LONDON'S GLOBAL UNIVERSITY

Influenza-Like-Illness Rate Surveillance Using Web Search Queries

Candidate Number: FRRT1[1]

MEng Computer Science

Supervisor: Vasileios Lampos

Submission date: 26/04/23

[1]**Disclaimer:** This report is submitted as part requirement for the MEng degree in Computer Science at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Abstract

In the field of epidemiological surveillance, the use of user-generated content has become increasingly prevalent in the early detection and estimation of infectious diseases, such as Influenza. Specifically, prior research has demonstrated the effectiveness of utilising online web search queries for both nowcasting and forecasting Influenza-Like Illness (ILI) rates in different geographical locations. Google Flu Trends represents the first real time Influenza surveillance system utilising such methodologies. Subsequent studies have built upon their efforts, suggesting improvements in experimental design and models for the prediction tasks.

In our study, we build on this existing work through exploring the predictive capabilities of deep learning models, addressing the task of nowcasting and forecasting ILI rates in England using online web search queries. We evaluate the predictive performance of our models over five consecutive flu seasons from 2014-15 to 2018-19, using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Bivariate Correlation (r).

Initially, we conduct a replication study of a proposed baseline Elastic Net model [1], in conjunction with exploring various feature selection methods for the nowcasting task. This encompasses a correlation-based approach, directly measuring data correlations between queries and ILI rates; using sentence embeddings to capture the textual semantics of the queries; and a hybrid method combining the two. Building upon these preliminary models and retaining the optimal hybrid features, we present a novel, minimal Feed Forward Neural Network (MFFNN) model that surpasses the Elastic Net in the predictive performance with an improvement of 10.56% in MAE, 13.86% in MAPE, and 2.1% in Bivariate Correlation. This effectively introduces a new baseline model for the nowcasting task.

This is then further succeeded by an exploration of deeper neural network architectures, including the incorporation of historical lagged features. Specifically, we propose a deeper FFNN (DFFNN) model, incorporating 14 historical lagged features, that achieves a highly competitive nowcasting performance of 1.25 MAE, 22.36% lower than our newly introduced baseline (MFFNN) and surpassing performances of more traditional machine learning methods, reported in similar studies [1], [2].

Continuing our analysis, we apply our neural network models in the forecasting task, at different horizons (γ), predicting ILI rates for $7, 14, 21$, and 28 days ahead. We observe that our DFFNN nowcasting architecture demonstrates strong and competitive forecasting performance across flu seasons, notably outperforming both our MFFNN and a baseline persistence model for longer forecasting horizons. These results once again surpass forecasting performances reported in similar studies [3], which employ more complex neural network architectures, albeit on three flu season compared to our five.

These findings not only underscore the competitive performance achieved by our DFFNN model, outperforming our linear baseline in both nowcasting and forecasting tasks, but also establish a benchmark against which more complex deep neural network architectures can be compared against, validating their performance and necessity over our simpler architectures.

Contents

Chapter 1

Introduction

1.1 Research Objective

User-generated content (UGC) on online platforms has become a focal point in a wide array of research fields, including Behavioural Sciences [4]–[6], Computational Linguistics, Social Science, and Health [7]. This growth is particularly evident in the field of epidemiology, where a growing body of research underscores the effectiveness of using online website and search engine logs for predicting Influenza-Like Illness (ILI) rates [8]–[10]. Notably, most influenza surveillance systems currently used by UK Health Security Agency (UKHSA) and other international public health organisations primarily rely on traditional methods of collating data stored by established health-care systems. This approach, however, may not comprehensively capture the full spectrum of the population affected by influenza. It tends to represent only a specific subset of individuals who actively seek medical attention.

The utilisation of online UGC can offer a more timely, accessible and less costly source of information about the health status of a population [10]–[12]. It also captures this lower segment of the disease population pyramid, consisting of those individuals exhibiting ILI symptoms who do not actively seek medical assistance. This is particularly beneficial in lower developed countries, where Influenza outbreaks can be more severe and established, affordable healthcare monitoring systems may not be as prevalent.

Google Flu Trends (GFT) emerged as the first real-time system to apply UGC-based methods across numerous countries over a significant timespan [10]. Although GFT faced considerable critique due to inconsistencies in its ILI predictions when compared to traditional data sources [13]–[15], subsequent research has built upon and refined these earlier methods [11], [16]. These primarily consist of regularised and auto-regressive traditional machine learning models. Recent research efforts have additionally demonstrated the significant potential of deep learning models in this domain in overseas locations [12], [17]. However the majority of these proposed models are overly complex without any simpler baseline models established, validating their performance and thus necessity in the prediction task. This motivates our study not only to explore the performance potential of deep learning models over more traditional approaches for influenza surveillance in England but also to examine the performance gains attainable whilst utilising relatively simple architectures.

When utilising UGC to infer the state of influenza or other health-related diseases in a region, a

key challenge involves discerning genuine influenza-related search queries from spurious or unrelated ones that negatively influence the predictive capabilities of the model [1], [18], [19]. Furthermore, while some queries may be flu-related, they might not necessarily indicate that an individual is suffering from flu-like symptoms. Such queries could be influenced by external factors such as media propaganda or public news events, leading individuals to search for flu-related terms out of curiosity or concern, rather than due to exhibiting any symptoms. Hence, these limitations must be taken into consideration when developing such surveillance systems.

In our study, we investigate the tasks of nowcasting and forecasting ILI rates in England through the analysis of Google search queries. This involves the prediction of five distinct flu seasons, spanning from 2014-15 to 2018-19. For each flu season prediction, we utilise all preceding flu season data from 2009-10 onwards as our training data. Our initial step involves conducting a replication study of recent research approaches in nowcasting, opting for an Elastic Net model as our baseline [1]. To address the issue of spurious queries in our dataset and obtain the most ILI predictive set of features, we explore various feature selection methods in conjunction with our baseline model. These methods include a correlation-based approach, which aims to measure direct correlations in the data between queries and ILI rates. Additionally, sentence embeddings are employed to assess the semantic relevance of queries. Lastly, we investigate a hybrid method, combining these two approaches in effort to leverage their respective strengths.

Advancing our research, we investigate a novel minimal Feed Forward Neural Network (MFFNN) model designed to better discern non-linear relationships within the data, enhancing the performance achieved by our baseline model. This aims to effectively introduce a new baseline benchmark for the nowcasting task, demonstrating the significant potential of deep learning models in this domain. Building upon this foundation, we further explore deeper FFNN architectures (DFFNN) to improve upon the performance of our newly introduced baseline. This involves architectural modifications, as well as the incorporation of historical lagged features in the dataset.

Following this, we shift our attention to the task of forecasting ILI rates, evaluating the predictive performance of our neural network nowcasting models against a baseline persistence model across different forecasting horizons γ. This involved changing our prediction task from predicting the ILI rate on day t, to predicting the ILI rate on day $t + 7, 14, 21$ and 28.

These experiments intend on achieving a robust and competitive performance for the nowcasting and forecasting tasks, outperforming our linear baseline, demonstrating the efficacy of deep learning models in the domain. This will additionally provide a benchmark upon which the more complex architectures in the field can be compared against. This comparison enables us to evaluate whether these advanced models yield further performance gains and if such gains are significant in justifying their increased complexity and training difficulty compared to our simpler models.

1.2 Project Aims and Goals

Aims:

1. Contribute to the field of epidemiological surveillance in England using UGC by achieving competitive nowcasting and forecasting performances with relatively simple deep learning models.

2. Enhance proficiencies in machine learning algorithms and practical training of models tai-

lored for the distinct prediction task, while gaining an in-depth awareness of the various experimental considerations essential for effective model training.

Goals:

1. **Feature selection exploration:** Explore different feature selection methods, including a correlation based approach, sentence embeddings, and a hybrid approach, in effort to mitigate the impact of spurious queries in the dataset, aiming for optimal predictive performance.

2. **Elastic Net baseline model (nowcasting):** Conduct the training and evaluation of an Elastic Net baseline model using all feature selection methods. This objective seeks to provide a baseline model and a comparative analysis of the different feature selection approaches for the nowcasting prediction task.

3. **Minimal FFNN model (nowcasting):** Train and evaluate the most streamlined FFNN model that outperforms the baseline Elastic Net, effectively introducing an improved new baseline for the nowcasting prediction task.

4. **Exploration of FFNN architectures (nowcasting):** Explore and compare deeper FFNN architectures, including the incorporation of historical lagged features. This aims to achieve a competitive predictive performance against existing work in the field.

5. **Evaluate performance of FFNN models in the forecasting task:** Evaluate and compare the performances of our FFNN architectures against a persistence baseline model, in the forecasting prediction task. Repeat this across different forecasting horizons predicting day $t + 7, 14, 21$ and 28.

1.3 Project Overview

The project began with an exploration of the problem domain and relevant literature to understand recent advancements in the prediction tasks. Subsequently, we conducted preliminary research on the various machine learning algorithms employed, their strengths, weaknesses, and experimental procedures for model training. These findings were consolidated into the Background Information chapter of the report, laying the groundwork for our study's experiments.

Before training machine learning models, we pre-processed the dataset to ensure its suitability for experiments. This involved interpolating, removing redundancy, and smoothing the data. We also conducted preliminary data analysis to understand its properties and relationships, motivating subsequent experimental decisions. This included observing the time series trend in ILI rates and identifying seasonal patterns.

The initial step in our prediction task involved exploring various feature selection methods, including correlation-based, sentence embedding, and hybrid approaches. We then trained an Elastic Net model for nowcasting ILI rates for each flu season using features selected by these methods, allowing a comparison of their performances and determining the optimal feature selection approach for this prediction task.

Next, we developed a minimal FFNN (MFFNN) model that outperformed our baseline in the nowcasting prediction task. We further explored deeper, more complex FFNN architectures, including historical lagged features, aiming for further competitive predictive performance.

Lastly, we evaluated the predictive performance of our FFNN architectures against a baseline persistence model for the forecasting task. This was repeated for different forecasting horizons γ, assessing the models' ability to predict the ILI rate on day $t + 7, 14, 21$ and 28.

1.4 Report Overview

The subsequent sections of the report are structured as follows:

The next chapter, titled Background Information, covers the problem domain, including concepts of UGC, Influenza, and time series nowcasting/forecasting. It also examines literature relevant to the use of UGC in epidemiological surveillance, shedding light on existing work in the field. This review motivates the goals of our study and informs the various models and experimental approaches we explore. Additionally, the chapter explains the different machine learning algorithms employed in our prediction tasks.

The Influenza-Like Illness Rate and Web Search Query Data chapter is organised into three key sections: Dataset, Data Preprocessing, and Data Analysis. The Dataset section provides a description of our data source and its contents. The Data Preprocessing and Data Analysis sections detail the processes implemented in preparing the data for the machine learning models and conducting preliminary analysis to better understand data properties, respectively.

The Nowcasting Influenza-Like Illness Rates With Regularised Linear Regression chapter details experiments conducted and provides an analysis of corresponding results related to feature selection and the training of the Elastic Net baseline model for the nowcasting task. The Nonlinear Regression Models For Predicting Influenza-Like Illness Rates chapter covers experiments pertaining to the training of different neural network architectures, specifically the Minimal FFNN and deeper FFNN architectures, for the nowcasting task. This is followed by an evaluation of these architectures in the forecasting task, predicting the ILI rate on days $t + 7, 14, 21$ and 28, against a persistence baseline model.

In the Conclusion chapter, a summary encapsulates the project's goals, activities, and findings. A critical evaluation of the results against the initial project objectives is provided in the Evaluation section. Challenges encountered during the project are discussed, and potential future work that could apply the findings of our models and expand upon the study is outlined. Additionally, the report includes an Appendix containing additional data analysis graphs and project deliverables, such as the project plan, interim report, and project's source code.

