¿Donald Trump o Hillary Clinton? Prediciendo quién es quién con tweets

Nerea Pérez Ruiz Jialian Zhou He José Jaén Delgado

1 Introducción

Hoy en día millones de usuarios utilizan las Redes Sociales (RRSS) como principal medio de información, jugando éstas un papel imprescindible en la creación de tendencias y opiniones. Debido a ello, no es de extrañar que cada vez más políticos se sirvan de las RRSS para comunicar sus ideas y lograr apoyos.

En este proyecto analizamos miles de publicaciones en Twitter de Donald Trump y Hillary Clinton al objeto de estudiar los rasgos lingüísticos propios de cada uno, distinguiendo la forma de expresarse de ambos para así predecir con precisión la autoría de diferentes tweets. Adoptando las perspectivas Frecuentista y Bayesiana al mismo tiempo que proponemos diferentes variables explicativas, presentamos un estudio de clasificación completo, acompañado de las interpretaciones y aclaraciones pertinentes para mayor claridad.

2 Descripción de los Datos

Para obtener los tweets de Trump y Clinton hemos recurrido a una de las páginas web más conocidas dentro de la comunidad Data Science: *Kaggle*. La base de datos con la que trabajaremos cuenta con un total de **6,444** publicaciones, de las cuales aproximadamente un 51% corresponden a la candidata demócrata, y el 49% restante a Donald Trump. Por tanto, nuestro estudio se centra en un problema de **clasificación binaria**, donde no hace falta recurrir a ténicas de remuestreo tales como el *Random Oversampling* o *Random Undersampling* puesto que los datos están equilibrados, esto es, el número de tweets de Trump y Clinton no difiere sustancialmente.

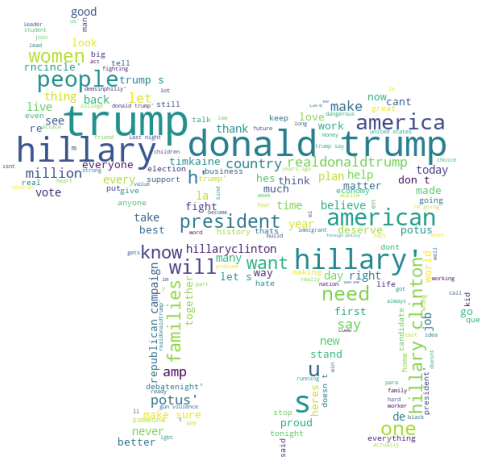
3 Metodología del Proyecto

A lo largo del presente estudio utilizaremos el software estadístico **R** y el lenguaje de programación **Python** para llevar a cabo tareas de Data Cleaning, Análisis Exploratorio de Datos y Modelización del Clasificador Bayesiano Ingenuo, al fin de preparar adecuadamente las variables explicativas que servirán de base para la obtención de resultados relevantes. Todas las operaciones siguen una lógica estadística que iremos exponiendo a lo largo del proyecto.

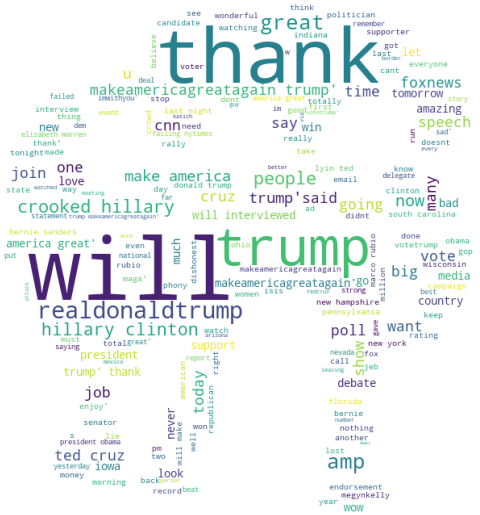
4 Data Cleaning y Análisis Exploratorio de Datos

Analizando más detenidamente las publicaciones con las que trabajaremos, encontramos que los tweets están repletos de símbolos, enlaces a páginas web y ciertos errores léxicos que dificultan nuestra tarea. Debido a ello, es necesario pulir todos los desperfectos para proseguir con nuestro proyecto.

En primer lugar hemos programado un filtro de URLs para eliminar los links de internet, ya que los candidatos son proclives a incluirlos en sus publicaciones de Twitter. A continuación, suprimimos números y puntuaciones, así como "stop-words" (conjunto de palabras utilizadas comúnmente en un lenguaje) y espacios en blanco. El resultado se muestra en las siguientes nubes de palabras:



(a) Hillary Clinton



(b) Donald Trump

Figure 1: Wordcloud de tweets

Tras el proceso de *Data Cleaning* es fácil observar que las palabras usadas por Hillary Clinton y Donald Trump son muy similares, dirigiéndose entre sí y mencionándose a ellos mismos. Además, los dos tratan de incitar al voto ("*Vote!*").

