez nos volvimos sedentarios, ya que nos ayuda a prepararnos para las cosechas en temporadas de sequía o en tiempos de abundante lluvia. La incertidumbre en la vida humana siempre ha sido un tema muy importante, tal vez tenga que ver con que somos seres con miedo y la mayoría del tiempo queremos saber que nos espera en el futuro.

Tenemos miedo de perder una economía estable, del comportamiento de fenómenos meteorológicos y/o sus consecuencias como la sequía extrema, el calentamiento global, ciclones tropicales, de la propagación de un virus nuevo y mortal como el SARS-CoV-2, etc. es decir existen muchas posibilidades de perder nuestras pertenencias incluyendo la vida por una u otra cosa. Estudiar el futuro nos ayuda interpretar que errores podemos evitar y tener un mejor futuro para la humanidad.

En la última década, la actividad en redes sociales se ha incrementado exponencialmente. A tal nivel que hoy día han superado y desbancado a la comunicación telefónica tradicional y a los medios de comunicación. Posiblemente después de la emergencia sanitaria por la que pasa el mundo en la actualidad, los viajes de negocios podrían ser sustituidos por vídeo conferencias de manera impresionante.

Las redes sociales cada vez abarcan más áreas de interacción humana, se han creado extensiones de compra y venta de artículos, extensiones adicionales para conocer usuarios nuevos, canales de entretenimiento, con el objetivo que el usuario este mas tiempo inmerso en la red y genere nuevos datos en beneficio de los propietarios de las redes sociales. Lo anterior es la razón por la que las redes sociales poseen una gran influencia en las ciencias económicas, especialmente en el área del mercado bursátil *[1],* así como en otras áreas.

Para el análisis de series temporales se utilizan métodos que tienen en común el supuesto de que múltiples variables muestran dependencia mutua a lo largo del tiempo. Es decir, una variable x1 en el momento t puede predecirse hasta cierto punto por esa misma variable en el momento t-1, o por otra variable x2 en el momento t -1. Por ejemplo, el indicador NASDAQ-100 puede ser predictivo de sí mismo en el momento t-1 y el análisis de emociones junto con la ansiedad social futura en el momento t de los consumidores de Netflix, Facebook y Microsoft. A su vez, también el análisis de emociones junto con la ansiedad social futura puede afectarse a sí mismas en t-1 y el indicador NASDAQ-100 en el momento.

En este caso aprovecharemos de la información generada por los usuarios de Twitter. A mediados de la década pasada *[4]* Jiang ZHU publicaba en un artículo que las redes sociales son una excelente fuente para generar un corpus de las emociones de las personas debido al crecimiento desmesurado de usuarios

en la red. Expresaba que una nueva manera de agrupar las sociedades que se estaban formando en las redes sociales como avatar era a través del estudio de las emociones de los usuarios *[2]*.

2. Estado del arte

Presentaremos el estado del arte considerando primero los trabajos relacionados con análisis de sentimientos, y posteriormente nos enfocaremos en las series de tiempo.

2.1. Análisis de sentimientos

El análisis de sentimientos es el proceso de analizar texto con herramientas como el aprendizaje automático para identificar y clasificar los textos de un conjunto de datos. Se utilizan diferentes fuentes de texto o corpus para realizar análisis de sentimientos. Un artículo que investiga el futuro de Twitter como herramienta para extraer metadatos a partir de la información textual publicada en Twitter, menciona que es una herramienta auxiliar considerable para pronósticos en las series de tiempo*[?][?][?]*.

Por otro lado, Nisar también menciona que Twitter sera una de las fuentes dominantes y más utilizadas para realizar análisis de sentimientos *[?]*. En la actualidad Twitter se considera un pilar clave de las redes sociales ya que se desempeña como un podio para que celebridades, deportistas, políticos, científicos, etc. expresen sus opiniones sobre un tema e influyan en las sociedades virtuales.

