Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Actuaría

Análisis Multivariado

Noticias

Equipo

- Jessica Fernanda Rodríguez Mondragón
- Jennifer Itzel García Carrillo
- José Maximiliano Manrique Arreola
- Andrea Nava Torres
- Francisco Adan Arias Moreno

Semestre 2020-2

Índice general

Resumen

Internet ha acelerado sin duda la velocidad a la que la información puede viajar por el mundo, ya sea para afectar a algunos o beneficiar a otros. Hoy en día es difícil determinar si un suceso de última hora es real o no en las primeras horas de haber acontecido. A continuación, se detallan los procesos de limpieza e ingeniería de datos para clasificar mediante algoritmos de machine learning de tipo supervisado y agrupar texto con la ayuda de un algoritmo de aprendizaje máquina de tipo no supervisado, proveniente de noticias falsas y noticias verdaderas, publicadas en sitios en línea y recabadas alrededor de la época de elecciones en Estados Unidos en 2016.

Introducción

En los últimos años, las redes sociales han revolucionado la forma en que la población y los gobiernos acceden, consumen y difunden noticias. Hoy en día, los usuarios de Twitter y Facebook, utilizan estas redes como su principal medio de información y comunicación. Según Allcott [1] durante las elecciones en Estados Unidos en el 2016 se encontraron los siguientes hallazgos: 1) de las noticias falsas conocidas que aparecieron en los tres meses anteriores a las elecciones, las que favorecian a Trump se compartieron un total de 30 millones de veces en Facebook, mientras que las que favorecian a Clinton se compartieron 8 millones de veces; 2) es mucho más probable que las personas crean historias que favorecen a su candidato preferido.

Las redes sociales, permiten que los usuarios sean productores y consumidores de una serie de contenidos a la vez, lo cual ha facilitado el alcance y propagación que tienen las fake news, ya que pueden llegar a replicarse miles de veces en cuestión de segundos. Si bien Facebook puede considerarse como la red principal de propagación de contenido falso, es en Twitter, donde se encuentra la mayor cantidad de noticias falsas, ya que por la naturaleza de la red social se expanden con mayor facilidad.

Según un estudio realizado por Twitter en 2018, junto con un grupo de investigadores del Instituto Tecnológico de Massachusetts (MIT), demostraron que las noticias falsas tienen un 70 % más de probabilidad de ser retuiteadas que las noticias verdaderas, es decir, las fake news se difunden significativamente más rápido y de una manera más profunda. Si bien, podríamos pensar que el uso de bots contribuye en el alcance de estas, por lo que en el mismo estudio se demostró que los seres humanos son los principales responsables de la propagación de las noticias falsas, ya que se encontró que los bots difunden noticias falsas y verdaderas por igual.

Pero, ¿cuál es la principal causa de que las fake news se divulguen?, si bien una de las causas es por el contenido novedoso y algunas veces creativo, lo que despierta el interés de las personas, y en segundo lugar, estas plataformas se ocupan de conocer a sus usuarios, es decir, utilizan un algoritmo que distribuye el contenido más relevante para cada usuario, de esta forma logran que la información que llega a cada usuario esté condicionada y filtrada respecto sus gustos e intereses.

Dado lo anterior, estas pueden llegar a influir en la toma de decisiones de las personas, tal es el caso de las elecciones presidenciales de 2016, en la campaña del actual presidente de Estados Unidos Donald Trump, ya que su victoria se atribuye a la gran difusión de fake news en las redes sociales, especialmente en twitter, con el fin de destruir la imagen pública de su oponente Hillary Clinton a partir de informacion falsa, aun cuando esta era desmentida, la información se propagaba en cuestión de segundos, logrando su objetivo.

Es por eso que en este proyecto, utilizando la base de datos obtenida de Kaggle, bajo el nombre de “Fake and real news dataset”, que contiene una lista de artículos considerados como “fake news” y “real news” desarrollaremos clasificadores para detectar si una noticia, respecto a su contenido es falsa o no, de forma que podamos explorar el contenido y la estructura de una noticia falsa, para esto tomaremos nuestra unidad muestral como ”Noticia”.

Aplicación al negocio

0.1. Problemática

La cantidad de información generada a diario en internet es exorbitante, resulta una tarea difícil para los humanos que con tan solo leer una noticia, podamos saber si se trata de un texto escrito con la moralidad debida o no. Actualmente todas las personas, o un gran porcentaje de la población, cuenta con acceso a internet, lo cual, nos posiciona en un escenario en donde es más fácil acceder a cualquier tipo de información, desde información sobre artistas, conciertos, hasta información política.

Hoy en día, el tema del covid-19 es algo que ha estado presente en todo el mundo, por lo que circula mucha información tanto falsa como certera en internet , incluso hasta en las noticias de televisión abierta. La gente, al no salir de sus casas, usa el internet como medio de comunicación e información, sobre todo las redes sociales. Día a día se publican millones de noticias en todo el mundo, por lo cual es muy fácil que nosotros caigamos en una noticia falsa, de cualquier índole, desde las noticias que aseguran la renuncia del presidente de México, hasta la muerte de un artista, o la cura del coronavirus.