En cuanto a las diferencias apreciables, siguiendo una línea provocadora Trump se refiere a su contrincante de manera despectiva al llamarla "*crooked Hillary*". Asimismo, se centra bastante en su eslogan "*Make America Great Again*", en el futuro del país y sus promesas ("*will*"), al igual que se muestra agradecido con sus simpatizantes ("*thank*"). En el caso de Hillary, sus tweets giran entorno a la población, las mujeres y las familias. Por otra parte, Trump hace alusión a otros políticos estadounidenses, como Ted Cruz o Bernie Sanders, e incide bastante en la temática de los medios de comunicación como "*Foxnews*".

Consecuentemente, se puede deducir que aquellos tweets atinentes a la ciudadanía tenderán a ser de Hillary Clinton, mientras que aquellos que traten sobre repercusión mediática y ataques directos serán de Donald Trump con mayor probabilidad.

5 Clasificador Bayesiano Ingenuo

Llevaremos a cabo nuestra tarea mediante el Clasificador Bayesiano Ingenuo, método que aplica el Teorema de Bayes bajo una serie de supuestos:

$$\Pr(\text{Candidato}_i|\text{Palabra}_j) = \frac{\Pr(\text{Palabra}_j|\text{Candidato}_i) * \Pr(\text{Candidato}_i)}{\Pr(\text{Palabra}_j)}$$

- Independencia entre palabras e irrelevancia del orden de los términos lingüísticos: Suposición fuerte y restrictiva, que no obstante se cumple para la mayoría de locuciones.

- Identificacion binaria de palabras: El centro de atención se traslada desde el número de ocurrencias de palabras a la presencia o ausencia de las mismas en los tweets.

Dividiendo nuestra base de datos en dos conjuntos diferentes: *training set* (75% de los tweets) y *test set* (25% restante), realizamos un ejercicio de Machine Learning por el cual nuestro modelo aprenderá a clasificar publicaciones utilizando el primer conjunto de datos, siendo posible examinar su rendimiento al comparar los resultados con el test set.

El Clasificador Bayesiano Ingenuo admite la adopción tanto de la perspectiva Frecuentista como Bayesiana. La diferencia entre ambos se resume en que la clasificación a lo Frecuentista se basa sólo en los datos que aparecen en las publicaciones de Twitter sin considerar ninguna externalidad o inconveniente fáctico, mientras que la Estadística Bayesiana sí tiene en cuenta estas limitaciones. En esta última es posible incluir la opinión experta de investigadores, ya que da la opción de añadir una función **a priori** (criticado por los frecuentistas al inducir subjetividad). En este caso, suponemos una distribución Uniforme [0,1], lo que es equivalente a una distribución Beta(1, 1).

El clasificador basado en la Inferencia Frecuentista dará problemas con aquellas palabras no utilizadas en los tweets de uno de los candidatos pero sí por otro. Por ejemplo, "*brilliant*" aparece únicamente en las publicaciones de Trump, de ahí que el filtro frecuentista la asocie automáticamente y con probabilidad 1 a Trump. Esto no resulta muy razonable, dado que dicho término puede ser perfectamente empleado por Clinton: simplemente no ha sido recogido así en los datos. Siguiendo esta lógica, si alguna palabra no aparece en ninguno de los tweets estaríamos ante probabilidades nulas para ambos candidatos, siendo igualmente inverosímil.

Este problema se ve solucionado gracias a la distribución a priori de la Inferencia Bayesiana, incluida gracias a la **Suavización de Laplace**. Recordamos que nuestra distribución a priori era una Beta (1, 1), es decir, suponemos que las palabras tienen la misma probabilidad de pertenecer a los tweets de Trump que a los de Clinton. Por lo tanto, eliminamos el problema comentado anteriormente. De esta forma cada palabra tendrá una probabilidad de ocurrencia basada en la distribución a priori y la función de verosimilitud. Esto se podría interpretar como que los Métodos Bayesianos añaden un tweet ficticio con todas las palabras usadas en ambos grupos, al igual que otro mensaje completamente vacío. En el ejemplo anterior, ahora sí habría una ocurrencia de la palabra "*birlliant*" en el grupo Clinton. Así, a medida que los parámetros de la distribución a priori se alejen de cero, mayor será la diferencia obtenida entre el clasificador bayesiano y el frecuentista. Ilustrativamente:

