Coding historical causes of death data with Large Language Models

2023-09-15 13:38:49

Authors

Bjørn-Richard Pedersen, Maisha Islam, Lars Ailo Bongo, Eilidh Garrett, Alice Reid, Hilde Sommerseth

Abstract

This paper explores the feasibility of using Large Language Models (LLMs) to automate the assignment of ICD10h codes to historical causes of death, a challenging task due to the possibility of complex narratives and multiple diseases or injuries involved. Traditionally reliant on domain experts, this study investigates whether LLMs can effectively perform this task. The data we use to assess various LLMs' ability to assign accurate ICD10h codes come from the town of Ipswich in England, 1871-1901. Ipswich was a thriving port town with a variety of industries during this era. The dataset consists of the causes of death recorded in the death registrations of 43,020 individuals and transcribed as seen. These have been subsequently tidied and coded to ICD10h (a historically sensitive variant of ICD10). Our focus is twofold: understanding the steps needed to prevent plausible yet incorrect code assignments by LLMs and establishing a reliable validation process. We compare LLMs against a FastText classifier and employ random manual validations for comprehensive analysis.

Keywords

Large Language Models, LLM, AI, Historical data, Causes of death, death, Historical causes of death, Ipswich, FastText, Validation