Y claro, esta situación es muy alarmante, ya que la mayoría de las personas se deja llevar por el encabezado de la noticia, sin antes leerla, tal es el caso de la notica que estuvo circulando hace unos meses en Facebook, la cual mencionaba que en Querétaro se construiría un Disney World en 2020, esta noticia solamente mostraba el encabezado a los usuarios de Facebook, pero al momento de leerla hasta el final advertían que era falsa, lo cual no fue impedimento para lograr un alcance impresionante en las redes sociales. Si bien muchas noticias falsas tienen fines políticos y monetarios, como el caso que se mencionó en la introducción, para lograr un objetivo y perjudicar personas, pero para fines de este trabajo no tocaremos ese tema. Esto tiene consecuencias de impacto económico y social[2] por ello últimamente se han empleado algoritmos de machine learning para llevar a cabo esta tarea, aunque no hay un modelo destinado específicamente.

0.2. Solución

Lo que buscaremos será crear un clasificador que nos diga cuándo una noticia es real, y cuando es falsa basada en su contenido, para esto, se implementarán algoritmos de machine learning supervisado, se analizarán y se comparará su eficacia al clasificar las noticias falsas de las reales, con el objetivo de asignar la etiqueta de verificación, antes incluso de que un humano pueda reportarlo.

También se implementará el proceso para obtener clusters basados en el texto, con el propósito de mostrar si el factor de la veracidad de las noticias influye en la agrupación de éste. Lo cual, sería ideal poder implementar en las redes sociales, ya que las noticias que nos presentan solamente aparece el encabezado, por lo que se creará un etiquetado basado en qué tan probable es que el contenido de esa noticia sea verdadero o falso o en su defecto que sea ambiguo y "engañoso", es por eso que nos basaremos en un modelo parsimonioso, para respetar el hecho de que su interpretación sea sencilla y comprensible para los usuarios, de una forma que no perdamos interpretabilidad de lo que muestra el modelo, lo cual esperamos que esté al alcance de todos los navegantes de internet.

Exploración de datos

0.3. Diccionario de datos

En esta sección, presentamos las variables de los datos con las que se trabajó, en forma de diccionario, con el fin de entrar en contexto y presentar resultados. En la tabla siguiente se muestran las variables originales de nuestro conjunto de datos.

Variable	Tipo de dato	Descripción
title	String	El título con el que está registrada la noticia
text	String	El texto que contiene la noticia
subject	String	Clasificación de la noticia
date	String	Fecha en la que la noticia se publicó

De las variables anteriores, con el propósito de poder trabajar con esta información, se les realizaron transformaciones y limpieza, así como también se crearon nuevas variables que se muestran a continuación:

Variable	Tipo de dato	Tipo de variable	Descripción
target	Integer	Entero	Nos indica el tipo de noticia: 1 = Noticia real , 0 = Noticia Falsa
Final _ Text	String	Texto	Título y texto de la noticia
Final _ clean	String	Texto	Título y texto de la noticia con filtro de limpieza
tokenized _ text	List	Lista	Texto de la noticia limpio con las palabras separadas por comas
word_ count	Integer	Entero	Numero de palabras que contiene la noticia; Longitud de la noticia
date	Date	Categórica	Fecha en la que la noticia se publicó

0.4. Visualización de Datos

Como ya mencionamos, la variable de target nos indica cuando una noticia del conjunto de datos es verdadera y cuando es falsa, teniendo esto en mente, 23,471 corresponden a ”fake news”, mientras que 21,417 a real news”, es decir, un total de 44,888 registros, y la proporción con la que contamos de noticias es la siguiente:

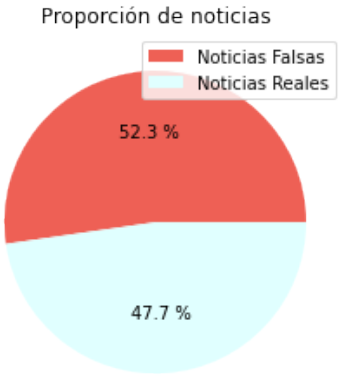

Figura 1:

De esta manera, observamos que las noticias falsas prevalecen en nuestros datos; en la siguiente gráfica vemos los temas que más controversia tienen en las fake news:

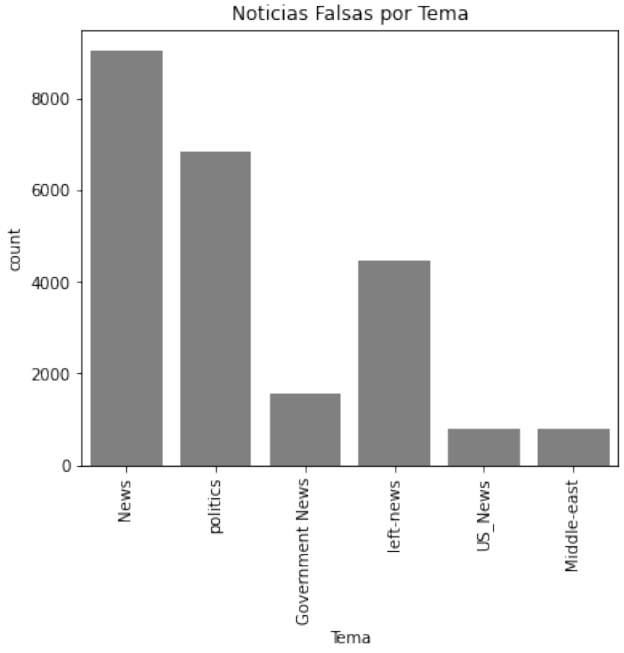

Figura 2:

