Tecnologías emergentes y datos abiertos:
PROCESAMIENTO DE LENGUAJE NATURAL

INSPIRE

Contenido elaborado por Alejandro Alija, experto en Transformación Digital y datos abiertos.

Abril 2020

INSPIRE

En esta sección veremos con más detalle algunos de los **casos de uso particulares** del **Procesamiento del Lenguaje Natural y sus aplicaciones prácticas**. Varios casos de uso descritos en el primer informe de la serie ***Awareness, Inspire y Action*** , titulado Tecnologías emergentes y datos abiertos: Inteligencia Artificial, tienen relación directa con las tecnologías de Procesamiento del Lenguaje Natural. Sin duda, la evolución presente y futura del NLP, depende en gran medida de los últimos avances en Inteligencia Artificial.

Algunos ejemplos de casos de uso ya han sido introducidos en la sección de *Awareness*. Ahora es el momento de profundizar en la predicción de texto, la clasificación de textos y la detección de fake news, tres interesantes ejemplos que nos dan una visión bastante amplia de las posibilidades de estas tecnologías.

Predicción de texto

Quizás uno de los avances más palpables en el campo del Procesamiento del Lenguaje Natural sea la **predicción de texto**. En los últimos meses hemos visto cómo las últimas actualizaciones de los principales clientes de correo electrónico y motores de búsqueda, traían consigo **una funcionalidad sorprendente**. Cuando escribimos las primeras letras (incluso antes) de un email o una búsqueda web, el motor de procesamiento de lenguaje natural, utiliza un modelo entrenado que predice las palabras que vienen a continuación en un ranking de probabilidad. Es sorprendente el buen funcionamiento de esta funcionalidad y nos confirma que los humanos funcionamos en base a patrones de comportamiento regular. Cada persona tiene una huella sobre cómo se expresa, las oraciones que utiliza con más frecuencia, los comienzos y finales de las conversaciones, etc.

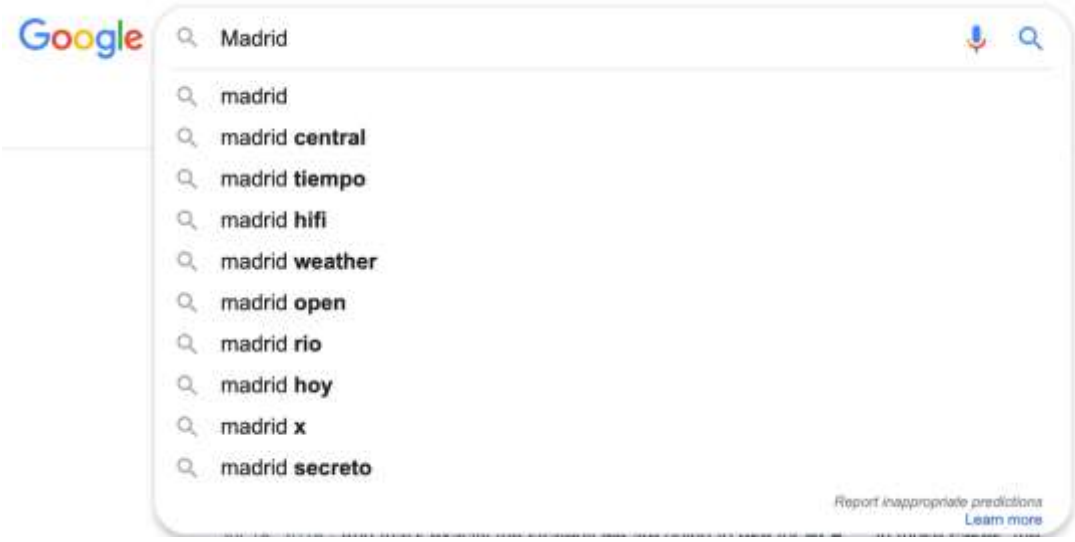

Figura 7. *Motor de palabras relacionadas de Google.*

En la sección *Action* veremos cómo el algoritmo RAKE (Rapid Automatic Keyword Extraction) permite extraer **conjuntos de términos** que aparecen juntos o correlativos con mayor probabilidad.

Clasificación de textos

La clasificación automática de textos es vital para muchas empresas y aplicaciones en la actualidad. La **detección de correos electrónicos fraudulentos** es un caso particular de la clasificación de textos. También lo son las **valoraciones de comentarios online**, por ejemplo, para valorar productos en base a los comentarios de las personas usuarias.

Los clasificadores de texto son tan útiles en la actualidad por varias razones:

1) **Son rápidos y sencillos.** Los algoritmos de clasificación de texto lineales son tan simples (en comparación con los modelos de Inteligencia Artificial más complejos como, por ejemplo, redes neuronales recurrentes) que se pueden entrenar rápidamente en un ordenador corriente. Además, la diferencia de precisión entre este tipo de algoritmos y otros mucho más complejos es casi nula, haciendo que no merezca la pena el sobreesfuerzo en el entrenamiento.