Chapter 2

Background Information

This chapter establishes the foundational background necessary for understanding the problem domain, covering concepts of User-Generated Content, Influenza, and time series nowcasting and forecasting. It also reviews existing literature relevant to the application of UGC in epidemiological surveillance, highlighting significant contributions to the field. This review informs the objectives of our study and motivates the methodologies employed in our experimental approach. The chapter concludes with a technical overview of the machine learning algorithms used in our prediction tasks.

2.1 User Generated Content

User-Generated Content (UGC) consists of a diverse array of content forms such as text, images, videos, and data that are created and shared by users across various online platforms. This includes social media posts, blog entries, online reviews, forum discussions, and web search queries. Each form of UGC can provide unique insights into public opinion, behaviours, and emerging trends.

UGC has gained significant attention in recent research, offering valuable insights across various fields such as Politics, Finance, Commerce, and Health. This body of research has demonstrated the extensive utility of UGC beyond its initial scope as components of social media platforms or search engines [4]. This versatility is highlighted through the predictive capabilities of platforms like Google Trends and Twitter, which have proven to be more than just channels for information dissemination in a news-like manner [5], [6].

The exploration of UGC in various contexts has unveiled its potential in mirroring and even predicting real-world phenomena. For example, Tumasjan revealed that Twitter is not just a medium for political deliberation but also a reflection of offline political sentiment. Using the LIWC text analysis software on 100,000 politically related tweets, they showed a direct correlation between the content of tweets and election results, suggesting that Twitter content can validly represent the political landscape [20]. Similarly, O'Connor's research demonstrated a significant correlation between public opinion, as expressed in surveys, and the sentiment expressed in contemporaneous Twitter messages. This finding suggests that social media can be a viable supplement, or even an alternative, to traditional polling methods in gauging public sentiment [21].

The predictive power of UGC extends beyond the domain of politics, even behaving as economic indicators. By analysing Twitter feeds with tools like OpinionFinder and GPOMS, Bollen investigated the relationship between Twitter mood states and the stock market. Specifically, their study established a link between public mood and the fluctuations in the Dow Jones Industrial

Average. This not only highlighted the predictive accuracy of UGC in financial markets but also underscored the nuanced nature of this data, where specific mood dimensions were more indicative of market changes than others [22].

The predictive nature of UGC is further evidenced in consumer behaviour studies. Research utilising web search query volumes has successfully forecasted outcomes in entertainment and retail sectors, such as predicting box-office revenues, video game sales, and music chart rankings. This highlights the capacity of search data to anticipate market trends, enhancing traditional forecasting models, showcasing the multifaceted applications of UGC [23].

Closer to the field of health, Choudhury demonstrated the use social media to identify early signs of major depressive disorder. By analysing changes in social media behaviour, such as decreased engagement and shifts in emotional expression, he developed a model that could potentially predict the onset of depression. This approach exemplifies how UGC can be a powerful tool in mental health surveillance, offering insights into individual well-being through digital footprints [7].

Collectively, these studies provide a comprehensive picture of UGC as a multi-functional resource that stretches beyond its original online context. Its ability to reflect and predict real-world trends and sentiments across various domains makes it as a valuable asset in contemporary research. As we pivot towards the specific application of UGC in epidemiology surveillance, the upcoming sections will explore how online UGC, specifically web search queries, has been utilised to predict and understand the dynamics of Influenza outbreaks. This will demonstrate the adaptability and potency of UGC in addressing complex public health challenges.

2.2 Influenza

Influenza is an infectious respiratory illness caused by influenza viruses. This condition spreads through respiratory droplets and is characterised by symptoms such as fever, cough, sore throat, and muscle aches. While most individuals recover within a week or two, influenza can be particularly severe among high-risk groups including the elderly, young children, and those with chronic health conditions. Monitoring the incidence and spread of influenza across populations can assist in the early detection of the illness and the implementation of public health measures which can significantly reduce the impact of outbreaks and save lives.

Influenza-Like Illness (ILI) encompasses a wider range of respiratory illnesses. This includes both viral and bacterial infections that exhibit symptoms similar, but not limited to influenza. The symptoms of ILI are often indistinguishable from those of influenza without precise laboratory testing, making ILI a valuable indicator for influenza surveillance.

The traditional mechanisms for influenza surveillance, primarily rely on data sourced from established healthcare providers, laboratories, and hospitals. Despite the importance of these facilities, these systems face several challenges, including delays in reporting and potential underestimation of actual case numbers. Furthermore, they may not fully capture the extent of the affected population, particularly those individuals who do not actively seek medical attention. This surveillance approach also suffers from geographical limitations in areas lacking established healthcare systems that can provide the necessary surveillance information. UGC provides solutions to these problems, as an accessible and cost effective source of information.

2.3 Influenza Surveillance Using Social Media

Similar to the political and social domains, in digital epidemiology, researchers have extensively explored the use of social media platforms like Twitter for influenza surveillance. Lampos et al. conducted a study during the H1N1 pandemic in the United Kingdom, analysing tweets containing self-diagnostic statements related to flu-like symptoms from June to December 2009. They developed a 'flu-score' based on collected tweets to reflect influenza prevalence and validated its predictive potential using linear regression, achieving a robust correlation of 92.34%. However, challenges included distinguishing actual flu cases from media-driven sensationalism on Twitter and potential biases from specific keywords [24].

Similarly, Lamb et al. categorised flu-related tweets into two groups: one for tweets that report actual infections, and another for those that express concerned awareness about the flu. They refined this categorisation to distinguish between self-reports of infection and reports about others, employing advanced natural language processing (NLP) techniques and a log-linear model with L2 regularisation for classification, aided by human annotations via Amazon Mechanical Turk [18].

Another noteworthy approach, undertaken by Broniatowski et al. developed an algorithm to filter relevant tweets tracking actual influenza infections, aiming to remove 'chatter' fuelled by media attention. The system was actively deployed during the 2012-13 influenza season, with a focus on both national and local levels, again using a log-linear model with L2 regularisation. Results were compared with traditional surveillance data, showing high correlation coefficients with the CDC (0.93) and NYC (0.88) data and an 85% accuracy in detecting weekly directional changes in influenza prevalence [19].

These studies collectively highlight the importance of precise differentiation of tweets related to influenza infections from general discourse, as opposed to relying solely on all flu-related tweets or simplistic keyword counting methods. This has shown to significantly enhance the precision of influenza surveillance when utilising Twitter.

2.4 Influenza Surveillance Using Web Search Queries

Closer to our study, several research efforts have also provided valuable insights into the utility of web search activity in Influenza surveillance and early detection.

Web search queries are terms and phrases users input into search engines to gather information, serving as a real-time reflection of public interests and concerns. This can often reveal the collective anxiety about symptoms, treatments, and general information related to ILI.

The emerging field of digital epidemiology involves analysing these web search queries, with the intuition that increases in searches related to flu symptoms or treatments indicate a growing prevalence of influenza amongst the population and often precede actual spikes in influenza cases. By closely monitoring and analysing these more accessible search trends, researchers can potentially detect influenza outbreaks earlier than traditional surveillance methods.

2.4.1 Early Research

An early study explored the potential of online web activity to track influenza levels in the population. It focused on the correlation between article accesses on influenza topics on Healthlink, a consumer health information website, and traditional influenza surveillance data from the CDC.

Throughout 2001, the research team monitored weekly access counts for 17 influenza-related articles on Healthlink, comparing them to two CDC benchmarks: the weekly regional number of ILI cases and positive influenza tests. The results showed a moderate correlation between article accesses and CDC surveillance data. However, the timeliness of web data compared to traditional methods varied, leading to inconclusive findings on its ability to detect influenza outbreaks earlier. Challenges included assigning spatial locations to web users and the influence of non-disease-related factors on web access patterns [8].

Despite these concerns around wider web activity, significant research has focused on the specific utility of web search queries.

A notable study by Polgreen et al. investigated the relationship between the frequency of internet searches for influenza on Yahoo! and the actual occurrence of the disease in the United States. They analysed Yahoo! search queries from March 2004 to May 2008, counting daily unique queries containing influenza-related terms and normalising them by the total number of searches. Using linear models with 1-10-week lead times, they predicted the percentage of cultures positive for influenza and deaths from pneumonia and influenza. The models successfully forecasted increases in positive cultures up to 3 weeks in advance and mortality up to 5 weeks ahead, demonstrating the real-time surveillance potential of internet searches. However, challenges remain in discerning credible information among numerous health-related sites, and reliance on a single search engine limits generalisability. Additionally, external factors like media coverage may influence search behaviours, potentially affecting model accuracy [9].

2.4.2 Google Flu Trends

Google Flu Trends (GFT) is the first real time system to employ online UGC, specifically web search queries for the task of Flu surveillance. They proposed a model that could estimate the likelihood of physician visits being related to ILI, based on corresponding search queries. The methodology involved a comprehensive analysis of Google web search logs spanning five years, encompassing hundreds of billions of individual search queries. The researchers compiled weekly counts of the most common 50 million search queries in the United States, maintaining separate aggregates for different regional states. These counts were normalised by dividing the weekly count for each query by the total number of online search queries from the same location, ensuring the data accurately reflected the relative frequency of ILI-related searches. A pivotal aspect of the study was the automated selection of ILI-related search queries. This process involved rigorously testing each of the 50 million candidate queries to identify those that most accurately modelled the CDC-reported ILI visit percentage in each region. This resulted in 45 high-scoring queries that best fit the ILI data across nine regions.

A linear model was employed to analyse the relationship between ILI physician visits and ILI-related search queries. This model was expressed as $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits at time t, $Q(t)$ represents the ILI-related query fraction at the same time, α is the multiplicative coefficient, and ε is the error term. The logit function, defined as $\ln(p/(1-p))$, was used to transform the probabilities into a linear scale suitable for modelling. This model demonstrated a robust correlation with CDC-reported ILI percentages, achieving a mean correlation of 0.90 across the nine regions. This strong correlation was further validated against state-reported ILI percentages from Utah, where a correlation of 0.90 was observed across 42 validation points. These results underscored the efficacy of leveraging web search queries for real-time influenza surveillance and its potential for application in international settings. Thus,

enhancing public health planning and intervention strategies [10].

Although GFT represented a significant advancement in utilising online UGC for influenza surveillance, subsequent studies have since outlined its limitations and provided insights into potential enhancements. Cook et al. focused on GFT's performance during the 2009 H1N1 pandemic, comparing the original model against a revised version that accounted for the pandemic's early data and included less frequent queries. Their findings highlighted that the original model, despite its effectiveness before the pandemic, consistently underestimated ILI activity during the outbreak. In contrast, the revised model significantly improved in accuracy, achieving a correlation of 0.95 with ILI activity data during the H1N1 peak, compared to the original model's correlation of 0.29. This highlighted the importance of model adaptability in response to new health trends [13].

Extending this assessment, Olson et al.'s decade-long review (2003-2013) of GFT critiqued its ability to capture the timing and magnitude of flu outbreaks at national, regional, and local levels. This study revealed GFT's failures to detect the initial wave of the 2009 H1N1 pandemic and its overestimation of the 2012/2013 season's severity, underscoring the challenges posed by changes in search behaviour, geographic and demographic variations, and the timing of model updates in accurately forecasting flu trends [14].

Lazer et al. discussed the broader issue of relying heavily on big data for health surveillance, a concept they termed 'big data hubris'. They pointed out the inherent risks in neglecting traditional epidemiological methods and the potential for overfitting search terms to the data, which led to GFT's misclassification of the 2009 H1N1 pandemic. This critique underscores the complexities of measuring and validating big data-driven models against traditional, robust epidemiological approaches [15].

These studies collectively highlight the necessity of continuous refinement and the integration of traditional epidemiological insights with digital surveillance tools. They advocate for the adaptability of models like GFT to changing health landscapes and search behaviours, demonstrating the importance of a balanced approach to leveraging big data in public health monitoring.

