Abstract

This research project focuses on exploring the research area of disinformation. This is not a new concept but it has become an increasingly discussed and researched topic over the last 5 years. Despite this there is still limited research on the topic and this leaves many areas unexplored.

One such area is the use of multiclass models to distinguish between different types of disinformation. Another research area growing in popularity is the investigation of explainable machine learning techniques. This is an important field of study for a variety of reasons and it is key to integrating machine learning techniques into more and more tasks which are currently done manually by humans. After extensive research, this project will attempt to bring together elements from many existing papers in order to to create a machine learning system which can distinguish reliable news from fake news and satire, using explainable machine learning methods.

A modified version of the standard machine learning methodology was implemented in this research with the following steps: data gathering, feature selection, model selection and implementation, evaluation of models based on their classification performance and explainability criteria, and an investigation into the defining features of each class of news. This process produced several sets of models which aim to investigate different aspects of the written disinformation problem.

The models developed in this research are compared directly with models from related literature, revealing improvements made by this research in some areas, as well as its potential limitations others. This research has made some major contributions to the field including the generation of a novel dataset, an investigation into the distinguishing features of real, fake, and satire news, as well as creating classification models that can compete with those currently in the state of the art.