Como podemos notar, la categoría ”News”, representa la mayor parte de la concentración de las noticas, dentro de esta se encuentran temas de caracter mundial, y una gran parte está conformada por temas de política, mientras que las demás categorías, nos hablan de noticias gubernamentales, de noticias de partidos políticos de izquierda, noticias de Estados Unidos, en general. Por otro lado veamos cuales son los temas en los que se concentran las noticias reales:

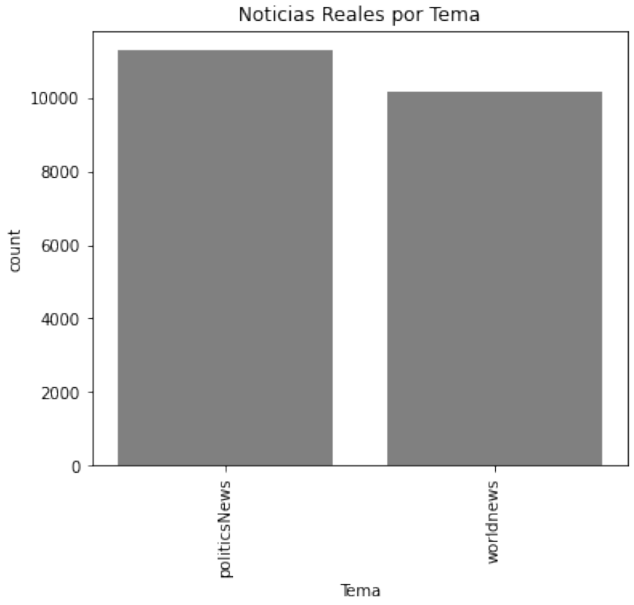

Figura 3:

En el gráfico anterior se observan solo dos grandes y amplias categorías, temas políticos y temas de caracter mundial, podemos notar que los temas en los que se presentan las noticias verdaderas conforman los temas de las noticias falsas, es decir, estos últimos están contenidos en los de real news.

Es importante recordar que los datos estaban separados en dos datasets distintos, y aunque las columnas que los formaban eran las mismas, aquí podemos caer en cuenta de que su la agrupación en la columna 'subject' fue distinta para los dos tipos de noticias.

0.5. Distribución de las noticias

Para efectos de modelación, y de entender como procesar la información, revisamos la distribución de algunas propiedades de las noticias. La longitud de palabras que tienen las noticias falsas y las noticias verdaderas, cuya distribución se muestra en la siguiente gráfica:

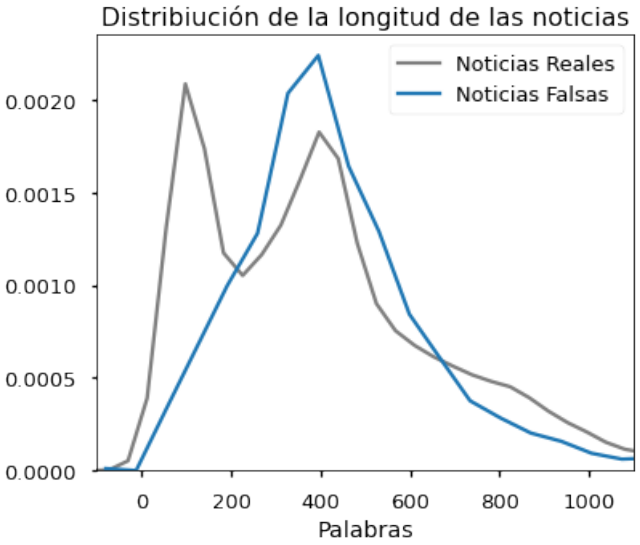

Figura 4:

En lo anterior, podemos notar que la mayor parte de las noticias reales están por debajo de los 200 caracteres, es decir nos habla de noticias concretas, sin embargo, cuando estas rebasan ese límite, se nota un cambio en la estructura de la densidad, mientras que la mayor parte de noticias falsas está concentrada en los 400 caracteres.

Por otro lado en la siguiente figura se muestra la distribución de el numero de noticias que nos encontramos por mes y la proporción para cada etiqueta.

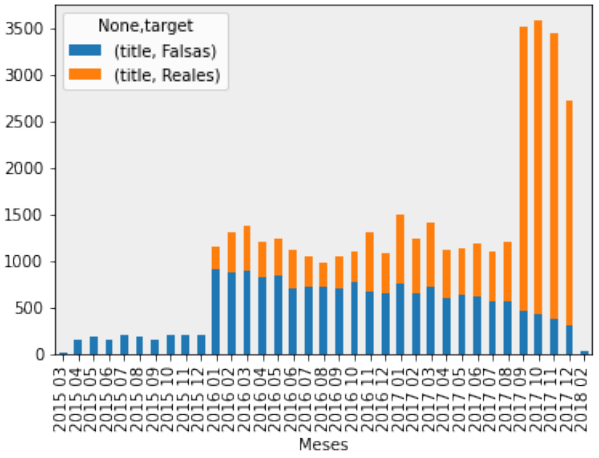

Figura 5:

Como podemos observar, no nos encontramos con la presencia de noticias de tipo real entes de enero de 2016 y tampoco después de diciembre de 2017.