Despite the significant critiques highlighted by these evaluation studies, subsequent research efforts have revisited the capabilities of GFT, suggesting that the errors associated with it may not be as substantial as previously thought. Kandula et al.'s study employed a methodological critique and refinement over five influenza seasons, contrasting GFT's weekly ILI estimates with actual ILI rates from ILINet and real-time rates (ILIp). By applying a random forest regression model that used historical ILI and GFT data, they significantly reduced GFT's error rate by 44%, notably achieving an 80% reduction during the 2012-13 season. Their study further explored forecasting flu trends using ARIMA models, integrating corrected GFT data with ILIp, which consistently improved the accuracy of one to four-week-ahead forecasts. This approach lowered the Mean Squared Error (MSE) across most seasons and regions, indicating the potential for corrected GFT data to enhance influenza forecasting when properly adjusted and integrated with traditional surveillance methods [25].

2.4.3 Nowcasting Influenza-Like Illness Using Web Search Queries

Time Series Nowcasting

Time series nowcasting is a technique used for predicting current, real-time values in a time series. This is different to traditional forecasting methods which rely on historical data to predict future

trends. This predictive task provides insights into the current prominence of Influenza amongst a population. In the context of the following studies, they employ time series nowcasting to predict the daily ILI of different flu seasons using the normalised daily frequencies of web search queries.

Traditional Machine Learning Models

Subsequent research endeavours have built on the foundations of GFT, exploring alternative approaches for nowcasting influenza using web search queries. Preis et al. developed an adaptive approach to 'nowcasting' influenza levels using Google search queries, moving beyond GFT's static model by employing dynamic nowcasting models that adjust to changing search patterns and flu trends over time. This approach utilises ARIMA models combined with historical ILI patient visit data, achieving more accurate real-time estimates before the release of official figures. The study demonstrated a reduction in the MAE of in-sample nowcasts by 14.4% over baseline models, with out-of-sample improvements between 16.0% to 52.7%, depending on the training interval used. Yet, the methodology faces challenges from non-specific search queries driven by media coverage or public concern, not directly indicative of ILI symptoms [16].

Kandula et al. further refined influenza nowcasting by focusing on subregional accuracy using web search trends, covering six seasons across the U.S. to assess patterns at a more localised level. Through integrating ARIMA with random forest models and Google Extended Trends data, their study evaluated the effectiveness of using more granular state-level ILI data compared to broader regional data. While the findings suggest these models offer enhanced estimates over traditional autoregressive models, their superiority over GFT is not uniform. However, incorporating subregional health data with search trends markedly improves performance. The study also notes the limitation of using search trend data for detailed geographic nowcasting due to the limited availability of granular data, highlighting the ongoing need for advancements in data collection and analysis methods to improve local nowcasting accuracy [26].

Another study conducted by Lampos et al.'s aimed to refine nowcasting of ILI rates in the United States, covering data from 2004 to 2013 and focusing on the last five seasons (2008-2013). The study aimed for a direct comparison with GFT by reapplying its methodologies to updated models and data. Initially, an Elastic Net linear model was utilised to sift through numerous search queries, selecting those most predictive of ILI rates, extending beyond GFT's aggregate frequency approach. These queries were then analysed using a nonlinear Gaussian Processes (GP) regression framework, capable of capturing complex relationships between search data and ILI rates. This analysis was further refined by clustering queries with a composite GP kernel based on similarity, offering a more detailed examination than GFT's single-variable method. Additionally, the study integrated autoregressive elements, specifically an ARMAX model from the Auto-Regressive Moving Average (ARMA) series, which includes regression on past ILI values, a seasonal component for yearly patterns, and was trained using maximum likelihood estimation. The GP model demonstrated superior accuracy, outperforming both the Elastic Net and GFT models. The Elastic Net model showed an 8.5% improvement over GFT, while the GP model further improved predictions by 1.1%. Incorporating the GP model into an autoregressive framework (AR+GP) yielded even better results, reducing prediction uncertainty and surpassing baseline and Elastic Net AR models in performance. Despite these advancements, the study acknowledged limitations such as potential data scarcity at more granular geographical levels and the challenge of spurious correlations from unrelated queries. The complexity of the nonlinear GP model also poses interpretability challenges

compared to simpler linear approaches [11].

Lampos et al. aimed to refine these nowcasting results in a subsequent study, through mitigating the impact of spurious search queries. The research explored three principal feature selection strategies to improve prediction accuracy: traditional correlation-based selection, using neural word embeddings, and a hybrid method that combines both. The correlation-based method selected features (search queries) based on their Bivariate correlation with ILI rates, applying different thresholds ($r > \rho, \rho \in [0, 1)$) to identify queries significantly related to ILI rates. The neural word embedding approach leveraged neural networks to infer the semantic context of queries, mapping them into a 512-dimensional space to quantify their semantic proximity to flu-related terms, with feature selection based on exceeding the mean similarity scores by a set number of standard deviations ($S > \mu_S + x\sigma_S, x \in [0, 1)$). The hybrid method aimed to harness the advantages of both strategies, starting with correlation-based selection and refining it through word embedding criteria, thereby capturing both direct correlations in the data and the semantic nuances of the queries. The study's results indicated that the optimal correlation-based model (with a threshold of $r > 0.4$) outperformed the best word embedding model, achieving a MAE of 2.137 compared to 3.006. However, employing the hybrid method, specifically with thresholds of $r > 0.30$ and $S > \mu_S + \sigma_S$, yielded the most effective results - a correlation coefficient of 0.913, an MAE of 1.880, and a MAPE of 36.23%. In comparison, the correlation-based approach alone (with $r > 0.40$) resulted in a Bivariate correlation of 0.876, an MAE of 2.137, and a MAPE of 47.15%. This investigation underscores the benefits of combining traditional statistical learning with semantic analysis via word embeddings for identifying flu-indicative queries, presenting a more comprehensive and statistically robust approach to predicting ILI rates. It acknowledges potential biases in search query data and emphasises the critical role of fine-tuning feature selection thresholds to maintain a balance between feature reduction and model performance [1].

Closer to the application of our study concerning public health in England, Flu Detector emerges as the sole online tool for Influenza surveillance employing UGC in England. Lampos et al. introduced Flu Detector as an online instrument that utilises Twitter and Google search data to estimate ILI rates in England. This tool relies on two primary data sources: Twitter Data: It collects tweets in England through Twitter's Streaming API. Google Search Data: Flu Detector accesses Google Trends outputs via a private Google Health Trends API, providing normalised frequencies of search queries. The tool employs supervised learning techniques to construct models for estimating flu rates based on Twitter and Google search data. It incorporates recent advancements in statistical natural language processing, such as the utilisation of neural word embeddings, as discussed earlier, to effectively process and interpret the feature data. The performance of the Google search-based model exhibits an average MAE of approximately 1.5 per 100,000 people across four flu seasons (2012/13 to 2015/16) when compared to the corresponding rates from the Royal College of General Practitioners (RCGP). The average Bivariate correlation coefficient is reported as 0.95. The paper concludes with an optimistic assessment of Flu Detector's potential as a supplementary source to traditional flu surveillance schemes. Nevertheless, the paper acknowledges certain limitations, including concerns related to data representation and stability due to the relatively modest portion of data sourced from Twitter. [2].

Deep Learning Models

The majority of research efforts discussed have employed linear models, integrating various forms of regularisation or autoregressive components. However, deep learning models have recently shown significant promise, especially in their ability to navigate the complex, nonlinear relationships that characterise the time-series data prediction task.

This potential was notably explored in a study nowcasting the 2015 flu season in the United States. This study employed a range of state-of-the-art deep learning models, including Convolutional Neural Networks (CNN) for their feature learning and classification capabilities, Recurrent Neural Networks (RNN) for handling sequential data effectively, and Sequence-to-Sequence (Seq2Seq) models for processing high-dimensional time-series data. It utilised data from Google Flu Trends, based on search queries across the United States, in conjunction with ILI occurrence data from the CDC, encompassing a comprehensive dataset from all 50 states. Prior to deploying deep learning methods, the data was subjected to a preliminary analysis using classical time-series forecast models such as ARIMA and the Prophet model to assess stationarity, seasonality, and trends. The study found that the RNN model, with a MSE 0.13, demonstrated superior performance over the CNN model, which had a MSE of 0.18. This suggests a greater suitability of RNN for time-series data analysis. Despite the study's limited scope, predicting only a single flu season, it highlights the potential of deep learning models, particularly RNN, in nowcasting ILI occurrences. However, concerns remain around the variability in search queries and the absence of feature selection, which hinders the model's reflection of the actual influenza incidence [17].

In a separate endeavour, He et al. developed a spatiotemporal ILI tracking method in Taiwan using an advanced Spatiotemporal Residual Network (ST-ResNet). This research aimed for accurate ILI tracking at a regional level, incorporating Google search data and public holiday information to augment the prediction of current influenza activity. The study processed ILI data from 19 cities and countries in Taiwan, transforming these into region-based heatmaps to effectively handle the irregular shapes and complex spatial correlations between regions. The extended ST-ResNet model is adept at capturing complex spatial and temporal structures and was compared against various predictive methods including ARIMA, ARIMAX, ARGO, ARGO2, and LSTM. The results showed that ST-ResNet outperformed other models in most cities/counties, evidencing its efficacy in modelling the complex dynamics of influenza transmission. These findings contrast with the previous study that highlighted the effectiveness of sequence-based models over complex CNN architectures. The study concluded that integrating offline ILI data with multisource external data, such as Google search data and holiday effects, leads to more precise influenza tracking. However, it also highlighted the limitations of static models like ST-ResNet and the performance deterioration due to time lag. Moreover, both of the presented models are quite complex architectures where there has been no validation of their performance gains and necessity over simpler architectures [12].

2.4.4 Forecasting Influenza-Like Illness Using Web Search Queries

Time Series Forecasting

Time series forecasting is a method used to predict future values in a sequence of data points. Unlike nowcasting, which focuses on estimating current values, forecasting looks ahead to predict outcomes at future time points, such as days $t + 7, 14, 21$, and so forth. This predictive capability can offer valuable insights towards detecting and preventing outbreaks in advance.

Traditional Machine Learning Models

An early study by Dugas et al. aimed to leverage real-time, accessible data, including GFT, to develop and Influenza forecasting model. Similar to the nowcasting studies they utilised a combination of Box-Jenkins, generalised linear models (GLM), and generalised linear autoregressive moving average (GARMA) methodologies to predict weekly influenza cases one week in advance. The most accurate model was found to be a GARMA(3,0) model that integrated Google Flu Trends data, achieving predictions within seven cases of actual numbers 83% of the time for seven out-of-sample outbreaks. This model underscored the significant improvement Google Flu Trends data provided over base models, making it a valuable external data source [27].

Deep Learning Models

Similar to the nowcasting task, studies have began exploring the potential of deep learning models in forecasting ILI rates.

Xu et al. was the first to integrate deep learning methodologies with Google search data and conventional sources to forecast influenza in Hong Kong. Unlike the nowcasting studies, Xu et al. assessed the predictive capabilities of the traditional regularised and autoregressive models, namely Generalised Linear Model (GLM), Least Absolute Shrinkage and Selection Operator (LASSO), and ARIMA, against a more simple Feedforward Neural Network (FFNN), as opposed to immediately proposing further complex, deeper architectures. They focused on forecasting the number of new cases of ILI in general outpatient clinics. Each model was tasked with predicting ILI cases one to two weeks ahead, leveraging a mix of Google search, weather, and historical ILI data. The FFNN model emerged as notably efficient, particularly in pinpointing the timing of ILI peaks, displaying superior accuracy over the other models, with a Root Mean Squared Error (RMSE) of 1.73 and a MAE of 1.30 for one-week-ahead predictions. By applying Bayesian Model Averaging (BMA) to integrate the forecasts from all models, the study saw an increase in predictive precision, with BMA achieving an RMSE of 1.53 for forecasts a week in advance. This not only underscores the individual strengths of deep learning models but also demonstrates the collective benefit of merging different predictive models in improving forecast accuracy [28].

