Project Proposal

Till Grutschus, Ricardo Lammert Zepeda, Jurek Sander

1 Research Question

With the increasing amount of scientific publications, it becomes more and more difficult to keep up with the latest advances. Therefore, it becomes imperative for different actors in the scientific community to quickly identify relevant publications. Researchers, publishers and readers are interested in the impact of scientific literature. While researchers might want to optimize their own writing to increase visibility, publishers may want to quickly gauge the potential impact of papers they are considering for publication. Readers may want to filter the new publications for relevance.

In this project, we want to address this problem by pondering the following question:

> Can we predict the (future) impact of a scientific publication based on its metadata?

By answering this question, we hope to provide a tool that can help researchers, publishers, and readers to quickly identify relevant publications.

2 Dataset

To address this question, we will use data from the arXiv repository found here: https://www.kaggle.com/Cornell-University/arxiv. Additionally, we will use citation data extracted from crossref.org. In the following we will briefly describe the two datasets.

arXiv Dataset The arXiv dataset contains metadata of 1.7+ million scientific publications from the arXiv repository. A detailed description of the dataset can be found on the dataset's Kaggle page. We expect the most relevant attributes for our project to be the title, abstract, authors, categories, and comments (e.g. number of pages, figures, tables) of the publication. A detailed feature analysis will be part of the project.

crossref.org Crossref offers a publicly available API for accessing citation data. Additionally, data dumps of the crossref database are available for download. We will use the number of citations to extract the ground truth of the impact of a publication.

3 Methodology

Dataset acquisition As described, we will use the crossref API to enrich the available arXiv dataset.

Data preprocessing Textual data will be preprocessed, tokenized and vectorized as learned in the assignments. Missing data will be handled appropriately. Subsequently, we will perform a feature analysis to identify the most relevant features for our project and potentially reduce the dimensionality of the dataset.

Data mining We will use the citation count to devise a discrete ordinal target variable for the impact of a publication. We will then evaluate different classification algorithms to predict the impact of a publication based on its metadata.