Módulo 2, Práctica 3

[**Descubrimientos**](#_7odaw0ddvymp) **2**

[Limpieza y Exploración de Datos](#_dxmo3id0zbsu) 2

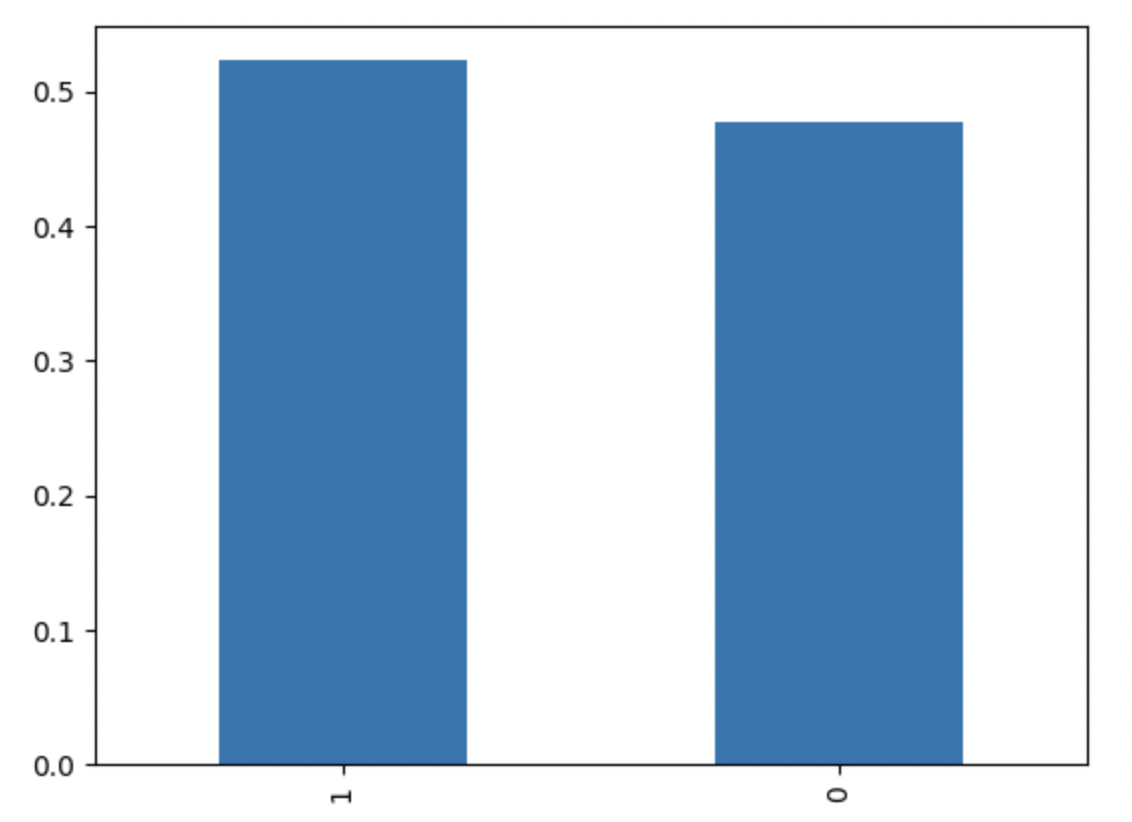
[**Recomendación**](#_3wvwr9jd12ai) **5**