Nalini *[?]* analizó los sentimientos primarios que se han generado durante la pandemia. Se centro en la reacción de las emociones de usuarios de Twitter al encontrarse con noticias falsas. El análisis textual se realizó mediante la herramienta de análisis textual *Bidirectional Encoder Representations from Transformers* (BERT). En este artículo se compararon los resultados de BERT con los de otros modelos tales como como regresión logística (LR), máquinas de vectores de soporte (SVM), y redes de memoria a corto y largo plazo (LSTM). Para nuestro caso de estudio se aprovecharán estos resultados para compararlos con otros modelos del estado del arte.

2.2. Series de tiempo

Una serie de tiempo es un conjunto de valores observados cronológicamente durante un intervalo temporal. Las series de tiempo nos ayudan a realizar pronósticos con el fin de orientar las decisiones en muchas áreas del mundo como los mercados, el transporte, la identificación de fallas, el clima, entre otros. Para pronosticar una variable se debe construir un modelo y estimar sus parámetros usando datos históricos, es decir,un pronóstico se logra a través de una caracterización estadística de los enlaces entre el presente y el pasado.

En nuestro trabajo de estudio se pretende buscar la existencia de correlación entre el comportamiento de indicadores bursátiles y análisis de emociones primarias de usuarios relacionados con algunas de las empresas con mayor influencia en los distintos indicadores bursátiles extra´ıdas de los comentarios en Twitter, debido al cierto determinista del comportamiento humano. Sin embargo, el alcance mostrado en este trabajo solo converge hasta el análisis de sentimientos y experimentación de la tendencia de la fluctuación de *tweets* en la red con etiquetas relacionadas con la pandemia o el virus SARS-Cov-2.

El análisis de emociones es afectado constantemente por las preferencias de consumidores, noticias falsas en las redes sociales, experiencia del cliente y comentarios de influencias sobre algún producto o servicio, por lo que el problema de predicción de indicadores bursátiles se considerara como un modelo con variables exógenas. El análisis de series de tiempo requiere ciertas técnicas que nos permitan distinguir cuánto de esa variabilidad en el tiempo depende de determinados factores y cuál es la correlación entre variables de tal manera que podamos explicarlo dentro de nuestros modelos.

Se efectuó un análisis de las emociones representadas en series de tiempo donde se describía el comportamiento emocional del usuario a través del tiempo. El resultado de este trabajo aporto que al analizar los sentimientos textuales en redes sociales las emociones fueron cambiando al paso del tiempo *[3].* . Tomando este articulo como referencia inicial, en este trabajo se pretende analizar las emociones de los usuarios de Twitter durante la pandemia generada a causa del virus SARS-Cov-2.

2.3. Diferencia de esta propuesta con respecto al estado del arte presentado anteriormente

Se ha visto en el estado del arte que el análisis de textos ha tenido buenos resultados al ser tratados con *transformers* empleando modelos BERT y modelos T5. Comparado con los resultados obtenidos en los modelos de regresión logística, *random forest*, *modelos bayesianos* y máquinas de soporte vectorial aplicados a la clasificación binaria y clasificación multiclase muestran sin duda una un área de mejoría.

En este trabajo, se realizó parte del preprocesamiento de datos como la tokenizacion y stemming con librerías como NLTK. Como trabajo futuro se pretende comparar los resultados del preprocesamiento de datos obtenidos con la librería Spacy y Stanford.

3. Desarrollo de la solución

3.1. Descripción de la solución al problema planteado

Inicialmente los datos se normalizan y estandarizan para reducir la complejidad de los datos y lograr un modelo más general.