Limpieza del Texto

En primer instancia, la columna 'date', contaba con diferentes formatos de texto, no todas las fechas estaban homogeneizadas en el formato de Date, en especial,para algunos registros, esta columna contenía hipervínculos, la inconsistencia se resolvió eliminando estos registros de carácter inválido y se creó una función para homogeneizar el formato de todas las fechas restantes y que así fuera más manipulable.

En segundo lugar, se consideró tomar el título de la noticia seguido del texto correspondiente lo cual nos creó la variable Final _Text, para contener todo el texto en una misma columna.

Title	Text	Final_Text
Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing'	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media.	Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media.

0.6. Stopwords y Lematización

Como podemos notar, en el ejemplo anterior, dentro de nuestro texto, existen diferentes caracteres, apóstrofes, puntuaciones, hashtags, las conocidas stopwords, palabras que por si solas no poseen un significado como pronombres, preposiciones, adverbios e incluso algunos verbos,como por ejemplo "the", "a", "an", "in", por lo que haciendo uso de la lista de stopwords del kit de herramientas de lenguaje natural o mejor conocido por sus siglas en inglés NLTK, se creó una función en Python, para eliminarlas de nuestra variable Final_Text.

Dentro de esta misma función se realizó el proceso de pasar todas las palabras a minúsculas,es decir normalizar el texto, ya que para poder darle un sentido, para nosotros es lo mismo tener 'DONALD' a 'donald y 'Donald', pero la máquina lo reconoce de manera distinta.

A pesar de que el texto final se redujó significativamente, nos encontramos con palabras diferentes en representación de una misma, por ejemplo, 'sends', 'send', 'sent', es decir, la misma palabra en diferentes conjugaciones, por lo que se procedió al proceso de lematización, el cual relaciona una palabra derivada con su forma canónica o lema. Después de todo el proceso anterior nuestro texto inicial quedó de la siguiente manera:

Final_Text	Final _clean
Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media.	donald trump send embarrass new year eve message disturb wish americans happy new year leave instead give shout enemies haters dishonest fake news media

Cabe mencionar, que una gran parte de las noticias, fueron extraídas de Twitter, por lo que algunas contenían los nombres de los usuarios, emojis, url y html, por lo que de igual manera se crearon dos funciones para eliminarlas, y así nuestra longitud de palabras se disminuyó aún más.

0.7. Tokenización

Como último paso, una vez que objutivimos el Texto limpio final, se procedió a convertir las palabras del texto en elementos de una lista, este proceso es conocido como tokenización, dejando nuestro texto de la siguiente manera:

Final _clean	tokenized _ text
donald trump send embarrass new year eve message disturb wish americans happy new year leave instead give shout enemies haters dishonest fake news media	['donald', 'trump','send', 'embarrass', 'new', 'year', 'eve', 'message', 'disturb', 'wish', 'americans', 'happy', 'new', 'year', 'leave', 'instead', 'give', 'shout', 'enemies', 'haters', 'dishonest', 'fake', 'news media']

0.8. Palabras frecuentes

Nuestro dataset contiene la fecha de las elecciones presidenciales de Estados Unidos en 2016, por lo que es de esperarse que una de las palabras más repetidas sea el nombre del presidente de Estados Unidos,ya que evidentemente aumentó el trafico de noticias reales, pero tambien el de noticias falsas [1]. A continuación se muestra la proporción en la que esta palabra aparece tanto en noticias falsas como verdaderas.

