Title: Towards Transparent Linguistic Analysis of Dutch Newspaper Article Genres using Machine Learning

Authors: Erik Tjong Kim Sang, Kim Smeenk, Aysenur Bilgin, Tom Klaver, Laura Hollink, Jacco van Ossenbruggen, Frank Harbers and Marcel Broersma

Systematic study of genre in newspapers sheds light on the development of journalism discourse. The genre conventions that can be discerned in a newspaper text signal the underlying discursive norms and practices of journalism as a profession. Historical newspapers are increasingly becoming available thanks to digital newspaper archives (in the Netherlands available through Delpher.nl), providing the opportunity for large-scale empirical research. However, the digital archives do not contain fine-grained genre information that is required for this purpose. Therefore, we use machine learning to automatically assign genre labels to newspaper articles.

Machine learning facilitates substantial improvements to the outcomes of existing research by providing increased amounts of enriched data. However, the decision-making process of the machine learning pipeline needs to be verified. Our previous findings (Bilgin et al., 2018) show that accuracy scores alone are not enough to assess the performance of these pipelines and that making an informed choice not only empowers optimal study of the historical development of genre, but also increases the trustworthiness of the results. This work shows that employing a transparent approach driven by model interpretability facilitates fair comparison as well as validation of the underlying decision-making criteria of the machine learning pipelines. The criteria are presented in the form of important features, creating insights on interactions between genre-related linguistic features and bag-of-words features.

Reference: Aysenur Bilgin et al., Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History. In Proceedings of the 14th IEEE eScience conference, 2018, http://arxiv.org/abs/1810.00968