EUR-443

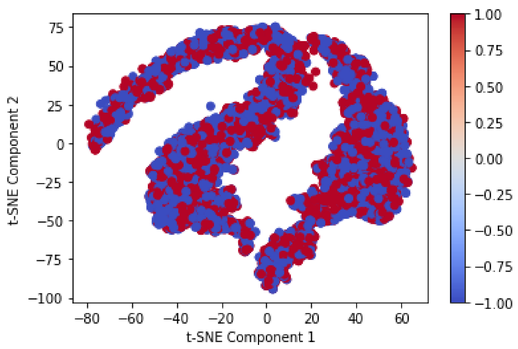
The Compression Connection: NCD and kNN in Law Enforcement Text Analytics

ABSTRACT

This study explores a text classification strategy utilizing k Nearest Neighbors (kNN) and gzip compression to generate normalized compression distances (NCD) for sentiment analysis.
Contrasting with complex models like BeRT, this minimalist approach was assessed on a sentiment-labeled dataset. The technique focused on core principles and the interpretability of NCD as features for kNN classification. Despite its simplicity, the method demonstrated a promising 70% accuracy, indicating potential for efficient, resource-light sentiment analysis.

METHODS

The approach utilized the k Nearest Neighbors algorithm for classification, with feature vectors derived from normalized compression distances (NCD) computed using gzip compression.
NCDs were calculated by comparing the compression lengths of text pairs, providing a basis for kNN to differentiate between sentiment classes. This method prioritizes computational efficiency and leverages statistical patterns in text compression for classification.



RESULTS

The investigation yielded a 70% accuracy in sentiment classification, showcasing that a basic kNN and gzip compression approach can effectively discern sentiment from text data.

Simplified sentiment analysis method using NCD kNN and NCD was tested, achieving 70% accuracy, revealing the viability of lightweight text classification techniques.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Gabriel Gillott
Dr. Hafiz Khan