Another study by Liu et al. also investigated the same task of forecasting influenza in Hong Kong using the same data. Their goal was to create a more accurate way to see when and how bad the flu season would be, which is important for planning healthcare responses. Their approach employed a Stacked Autoencoder (SAE) to understand the large amount of Google search data and a Variational Mode Decomposition (VMD) to break down flu case data into simpler patterns. Using this data, future Flu forecasts were then predicted using an Artificial Neural Networks (ANN). Their approach was more accurate in predicting the flu, even during the COVID-19 pandemic, against traditional autoregressive methods [29].

Despite the robust performances of both proposed methodologies, the researchers acknowledge the increased complexity of these deep learning models and thus the potential overfitting and limitations of their approach in generalising to other locations and data.

Shifting to the challenge of predicting flu trends in England, Morris et al. developed a more advanced deep learning model that combines Bayesian Neural Networks (BNNs) with data from web searches to estimate how widespread ILI is. The intuition behind using BNNs was to tackle the uncertainties often seen in disease forecasting, which come from both the data we have and the models we use. Through analysing 14 years of data from England, specifically the last four flu

seasons, the study compared how traditional neural networks stack up against BNNs that consider these uncertainties. They used a Long Short-Term Memory (LSTM) model, known for its efficacy on time series data, alongside simpler feed-forward networks. The LSTM model was more accurate in forecasts that looked more than a week ahead, especially when it took into account uncertainties from both data and the model itself. For forecasts made a week in advance, this approach had a MAE of 2.29 and a Root Mean Squared Error (RMSE) of 3.82. While it was slightly less accurate in the short term, it was much better for longer-term forecasts, showing the benefits of including uncertainty in the models. Interestingly, the simpler FFNN had a lower MAE of 1.62 and RMSE of 2.5 for the same period but didn't model uncertainty like the BNNs did. This study revealed an important balance in forecasting ILI: adding uncertainty might reduce accuracy for short-term forecasts but greatly improves the reliability and value of forecasts over the long term. It also demonstrates the efficacy of simpler FFNN architectures. This is beneficial for public health planning, where certainty of forecasts can lead to better decisions in response to flu outbreaks [3].

Building on these insights, Morris et al. further explored ILI forecasting in the United States, utilising an Iterative Recurrent Neural Network (IRNN) architecture. This model incorporates Bayesian layers, significantly enhancing forecast accuracy by providing associated uncertainty intervals with their predictions, achieving a 10.3% improvement in MAE and a 17.1% increase in forecasting skill across four flu seasons. The IRNN model's capability to issue early warnings of flu outbreaks represents a significant advancement in employing BNNs for public health forecasting [30]. Despite these strides, the studies addressed the limitations of their work, acknowledging the potential challenges in data reliability and the model's effectiveness across various locations. Collectively, these studies underscore the efficacy of uncertainty modelling, showcasing its role in achieving superior predictive performance in the forecasting task.

2.4.5 Summary

The collective evidence gathered from the various studies demonstrates the effectiveness of UGC in predicting Influenza outbreaks. Specifically, deep learning models have shown promise in this area, outperforming traditional machine learning methods. Whilst there hasn't been an exploration of deep learning models for nowcasting ILI rates in England, an additional challenge arises in the absence of well-established simpler deep learning models that can serve as reliable benchmarks for comparison in both prediction tasks. Currently, there's a trend in the field towards overly complex models, with researchers aiming for maximum performance without clarity on whether such complexity is truly necessary for significant improvements.

This invites further investigation into exploring the predictive abilities of simpler neural architectures, in our given prediction tasks, whilst incorporating advanced feature selection techniques, known to help reduce the influence of spurious queries[1]. Our experiments aim to harness the potential of these simpler models to achieve competitive performances in predicting ILI rates in England, outperforming the more traditional machine learning approaches [1], [2]. This approach also seeks to establish robust benchmarks, where the performance and necessity of more complex architectures, similar to those proposed in the discussed studies, can be evaluated and validated against.

2.5 Review of Employed Machine Learning Algorithms

The following sections will cover the machine learning algorithms our study employs for the prediction task.

2.5.1 Linear Models

Linear Regression

Linear regression is the simplest algorithm for a regression-based supervised learning task. It is a statistical method that aims to establish a linear relationship between a dependent target variable and one or more independent feature variables [31]. In a typical machine learning setting, you start with a dataset comprising n samples, each denoted as $(\mathbf{x}_i, y_i)$. Here, $\mathbf{x}_i$ represents an m-dimensional feature vector (belonging to $\mathbb{R}^m$) and y_i is the corresponding target value in $\mathbb{R}$. This dataset can be expressed as an $n \times m$ design matrix $\mathbf{X}$ and an n-dimensional target vector $\mathbf{y}$, as shown in Equation 2.1:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{2.1}$$

Linear regression seeks to learn a linear function $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^m$ is the weight vector and b is the bias term. The objective is to find the weight vector $\mathbf{w}$ and bias b term that minimises a loss function over all the training samples. The Ordinary Least Squares (OLS) method, for example, aims to minimise the Mean Squared Error (MSE), which is the sum of squared differences between actual values and predictions. This is defined by the following optimisation task:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b))^2, \tag{2.2}$$

where argmin indicates we are minimising the function with respect to the parameters w and b. This can also be expressed in matrix form shown in equation 2.3:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} ||\mathbf{y} - (\mathbf{X}\mathbf{w} + b)||^2. \tag{2.3}$$

In OLS, we are guaranteed a solution because of the optimisation problem is at least convex and bounded. However, the uniqueness of the solution depends on the rank of the hessian and thus the nature of the objective function's convexity.

Underfitting and Overfitting

In machine learning, underfitting occurs when a model is too simplistic to capture the data's complexity, leading to high training and test errors [32]. This often happens in models lacking necessary complexity, like linear models used for non-linear data. Overfitting, on the other hand, arises from overly complex models fitting noise in the training data, resulting in low training errors but poor generalisation to new data [32].

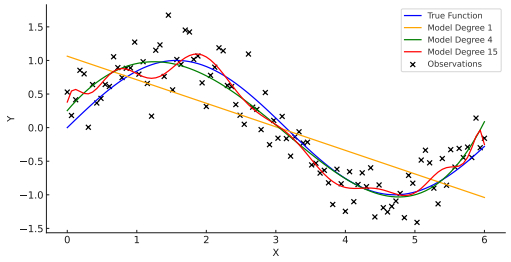

Figure 2.1: Training Data Fit By Polynomial Degree

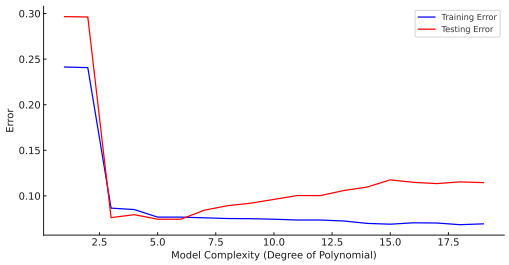

Figure 2.2: Underfitting and Overfitting

The balance between underfitting and overfitting is known as the bias-variance tradeoff. Bias stems from oversimplified models, while variance arises from excessive complexity. The optimal model finds a balance between these errors, minimising validation error and achieving optimal generalisation to new data. Figures 2.2 and 2.3 illustrate underfitting and overfitting, with underfitting represented by high errors for low model complexity and overfitting by close fitting to training data but poor performance on test data.

Regularisation

A key technique used to address the challenge of overfitting is regularisation. This involves introducing a penalty on the complexity of the model. This approach modifies the optimisation function to minimise not only the error but also to keep the model weights as constrained as possible, as these are directly correlated to model complexity, thus preventing overfitting and enhancing the model's generalisation capabilities.

The general form of the regularisation-modified optimisation function is shown in equation 2.4:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - (\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b))^2 + \lambda \cdot \text{Penalty}(\mathbf{w}) \right], \tag{2.4}$$

where λ is the regularisation hyper-parameter controlling the penalty's strength, and Penalty($\mathbf{w}$) is a function that imposes a cost on larger weights. There are three commonly used regularisation techniques:

1. **Lasso Regression (L1 Regularisation):**
   Lasso (Least Absolute Shrinkage and Selection Operator) introduces a penalty equal to the absolute sum of the weights. It can lead to some coefficients shrinking to zero, essentially performing feature selection [33].

$$\text{Penalty}(\mathbf{w}) = ||\mathbf{w}||_1 \tag{2.5}$$

2. **Ridge Regression (L2 Regularisation):**
   Ridge adds a penalty equivalent to the square of the sum of the weights. This method reduces the coefficients' magnitude but does not zero them out, meaning it does not perform feature selection [34].

$$\text{Penalty}(\mathbf{w}) = ||\mathbf{w}||_2^2 \tag{2.6}$$

3. **Elastic Net:**
   Elastic Net is a hybrid of L1 and L2 regularisation, incorporating both penalties. It is

particularly useful when dealing with multiple correlated features [35].

$$\text{Penalty}(\mathbf{w}) = \alpha||\mathbf{w}||_1 + \frac{(1-\alpha)}{2}||\mathbf{w}||_2^2 \tag{2.7}$$

Here, α is a mixing parameter that balances the L1 and L2 penalties.

Model Selection

Cross-validation is a technique in machine learning used to evaluate a model's generalisation ability to optimise model parameters such as the regularisation hyper-parameters in Lasso, Ridge and Elastic Net. It involves partitioning the dataset into multiple subsets and then training each model hyperparamter setting on a combination of these subsets, followed by testing on the remaining subset. This process is repeated several times with varying training and validation subsets. The most common variant is k-fold cross-validation, where the training dataset is divided into k equal parts, known as 'folds'. A given model setting is trained and validated k times, each time using $k-1$ folds for training and one fold for validation. The results from each validation fold are averaged to yield a comprehensive performance measure, where the best performing model setting is then carried forward for training on the entire training data. This approach ensures every data point is utilised for both training and validation, providing a robust evaluation of the model's performance.

2.5.2 Neural Networks

Neural Networks represent a significant leap from traditional models, offering a more flexible architecture to better model complex, non-linear relationships in data.

Feed Forward Neural Network

A Feed Forward Neural Network (FFNN), serves as the foundation for more advanced neural architectures [36]. An FFNN consists of an input layer, multiple hidden layers, and an output layer, all fully connected. Each unit in the input layer, denoted as x_1 to x_5 in Figure 2.4, links to each unit in the next hidden layer through weighted connections.

The input layer receives data, where each unit corresponds to a feature of the input vector $\mathbf{x}_i \in \mathbb{R}^n$. Neurons in the hidden layers compute a weighted sum of the previous layer's outputs, adding biases and applying a non-linear activation function σ to enable non-linear data modelling. Various activation functions cater to different data types and predictive tasks. For instance, the ReLu function is commonly used for its efficiency and ability to mitigate the vanishing gradient problem [37].

Each hidden layer refines the data transformation, enhancing the network's depth and complexity. The output layer's structure varies with the task: a single neuron for regression or multiple for classification. It combines the data processed by the last hidden layer with weights and biases to produce the final outputs or predictions.

Training

Training a FFNN involves several key steps: weight initialisation, forward and backward passes, and optimisation, each crucial for effective learning.