Posteriormente se realizó un análisis del procesamiento del lenguaje natural, el primer ejercicio en esta etapa de experimentación fue utilizar los modelos de

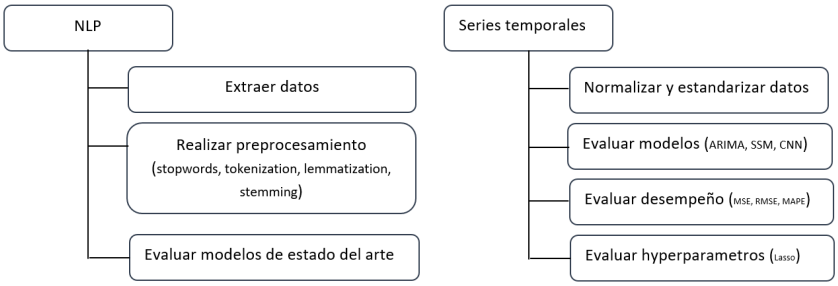

Fig. 1. Diagrama de flujo del proceso para clasificar sentimientos.

clasificación de regresión logística, *random forest*, *modelos bayesianos* y máquinas de soporte vectorial para clasificación multiclase, para realizar una clasificación binaria y experimentar con una clasificación multiclase, con esta misma familia de modelos de *machine learning*. Posteriormente se utilizó el mismo conjunto de datos para trabajar en un modelo ARIMA para predecir la tendencia de la fluctuación de *tweets* relacionados con la pandemia generada a causa de virus SARS-Cov-2.

Las métricas de desempeño utilizadas se dividirían en dos rubros. Para los modelos de procesamiento de lenguaje natural usaremos las métricas *recall*, *accuracy*, *precision* y *f1-score*. Y para el análisis de series temporales la métrica *MSE*.

En un trabajo futuro, se pretende cambiar la familia de modelos de *machine learning* por modelos relacionados con *transformers* como los modelos BERT y modelos T5. Así mismo, se pretende generar una serie temporal de la fluctuación de *tweets* a partir de etiquetas relacionadas con la pandemia y el virus SARS-Cov-2 pero con un intervalo cercano a un año. Posteriormente, se pretende buscar correlación con otras series de tiempo relacionadas con el sector financiero.

3.2. Novedad científica

Aprovechando la riqueza brindada en el trabajo del estado del arte en el que se encuentran trabajando muchos de nuestro colegas y mentores. Y en consecuencia gracias a la evaluación y aplicación de diferentes tipos de vista de AA para series de tiempo y análisis de textos nos da ya mejor percepción de nuestro estudio.

La novedad de este trabajo se centra en la búsqueda y análisis de la relación de las emociones de los usuarios de twitter durante la pandemia generada por el SARS-Cov-2. Y su correlación en entre las emociones como la tristeza, alegría, temor e ira. En la figura 1 se muestra el diagrama de flujo del proceso para clasificar sentimientos.

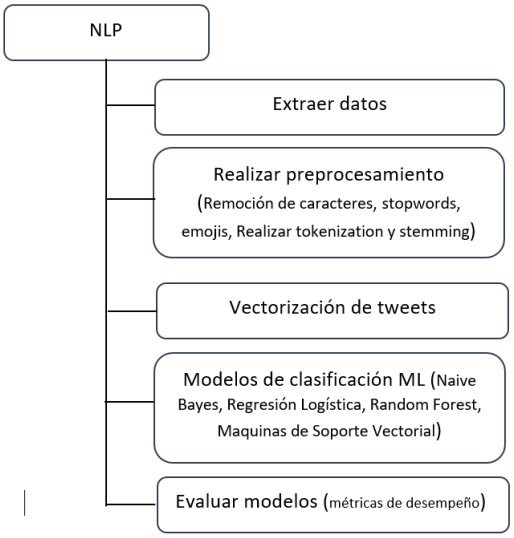

Fig. 2. Diagrama de flujo del proceso para clasificar sentimientos.

3.3. Evaluación

En este trabajo se evaluaron los experimentos realizados para la clasificación de emociones y para la predicción de series temporales. Primeramente, se evaluó el desempeño de los algoritmos implementados para la clasificación binaria y clasificación multiclase. Posteriormente, se evaluó un experimento de series temporales creado a partir del flujo de los *tweets* en la red registrados en el mismo conjunto de datos empleado para el análisis de sentimientos.

