Paper 5 Summary

In "Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches," the authors systematically evaluate different approaches for text classification of unseen classes. The study focuses on two main categories: similarity-based approaches and zero-shot text classification (0SHOT-TC). The experiments are benchmarked on four text classification datasets, including a new dataset from the medical domain. The authors propose novel baselines and a similarity-based approach, Lbl2TransformerVec, which outperforms previous state-of-the-art approaches in unsupervised text classification. The results show that similarity-based approaches generally outperform zero-shot approaches in most cases, and using advanced similarity-based methods such as SimCSE or SBERT embeddings further increases classification results. However, the authors note that zero-shot entailment approaches yield promising results in predicting instances of unseen classes, particularly for domain-specific datasets. The study also explores the impact of the length of text documents on the performance of SimCSE and SBERT-based approaches, concluding that larger Pretrained Language Models (PLMs) yield better results for 0SHOT-TC, while their performance is highly domain-dependent. Additionally, the authors conduct a correlation analysis to measure the relationship between the average number of document words per class and F1-scores, finding no statistically significant correlation trend. The study highlights the importance of advanced similarity-based approaches for unsupervised text classification, particularly in the context of predicting instances of unseen classes, and provides valuable insights for future research in this area.

The study systematically evaluates different approaches for text classification of unseen classes. It focuses on two main categories: similarity-based approaches and zero-shot text classification (0SHOT-TC), benchmarking them on four text classification datasets, including a new dataset from the medical domain. The authors propose novel baselines and a similarity-based approach, Lbl2TransformerVec, which outperforms previous state-of-the-art approaches in unsupervised text classification. The results show that similarity-based approaches generally outperform zero-shot approaches in most cases, and using advanced similarity-based methods such as SimCSE or SBERT embeddings further increases classification results. However, the authors note that zero-shot entailment approaches yield promising results in predicting instances of unseen classes, particularly for domain-specific datasets. The study also explores the impact of the length of text documents on the performance of SimCSE and SBERT-based approaches, concluding that larger Pretrained Language Models (PLMs) yield better results for 0SHOT-TC, while their performance is highly domain-dependent. Additionally, the authors conduct a correlation analysis to measure the relationship between the average number of document words per class and F1-scores, finding no statistically significant correlation trend. The study highlights the importance of advanced similarity-based approaches for unsupervised text classification, particularly in the context of predicting instances of unseen classes, and provides valuable insights for future research in this area.