The process begins with weight initialisation, setting the initial values for the neuron connections. Proper initialisation, using techniques such as Xavier and He initialisation, is important to

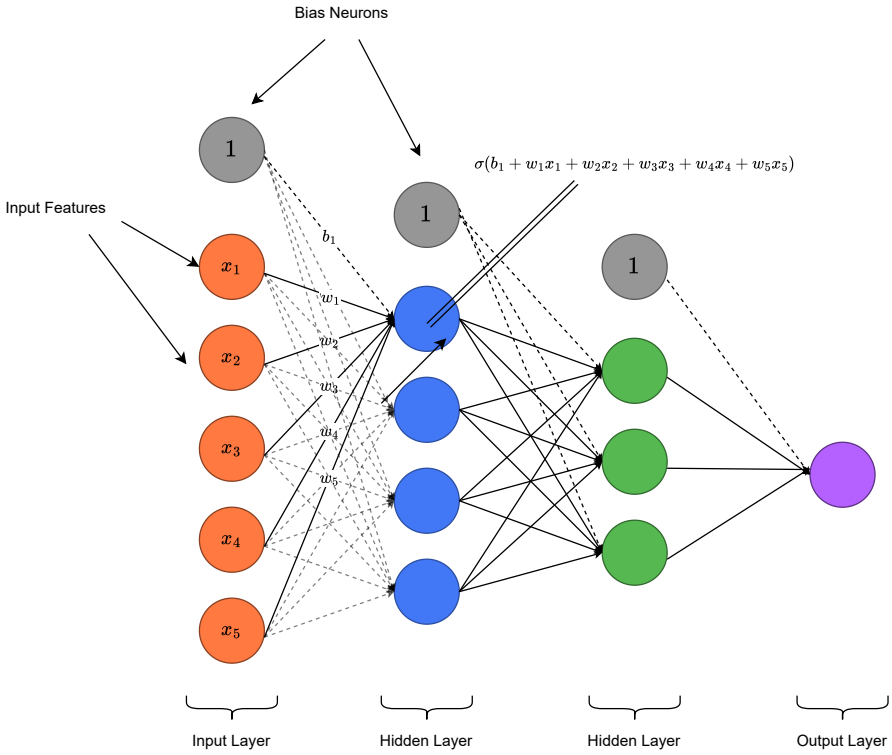

Figure 2.3: Architecture of a vanilla Feed Forward Neural Network (FFNN)

avoid issues like gradient vanishing or exploding during early training [38], [39].

During the forward pass, input data is processed through the network, with each neuron calculating a weighted sum of its inputs, applying an activation function, and passing the output onward. This culminates in an initial prediction at the final layer. To refine these predictions, backpropagation adjusts the weights based on errors, calculated by a loss function. This involves a backward pass through the network, updating each weight based on its error gradient. The most common optimisation technique used is Gradient Descent, which updates weights in the direction opposite to the gradient to minimise error as shown in Equation 2.8:

$$\mathrm{w_{t+1}} = \mathrm{w_t} - \eta \cdot \nabla_{w_t} l, \tag{2.8}$$

where w_{t+1} is the updated weight at step t+1, w_t is the weight at step t, η is the learning rate, and l is the loss function.

Gradient Descent has several variations: batch, which uses the entire dataset for each update; Stochastic Gradient Descent (SGD), which updates weights after each example, offering faster convergence at the cost of higher variance (shown in Figure 2.6); and Mini-Batch Gradient Descent, which strikes a balance by using subsets of the dataset, combining rapid convergence with lower variance [40].

The learning rate η is a critical hyperparameter in gradient descent, influencing the step size during weight updates. Incorrectly set, a high rate may cause divergence, while too low a rate can slow convergence and risk getting stuck in a local minima.

Among optimisation algorithms, AdaGrad adjusts the learning rate for each parameter based on the historical squared gradients, enhancing updates for infrequent parameters while diminishing

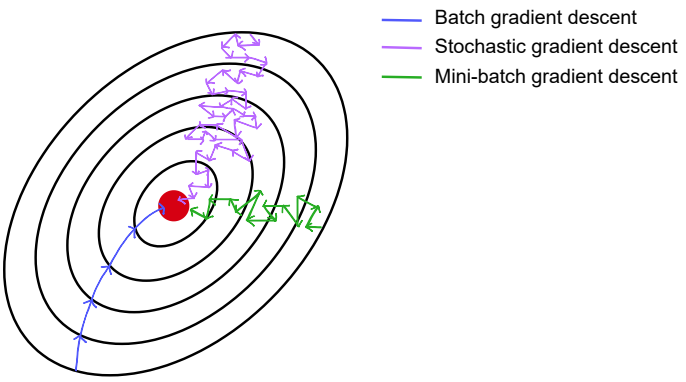

Figure 2.4: Gradient descent convergence

updates for frequent ones. However, its accumulative nature leads to a continuously decreasing learning rate, which can stall training in deep networks [40].

To address the diminishing learning rate issue, RMSProp modifies the AdaGrad approach by using a decaying average of squared gradients, allowing for more balanced updates and preventing the learning rate from reducing too swiftly. While RMSProp helps stabilise training, it does not incorporate momentum, which can affect convergence rates by smoothing the optimisation path [40].

For our models, we employ the Adam optimiser, which combines the benefits of AdaGrad and RMSProp and includes momentum. Adam adjusts learning rates by estimating both the first (mean) and second (uncentered variance) moments of the gradients, incorporating a bias correction for more accurate updates early in training [40]. This optimiser enhances convergence speed and efficacy.

After completing a full forward and backward pass through the entire training data and subsequently updating the weights through gradient descent, a single epoch of training has been completed. Typically, multiple epochs are involved in training, with the network repeatedly computing its loss and updating its weights accordingly, improving performance with each epoch until it converges to a stable solution or reaches a predetermined number of epochs. The final output neuron result obtained at the conclusion of all forward and backward passes represents the model's prediction.

Chapter 3

Influenza-Like Illness Rate and Web Search Query Data

This chapter begins by describing the dataset utilised for our tasks of nowcasting and forecasting ILI rates in England. It then outlines the comprehensive data preprocessing steps undertaken to ensure the data was adequately prepared for model training. This is followed by an analysis section, providing insights into the properties and trends present in the data, informing our subsequent experimental decisions.

3.1 Dataset

In this study, we utilised a time series based dataset spanning from January 2009 to September 2019. This dataset encompasses both weekly ILI rates and daily web search query frequencies.

The ILI rates were sourced from the Royal College of General Practitioners (RCGP) and UK Health Security Agency (UKHSA), representing the incidence of doctor consultations for ILI symptoms per 100,000 individuals in England. Figure 3.1 displays the interpolated daily ILI rate across various flu seasons, with the five colour-coded plots corresponding to the flu seasons predicted in our nowcasting and forecasting tasks. Additionally, the 18,676 web search query frequencies, were obtained through an academic API provided by Google Health Trends for academic research purposes with a health oriented focus. These are a non standardised version of the publicly available Google Trend outputs. Examples of these queries are provided in Table 3.1. The search query frequencies underwent a normalisation process, scaled within a range from 0 to 1. Specifically, the normalised frequency of a given query q on day t is represented as:

$$q_t = \frac{\#\text{ searches of } q \text{ on day } t}{\#\text{ searches on day } t}. \tag{3.1}$$

This normalisation adjusts for the daily variations in overall search volume, ensuring a consistent and reliable comparison analysis.

3.2 Data Preprocessing

The initial phase of our study involved a series of preprocessing steps to prepare the data for the machine learning models used in predicting ILI rates. Our first step addressed the discrepancies in

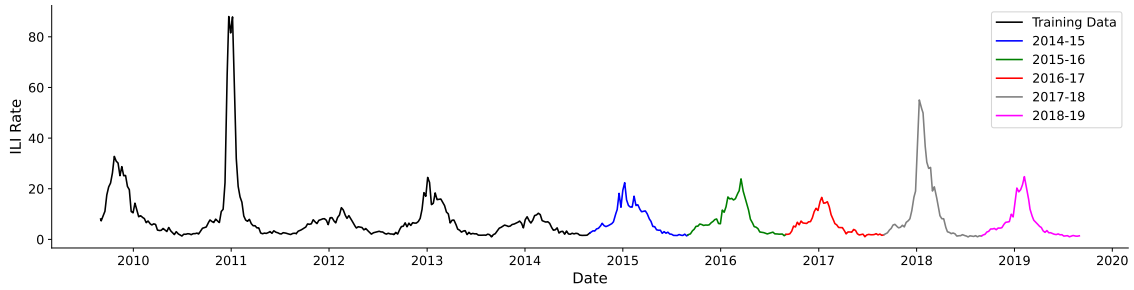

Figure 3.1: Daily influenza-like illness (ILI) rates in England from September 1, 2009 - August 31, 2019, obtained by RCGP and UKHSA. Training and test periods are denoted with different colourings.

data granularity between the weekly ILI rates and the daily normalised search query frequencies. We began by interpolating the weekly ILI rates into daily rates. This was achieved by taking each weekly ILI rate to represent the ILI rate for the Thursday of that week. For days falling between two consecutive Thursdays, we employed linear interpolation. For a given week with an ILI rate I_w corresponding to the Thursday, the daily ILI rate I_d for days between this and the next Thursday (Friday of week w to Wednesday of week $w+1$) was computed using:

$$I_d = I_w + \left(\frac{I_{w+1} - I_w}{7}\right) \times d, \tag{3.2}$$

where d represents the index of the day for which we are interpolating the rate. Specifically, Friday is indexed as 1, Saturday as 2, up to Wednesday as 6. Following this, we rectified minor discrepancies in the dates between the ILI rate data and the normalised web search query frequencies by aligning the datasets to ensure synchronisation of dates across both sets.

Our study aims to predict the ILI rates for five distinct flu seasons, spanning from 2014-15 to 2018-19. We define each flu season, y to $y+1$, to cover the period from 01-09-y to 31-08-$y+1$. Daily search query frequencies are utilised as predictors for the corresponding daily ILI rates. The training data for predicting a specific flu season includes all available data from previous seasons, beginning with the 2009-10 season. Therefore, for the first flu season prediction - 2014-15, the training dataset encompasses data ranging from 01-09-2009 to 31-08-2014. The training data for subsequent flu season predictions would just add the test data from the previous flu season prediction.

An important aspect of our preprocessing involved the examination of search query frequencies for potential redundancy. This process focused on identifying and removing queries that consistently showed zero normalised frequency throughout the training period of each flu season prediction. Given that the training data for our first flu season prediction (01-09-2009 to 31-08-2013) forms a subset of the training data for subsequent season predictions, it is impossible for a query to transition from having a single non-zero frequency in the initial season to a 'only' zero frequency in later seasons. Consequently, we eliminated 1160 queries that had a zero normalised frequency throughout the entire duration of the first flu season prediction's training data, implicitly removing such queries in subsequent flu season prediction training data's. This removal of non-informative features was designed to refine our dataset, thereby potentially augmenting the efficiency and effectiveness of our model's learning process for each season prediction.

Following this, their still existed some redundancy in our data in the form of duplicate query features. We discovered several instances of queries composed of identical words but in varying orders such as 'flu nhs' versus 'nhs flu' or 'flu jab nhs' versus 'nhs flu jab'. These variations, while syntactically distinct, did not contribute additional value as separate features within our analysis. To streamline our dataset and enhance the accuracy of our predictive models, we opted to combine these duplicates into a single normalised query frequency. This was achieved by summing the normalised frequencies of the duplicate queries for each day, combining their influence on the ILI rate predictions whilst preserving the integrity of the query normalisation. As a result, 316 duplicate queries were removed leaving us with a dataset comprising of 17200 queries.

The final stage of our data preprocessing involved the smoothing of the normalised query frequencies. This process was essential to reduce noise and variability in the query data, potentially impacting the predictive accuracy of our models. To achieve a smooth temporal profile for each query, we applied a weighted moving average over a 14-day window. This window size was chosen as it effectively balances capturing the relevant temporal trends, including the typical incubation period of influenza [41] and weekly search patterns, while ensuring computational efficiency. Specifically, the smoothed normalised frequency of a query on a given day t, denoted as q_t^{smooth}, was computed using the weighted average of its frequencies from the current day and the preceding 13 days.

$$q_t^{\text{smooth}} = \frac{1 \times q_t + \frac{13}{14} \times q_{t-1} + \frac{12}{14} \times q_{t-2} + \ldots + \frac{1}{14} \times q_{t-13}}{1 + \frac{13}{14} + \frac{12}{14} + \ldots + \frac{1}{14}} \tag{3.3}$$

This weighted averaging approach assigns a gradually decreasing weight to the query frequencies as we move backward in time, with the most recent day (day t) having the highest weight. The rationale behind this weighting scheme is to give more prominence to recent query activities while still capturing the trend over the preceding two weeks. By smoothing the data in this manner, we aimed to create a more stable and representative time series for each query, which is expected to enhance the predictive performance of our machine learning models.