Las medidas de desempeño que fueron empleadas para los modelos de procesamiento de lenguaje natural fueron las métricas *recall*, *accuracy*, *precision* y *f1-score*. Y la métrica empleada para las series de tiempo fue *MSE*, en trabajos futuros, se pretende realizar un análisis de series temporales profundo, en el cual se considera emplear adicionalmente las medidas de desempeño *MSPE* y *MAPE*.

4. Experimentos y resultados

En este trabajo se analizaron *tweets* recopilados durante el inicio de la pandemia. Específicamente del bimestre marzo y abril de año 2020. Como bien se mencionaba con anterioridad, en esta experimentación se trabajó con modelos de clasificación para predecir el sentimiento de los *tweets* relacionados con el virus SARS-Cov-2.

Como se mencionaba en la sección introductoria, el número de casos acelerado a nivel mundial se convirtió en pánico, el miedo y la ansiedad entre la sociedad.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   UserName       41157 non-null  int64
 1   ScreenName     41157 non-null  int64
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.9+ MB
```

Fig. 3. Información de conjunto de datos tomada de Python.

Y uno de los medios donde se ha reglado es en las herramientas digitales de conexión social como Tweeter.

Los datos utilizados para esta parte experimental se refieren a un conjunto de datos estructurado a partir de 5 variables, una etiqueta de clasificación de sentimientos y 41157 registros.

De este conjunto de datos podemos ver que la mayor participación porcentual de *tweets*, emergen de ciudades grandes como Londres, Nueva York y Los Angeles. Para esta etapa de pruebas se ha recopilado únicamente *tweets* en idioma inglés, es por ello que no aparece aún alguna ciudad de México. La recopilación de información en español, así como la información de un mayor número de meses está considerada para la siguiente etapa de esta investigación.

Para esta etapa experimental, se analizaron cinco clases de sentimientos a partir del texto contenido en los *tweets* reflejados en la etiqueta *Sentiment*. Esta etiqueta tiene cinco tipos de sentimientos como se mencionaba previamente y son los siguientes, extremadamente negativos, negativos, neutrales, positivos y extremadamente positivos.

A continuación, podemos encontrar en la siguiente gráfica de anillo la participación porcentual las ciudades a partir de la agrupación de los *tweets* a nivel global.

En la Figura 5, podemos ver la estructura del conjunto de datos utilizado en esta etapa de experimentos y resultados. Inicialmente en la columna OriginalTweet, se encuentra el contenido textual tal y como se recopilo en la Tweeter. En la parte del procesamiento de la información se normalizará el texto. Es importante mencionar que para esta experimentación los caracteres especiales y emojis fueron removidos. Sin embargo, se tiene presente que estos caracteres deben de evitar ser removidos durante la limpieza de texto para darle mayor importancia al análisis de texto en el campo de procesamiento del lenguaje natural. Debido a que no dejan de ser un lenguaje y la consecuencia de ser removidos es que con ellos se remueve información valiosa.