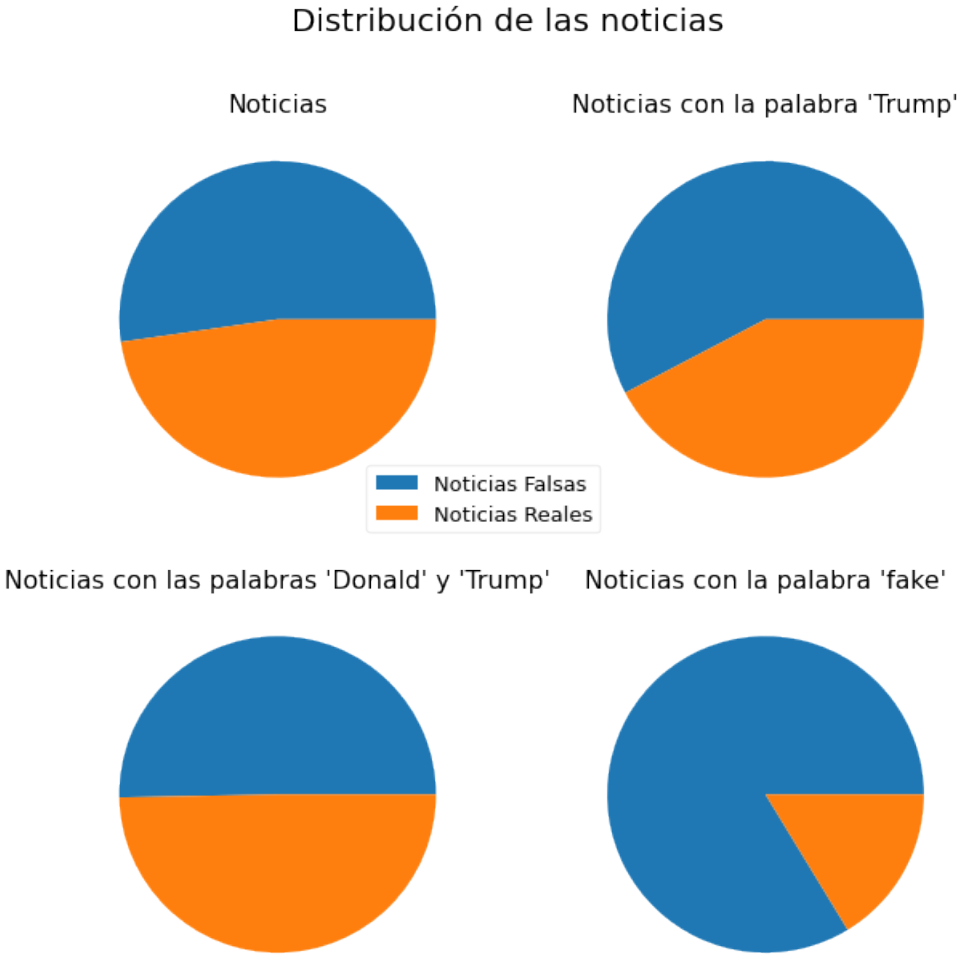

Figura 6:

Como podemos ver, tenemos alta presencia de los términos Donald y Trump, además de la palabra 'fake' en las noticias, lo cual puede afectar el desempeño de nuestro modelo, haciendo que el modelo solamente aprenda a identificar si cualquiera de las palabras 'Donald', 'Trump' o 'fake', están o no, ya que como se muestra en la gráfica, la presencia de la palabra 'fake' es mayor en las noticias falsas, por lo que se procedió a eliminar las palabras que más se repetían en nuestro texto, tanto en noticias falsas como en verdaderas, haciendo uso de la función replace().

Modelación

Antes de empezar a proponer y evaluar modelos, se realizó un tratamiento a nuestra tabla de datos original, sin haber pasado por el filtro de limpieza, el cual fue particionar nuestros datos en 60 % para entrenar al modelo, 20 % para validar y 20 % para probarlo, de esta forma, se entrenaron los modelos y se validaron respecto a su porcentaje correspondiente, dandoles un tratamiento por separado, ya que algunos tratamientos consideran datos obtenidos a través de la distribución de la muestra y al final se probaron los modelos con datos totalmente desconocidos, que fueron los datos correspondientes a test.

Por otra parte, a nuestra variable Final_clean, el texto de las noticias una vez pasado por el filtro de limpieza, se procedió con aplicarle Tf-idf , del inglés Term frequency- Inverse document frequency, que significa frecuencia de término- frecuencia inversa de documento, el cual es el producto de estas dos medidas, es decir, son los componentes de las puntuaciones resultantes asignadas a cada palabra, que resaltan las palabras que tienen una mayor significacia en el texto.

- Term frequency: resume la frecuencia con la que aparece una palabra en un documento.
- Inverse document frequency: reduce las palabras que aparecen mucho en los documentos.

Dado lo anterior, la función TfidfVectorizer() de python, tokeniza el texto, aprende el vocabulario y la ponderación de frecuencia de los textos inversa, lo cual le permite codificar nuevos textos, lo que permite manejar el hecho de que algunas palabras aparecen más en las noticias que en otras.

Para los siguientes modelos se ajustó el TfidfVectorizer con los parámetros df_min=.02 y df_max=.5, lo que quiere decir es que va a ignorar aquellas palabras que tienen presencia en mas del 50 % y menos del 2 % de los documentos.

0.9. Modelación no Supervisada

Con motivo de asignarles un perfilado a las noticias respecto al texto que las conforma, se realizó un modelo de n vecinos más cercanos, con n=4.

Se concluye el siguiente perfilado:

Figura 7: Encontramos palabras "llamativas"para entrar a un link o notas amarillistas

Figura 8: Noticias sobre conflictos entre corea del Norte y nuestro país vecino.

Figura 9: En este cluster se encuentran palabras típicas de una noticia cuando la elección del 2016 tuvo lugar en USA.

Figura 10: Son noticias de carácter mundial, en las que Estados Unidos suele estar involucrado,y en general.

0.10. Modelación Supervisada

0.10.1. Regresión Logística

La regresión logística, al igual que otras técnicas estadísticas multivariadas, es muy utilizada en la industria, ya que es usada para modelar la relación entre una variable de respuesta binaria, en este caso la variable target, y un conjunto de variables predictoras. Teniendo el texto de las noticias, entrenamos el modelo solo con el 60 % de los datos que pertenecen a la parte de train.

Una vez que entrenamos el modelo, nos basamos en dos métricas para comprobar que nuestro modelo generaliza de una manera correcta, las cuales son las siguientes:

- Accuracy Score: Proporción de las clases correctamente clasificada.
- F1 Score: considera tanto la precisión p como el recall r para calcular la puntuación $\frac{q}{p}$, número de resultados positivos correctos dividido por el número de todos los resultados positivos devueltos por el clasificador, y r es el número de resultados positivos correctos dividido por el número de todas las muestras relevantes.