3.3 Data Analysis

Following the data preprocessing, we conducted some preliminary analysis on our data in effort to understand the properties and relationships present, motivating subsequent experimental decisions.

3.3.1 Influenza-Like Illness Rate Trend

Figure 3.1 illustrates the seasonal variations and peak ILI periods across multiple flu seasons. The graph consistently reveals a marked increase in ILI rates during the winter months, followed by a gradual decline as the warmer seasons commence, aligning with anticipated seasonal health trends. This cyclical nature of ILI occurrences is further explored in Figure 3.2, where ILI rates are normalised for each flu season prior to calculating the daily mean ILI rate across all seasons. The resultant mean curve parallels this seasonal trend, while the shaded region, representing one standard deviation above and below the curve, conveys a $\tilde{68}\%$ confidence interval for this trend, highlighting the variability in ILI rates across different flu seasons.

Despite the general seasonal trend, the pattern is not uniform across all flu seasons. Specifically, Figure 3.1 indicates that certain seasons, such as 2009-10, 2011-12, and 2018-19, experience a sharp

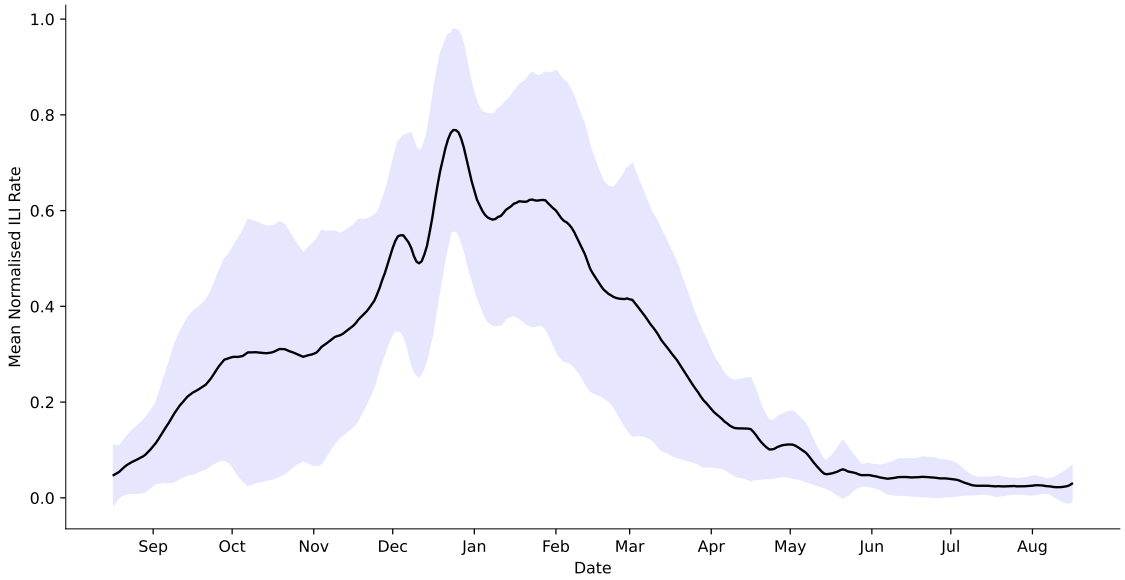

Figure 3.2: Mean daily normalised influenza-like illness (ILI) rates, in England across flu seasons, ranging from 2009-10 to 2018-19. The graph features a shaded region representing one standard deviation above and below the mean curve, which serves as a 68% confidence interval for the ILI rate across flu seasons. The ILI rates for each flu season were normalised independently in effort to compare seasonal variations in the trend across flu seasons.

increase to a distinct peak in ILI rates, while others, such as 2012-13 and 2014-15, exhibit additional smaller, delayed peaks instead of a straightforward decline. Furthermore, the seasons of 2015-16, 2016-17, and 2018-19 present additional peaks prior to the largest spike. The occurrence timing of these peaks also varies; for instance, the 2009-10 peak transpires towards the end of the year, whereas in other seasons, it aligns with the transition to a new year (2015-16 and 2017-18) or spans the end and beginning of adjacent years, as seen in 2010-11. This diversity in peak timing and frequency is evidenced by the mean curve in Figure 3.2, where the shaded region, indicating variability, is most extensive around the periods preceding the peak, in October-November, and following it into February-March.

In addition to these observed variations within the ILI trends, Figure 3.1 also highlights significant differences in the actual ILI rate values reported across flu seasons. Certain seasons, such as 2010-11, exhibit peak ILI rates significantly higher than the average, with values above 80, largely due to the H1N1 pandemic's impact. Similarly, the 2017-18 season peak, which exceeds 60, presents challenges to the model's predictive accuracy during unusual events. These findings underscore the need for our models to not only capture general seasonal patterns but also accurately predict ILI rates by adapting to the unique characteristics of each season, particularly in the face of atypical spikes.

3.3.2 Web Search Queries

As mentioned, our dataset comprises 18,676 web search queries. Table 3.1 illustrates some examples of queries present in our dataset. These are categorised into 3 groups based on their Bivariate correlation with the ILI rate: a strong positive correlation, in the top 1% of correlated queries, almost no correlation, from a middle percentile and a strong negative correlation, in the bottom 1% of correlated queries. Ten queries from each category were randomly selected to showcase the

Strong Positive Correlation	Almost No Correlation	Strong Negative Correlation
flu symptoms	hand foot mouth virus	sun allergy
laryngitis	spot on lung	leg cramp
flu epidemic	diabetes wiki	treatment for sunburn
bad flu	caffeine and breastfeeding	hayfever eyes
starve a cold	how to relieve back pain	varicose veins
colds and flu	fat flush diet	foot blisters
symptoms of the flu	why does my chest hurt	symptoms of hayfever
treat flu	hepatic	how many calories in watermelon
get rid of a cough	prognosis definition	a stye
chest infection symptoms	surgeon games	ingrown

Table 3.1: Examples of web search queries in our dataset, obtained from Google Health Trends. These are categorised based on their correlation levels with the ILI rate: Strong Positive Correlation, encompassing 10 randomly selected queries ranking in the top 1% of correlated queries; Almost No Correlation, featuring 10 randomly selected queries chosen from the middle percentile and Strong Negative Correlation, which includes 10 randomly selected queries positioned in the bottom 1% of correlated queries.

range of queries present in our dataset and their diverse statistical relationships with ILI rates.

Queries with a strong positive correlation typically bear a semantic relation to influenza, as they relate to symptoms or treatments associated with the flu. This category spans straightforward keyword searches such as 'flu symptoms' and 'bad flu', to more nuanced, information-targeted queries like 'get rid of cough'. These queries are especially valuable for our prediction task because they suggest a direct search for information on flu-like symptoms, thereby more likely to correspond with actual influenza incidence within the population.

Figure 3.3 illustrates the trend of search query frequencies across various flu seasons, showcasing an example from each category of correlation. The highly correlated query, 'flu symptoms', mirrors the seasonal trend observed in ILI rates, with its frequency rising during the winter peak of flu activity and decreasing in warmer months. Notably, this query's frequency pattern captures the subtle seasonal characteristics of each individual flu season, including multiple peaks and their varying timing. Furthermore, the frequency of 'flu symptoms' searched aligns with significant ILI outbreaks in the 2010-11 and 2017-18 seasons, indicating public concern during the H1N1 pandemic, in the case of the former. This alignment underscores the query's significant potential for predicting influenza trends.

Queries demonstrating almost no correlation encompass a diverse range of searches. These include queries entirely unrelated to flu, such as 'surgeon games', alongside those associated with medical conditions that, while suggestive of illness, are not specifically linked to the flu, like 'hand foot mouth virus'. These searches exhibit some degree of seasonality, as depicted in Figure 3.3, yet their lack of a robust, consistent pattern, that also aligns with the trend in ILI rates diminishes their relevance for predictive task.

Queries with a strong negative correlation highlight the dataset's breadth, including searches that are not related to ILI, however are more pertinent to opposite seasons, such as 'sun allergy' and 'treatment for sunburn'. The frequencies of these queries peak during the summer, reflecting their semantic relation the sun and hot weather, presenting an inverse relation to ILI rates. This illustrates how search behaviours shift with the seasons and reinforces the need to differentiate between related and unrelated queries for our given prediction task.

This analysis of various query types highlights the extensive range of queries available in our dataset. It also emphasises the importance of feature selection in refining our predictive model's

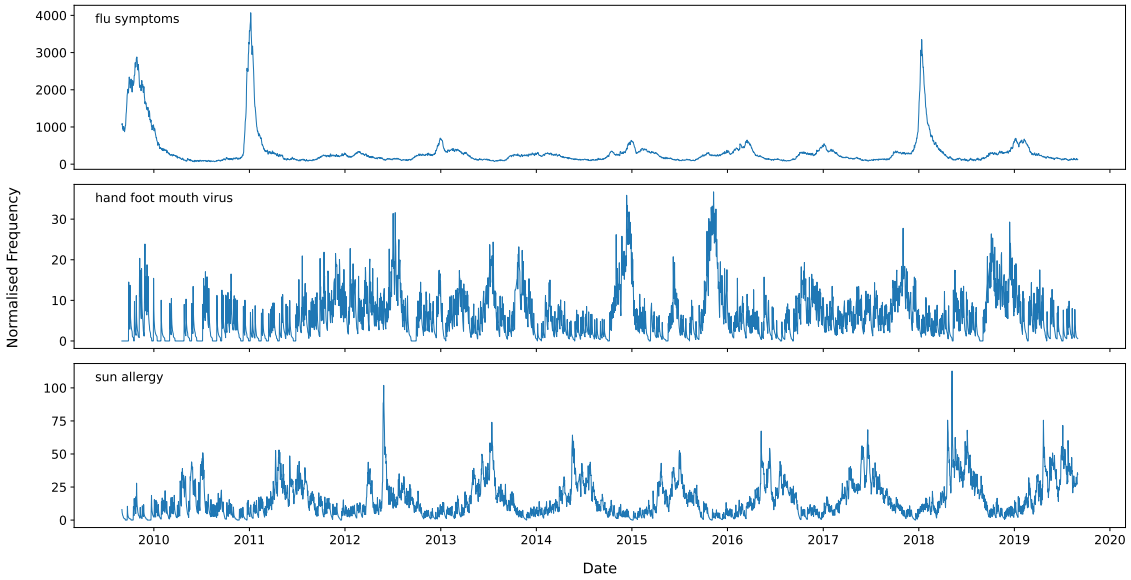

Figure 3.3: Daily web search query frequencies of three queries in England from September 1, 2009, to August 31, 2019, obtained by Google Health Trends. The queries, 'flu symptoms' (strong positive correlation), 'hand foot mouth virus' (almost no correlation), and 'sun allergy' (strong negative correlation), exhibit varying degrees of correlation to the ILI rate.

input by identifying queries relevant to Influenza. This process helps mitigate the impact of spurious queries on our model's learning, increasing the precision in which it can nowcast and forecast ILI rates.

3.3.3 Web Search Query and ILI Rate Correlation

As previously observed, various flu seasons exhibit their own individual characteristics, including distinctive peaks in ILI rates, often attributed to factors like occasional pandemics. These variations suggest that different flu seasons may feature varying prominence of specific symptoms and flu-related illnesses. Consequently, certain web search queries might demonstrate a stronger association with flu and thus a higher correlation with ILI rates during specific flu seasons compared to others. In light of this, our aim was to investigate how the correlation between search queries and ILI rates varies across flu seasons, informing the methodologies employed during our feature selection.