# Descubrimientos

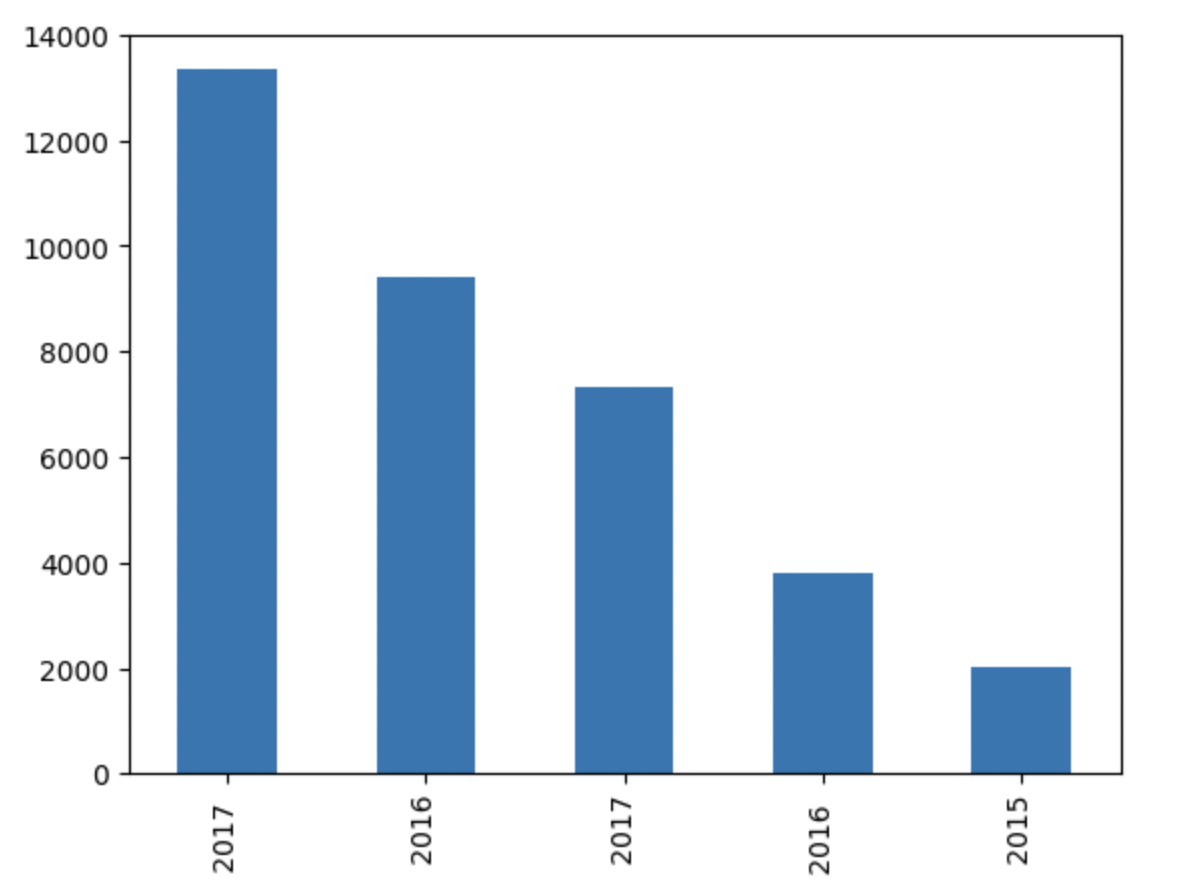
El objetivo de esta práctica es poder crear un modelo de clasificación de texto, el cual nos permitiese saber qué noticias son verdaderas o falsas dado un cierto grupo de datos.