2) **Son casi independientes del idioma.** Para un algoritmo de clasificación no importa en qué idioma está el texto siempre que pueda separarse en palabras y medir los efectos de esas palabras.
3) **Su precisión es muy alta,** casi comparable con los seres humanos. El número de falsos positivos en detección de correo fraudulento o en identificación del género de una película o libro es muy bajo, casi despreciable.

Figura 8. *Flujo simplificado de un motor de detección y clasificación de correo fraudulento.*

La forma más sencilla de crear tu propio clasificador de textos es utilizar la herramienta de código abierto fastText, creada originalmente por Facebook.

Fake News

No todas las aplicaciones del Procesamiento del Lenguaje Natural tienen el objetivo de servir a un buen propósito. Recientemente estamos asistiendo a la explosión de las **fakes news o noticias falsas**: información falsa creada de forma deliberada y publicada a través de las redes sociales o diarios electrónicos con el objetivo de polarizar la opinión pública en un determinado sentido u orientación.

Es ya un hecho que las fake news se han utilizado de forma masiva en campañas de desprestigio político. Además de las fake news, recientemente, han salido a la luz los conocidos **deep fakes o videos falsos** que aplican la misma fórmula para modificar videos en vez de textos. En estos vídeos, se sustituye el rostro y la voz del verdadero protagonista por otra persona, habitualmente famosa, de la que existe gran cantidad de fotografías y audios en Internet. Si bien la calidad de los deep fakes está lejos de ser perfecta, en el caso de las fake news escritas es prácticamente imposible distinguirlas de las noticias reales salvo que se sepa específicamente que el contenido es falso.

En febrero de 2019, OpenAI anunció una nueva arquitectura de Procesamiento del Lenguaje Natural llamada **GPT-2**. GPT-2 es capaz de **generar fragmentos de texto absolutamente realistas** a partir de tan solo un par de palabras cómo comienzo de la frase.

Los resultados de GPT-2 son tan sorprendentes como inquietantes. Sus propios creadores dicen en su página web *Politicians may want to consider introducing penalties for the misuse of such systems, as some have proposed for deep fakes. (*Los políticos pueden considerar introducir sanciones por el mal uso de tales sistemas, como algunos han propuesto para los deep fakes).

Veamos un ejemplo de lo que GPT-2 es capaz de hacer. Adam Geitgey, en su blog de Medium, nos enseña un ejemplo muy ilustrativo.

Si utilizamos GPT-2 con un fragmento de texto inicial como **Abraham Lincoln** se genera una oración que se ajusta a este personaje histórico:

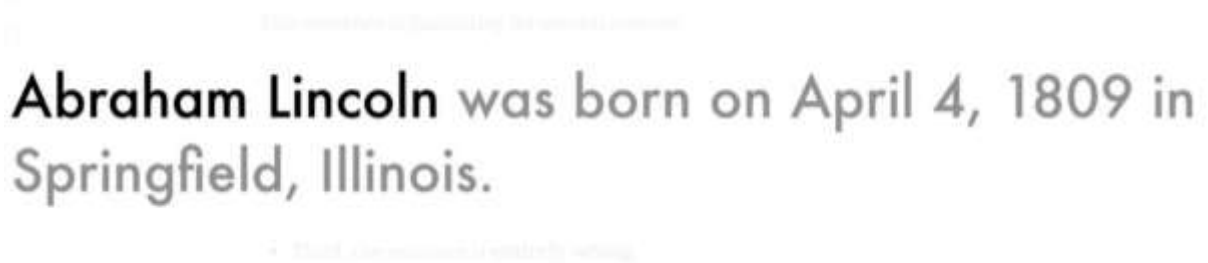

Figura 09. *Frase generada automáticamente por el algoritmo GPT-2.*

La frase es fascinante por varias razones:

- **Primero**, muestra que el modelo ha *entendido* que Abraham Lincoln era una persona (figura histórica) en Estados Unidos que nació en 1809.
- **Segundo**, la oración está perfectamente escrita e indistinguible de algo escrito por un ser humano.
- **Tercero**, la afirmación es completamente falsa.

Abraham Lincoln nació en Estados Unidos el 12 de abril de 1809 en Kentucky y no el 4 de abril en Illinois. Pero, sinceramente, yo no lo sabía, lo he tenido que consultar en Wikipedia y en ningún momento me ha parecido sospechoso ni alarmante la afirmación.

A pesar de los usos maliciosos que se puedan derivar del uso de esta tecnología, los algoritmos que subyacen son siempre un arma de doble filo. De la misma forma que generan noticias falsas con un asombroso realismo, se pueden reorientar a la detección de noticias falsas y la lucha contra las emergentes fake news.