Our approach commenced through calculating the Bivariate correlation between the normalised frequencies of search queries and ILI rates for each flu season, filtering the top 20 correlated queries. Our analysis extended beyond individual flu seasons to a comparative examination, comparing the correlations of a given top 20 queries across adjacent flu seasons. This inter-seasonal perspective allows us to discern the evolution of query correlations' strength over a short time. Adding a quantitative dimension to our analysis, we were also interested in the frequency in which each query appeared in the top 20 correlated list across flu seasons. Hence, Figure 3.5, additionally plots the distribution of frequencies in which a query appears in the top 20 correlated list across flu seasons. This quantitative representation offers clear insights into the consistency in which queries are deemed highly correlated across flu seasons.

A subset of the most informative results are presented in Figure 3.4, while the complete series of graphs can be found in the appendix. Figure 3.4 specifically illustrates the aggregated data for the earlier flu seasons of 2010-11 and 2011-12 along with the later flu seasons 2016-17 and 2017-18.

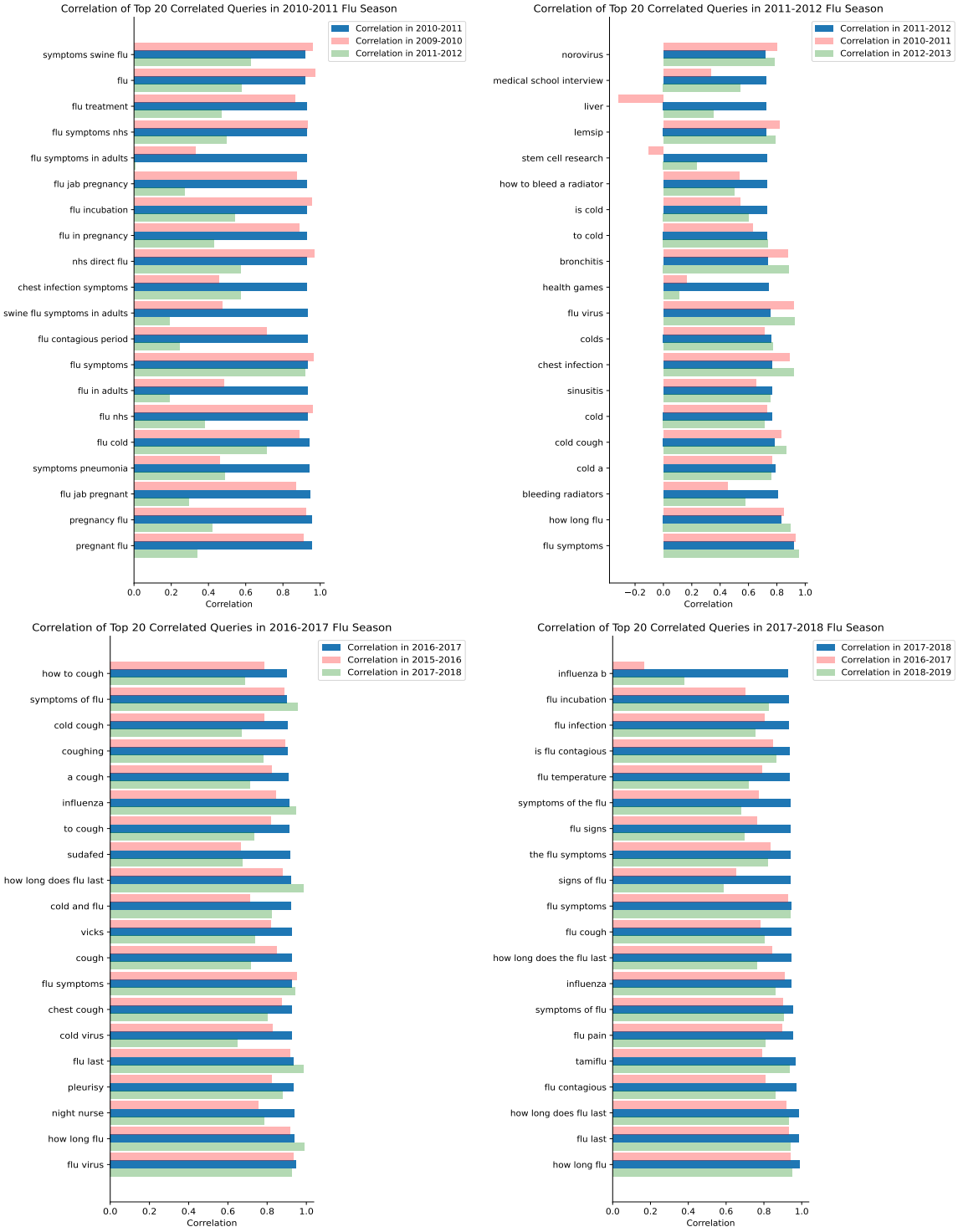

Figure 3.4: Correlations of the top 20 correlated web search queries in a given flu season $y-y+1$. Correlation values of these queries are reported for $y-y+1$ and adjacent flu seasons. The graph highlights fluctuations in the top 20 correlated queries across flu seasons, focusing particularly on correlations between adjacent seasons.

During the initial flu seasons of 2010-2011 and 2011-2012, a significant variation in the correlation of search queries with ILI rates is evident. In 2010-2011, the 'symptoms swine flu' query registers

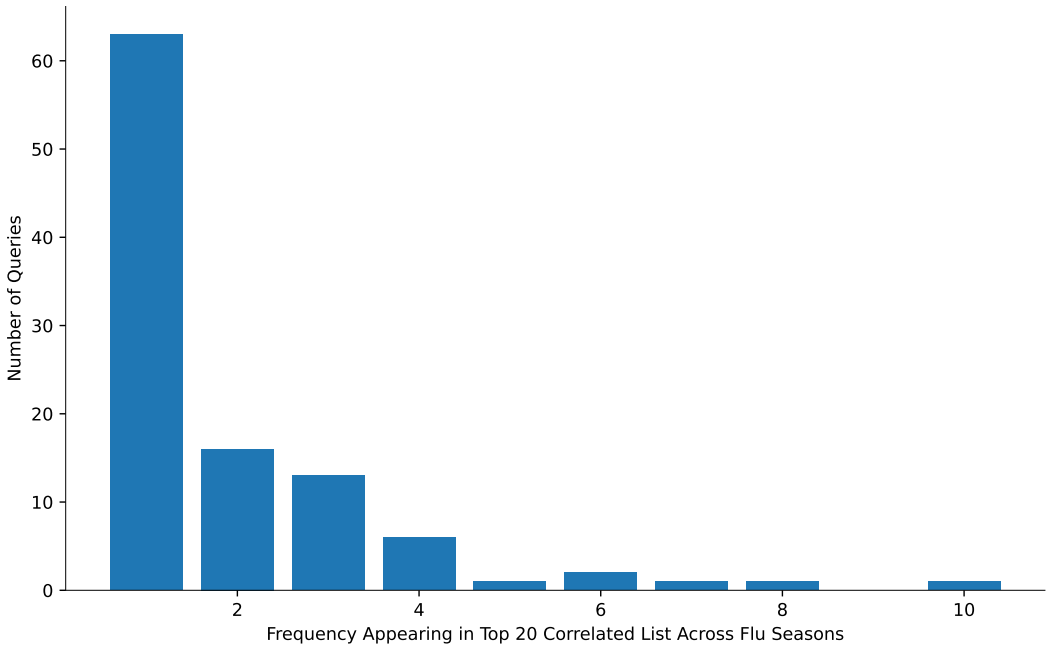

Figure 3.5: Frequency distribution of queries within the top 20 correlated list across flu seasons. The graph illustrates the count of queries at different frequencies within the top 20 correlated list across multiple flu seasons. The distribution highlights the variability in query correlations across flu seasons, with a notable proportion of queries appearing infrequently in the top 20 list.

a near-perfect correlation, reflecting public attention during the H1N1 pandemic. However, this query's correlation drops to around 0.6 in the following season and exits the top 20 list, illustrating a decrease in concern as the pandemic's immediacy recedes. This pattern can be observed in the majority of queries present in this top 20 list including pregnancy and adult related terms such as 'pregnancy flu', 'flu jab pregnant', and 'flu symptoms in adults'. These also show substantial correlations initially but later exhibit a decline in both correlation and frequency. Such a trend indicates a shift from specific searches to a broader range of flu-related concerns as the public health landscape stabilises.

In the 2011-2012 season, the fluctuation in query correlations is more pronounced, with some queries in the top 20, showing negative correlations in subsequent seasons. Examples such as 'liver' and 'stem cell research', despite their high correlation in one season, do not appear flu-related, suggesting an incidental or anomalous correlation possibly tied to overlapping seasonal activities outside the realm of influenza.

The later flu seasons of 2016-2017 and 2017-2018 exhibit a slightly different pattern. Although there is some variability in the top 20 correlated queries between these seasons, the correlations of these queries are more consistent. Most notably, the correlation of the majority of queries in these lists varies by less than 0.3 across adjacent flu seasons. This variability is significantly reduced compared to the pronounced correlation differences observed between adjacent seasons in earlier flu seasons. This stability may indicate a period without significant pandemic events, leading to more consistent public search behaviour that closely aligns with perennial flu symptoms and concerns.

In summary, the examination of search query data from earlier seasons emphasises the influence of specific health events on search behaviour. Dramatic shifts in correlations and the composition of top queries between seasons are evident, particularly during periods marked by pandemics or other

public health emergencies. Conversely, the data from later seasons display greater consistency, with a set of queries that maintain a steadier correlation with ILI rates over time. Consequently, for predicting ILI rates in a given flu season, a correlation feature selection strategy that prioritises the most recent flu season data will mitigate the risk of incorporating outdated queries that were highly correlated due to past health crises but are no longer relevant. This approach ensures the model is reflective of recent health concerns and patterns, thereby enhancing the predictive accuracy of the model.

Nevertheless, the analysis also uncovers that even in later seasons, not all queries maintain a consistent correlation from one season to the next, with the number of queries appearing in the top 20 list of other seasons fluctuating significantly. As illustrated in Figure 3.5, a substantial majority of queries (over 60) ranked in the top 20 for only one flu season. The number of queries in the top 20 diminishes progressively between two to four appearances, with a mere six queries featuring in more than four seasons' top lists. This observation underscores a significant variability in the top 20 correlated queries across different seasons, with 'flu symptoms' being the rare consistent presence. This persistent variability underscores the existence of individual seasonal differences that can influence search behaviours, irrespective of whether the data is from earlier or later seasons. Such individual variations imply that relying solely on this recent correlation-based feature selection approach, might not capture the full spectrum of relevant queries.

This may necessitate a more sophisticated feature selection methodology that can account for the nuanced semantic context of search queries. This is where recent research efforts have been directed, investigating methodologies that explore the semantic context of search queries [11]. Approaches employing sentence embedding provide a means to understand the textual meaning of queries and thus their semantic relevance to Influenza, while hybrid methods aim to integrate this semantic analysis with traditional correlation techniques. The exploration of these advanced methods is expected to refine the selection of features, ensuring that the model captures the most informative and indicative signals for ILI trends, leading to reliable and accurate nowcasting and forecasting predictions.

Chapter 4

Nowcasting Influenza-Like Illness Rates With Regularised Linear Regression

This chapter details the nowcasting experiments conducted and corresponding analysis of results regarding the baseline Elastic Net model. This begins with the exploration of different feature selection methods before training the model on each of these feature sets in an effort to determine the most effective method and the best baseline nowcasting performance.

4.1 Feature Selection

From the 17,200 web search queries present in our preprocessed dataset, accurately identifying those that truly reflect ILI rates is an important aspect of the prediction task. As discussed in earlier chapters, a notable challenge when using UGC, specifically web search queries, for flu surveillance, is discerning irrelevant, spurious queries from those genuinely indicative of ILI symptoms. Recent studies have demonstrated the effectiveness of employing diverse feature selection methods to mitigate the effect of these irrelevant queries, thus improving predictive performance [1]. Taking motivation from these studies, and our analysis of the data, we employ three feature selection techniques in an effort to obtain the most relevant set of features for our models. Subsequent sections will detail the implementation of each feature selection method and the resulting feature sets derived from them.