## Limpieza y Exploración de Datos

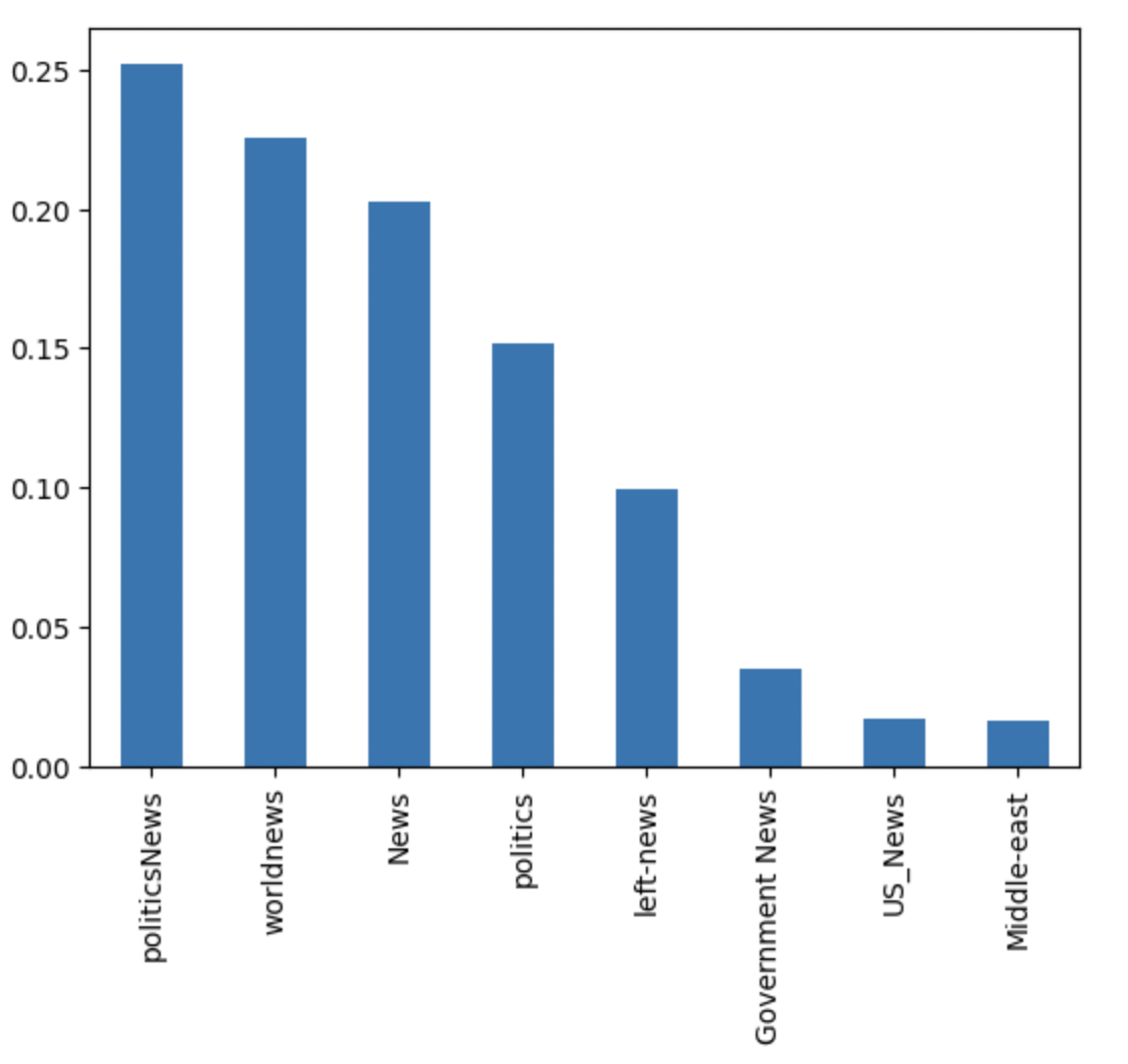
Para poder realizar este modelo, tenemos que realizar una limpieza de nuestro conjunto de datos de entrenamiento. Para ellos, al momento de ver cuántas noticias verdaderas/falsas hay, nos dimos cuenta que hay 18,778 fake news y 17,140 noticias reales, esto nos dice que más del 50% de nuestras noticias son falsas a una relación de ~52% y ~48% respectivamente.



De igual manera, revisamos los año en los cuales las noticias salieron, esto para saber si había un cierto sesgo o algún motivo detrás de ello; a nuestra sorpresa, vimos que hay cerca de 20,000 registros sólo del 2017, mientras que el siguiente año con más registros es 2017 con cerca de la mitad.