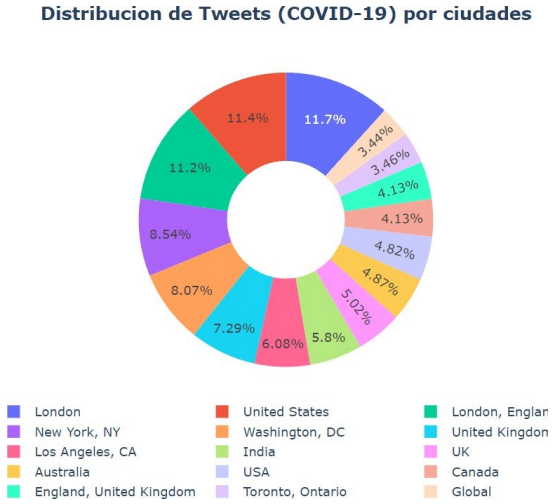

Fig. 4. Participación porcentual de tweets por ciudades.

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/l2NlzdxNo8	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate su...	Positive
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P	Positive
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is empty...\r\n\r\n\r\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need. \r\n\rStay calm, stay safe.\r\n\r\n...	Positive
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COVID19 outbreak.\r\n\r\n\r\nNot because I'm paranoid, but because my food stock is litteraly empty. The #coronavirus is a serious thing, but please, don...	Extremely Negative

Fig. 5. Estructura del conjunto de datos.

4.1. Limpieza de texto

Como podemos ver, hasta ahora se ha comenzado a trabajar con una análisis y exploración de datos. A continuación, se hablara sobre la parte del preprocesamiento de los datos. En esta sección, lo primero que se realizo fue la remoción de usuarios en los *tweets*. Es decir, se removieron los @usuario del tweet.

Posteriormente se removieron las direcciones HTTP y URLs. Para este paso se utilizó la librería 're' de Python, por sus siglas en ingles de expresión regular.

4.2. Stopwords

Los siguientes pasos que se llevaron en la experimentación fue continuar con la limpieza de datos, removiendo la puntuación, los números, caracteres especiales

Fig. 6. Nube de palabras de conjunto extremadamente positivo.

y *stopwords*. Las *stopwords* son las palabras más comunes en el lenguaje natural, que para el procesamiento del lenguaje natural no aportan mucho valor.

4.3. Tokenización

Posteriormente se llevó a cabo la *tokenización* de los *tweets* con la librería NLTK. La *tokenización* es una forma de separar un fragmento de texto en unidades más pequeñas llamadas tokens. En el procesamiento del lenguaje natural los tokens pueden ser palabras, caracteres o n-gramas.

4.4. Stemming

De manera continua se le dio un tratamiento de *stemming* al texto con la librería NLTK. El *stemming* es el proceso de eliminación de una parte de una palabra, o la reducción de una palabra para su raíz.

Posteriormente, se volvió a trabajar en un análisis exploratorio de los datos a partir de la información procesada. Con la intención de buscar tendencias en el conjunto de datos. Se llevó a través de una nube de palabras generada a partir de cada subconjunto de *tweets* extremadamente negativos, negativos, neutrales, positivos y extremadamente positivos.

En cuanto a la frecuencia de la etiqueta de los sentimientos, nos muestra una distribución en la que a pesar de estar en inicio de pandemia la mayoría de las personas tienen mayormente sentimientos positivos, sobre sentimientos negativos. Se muestra en ese bimestre optimismo.