Los resultados del entrenamiento del modelo son los siguientes:

Score	
Accuracy Score	0.9897
F1 Score	0.9893

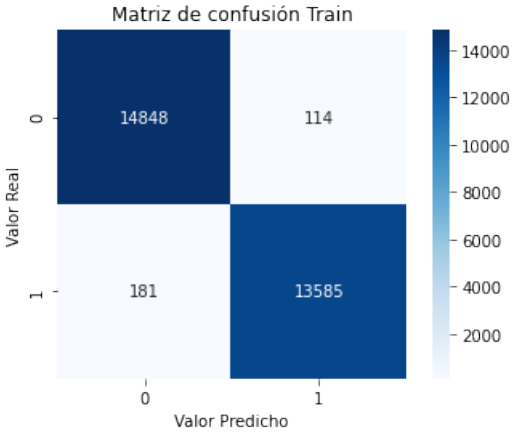

Figura 11:

En especial buscaremos mantener el F1 Score igual para validación y test, ya que como podemos notar, al entrenar el modelo, lo ideal que esperariamos es que las noticias que son falsas las clasificara como falsas, es decir, mantener la parte de falsos positivos relativamente baja. En esta ocasión,nuestro modelo clasificó 181 Falsos Positivos, lo cual es un grado de error bajo para el total de los datos presentados.

mientras que para validación:

Score	
Accuracy Score	0.9827
F1 Score	0.9820

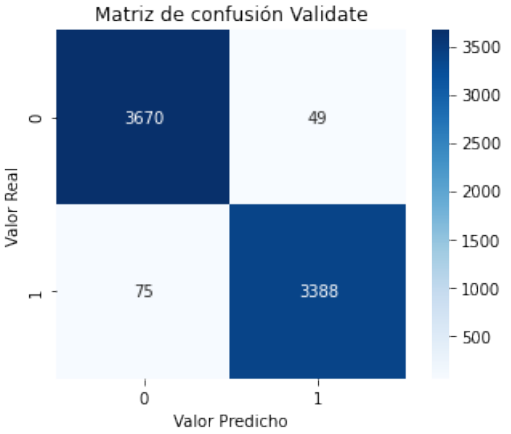

Figura 12:

Observamos que el F1 Score se mantiene en .9820, y por último probaremos nuestro modelo de regresión logística con datos completamente desconocidos, los cuales son los correspondientes a test de la partición de nuestro dataset, y obtenemos lo siguiente:

Score	
Accuracy Score	0.9867
F1 Score	0.9861

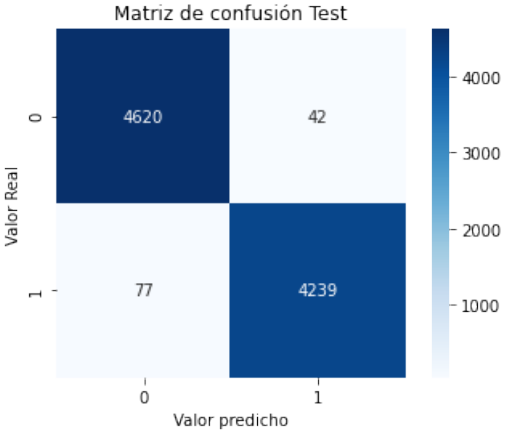

Figura 13:

Basados en los resultados anteriores nuestro modelo de regresión logística generalizó de una manera correcta, ya que al probar con noticias totalmente desconocidas seguimos obteniendo un grado de F1 Score de 98 %, lo cual es de suma importancia, ya que para ponerlo en práctica nos interesa que los modelos propuestos se equivoquen en menor cantidad de veces al clasificar una noticia falsa como verdadera, porque de no ser así seguiriamos cayendo en el círculo de la propagación de desinformación que existe en internet.

Para este caso, nuestro modelo generaliza de una forma adecuada, por lo que se propusieronetiquetas basadas en la probabilidad de que cuando nosotros probemos el modelo con las noticias desconocidas, este lo clasifique como una noticia falsa, como se muestra a continuación en el ejemplo:

ID Noticia	Prob. de falsa	Prob. de verdadera
1	0.309958	0.690042
2	0.999658	0.000342
3	0.997645	0.002355

Principalmente nos interesa que las fake news se clasifiquen como falsas y no verdaderas, esto con el fin de disminuir el impacto negativo que tienen las fake news, por lo que procedimos a realizar el siguiente etiquetado basados en los rangos de probabilidad y en el contenido de las noticias:

Tipo de contenido en noticias	Etiqueta	Descripción
Información Falsa	Rojo	Son todas aquellas noticias que poseen información absolutamente falsa
Información no verificada	Amarillo	Noticias con contenido que no es totalmente falso
Información verdadera	Verde	Noticias con información verdadera

El etiquetado anterior aplicado a nuestros datos nos indica que efectivamente el modelo clasificó bien las noticas falsas, y verdaderas y hay un porcentaje mínimo de noticias, que ya sea por la naturaleza de estas o por su contenido en las que el modelo se equivocó.