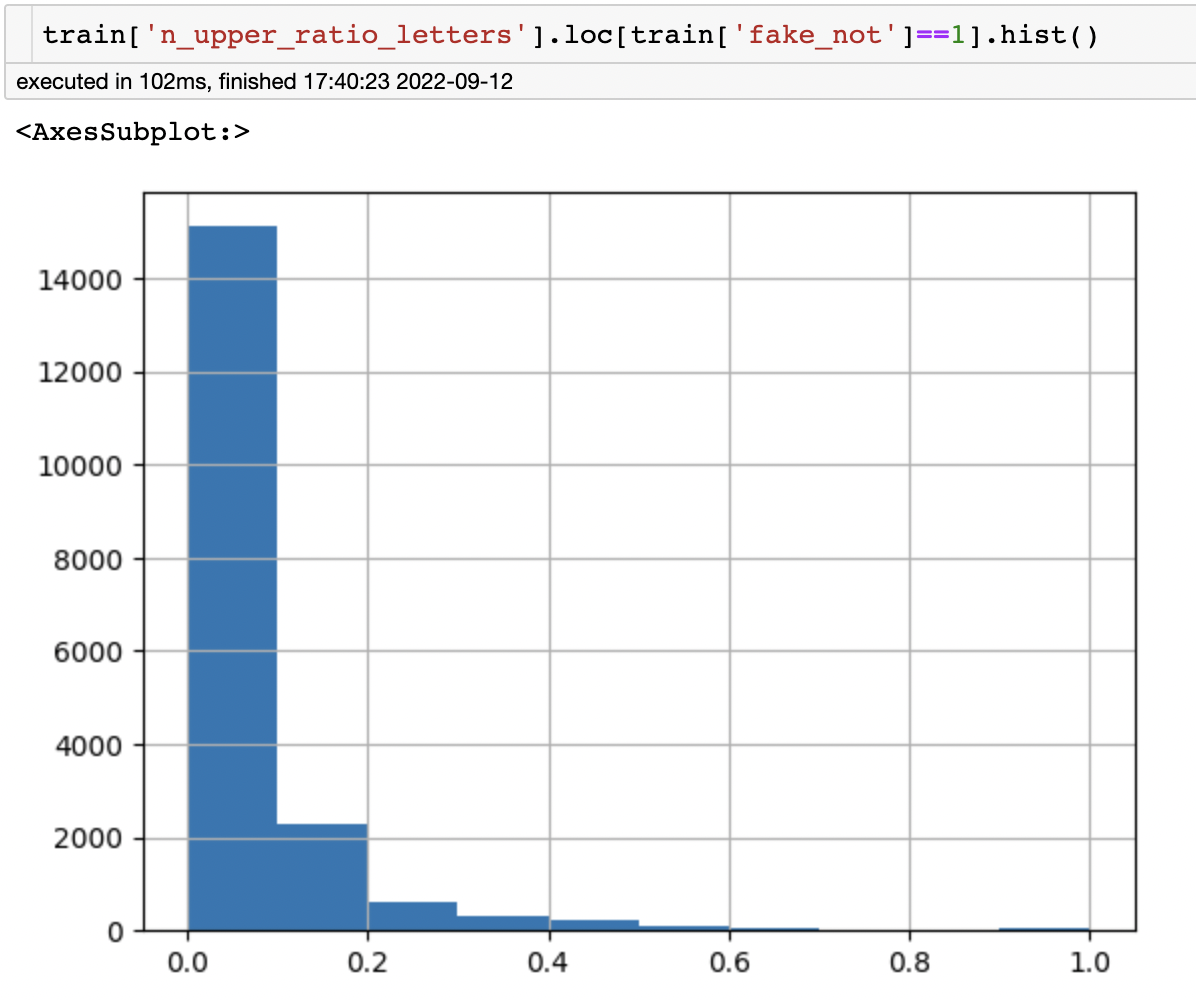