Dada la variable aleatoria P_T (una palabra T es utilizada por Trump), y en el caso concreto de que dicha palabra se encuentre en tres mensajes de Trump y ninguno de Clinton (tomando la versomilitud la forma $f(\text{datos}|\theta_{P_T}) = \theta_{P_T}^3(1-\theta_{P_T})^0$):

Aplicando el Teorema de Bayes:

$$
\begin{aligned}
f(\theta_{P_T}|\text{datos}) &\propto f(\text{datos}|\theta_{P_T}) * f(\theta_{P_T}) \\
&\propto \theta_{P_T}^{3}(1-\theta_T)^0 * 1 \\
&\propto \theta_{P_T}^{4-1}(1-\theta_{P_T})^{1-1} * 1 \\
&= \frac{1}{\text{B}(4,1)}\theta_{P_T}^{4-1}(1-\theta_{P_T})^{1-1}
\end{aligned}
$$

Por tanto, $f(\theta_{P_T}|\text{datos}) \sim \text{B}(4,\ 1)$, y descartamos escenarios con probabilidades extremas (iguales a cero o uno).

6 Modelo de Clasificación TF-IDF

En esta sección proponemos una variable explicativa alternativa a la presencia de palbras por tweet denominada *Term Frequency - Inverse Document Frequency* (TF-IDF). Dicho método permite reflejar la importancia de cada palabra, compensando el número de veces que una palabra aparece con la frecuencia de esta misma en el corpus.

$$
\text{TF-IDF} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} * \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)
$$

Se puede apreciar cómo la primera parte del producto refleja el *Term Frequency*, el cual es ajustado por el logaritmo de la ratio entre el número de tweets totales y el conjunto de tweets en los que se encuentra dicha palabra (*Inverse Document Frequency*).

Introduciendo el TF-IDF como regresor en un modelo de clasificación logística estimado mediante Máxima Verosimilitud y regularizado por *Elastic Net*, obtenemos un rendimiento claramente superior al de los clasificadores anteriores.

Utilizar el porcentaje de categorías que los clasificadores predicen correctamente (*Accuracy*) no es suficiente, por lo que empleamos otras medidas de rendimiento como *Precision* (probabilidad de que un tweet recuperado al azar sea relevante), *Recall* o *Exhaustividad* (probabilidad de que un tweet relevante sea recuperado en una búsqueda) y *F1 Score* (media armónica de Precision y Recall del modelo).

Nótese que incluimos TF-IDF en el apartado de Estadística Frecuentista por convención, en realidad no puede compararse en términos de perspectivas estadísticas con el Clasificador Bayesiano ingenuo al tratarse de una técnica diferente.

Table 1: Comparación del Rendimiento de Perspectivas y Modelos

	Estadística Frecuentista		Estadística Bayesiana
	Bayesiano Ingenuo	TF-IDF	Bayesiano Ingenuo
Accuracy	0.88	**0.929**	0.92
(Accuracy Five Word)	(0.89)	-	(0.9)
Precision	0.89	0.925	**0.927**
(Precision Five Word)	(0.885)	-	(0.894)
Recall	0.83	**0.937**	0.916
(Recall Five Word)	(0.9)	-	(0.912)
F1 Score	0.88	**0.931**	0.921
(F1 Score Five Word)	(0.89)	-	(0.9)

Nota: Los números entre paréntesis hacen referencia a los modelos que exigen al menos cinco ocurrencias de palabras por tweet. En negrita el rendimiento máximo

7 Conclusión

El Clasificador Bayesiano Ingenuo empleado en este proyecto es un método que se ha demostrado efectivo para predecir la autoría de los tweets de los candidatos Trump y Clinton. Descansando sobre supuestos aparentemente restrictivos, encontramos que en la práctica no resultan del todo irrealistas.

Encontramos que el tipo de perspectiva metodológica adoptada es asimismo imperativa. Mientras que la Estadística Frecuentista se muestra incapaz de salvar obstáculos tales como los valores extremos de probabilidades estimadas (se necesitaría una base de datos considerablemente grande para que no se diera este inconveniente), la Inferencia Bayesiana aporta soluciones interesantes y flexibles para dar respuesta a este problema.

La clave se encuentra en la **Suavización de Laplace**, que en la práctica funciona como mensajes artificiales que se introducen en el corpus para evitar que las palabras sin registrar no se queden sin asociar a ninguna categoría o se clasifiquen erróneamente al único grupo en el que aparecen.

Por último, hemos propuesto un modelo de clasificación alternativo a fin de obtener resultados más precisos. El TF-IDF supera en este caso concreto al Clasificador Bayesiano Ingenuo, al ajustar la importancia de las palabras en relación a sus ocurrencias y relevancia lingüística.