Respecto a las etiquetas asiganadas a las noticias desconocidas, podemos notar que el 51.8 % de la información representan noticias con una alta probabilidad de ser falsas, mientras que el 44.5 % poseen un grado de probabilidad menor de ser fake news y por otro lado existen las noticias que cuentan con una probabilidad intermedia de pertenecer a ambas:

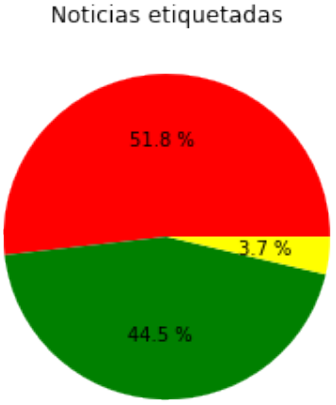

Figura 14:

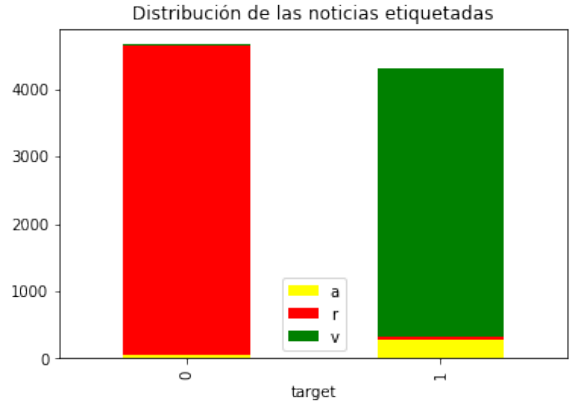

Figura 15:

Como se muestra en la figura 15, el porcentaje de error al asignar noticias falsas como verdaderas es mínimo, en cambio cuando estamos bajo un escenario de noticias verdaderas, la etiqueta amarilla se encuentra en mayor proporción, esto quiere decir que existe una probabilidad de que en las noticias verdaderas se encuentre información no verificada.

0.10.2. Perceptrón

El algorítmo Perceptrón,es parte de los algorítmos en línea que son impulsados por error. Por lo que en el caso ideal garantiza un error prácticamente nulo al momento del entrenamiento del modelo, y en el peor de los casos no se tiene garantizado su comportamiento. Por este motivo, decidimos implementar este algorítmo y obtuvimos resultados no muy favorables:

En el caso de los datos de entrenamiento:

Score	
Accuracy Score	0.9953
F1 Score	0.9951

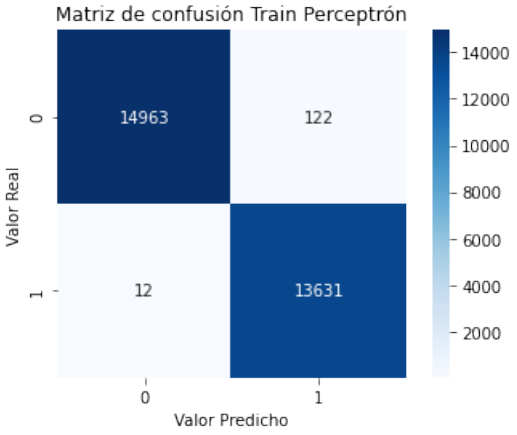

Figura 16:

Modelo con datos de validación:

Score	
Accuracy Score	0.9831
F1 Score	0.9818

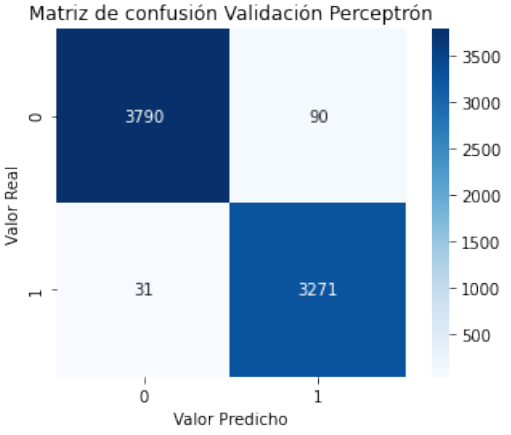

Figura 17:

Desde este resultado, podemos notar que el modelo sufre de sobreajuste, no está generalizando, veremos que tal nos va cuando lo probamos con los datos desconocidos:

Score	
Accuracy Score	0.9849
F1 Score	0.9841

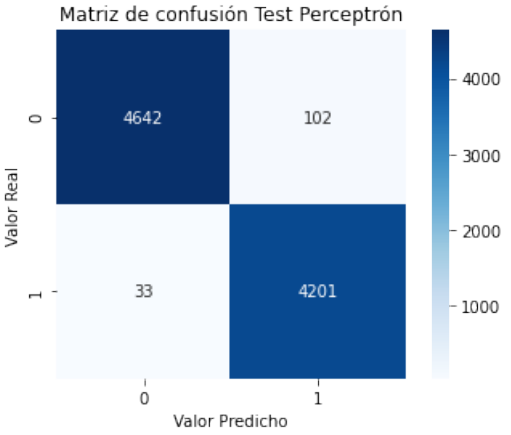

Figura 18:

Como nos esperabamos, el modelo sobreajusta, lo cual no nos es funcional, por lo que seguimos preferiendo la regresión logística, que hasta ahora funcionó de manera adecuada.