En este trabajo, se extrajeron los *tweets* a partir de *hastags* intuitivos de manera empírica. Como *#COVID2019*, *#Coronavirust*, *#confinement* principalmente. Sin embargo, en el análisis exploratorio, se presentó la idea para trabajos futuros de realizar una nueva extracción de *tweets* a partir de los *hastags* con mayor frecuencia presentados a lo largo de la pandemia.

A continuación, se presenta la distribución que se obtuvo de las etiquetas o *hastags* con mayor frecuencia de los conjuntos de sentimientos, positivos, neutros y negativos.

Fig. 7. Nube de palabras de conjunto positivo.

Fig. 8. Nube de palabras de conjunto extremadamente negativo.

Fig. 9. Nube de palabras de conjunto negativo.

4.5. Conjunto de entrenamiento y conjunto de prueba

Posteriormente a este pequeño análisis exploratorio, se generó un conjunto de entrenamiento y un conjunto de prueba con la librería de *machine learning* scikit-learn. Para el ejercicio de clasificación multiclase. En esta parte de experimentos y resultados se evaluaron los modelos de clasificación de regresión logística, *random forest*, *modelos bayesianos* y máquinas de soporte vectorial.

Fig. 10. Nube de palabras de conjunto neutro.

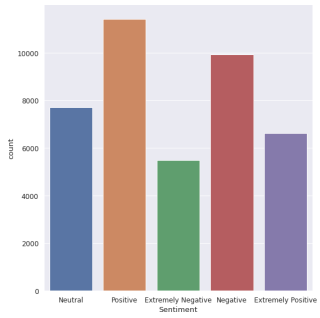

Fig. 11. Distribución de sentimientos en tweets.

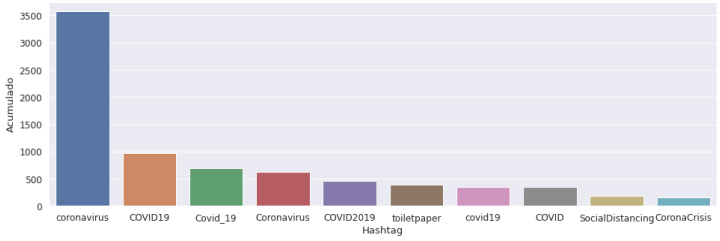

Fig. 12. Top 10 de hastag con mayor frecuencia en tweets positivos.

4.6. CountVectorizer

Posteriormente se le aplico la función de *CountVectorizer* de *scikit-learn*. La función *CountVectorizer* es un proceso de extracción de características o vectorización y es empleado para convertir documentos de texto en un vector o matriz de recuentos de tokens de cada palabra que aparece en todo el texto.

Este proceso de *CountVectorizer*, nos funcionara también para los *transformers* empleados en la siguiente etapa de la investigación. El trabajo con modelos BERT, T5 text-to-text-transfer-transformer.

A continuación, se muestra un ejemplo ilustrativo.

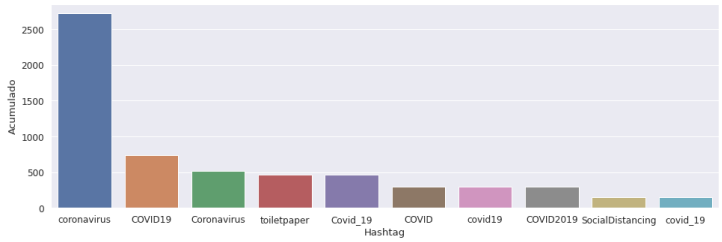

Fig. 13. Top 10 de hastag con mayor frecuencia en tweets neutros.

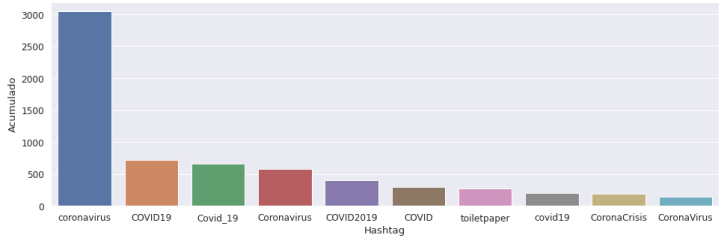

Fig. 14. Top 10 de hastag con mayor frecuencia en tweets negativos.

	at	each	four	help	helps	many	one
document[0]	0	0	0	0	1	0	0
document[1]	0	0	1	1	0	0	0
document[2]	1	1	0	0	1	1	0

Fig. 15. Ejemplo de proceso de CountVectorizer.

El primer ejercicio en esta etapa de experimentación fue utilizar los modelos de clasificación de regresión logística, *random forest*, *modelos bayesianos* y máquinas de soporte vectorial para clasificación multiclase debido a las cinco etiquetas de nuestro conjunto de datos (extremadamente negativos, negativos, neutrales, positivos y extremadamente positivos).

Los resultados del desempeño de los indicadores *precision*, *recall* y *f1-score* en los clasificadores empleados son los siguientes:

En cuanto a los resultados obtenidos, se puede ver que indudablemente el modelo de regresión logística es el que tiene un mejor desempeño, aun cuando su desempeño es sin duda mejorable, utilizando algún método más sofisticado o un método del estado del arte.

