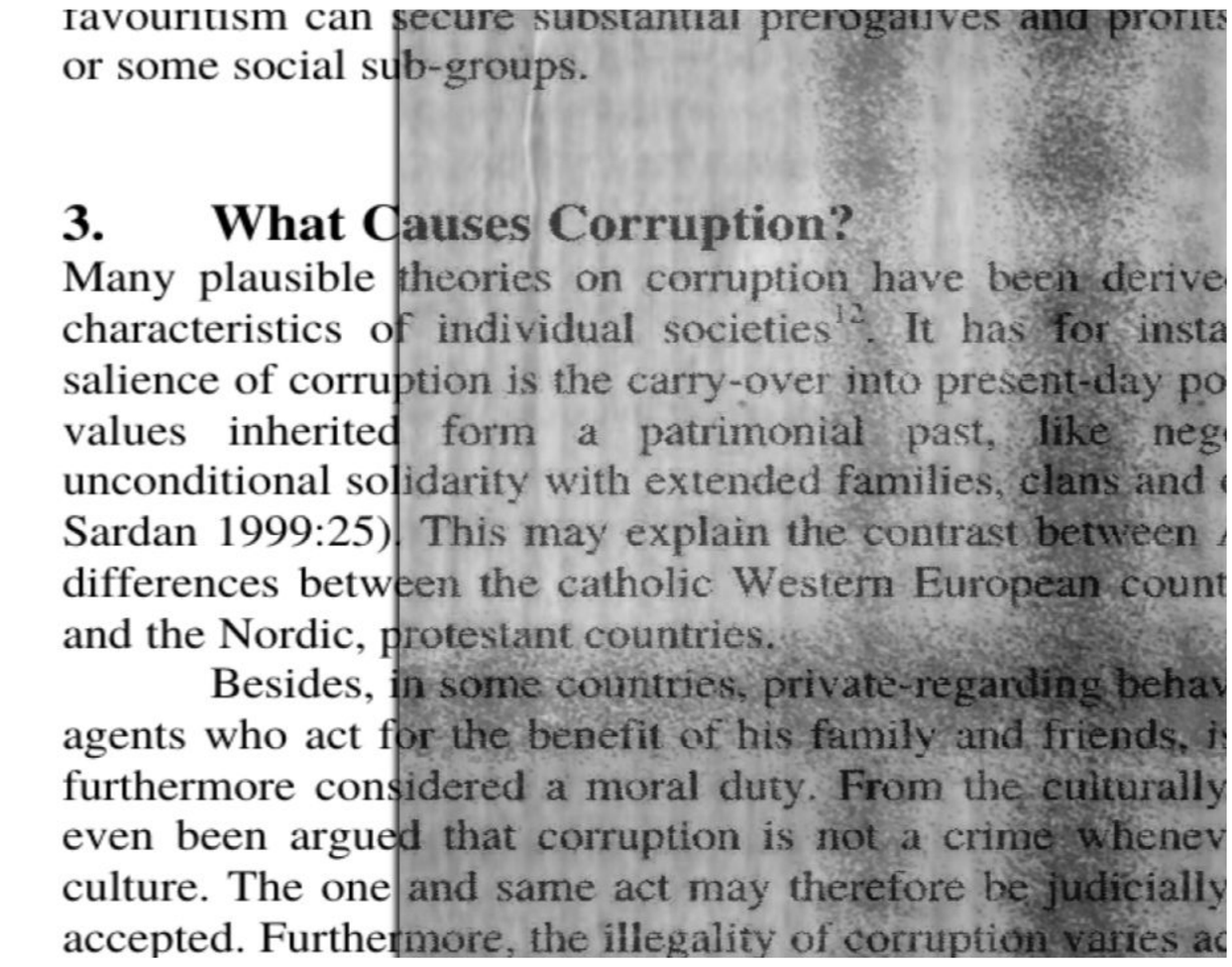
ShabbyPages: A Robust Corpus for Training Document Image Models

Alexander Groleau,, Kok Wei Chee, Stefan Larson, Jonathan Boarman

Sparkfish

github.com/sparkfish/shabby-pages


favouritism can secure substantial prerogatives and profits
or some social sub-groups.

3. What Causes Corruption?

Many plausible theories on corruption have been derive
characteristics of individual societies[12]. It has for insta
salience of corruption is the carry-over into present-day po
values inherited form a patrimonial past, like neg
unconditional solidarity with extended families, clans and
Sardan 1999:25). This may explain the contrast between
differences between the catholic Western European count
and the Nordic, protestant countries.

Besides, in some countries, private-regarding behav
agents who act for the benefit of his family and friends, i
furthermore considered a moral duty. From the culturally
even been argued that corruption is not a crime whenev
culture. The one and same act may therefore be judicially
accepted. Furthermore, the illegality of corruption varies ac

ShabbyPages is a corpus of born-digital document images with both ground truth and distorted versions appropriate for use in training and evaluating document denoising models. This new dataset with synthetically-generated real-world representations can be used to improve document layout detection, text extraction and OCR processes that depend on denoising and binarization preprocessing models.

ShabbyPages was created using the Augraphy data augmentation library. The image at left shows clean and noisy document versions.

In the figure at right, we evaluate a denoising model trained on the *NoisyOffice* dataset on two samples from the *ShabbyPages* dataset (bottom two rows). We observe that the *NoisyOffice* model fails to denoise *ShabbyPages* completely.