Alumno: Jorge de Jesús Cazares Duran

ESTUDIO SOBRE ALGORITMOS DE CLASIFICACIÓN DE TEXTOS: DEL TEXTO A LAS PREDICCIONES

La clasificación de texto (TC) es una técnica popular en el procesamiento del lenguaje natural (NLP), que se usa para asignar etiquetas a textos y usada en una variedad de metodologías especializadas. De acuerdo con lo reportado por Gasparetto y colaboradores (2022), su uso se ha ampliado debido a la gran cantidad de datos textuales que se generan diariamente. Entre las tareas comunes en NLP se encuentran el análisis de sentimientos (SA), la asignación de temas (TL), la clasificación de noticias (NC), la respuesta a preguntas (QA), el reconocimiento de entidades nombradas (NER) y la inferencia de lenguaje natural (NLI). Estas tareas pueden variar dependiendo del contexto.

El preprocesamiento implica convertir texto no estructurado en un formato que las máquinas puedan comprender, permitiendo a los modelos hacer predicciones. Los primeros métodos dependían de la ingeniería manual de características, mientras que los enfoques más recientes extraen automáticamente las características del texto. Los enfoques clásicos de aprendizaje superficial, como los métodos tradicionales de aprendizaje automático, son útiles cuando se cuenta con recursos limitados y la ingeniería de características es costosa. Por su parte, los modelos de aprendizaje profundo pueden extraer representaciones semánticas más complejas sin necesidad de diseñarlas manualmente.

El primer paso del preprocesamiento es la tokenización, que implica dividir el texto en unidades más pequeñas llamadas tokens. Tradicionalmente, estos tokens han sido palabras, pero los métodos recientes descomponen el texto en unidades más pequeñas, como n-gramas de caracteres o bytes. Los métodos de tokenización modernos, basados en datos, proporcionan mejores resultados que los métodos tradicionales basados en reglas. Además de la tokenización, se pueden aplicar otras operaciones como la eliminación de caracteres innecesarios o palabras vacías, aunque estas operaciones deben realizarse con precaución, ya que pueden afectar negativamente a los modelos de aprendizaje profundo.

Los enfoques clásicos de TC pueden beneficiarse de procesos como la lematización, que simplifica las palabras reduciéndolas a una forma natural. Aunque estos procesos mejoran el rendimiento, tienen limitaciones, como no distinguir entre palabras con diferentes significados que comparten la misma raíz. También pueden aplicarse otras operaciones como el etiquetado de partes del discurso (PoS), que implica separar una oración en sus componentes, aunque los errores en esta tarea pueden afectar el rendimiento general.

La tokenización avanzada mejora la eficiencia y capacidad de los modelos ya que divide el texto en unidades más pequeñas y manejables. Existen distintos tipos de tokenización avanzada. El Byte Pair Encoding (BPE), tokeniza y posteriormente combina los pares de caracteres más frecuentes, hasta formar sub-palabras, esta reducción del vocabulario permite que el modelo maneje mejor las palabras. El WordPiece trabaja de forma similar a BPE, pero usa un enfoque diferente ya que identifica las palabras que son extensiones que se pueden aplicar a otras, permitiendo que el modelo pueda manejar tanto palabras conocidas como desconocidas. En contraste, UnigramLM combina unidades pequeñas, iniciando con un conjunto más amplio y descartando las que no aportan valor al contexto. Finalmente, la tokenización SentencePiece es una herramienta que puede trabajar tanto con BPE como UnigramLM pero sin la necesidad de una tokenización inicial basada en idioma y sin depender de espacios.

La proyección en el espacio de características (transformación del texto en vectores) es un paso fundamental para que los algoritmos puedan trabajar con datos textuales de manera eficiente. La

evolución desde métodos simples como Bolsas de Palabras (BoW) hasta los sofisticados word embeddings ha permitido a los modelos de aprendizaje capturar no solo la frecuencia de las palabras, sino también su significado y las relaciones semánticas entre ellas, lo que resulta en modelos mucho más poderosos y efectivos para tareas como la clasificación de texto, traducción automática y análisis de sentimientos.

Los modelos gráficos probabilísticos, como Naïve Bayes son populares porque son simples y efectivos, aunque están limitados por suponer que todas las partes de los datos son independientes entre sí, algo que no ocurre con el texto. La clasificación de textos con k-nearest neighbours (k-NN) identifica las instancias más similares para asignar categorías, pero su rendimiento depende de la función de distancia elegida y se ve afectado en espacios de alta dimensionalidad. Las máquinas de vectores soporte (SVM) son métodos de predicción que convierten tareas de clasificación en problemas binarios, mapeando datos a un espacio de mayor dimensión y mejorando la separación de categorías. Los árboles de decisión son clasificadores intuitivos que dividen jerárquicamente el espacio de datos, pero pueden ser sensibles al sobreajuste. Los bosques aleatorios, por otro lado, combinan múltiples árboles, ofreciendo un mejor rendimiento. La regresión logística (LR) es un clasificador lineal que predice probabilidades de clases y se adapta mejor a la clasificación binaria, aunque puede extenderse al caso multinomial. Los bagging y boosting integran resultados de múltiples algoritmos para mejorar el rendimiento e interpretación.

A pesar de que el aprendizaje profundo domina en el procesamiento del lenguaje natural, modelos poco profundos como FastText y GHS-NET han logrado resultados competitivos en clasificación de textos, utilizando n-gramas y combinaciones de redes neuronales convolucionales y recurrentes para tareas específicas.

Reseña

El procesamiento del lenguaje natural se ha visto revolucionado en los últimos años gracias al desarrollo de modelos que permiten realizar tareas de manera más eficiente. Uno de los pilares de este crecimiento es la tokenización avanzada, que ayuda a los modelos profundos a evitar la necesidad de almacenar o procesar vocabularios extremadamente grandes. Otro aspecto analizado de suma importancia son los word embeddings, que son herramientas poderosas para mejorar la comprensión y el manejo del lenguaje por parte de las máquinas, facilitando una variedad de aplicaciones en el ámbito del procesamiento del lenguaje natural.

El artículo analiza los métodos de clasificación de aprendizaje superficial más estudiados en años recientes, permitiendo identificar las ventajas y desventajas para la clasificación de texto. Los modelos Ensemble, que integran diferentes modelos, podrían considerarse como los más efectivos en ciertos casos debido a su capacidad para mejorar la precisión de las predicciones. Sin embargo, existen desafíos por mejorar como la diversidad de idiomas y la integración de múltiples tipos de datos.

Referencia

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, 13(2), 83.