Reporte: A survey on Text Classification Algorithms: From Text to Predictions

Eber David Gaytán Medina

8 Oct 2024

El artículo ofrece una revisión a los modelos de clasificación de texto actuales. La clasificación de texto (TC - por sus siglas en inglés) es un área que ha estado ganando importancia en el procesamiento del lenguaje natural (NLP). No solo se utiliza como clasificador o como extractor de información, tiene más usos que convierten a los TC en una herramienta con aplicaciones muy prácticas. Por ejemplo: análisis de sentimientos (SA), etiquetado de temas (TL), clasificación de noticias (NC), respuesta a preguntas (QA), inferencia de lenguaje natural (NLI), reconocimiento de entidad nombrada (NER) y análisis sintáctico (SP).

Una parte muy importante en la clasificación de texto es la representación de este texto en términos computacionales. El artículo está dividido entre los métodos más viejos, que suelen ser manuales y son llamados Shadow Learning Approaches (SLA), mientras que los últimos métodos, Deep Learning Approaches a base de deep learning suelen ser automáticos en aspectos como la extracción de características.

Los Shallow Learning Approaches o Enfoques de Aprendizaje Superficial son aquellos que preceden a las redes neuronales y dependen de las predicciones diseñadas manualmente. Mientras que, los Deep Learning Approaches o Enfoques de Aprendizaje Profundo son aquellos modelos más nuevos a base de redes neuronales que nos permiten extraer características complejas sin necesidad de hacer el diseño manualmente.

Una parte importante de los métodos de clasificación de texto, son las operaciones de preprocesamiento. Estas son divididas dentro del artículo con dos vertientes, las que usan una metodología estándar, utilizadas en extracción de características de manera manual y las utilizadas por modelos profundos. Los utilizados en el primer ramo son: Standard Preprocessing Operations, Tokenización, Stopword y Noise Removal. Standard Preprocessing Operations son operaciones que limpian y normalizan datos. La Tokenización es el proceso de descomponer, en este caso texto, en sus unidades atómicas de tal manera que tengan un valor semántico pero sin necesidad de una motivación lingüística o explicación. Stopword y Noise Removal aseguran que después de la tokenización se eliminen elementos innecesarios y engañosos, tales como símbolos y caracteres especiales que no aportan información semántica. Otro proceso es lla-

mado Further Standardization of Text, que consta de estandarizar el texto mediante legalización u otras técnicas, es decir, "niños" es el plural de "niño" por lo que se cambia a singular para obtener la forma canónica y estandarizada.

Por otro lado, el procesamiento con Deep Models hace uso de las técnicas antes mencionadas añadiendo mejoras sustanciales a la tokenización, enfocadas a reducir el tamaño de las matrices generadas al procesar los textos. Existen otras estrategias para abordar estas tareas y que merecerían sus propios reportes como: Byte Pair Encoding, WordPiece, UnigramLM o SentencePiece, todas ellas aportando capacidades más puntuales y dependientes de las necesidades.

La transformación de texto en listas de tokens estandarizados se hace mapeando palabras a un vocabulario indexado. Técnicas como Bag-of-Words (BoW), Modelos de Lenguaje, Word Embeddings como Word2Vec o GloVe, o también FastText, permiten a los métodos de clasificación de texto mejorar la capacidad de aprender, predecir o entrenar a estos modelos.

La lista de modelos de métodos de clasificación de aprendizaje superficial es: Probabilistic Graphical Models (PGM), k-Nearest Neighbours (k-NN), Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR), métodos de aprendizaje en conjunto y métodos basados en neuronas.

Desde el punto de los métodos de aprendizaje profundo (Deep Learning) para la clasificación de texto, se empieza explicando cómo el deep learning ha supuesto un cambio de paradigma al permitir la extracción automática de características importantes de los datos, superando las limitaciones de las ya mencionadas "superficiales". La lista de estos es la siguiente: Multilayer Perceptrons (MLP), Recurrent Neural Neteorks (RNN), Convolutional Neural Networks (CNN), también, modelos de lenguaje profundo para clasificación como: RNN Encoder-Decoders, arquitecturas Transformer, BERT, GPT, modelos Transformer de lenguaje recientes, redes neuronales con grafos y contextualizadores.

El artículo es robusto y extensivo, la cobertura de información es exhaustiva, se destaca el análisis detallado de los conceptos primordiales como tokenizacion y normalización de datos. Además al confrontar los métodos deep y shallow learning nos permite contrastar cada modelo y sus capacidades, esto termina siendo valioso para los investigadores y profesionales que buscan una comprensión integral de cómo preparar y representar el texto para diferentes algoritmos.

Sin embargo, no hay ejemplos concisos, la evaluación depende de los benchmarks y deja de lado la información empírica. Existen otras métricas como el costo computacional que aunque son mencionados no se encuentran representados en datos duros, siendo esto realmente importante en el diseño de nuevos experimentos.