**Methodology Overview**

A diagram of a diagram

Description automatically generated

Figure 1 provides an overview of our research’s case study for investigating Reddit for disease forecasting.

It is designed to aid the understanding of how our case study was performed, highlighting the structure if our investigation, helping summarise what this project did.

Not only does this figure help visualise and break down the steps of our research, but it addresses where/what stage of our research is used to answer each of our research questions, where our research questions related to the feasibility of extracting information from Reddit for forecasting, whether text classification can effectively classify these texts, if relationships to Covid-19 are present, and our forecasting exercise.

Figure 1 reveals that following data collection and pre-processing, we label sampled Reddit texts along selected categories relating to Covid-19 infections, addressing RQ1, determining feasibility. These labelled samples then validate/optimise our BERT models before classifying our Reddit dataset, and BERTs performance answers RQ2. Subsequently, classified text is transformed into time-series data, with correlational and predictive analysis addressing RQ3, mirroring related research [5, 8, 31].

Our final stage assesses Reddits predictive utility, evaluating Reddit-enhanced RFR Covid-19 predictions against baseline models following RFR validation/optimisation. For this forecasting exercise we employ a rolling-window approach for generating training and test sets, a walk-forward testing approach, and lagged Covid-19 features to forecast future incidence. This produces an extended testing period, regular model updates, mimicking real-world forecasting scenarios in-line with comparable disease forecasting research [15, 31, 53].