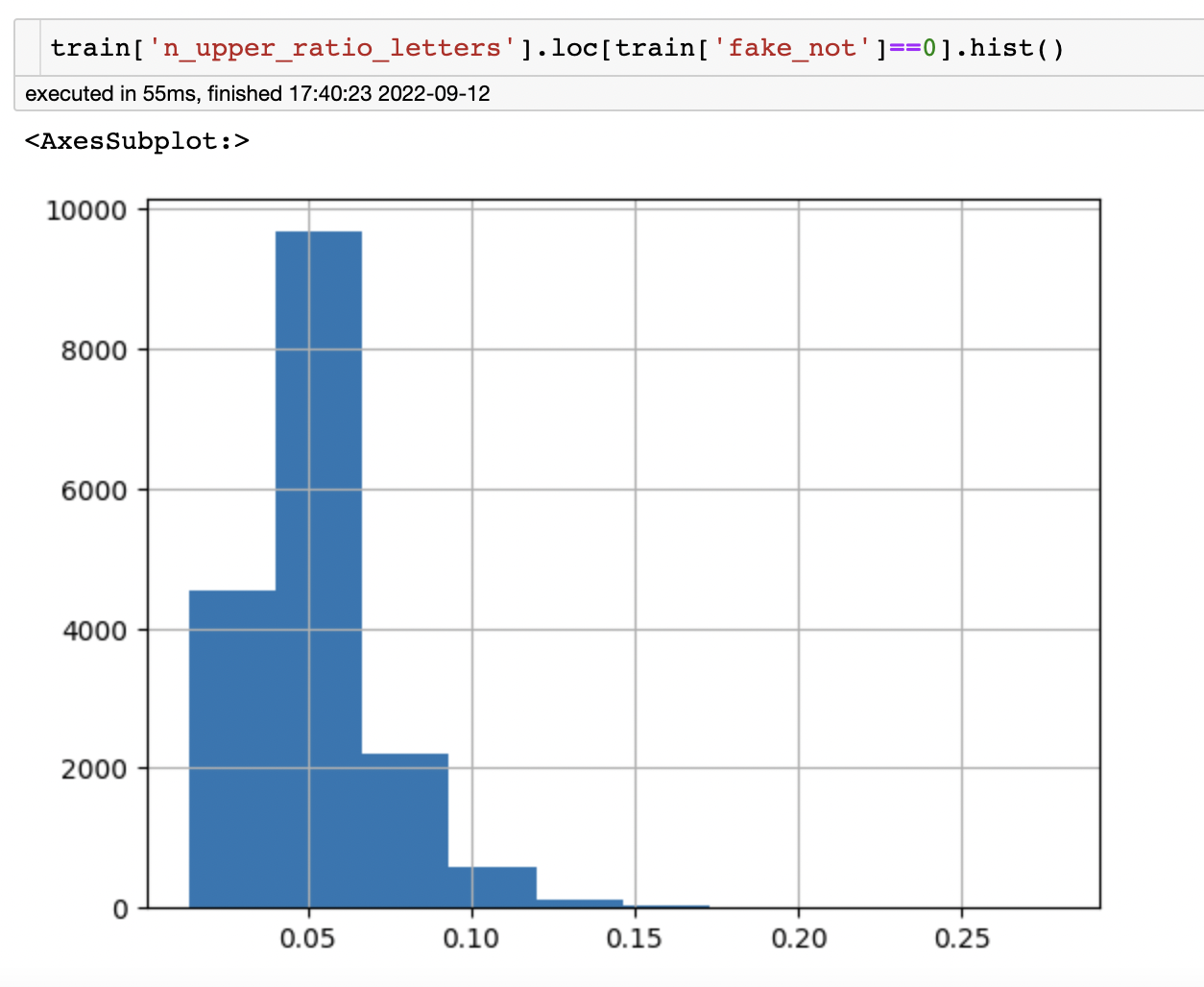