Para este grupo de medidas de desempeño, el siguiente modelo que podemos ver que ofrece un segundo mejor resultado es el modelo de máquinas de soporte vectorial.

Por otro lado, si nos guiáramos únicamente por el desempeño calculado por *accuracy* , encontraríamos que nuevamente el modelo con mejor desempeño en

Maquinas de Soporte Vectorial

	precision	recall	f1-score
Extremely Negative	0.48	0.71	0.57
Extremely Positive	0.53	0.78	0.63
Negative	0.58	0.55	0.56
Neutral	0.71	0.64	0.67
Positive	0.67	0.55	0.61

Regresión Logística

	precision	recall	f1-score
Extremely Negative	0.62	0.68	0.65
Extremely Positive	0.62	0.71	0.66
Negative	0.54	0.56	0.55
Neutral	0.72	0.64	0.68
Positive	0.62	0.58	0.59

Fig. 16. Resultados de precision, recall y f1-score.

Modelo	Accuracy
Regresión Logística	0.859451
Random Forest	0.828231
Naive Bayes	0.791667
Maquinas de Soporte Vectorial	0.607264

Fig. 17. Resultados de accuracy.

este ejercicio es el modelo de regresión logística, solo que en este caso es seguido del modelo *random forest*.

Otro experimento que se realizó en este trabajo, fue emplear la misma familia de modelos de *machine learning* pero para una clasificación binaria extremadamente negativos + negativos= negativos, extremadamente positivos + positivos =positivos.

Y los resultados son los siguientes:

En estos resultados, se pudo observar que nuevamente el modelo de regresión logística fue la que tuvo el mejor desempeño, sin embargo, no estuvo muy alejado del resto de los modelos. Por otro lado, podemos observar que el desempeño mejoro ligeramente en los clasificadores binarios contra los clasificadores multiclase.

El orden de los resultados en el *accuracy* para los clasificadores binarios fue similar al de los clasificadores multiclase.

Naive Bayes

	precision	recall	f1-score
0	0.69	0.74	0.71
1	0.85	0.82	0.84

Regresión Logística

	precision	recall	f1-score
0	0.77	0.84	0.80
1	0.92	0.87	0.89

Fig. 18. Resultados de precision, recall y f1-score.

Random Forest

	precision	recall	f1-score
0	0.70	0.82	0.75
1	0.91	0.83	0.87

Maquinas de Soporte Vectorial

	precision	recall	f1-score
0	0.69	0.87	0.77
1	0.94	0.84	0.88

Fig. 19. Resultados de precision, recall y f1-score.

Modelo	Accuracy
Regresión Logística	0.859451
Maquinas de Soporte Vectorial	0.845603
Random Forest	0.828474
Naive Bayes	0.791667

Fig. 20. Resultados de Accuracy.

Si bien aún no se han realizado experimentos con datasets relacionados directamente con series temporales como el dataset HistoricalData, el cual contiene información del indicador bursátil NASDAQ-100.

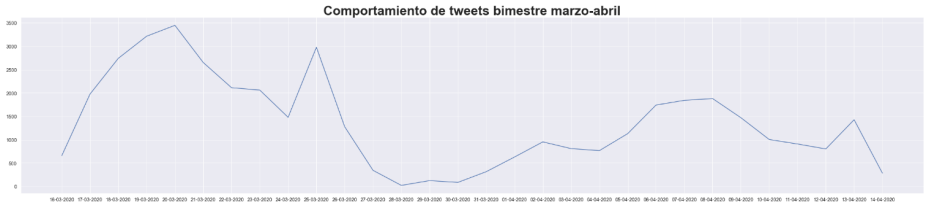

Fig. 21. Comportamiento de tweets marzo-abril, 2020.

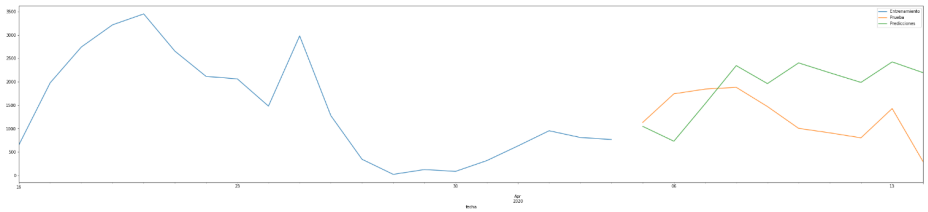

Fig. 22. Predicción de tweets marzo-abril, 2020.

Fig. 23. Comportamiento de tweets positivos marzo-abril, 2020.

