**Keyword labelling of works of art**

Art history is the discipline of classifying, critiquing and compiling the history of artworks from different perspectives. The labelling of artworks is also something that galleries have been experimenting with, and this project seeks to keyword label artworks using semi-supervised learning. I will focus on how to classify the large amount of text generated as introductory text for artworks, a process that can be built on a scalable premise, with the hope of building a training model that can automatically tag keywords based on introductory or evaluative text for artworks. In the future this could be used in conjunction with visual tagging of art for more precise and sophisticated search purposes.

This paper uses my graduation work research as a case study, focusing on tagging contemporary artworks. Archives in the contemporary art industry often include a paragraph written by the artist about the content of the work, and one or more critical essays about this work. As supervised learning requires a certain amount of labelled text, with a limited number of databases. I decided to use self-training semi-supervised learning, combining both supervised and unsupervised learning processes, to optimise the model.

The following is a partial list of my keywords for common themes in contemporary art.

*Globalisation*

*Rural life*

*Economy/Currency*

*Blockchain/bitcoin*

*Artificial Intelligence*

*Algorithms*

*"Post-human"*

*Advertising and branding*

*Consumerism*

*Internet culture*

*Subculture*

*Social networks*

*About Games*

*Copycats*

*imitation*

*"Borders"*

*Immigration*

*Language/translation*

*Ethnic and national identity*

*Gender identity*

*Western Perspective/Western Centrism*

*"The Other"*

*Bio-related*

*Medical/Health*

*History as text*

*Architecture related*

*Artist ethnography*

*"Systems"*

*Natural Landscapes*

*Environment and Ecology*

*Religion*

*Political Spectrum*

*Mythology*

*Interculturalism*

*Contemporary Archaeology*

*Memory*

*Time*

*Disaster*

*Feminism*

*Fertility/pregnancy/birth*

*Figures in the arts*

*Utopia*

*Heterotopia*

*Portraits*

*Humour*

*Fictional reality*

*Institutional Criticism*

*Censorship*

*Surveillance*

*Political events*

*War-related*

*Crime*

*Light*

*"Alienation"*

*Migration*

*Death*

*Mourning and Remembrance*

*Mysticism*

The labeled data is first trained with a supervised learning algorithm, using SVM as an example. The initial classifier is then used to predict the unlabelled data. The higher the probability of the predicted category, the higher the confidence level of the classification achieved, and the text with a higher confidence level is added to the labelled data together with its classification label, and the classifier is retrained (each time subtracting the selected samples from the unlabelled samples), iteratively until the unlabelled samples are the empty set.

As the self-training algorithm requires the use of high confidence unlabelled samples, the training set is expanded. If the unlabelled data predicts the wrong category, then each iteration will result in an accumulation of errors, causing a degradation in the performance of the classifier, so we repeatedly label the unlabelled samples to reduce the accumulation of classification errors in the training process and thus improve the classification results.