Finalmente, buscamos cuál es el tema con más registro dentro de nuestro conjunto de datos; a lo cual, nos dimos cuenta que la mayor cantidad de datos es referente a temas políticos con más del 25%.



Luego de esta exploración inicial, procedimos a limpiar nuestros datos. Para ello, vimos que no tenemos datos faltantes (missings), luego empezamos a normalizar los textos al quitar acentos y poner todo en minúsculas.

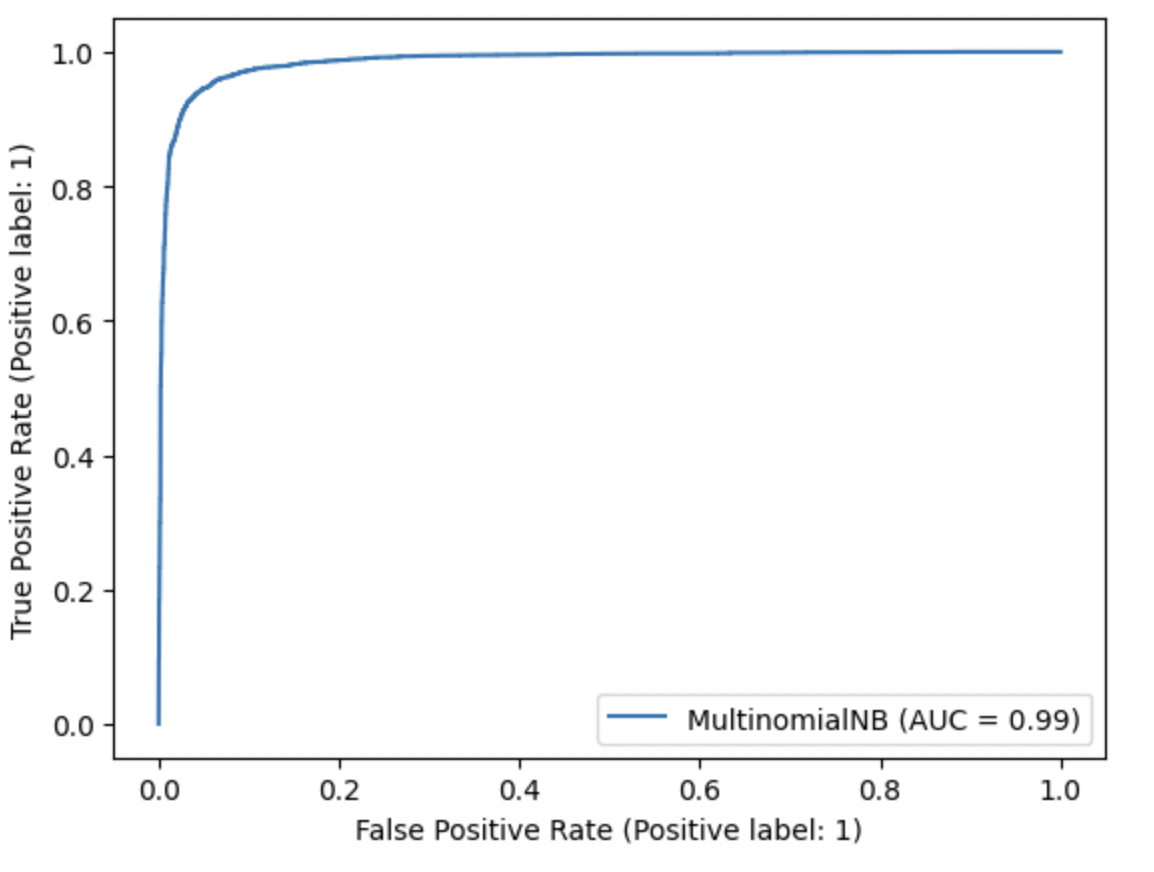
Después, sacamos el ratio de letras en mayúsculas vs letras en minúsculas. Al hacer esto, podemos ver el comportamiento con las noticias falsas. En este ejemplo, podemos darnos cuenta que las noticias reales (imagen izquierda) tienen menos varianza que las noticias falsas, las cuales incluso pueden tener valores cercanos a 1 (todas las letras con mayúsculas).



Después de un proceso de normalización, procedimos a seleccionar nuestras mejores variables para el modelo mediante el método *SelectKBest,* la cual nos arrojó las siguientes como mejores variables:

* N\_lower\_ratio\_letters
* N\_upper\_ratio\_letters
* Even
* Know
* like
* reuters
* said
* us
* via
* video

Luego, creamos nuestro modelo con Bayes, el cual nos entrenó el modelo, resultando esta curva ROC.



Al momento de utilizar el modelo con nuestros datos de testing, conseguimos 4,468 noticias reales y 4,512 noticias falsas, el cual nos da una relación cercana a 50/50.

# Recomendación

Con este modelo, se puede apoyar el blog de noticias y hacer “double check” con algunos criterios ya reconocidos en las noticias falsa; por ejemplo, la relación de letras mayúsculas vs total de letras.