Text Mining en Social Media. Master Big Data Analytics 2022-2023

Autores
Jonatan Victor Pace, Manuel Martín Castrillo
ChatGPT (Co-autor)

Abstract

La escritura generada por inteligencia artificial (IA) ha transformado la forma en que trabajamos, aprendemos y redactamos. En este estudio, abordamos el desafío de distinguir textos escritos por humanos de aquellos generados por inteligencia artificial (IA). Utilizando técnicas de vectorización y clasificación supervisada, desarrollamos un modelo que logra una precisión del 68% en la detección de estilos de escritura.

Si bien estos resultados iniciales son alentadores, reconocemos que aún queda trabajo por hacer. La precisión obtenida muestra que hay espacio para mejorar y continuar investigando en este campo. La capacidad de diferenciar entre escritura humana y generada por IA tiene implicaciones significativas en áreas como la autenticidad del contenido y la protección de la propiedad intelectual.

1 Introducción

El problema que abordamos en este informe se basa en identificar si un texto fue escrito por un humano o por una máquina entrenada en la generación de lenguaje natural. Con el objetivo de resolver esta tarea empleamos técnicas de minería de datos y aprendizaje supervisado con el fin de comprender qué características y diferencias tienen los textos analizados, entrenar un modelo de clasificación capaz de capturar estas particularidades y posteriormente poder utilizarlo para predecir si una nueva entrada de texto es de origen humano o generado por una máquina.

2 Dataset

El corpus utilizado para el análisis y entrenamiento de modelos consiste en 32.062 entradas de textos con las siguientes características:

- id: identificador único de cada entrada en el dataset.
- prompt: descripción o categoría de la entrada. Puede ser ”NO-PROMPT” (sin descripción) u otra descripción específica.
- text: texto o contenido de la entrada.
- label: indica si el texto fue generado por algún modelo de lenguaje (”generated”) o fue escrito por un ser humano (”human”). Esta característica es la clase a predecir.
- model: indica el modelo utilizado para los casos en que los textos son generados por un modelo de lenguaje.
- domain: dominio temático al que pertenece la entrada, como ”legal”, ”wiki” (Wikipedia) o ”tweets”.

El dataset se considera balanceado dado que las diferentes clases tienen una distribución similar:

Label	Total de registros	% sobre el total
generated	16.275	50.7
human	15.787	49.3

Así mismo disponemos de un conjunto de datos de prueba de 20.129 entradas de texto sin etiquetar. Este conjunto de datos nos permite validar qué tan buena son nuestras soluciones propuestas en un entorno similar al de producción.

3 Propuesta del alumno

Hemos abordado el problema de clasificación centrándonos en el procesamiento del atributo text, dejando fuera del análisis el resto de características (id, prompt, model y domain). Para

procesar este atributo hemos aplicado dos técnicas de vectorización diferente, la primera basada en contar la frecuencia de ocurrencia de cada palabra presente en el corpus y la segunda basada en calcular un peso para cada término teniendo en cuenta su frecuencia en el documento y su rareza en todo el corpus de texto.

Además, hemos aplicado diferentes tipos de n-gramas tanto de palabras como de caracteres con el fin de observar las diferencias de rendimiento de los modelos de clasificación al intentar capturar el tema o el estilo de escritura.

Luego del procesado de los textos, hemos dividido el corpus en dos conjunto de datos, uno para el entrenamiento y otro para la validación del modelo. Para este estudio hemos utilizado dos modelos de clasificación: regresión logística (LR) y support vector machine (SVM).

Finalmente hemos seleccionado el mejor modelo teniendo en cuenta el valor de la métrica f1-macro en fase de entrenamiento-validación. En dicha fase hemos utilizado la técnica de validación cruzada y búsqueda de hiper parámetros. Luego, hemos entrenado nuevamente el modelo final con todo el corpus etiquetado disponible y utilizando los mejores parámetros obtenidos.

4 Resultados experimentales

Para la realización de los experimentos hemos utilizado las librerías CountVectorizer (Cvect) y TfidfVectorizer (Tvect) de scikit-learn para la vectorización de los textos utilizando variantes en la parametrización de las mismas. Para todas las variantes hemos mantenido las primeras 5.000 características. Así mismo hemos aplicado n-gramas de palabras (para el desarrollo de un modelo capaz de capturar tópicos o temas) y n-gramas de caracteres (modelo para capturar estilos de escritura).

A continuación se muestra un resumen de las variantes aplicadas:

Vectorización	Analizador	Rango n-gramas
Cvect	word	1-1
Cvect	char	1-2
Cvect	char	1-3
Tvect	char	1-2
Tvect	char	2-3

Los resultados obtenidos al utilizar diferentes clasificadores han sido los siguientes:

Clasificador	Vectorización	f1-macro
SVM	Tvect char 2-3	**0.832**
LR	Tvect char 1-2	0.794
LR	Cvect char 2-3	0.788
LR	Cvect word 1-1	0.784
LR	Cvect char 1-2	0.728
LR	Tvect char 1-2	0.727

Si bien el mejor modelo obtenido en fase de entrenamiento-validación se ha logrado utilizando support vector machine hemos optado por el segundo mejor modelo (regresión jogística) para avanzar con la búsqueda de hiper parámetros, ya que el tiempo de entrenamiento era considerablemente menor que el SVM.

Los modelos finales que hemos utilizado para procesar el conjunto de datos de prueba fueron los siguientes:

1. Support vector machine con vectorización TF-IDF entrenado con el 70% del corpus etiquetado.

2. Regresión logística con vectorización TF-IDF, entrenado y parametrizado con validación cruzada y luego reentrenado con el 100% del corpus disponibles con un umbral de predicción del 50% para clasificar las instancias como generadas.

3. Regresión logística con vectorización TF-IDF, entrenado y parametrizado con validación cruzada y luego reentrenado con el 100% del corpus disponibles con un umbral de predicción del 80% para clasificar las instancias como generadas.

Los resultados obtenidos luego de procesar el conjunto de datos de prueba han sido los siguientes:

Modelo	f1-macro
LR (umbral 80%)	**0.689**
LR (umbral 50%)	0.628
SVM (umbral 50%)	0.509

5 Conclusiones y trabajo futuro

Los mejores resultados obtenidos sobre el conjunto de datos de prueba se han logrado con un modelo de regresión logística vectorizado con un rango de caracteres de 2 a 3 y con el umbral

de predicción modificado para etiquetar los textos como generados si la probabilidad de dicha clase es mayor al 80%. Esta modificación ha dado buenos resultados ya que ha logrado balancear las clases predichas en producción en la misma proporción que hemos observado en la fase de validación de los modelos. En contraposición, el modelo con el umbral al 50% ha predicho como textos humanos el 20% del total del conjunto de datos de pruebas, al igual que el support vector machine, este desbalanceo ha resultado que el modelo no funcione bien en un entorno de producción.

Consideramos que detectar con alta confiabilidad si un texto ha sido generado por un máquina es una tarea compleja que requiere más investigación. Como trabajos futuros podemos centrarnos en aplicar una búsqueda de hiper parámetros sobre el modelo de support vector machine y predecir el conjunto de prueba utilizando un umbral del 80%. También podemos experimentar aumentando el número de características utilizadas en la vectorización para ver si es posible capturar mejor el estilo de escritura en los textos como así también utilizar en forma conjunta la vectorización por palabras y caracteres.