0.10.3. Pasivo-Agresivo

Los clasificadores lineales, debido a la naturaleza esparcida de los datos, obtienen muy buenos resultados , al ser más simples y menos sensibles al ajuste de parámetros. El algoritmo Passive-Aggressive, es familia de los algoritmos de aprendizaje a gran escala,no requiere un factor de aprendizaje, por lo que ahora proponemos este clasificador lineal con nuestros datos, y los resultanos obtenidos son los siguientes:

Para el caso del entrenamiento del modelo:

Score	
Accuracy Score	0.9998
F1 Score	0.9998

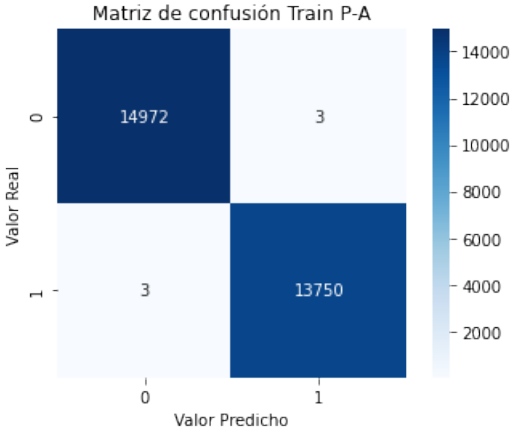

Figura 19:

Para la validación:

Score	
Accuracy Score	0.9906
F1 Score	0.9900

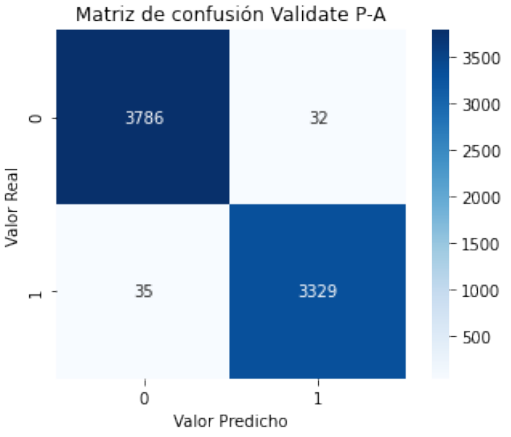

Figura 20:

Y por último para test:

Score	
Accuracy Score	0.9919
F1 Score	0.9916

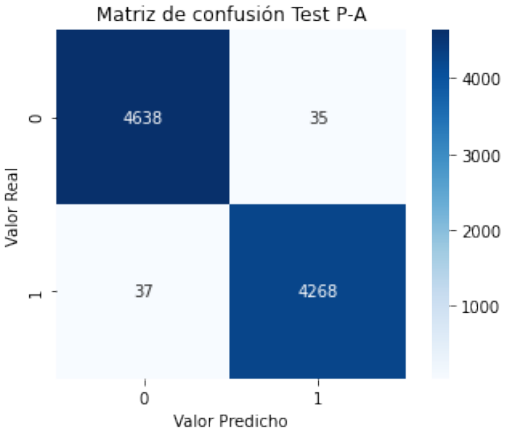

Figura 21:

Como podemos notar, el modelo está sobreajustado en los datos de entrenamiento, lo cual es un problema para los datos de validación, ya que al no generalizar bien, tendrá un rango de error mayor, como se nota el crecimiento del error en la matriz de confusión, por lo tanto, descartaremos este modelo.

Conclusión

Como primer conclusión, el trabajar con texto no es una tarea fácil, en particular por el tema de la dimensionalidad, de igual manera, por este tema dos de los modelos propuestos no generalizaron bien, por lo que se eligió implentar el modelo de Regresión Logística, con el cual además de ser parsimonioso, se obtuvieron buenos resultados, como obtener muy pocos Falsos Positivos, es decir que el error de clasificar una noticia falsa como verdadera fuese muy bajo, además de que este modelo generalizó de manera adecuada.

Se decidió optar por este modelo y realizar etiquetas con el fin de implementarlas en noticias de redes sociales, principalmente en los encabezados de estas, ya que el encabezado es la primer imagen que se presenta de una noticia en internet, y muchas personas se dejan llevar por esta primera impresión.

Por otra parte, en el proceso de limpieza de texto, se lograron identificar ciertos patrones que se presentan en las noticias falsas, tales como hipervínculos en el título y texto de la noticia, diferentes faltas de ortografía dentro de ellas, sin embargo también habían textos muy bien escritos que pertenecían a fake news, además del uso excesivo de mayúsculas en el encabezado de la noticia, con el fin de captar la atención de los receptores, mientras que en las noticias verdaderas, sí había presencia de estos patrones de comportamiento, sin embargo eran en menor cantidad.

Como conclusión final, si bien implementar el modelo no es una tarea difícil, sin embargo muchas de estas noticias falsas cumplen con un objetivo específico, la mayoría de las veces son temas políticos, ya sea para ganar las elecciones o no causar pánico en la sociedad, pero existen otro tipo de noticias falsas que leemos día a día en internet, por lo cual la implementación del modelo nos evitaría caer en manos de la desinformación.

Bibliografía

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, May 2017.

[2] N. Kshetri and J. Voas. The economics of "fake news". *IT Professional*, 19(6):8–12, 2017.