En este trabajo se realizaron experimentos iniciales, en el cual se buscó entender y observar la tendencia del comportamiento de la fluctuación de los *tweets* dentro de la red. Para la predicción del modelo ARIMA el error fue grande *MSE* 1121060.09. En trabajos futuros se pretende buscar mejorar este error a partir de enriquecer el conjunto de datos y probar con modelos *CNN*, como se mencionaba con anterioridad.

Inicialmente, se representó gráficamente el comportamiento general del tráfico de *tweets* en la red. Es decir, se consideraron las cinco etiquetas como una sola, para tener un agrupamiento global de la presencia de *tweets*. Distinguiéndolos a partir de la fecha de emisión en un formato de año-mes-día.

A continuación, se muestra la tendencia de la fluctuación de *tweets* en este bimestre marzo-abril, 2020. Se pudo observar que en consecuencia de que los *tweets* fueran mayoritariamente positivos y negativos. La tendencias global se vio marcada con un comportamiento muy similar.

Fig. 24. Comportamiento de tweets negativos marzo-abril, 2020.

5. Conclusiones y trabajo futuro

En este trabajo se ha propuesto una solución posible para identificar sentimientos en *tweets* relacionados con la enfermedad de la pandemia causada por el virus SARS-CoV-2. Se ha revisado el estado del arte, y se han identificado diversas soluciones para resolver esta tarea. Presentamos resultados utilizando modelos bayesianos, regresión logística, *random forest*, y máquinas de soporte vectorial, así como una propuesta de utilización para el estudio de comportamiento en noticias en redes sociales como series de tiempo. El desempeño es similar al obtenido en el estado del arte, aunque existe todavía margen de mejora. Como trabajo futuro, hemos propuesto utilizar BERT y modelos T5 para obtener un mejor desempeño, para la parte de procesamiento de lenguaje natural. Para la parte de las series temporales, se pretende utilizar modelos CNN y comparar los resultados con otro modelos como ARIMA y SSM. En el experimento de tratar los datos de la fluctuación de *tweets* en la red se pudo encontrar que los datos empleados para la etapa experimental, deben ser más grande. Lo que nos llevara a abordar otro reto asociado con el *big data*.

Agradecimientos. Este trabajo fue apoyado por el CONACyT, COFAA y el Instituto Politécnico Nacional, IPN-SIP gracias a los proyectos SIP 2083 y SIP 20210189. J.H.P.: Agradezco la atención y el apoyo de mis asesores, por esta oportunidad. Así mismo les agradezco por incentivar la curiosidad y el desarrollo científico constante que me han brindado.

Referencias

1. Chen, C. Y.-H., Härdle, W. K., Klochkov, Y.: SONIC: SOcial Network analysis with Influencers and Communities. Journal of Econometrics, (2021)
2. Li, J., Yang, G.: Network embedding enhanced intelligent recommendation for online social networks. Future Generation Computer Systems, vol. 119 (2021)
3. Sul, H. K., Dennis, A. R., Yuan, L.: Trading on twitter: Using social media sentiment to predict stock returns., (2017)
4. Zhu, J., Wang, B., Wu, B.: Social network users clustering based on multivariate time series of emotional behavior. The Journal of China Universities of Posts and Telecommunications, vol. 2014 (2014)