Jesús Javier Gortarez Pelayo

Resumen: A Survey on Text Classification Algorithms: From Text to Predictions (Sección 1 a 4)

Debido al aumento en la cantidad de documentos digitales y al ser la clasificación de texto una de las principales tareas en el procesamiento del lenguaje natural (PNL), se ha impulsado el desarrollo de algoritmos de aprendizaje automático para automatizar la clasificación de texto. En particular, los avances en aprendizaje profundo han permitido la creación de algoritmos que extraen automáticamente características del texto, mejorando la precisión y eficiencia en la clasificación.

Este artículo revisa de manera breve los modelos recientes de clasificación de texto, destacando el flujo de datos desde el texto crudo hasta las etiquetas de salida. Se comparan métodos tradicionales con los basados en aprendizaje profundo, tanto en su funcionamiento como en la transformación de los datos.

Preprocesamiento en la Clasificación de Texto

El preprocesamiento del texto es fundamental en la clasificación, ya que transforma el texto sin procesar en un formato que los algoritmos de aprendizaje automático puedan manejar eficientemente. Este proceso implica limpiar y normalizar los datos, lo que mejora la precisión y la eficiencia de la extracción de características.

Entre las operaciones más comunes se encuentran la *tokenización* (dividir el texto en unidades más pequeñas), la *eliminación de palabras vacías* ("stop words") que no aportan significado, la *eliminación de ruido* (caracteres especiales, puntuación, etc.), y la *estandarización*, que incluye la derivación (reducir las palabras a su raíz) y la lematización (convertir las palabras a su forma canónica). Todas estas técnicas contribuyen a que los algoritmos puedan analizar el texto de manera más efectiva y obtener mejores resultados en la clasificación.

Sin embargo, gracias al deep learning, se tienen otras técnicas de tokenización como *BPE* (combina tokens frecuentes para crear subpalabras), *WordPiece* (similar a BPE, pero usa la probabilidad de las subpalabras), y *UnigramLM* (asigna probabilidades a las subpalabras) permiten crear subpalabras, lo que ofrece un mejor equilibrio entre el tamaño del vocabulario y la capacidad de representar palabras raras.

Proyección al Espacio de Características en la Clasificación de Texto

Tras el preprocesamiento, se necesita la proyección al espacio de características. Este proceso transforma las palabras en representaciones numéricas para un mejor manejo. Existen diversos métodos para lograrlo, cada uno con sus ventajas y desventajas.

2024/10/08 Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. Information, 13(2), 83. https://doi.org/10.3390/info13020083

Bag-of-Words (BoW): Este método representa los documentos como colecciones de palabras sin orden, basándose en la frecuencia de aparición. Para mejorar la representación, se utilizan técnicas como TF (frecuencia de una palabra en un documento) e IDF (importancia de una palabra en un conjunto de documentos).

Modelos de Lenguaje: Estos modelos, como los n-grams, asignan probabilidades a secuencias de palabras, capturando información sobre el orden y la estructura del lenguaje.

Word Embeddings: Representan las palabras como vectores densos en un espacio continuo, capturando relaciones semánticas y sintácticas. *Word2Vec* (usa una red neuronal para aprender vectores), *GloVe* (se basa en la co-ocurrencia de palabras) y *FastText* (considera subpalabras para capturar información morfológica) son ejemplos de word embeddings.

Métodos de Clasificación de Aprendizaje Superficial

Si bien el deep learning ha ganado popularidad, los métodos de aprendizaje superficial ("shallow learning") siguen siendo relevantes en la clasificación de texto. Su simplicidad, interpretabilidad y eficiencia computacional los hacen atractivos para diversas aplicaciones.

Algoritmos clásicos como *Naive Bayes* (un clasificador probabilístico que asume independencia entre las características), *K-Nearest Neighbors* o *KNN* (que clasifica según los vecinos más cercanos), *Máquinas de Vectores de Soporte* o *SVM* (que buscan un hiperplano óptimo para la separación de clases), *Árboles de Decisión y Bosques Aleatorios* (modelos en forma de árbol que se pueden combinar para mayor precisión), y *Regresión Logística* (que predice la probabilidad de pertenencia a una clase) son ejemplos de técnicas que aún se utilizan.

Para mejorar el rendimiento, se pueden utilizar **métodos de Ensemble**, que combinan múltiples clasificadores. Entre ellos *Bagging*, que crea varios conjuntos de entrenamiento para entrenar un clasificador en cada uno, y *Boosting*, que entrena clasificadores secuencialmente, donde cada uno se centra en los errores de los anteriores.

Aunque el deep learning se basa en redes neuronales, también existen ejemplos de **redes neuronales superficiales** utilizadas en la clasificación de texto, como *FastText* y *GHS-NET*. La elección del método siempre dependerá de las características del conjunto de datos, los requisitos de rendimiento y los recursos computacionales disponibles.

El artículo brinda una enfoque general sobre los algoritmos de clasificación de texto, pero podría mejorarse al incluir una crítica más profunda de los modelos, destacando sus ventajas y desventajas en diferentes contextos. Ademas, una discusión sobre las limitaciones actuales, como los sesgos en los datos de entrenamiento o la dificultad para manejar información contextual compleja.

2024/10/08 Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. Information, 13(2), 83. https://doi.org/10.3390/info13020083