SEGUNDA ENTREGA:
PREDICCIÓN DE LA VERACIDAD DE PROPUESTAS LABORALES

UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

EQUIPO DE TRABAJO

SAMUEL A. ARISTIZÁBAL GONZÁLEZ

JULITZA DAZA ZAPATA

JUAN DAVID LOPEZ AGUIRRE

PROFESOR ENCARGADO: RAUL

RAMOS POLLAN

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS Y LAS INGENIERÍAS

MEDELLÍN

2023

Predicción de la Veracidad de Propuestas Laborales

Samuel A. Aristizábal González, Julitza Daza Zapata Juan David López Aguirre; Estudiantes U de A

I. Problema por resolver

De un dataset con dieciocho mil descripciones laborales de las cuales ochocientas son falsas se espera realizar un modelo capaz de clasificar las propuestas laborales falsas de las auténticas.

II. Dataset

El dataset escogido fue tomado de la plataforma Kaggle con el nombre de "Real / Fake Job Posting Prediction". Las muestras tienen como columnas/variables las siguientes características:

1. **Job_id**
2. **title** - Título de la publicación
3. **location** - Ubicación del lugar de trabajo
4. **department** - Departamento de trabajo
5. **salary_range** - Rango de salario
6. **company_profile** - Perfil de la compañía
7. **description** - Descripción del trabajo
8. **requirements** - Requerimientos para el trabajo
9. **benefits** - Beneficios con el empleo
10. **telecommuting** - 0 o 1 si se realiza teletrabajo
11. **has_company_logo** - Si tiene logo de la empresa (0 o 1)
12. **has_questions** - Si tiene preguntas (0 o 1)
13. **employment_type** - Tipo de empleo (tiempo completo, medio tiempo, entre otros)
14. **required_experience** - Experiencia requerida
15. **required education** - Educación requerida
16. **industry** - industria en donde se desarrolla el trabajo
17. **function** - Función del trabajo
18. **fraudulent** - Característica de salida (0 o 1)

Con el objetivo de llenar las 30 columnas mínimas pedidas en el enunciado se llenaron 12 variables más con las librerías pandas y random:

19. **country –** País de origen
20. **special_characters –** (0 o 1) si tiene caracteres especiales
21. **phone_number –** (0 o 1) si tiene número de contacto
22. **website –** (0 o 1) si tiene pagina web
23. **longevity** – Cuantos años lleva la empresa registrada en la plataforma
24. **vacants –** Numero de vacantes
25. **contract_type -** Definido o indefinido
26. **comercial_id -** (0 a 1) Existe un número de registro de la empresa en base de datos gubernamentales o no
27. **multinational –** (0 o 1) Es multinacional o no
28. **number_employees** – (0 a 1) Informa sobre el número de empleados actuales o no
29. **calification** – (0 a 1) Existen reseñas de los clientes en otras plataformas virtuales o no
30. **deadline –** (0 a 1) Hay fecha límite de la inscripción para la oferta

III. AVANCES

El código presentado realiza una serie de operaciones para limpiar y preprocesar un conjunto de datos que contiene información sobre ofertas de trabajo. El objetivo principal es convertir los datos al formato, adecuado para su posterior análisis y modelado.

Primero, el conjunto de datos de producción se carga en un objeto Pandas DataFrame usando la función read_csv(). Luego se crea una copia del registro original para evitar modificar los datos originales. Luego, los datos se escanean para identificar los objetos que tienen más del 60% de valores nulos y se han eliminado del conjunto de datos original. Esto se hace para reducir el tamaño del conjunto de datos y centrarse en funciones para las que hay una cantidad significativa de datos disponibles. Luego, los valores nulos se completan con una cadena vacía usando la función fillna(). Esto se hace para garantizar que los valores nulos no afecten el análisis posterior del conjunto de datos.

Algunas columnas que no son útiles para análisis posteriores se eliminan usando la función drop() En este caso, las columnas que se eliminaron son "teletrabajo", "has_company_logo", "has_questions" y "job_id". Luego se combinan varias columnas del conjunto de datos en una sola columna llamada "Texto". Esto se hace para recopilar todas las funciones necesarias en un solo lugar y facilitar el procesamiento previo del texto posterior.

Luego se realizan una serie de operaciones de limpieza en el texto. Primero, los saltos de línea y los espacios, que son caracteres de tabulación, se reemplazan con espacios vacíos. Luego, los números y caracteres especiales se eliminan utilizando expresiones regulares y la función apply(). Finalmente, convierta todo el texto a minúsculas usando la función apply().

En general, el código representa intentos de limpiar y preprocesar datos antes de analizarlos y modelarlos. Eliminar características con muchos valores nulos, fusionar características relacionadas en una sola columna y eliminar caracteres especiales y números son métodos comunes para limpiar el texto antes de realizar un análisis posterior.

IV. CONCLUSIONES

En conclusión, el código presentado realiza una serie de operaciones de limpieza y preprocesamiento de un conjunto de datos que contiene información sobre publicaciones de trabajo. La limpieza se realiza de esta manera debido a que los datos son en lenguaje natural, lo que significa que pueden contener una gran cantidad de ruido y características irrelevantes que pueden afectar la calidad del análisis posterior. La eliminación de características con muchos valores nulos, la combinación de características relevantes en una sola columna, la eliminación de caracteres especiales y números, y la conversión de todo el texto en minúsculas son técnicas comunes para limpiar el texto antes del análisis posterior. Estas técnicas permiten reducir el ruido en el conjunto de datos y asegurar que las características relevantes se identifiquen con precisión. Es importante tener en cuenta que la limpieza y preprocesamiento de datos es una parte esencial del proceso de análisis de texto y modelado. Si no se realiza una limpieza adecuada de los datos, los modelos resultantes pueden ser imprecisos y no proporcionar resultados útiles. Por lo tanto, es importante seguir buenas prácticas de limpieza de datos para garantizar que los modelos de análisis de texto sean precisos y útiles para la toma de decisiones.