4.1.1 Correlation

The initial approach we explored was a correlation-based feature selection. This method measures the Bivariate statistical correlation r, between ILI rates and the normalised frequencies of web search queries within the dataset, gauging their direct predictive capabilities. We established a threshold of $r \geq 0.5$ to select approximately the top 400 correlated queries for model training. This threshold was chosen to balance between capturing highly correlated queries and maintaining a sufficient number of features for the model to train on. Following the findings of our preliminary data analysis, we applied the correlation-based feature selection on a window consisting of the most recent flu seasons in the training data. This enabled us to mitigate the influence of significant fluctuations in query correlations from earlier to later flu seasons. After some preliminary experimentation with the data, a 5-year window was found to provide the best results, otherwise, the number of samples may become too small, overfitting the specific trends present in recent flu

seasons. Hence when predicting a given flu season y-y+1, the correlation-based feature selection is applied to the subset 01-09-y-5 to 31-08-y of the training data.

This method efficiently captures queries that not only exhibit this robust correlation but also are semantically related to influenza, demonstrating their predictive utility for our models. Examples include 'flu symptoms', 'flu treatment', and 'flu prevention'. However, this method also includes certain queries where their correlation might not directly indicate an actual increase in flu incidence. For example, the query 'flu jab side effects' might show significant spikes in search frequency during vaccine rollout periods, reflecting public interest or concern rather than an actual uptick in flu cases. Similarly, 'hot water but no heating' tends to be more frequently searched during colder months, which coincidentally align with flu seasons, but does not directly relate to flu activity. These examples underscore the potential drawbacks of relying solely on statistical correlation. Wile statistically significant, they require careful contextualisation to avoid misleading interpretations in flu surveillance. These findings align with our preliminary analysis, which additionally indicated that fluctuations in query correlations, even present in the later flu seasons, potentially render the correlation-based method alone insufficient for selecting the most predictive features. This illustrates the need of also considering the semantic meaning of queries in effort to provide greater context to our feature selection process, thus refining the predictive capabilities of our models.

4.1.2 Sentence Embedding

Aligning with recent advancements in the field [1], we utilise sentence embeddings to understand the textual semantics of queries more deeply, discerning their relevance to flu beyond mere correlations. Specifically, we employ all-MiniLM-L12-v2, a pre-trained sentence transformer model from the SentenceTransformers Python library. This is a fine-tuned variant of the microsoft/MiniLM-L12-H384-uncased model, utilised in a wide range of studies and was extensively trained across a range of datasets, including healthcare-related terminology. More information about this model is provided in the Software and Tools Appendix. This feature selection approach involved utilising the pre-trained model to encode each query into a 384-dimensional vector. We then calculated the cosine similarity of these vectors with the vector encodings of manually selected base queries that are closely related to influenza. This provided a similarity measure based on how closely query's are related in regards to their semantic meaning.

Through conducting some preliminary testing with a range of queries, we curated a set of base queries spanning various flu-related topics, providing a comprehensive domain coverage in capturing the most ILI indicative set of selected features. These queries included direct illness references like 'flu' and 'influenza', alongside terms associated with symptoms and treatments. Recognising the nuanced needs of flu surveillance, we also considered queries tied to higher-risk groups such as babies or pregnant women, acknowledging that searches within these demographics often indicate a heightened concern and probable association with ILI symptoms, thus likely to seek medical advice. Hence these queries may provide more predictive potential in the prediction tasks. The final selection of base queries was 'flu', 'flu NHS', 'influenza', 'pregnancy flu', 'baby flu', 'flu symptoms', 'how to get rid of flu', 'flu vaccine NHS', and 'flu medicine'. By averaging the cosine similarities computed against each of these base queries, we derived a final similarity score for each query in our dataset. Rather than employing a numerical threshold as done in the correlation-based approach, we implemented a hard filter to select the top 400 queries, establishing a semantically rich filter for query selection.

Similar to the correlation approach, this method identifies a set of queries that are semantically related to influenza and exhibit a strong correlation with the target variable. It also includes semantically related queries that, while they may have slightly lower correlations, possess greater semantic relevance and potentially offer better predictions of ILI rates. For instance, this approach selects queries like 'flu treatment options' and 'ways to prevent flu', which are overlooked by the correlation method. These queries contribute to a more comprehensive understanding of public discussions about flu, compared to other queries mentioned, that despite higher correlations, are not truly indicative of flu trends.

However, this method can also include queries like 'man flu meme' or 'flu jokes', which, while containing the term 'flu', are misleading in their semantic relevance to our specific context of influenza surveillance. These terms do not accurately relate to the actual medical condition of influenza or reflect relevant public health concerns, thus failing to correlate with ILI rates. This inclusion of irrelevant queries illustrates a significant limitation of relying solely on semantic similarity: it can sometimes capture queries that are neither semantically valid within the correct context nor statistically correlated with ILI trends. This can reduce the predictive accuracy of our models. This demonstrates the necessity to additionally explore a more nuanced approach that can filter out such misleading data, underscoring the potential of a hybrid method combining both semantic depth and statistical relevance.

4.1.3 Hybrid

In our pursuit to refine the feature selection process further, we devised a hybrid approach that leverages the strengths of both correlation and sentence embedding approaches. This strategy initially casts a wider net by selecting the top 1000 queries based on their cosine similarity scores, as determined by the sentence embedding method previously outlined. Upon establishing this expansive set of semantically relevant queries, we then apply a more lenient correlation-based analysis. Specifically, we impose a correlation threshold of ≥ 0.3 to sift through these semantically related queries, aiming to identify those that not only share a meaningful semantic link to influenza but also exhibit a reasonable correlation with ILI rates. This dual-filtering process allows us to narrow the initial selection down to again approximately 400 features that are now both semantically pertinent and statistically correlated in relation to ILI rates.

The hybrid method not only captures queries recognised by both the correlation and sentence embedding methods, such as 'flu symptoms', 'flu treatment', and 'influenza', but also efficiently integrates queries selected by each individual method. This includes queries like 'flu vaccination side effects', selected due to its statistical relevance in the correlation-based method, and 'best flu medicine uk', selected due to its semantic richness. This ability to integrate selections from both standalone methods while maintaining those recognised by both exemplifies the dual focus of the hybrid approach, ensuring a comprehensive and reliable indicator set for flu activity predictions.

Moreover, the hybrid method excels in identifying queries that were not captured by either standalone method. For instance, question based queries like 'how long to recover from flu' and 'how long does a flu last' did not rank high enough on either individual method's criteria but are clear indicators flu amongst the population. By including such queries, the hybrid method fills critical gaps in our dataset, offering a more nuanced understanding of flu trends across various query types and scenarios.

Additionally, the hybrid approach strategically excludes less relevant queries that were previously discussed. For example, it omits 'hot water but no heating', a query captured by the correlation approach due to its seasonal trends, as well as 'flu jokes' and 'flu game', which were selected by the sentence embedding method because of their loose semantic connections to flu. Although these terms are correlated or semantically linked, they are misleading and do not serve as robust indicators of influenza in our context. By excluding these queries, the hybrid method avoids diluting the predictive accuracy with irrelevant data, ensuring more actionable insights into flu trends and severity.

This refined selection underscores the hybrid method's capability to strike a balance between semantic insights and statistical correlations. It adeptly captures a wide range of relevant queries by harnessing the advantages of both the underlying methods while mitigating their respective drawbacks, reducing the inclusion of spurious or irrelevant queries. This sets a solid foundation for us to investigate the predictive capabilities of these various feature sets by training our baseline Elastic Net model with each of the feature selection approaches, in effort to obtain the best baseline nowcasting performance.

4.2 Elastic Net

Following the implementation of the feature selection methods, we then proceeded to train the baseline Elastic Net model.

To recall from chapter 3, we assess the predictive performance of our models on 5 distinct flu seasons: 2014-15 through 2018-19. The data for each flu season, $y - y + 1$, spans from 01-09-y to 31-08-y+1, facilitating a comprehensive capture of seasonal variations. The training data for each season prediction consists of all preceding flu season data from the 2009-10 flu season onwards, specifically ranging from the date 01-09-2009 to 31-08-y. The performance of the models are assessed via 3 metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Bivariate correlation (r), between the predicted and target variables.

4.2.1 Methodology

For each flu season prediction, we divided our dataset into training (01-09-2009 to 31-08-y) and test (01-09-y to 31-08-y+1) data before applying the given feature selection method, deriving our $4\tilde{0}0$ most relevant queries. We then standardised both the training and test data, with respect to the training data through z-scoring. This transformed the data to a mean of 0 and standard deviation of 1, ensuring a uniform scale. This is an essential step for models like ours that are sensitive to data scale variance, ensuring consistency and reliability in the learning process.

Following this, we then train an Elastic Net model on the features selected in our standardised training data, optimising the objective function outlined in Equation 4.1:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b) \right)^2 + \lambda \cdot \left(\alpha ||\mathbf{w}||_1 + \frac{(1-\alpha)}{2} ||\mathbf{w}||_2^2 \right) \right], \tag{4.1}$$

where λ is the regularisation hyper-parameter controlling the penalty's strength and α is a mixing parameter that balances the L1 and L2 regularisation terms, λ_1 and λ_2 respectively. Formally, $\lambda_1 = \lambda \cdot \alpha$ and $\lambda_2 = \lambda \cdot \frac{(1-\alpha)}{2}$. After some initial exploration we fixed the regularisation ratio α to 0.7, balancing the dual need to induce both sparsity in feature selection via Lasso, while also leveraging Ridge regression's capability to minimise coefficient size.

Determining the optimal regularisation strength (λ) involved employing a manual k-fold cross-validation process, uniquely structured where each fold corresponds to a distinct flu season within the training dataset. This approach aimed to closely mirror the actual prediction setting of our model, allowing us to assess the generalisability of hyperparameters in distinct flu seasons, in preparation for the predicting the actual test flu season. However, this still slightly deviates from the real prediction task as for certain iterations, we validate the hyperparameters on a single fold and flu season, whilst training the on subsequent flu seasons in the future. For the scope of our study we retain this general cross validation approach, with minor iterations as discussed in the subsequent section, as oppose to an extensive exploration of different methods.

A hyperparameter grid was constructed to explore the regularisation strength, ensuring a comprehensive examination across the full range of feature densities. This grid spanned from very sparse (retaining $\tilde{5}$% of features) to denser (retaining $\tilde{98}$% of features) feature sets, aiming to identify the λ value that achieves the optimal balance between model complexity and predictive accuracy. The grid consisted of 202 values spanning from 0.001 to 10 with varying granularities in between. As a result, the ranges for λ_1 and λ_2 are defined within [0.0007, 7] and [0.00015, 1.5], respectively.

The λ values in the hyperparameter grid were scored via their MAPE as this gave a more relative measure of their error across flu seasons compared to MAE. The λ value that resulted in lowest MAPE during cross validation was then selected as the optimal hyperparamter and was used in conjunction with the α value of 0.7 to train the Elastic Net model on the entire training dataset. The trained model was subsequently used to predict the target ILI rates for the test flu season, evaluating the predictive performance via the MAE, MAPE and the Bivariate correlation. These results were recorded for each flu season prediction, following a calculation of the mean and standard deviation across all flu seasons. We additionally recorded the λ_1 and λ_2 regularisation terms obtained with the corresponding number of non-zero weights retained (NZWR) by the model for each flu season prediction to gauge how the cross validation approach employed behaves.

4.2.2 Results

Cross Validating on All Flu Seasons in the Training Data

This methodology was employed for each of the feature selection approaches discussed above, with the results illustrated in Table 4.1.

The baseline model's performance across different feature selection methods reveals a consistent pattern of underperformance. Across all approaches, the average MAE remained above 2.34, the average Bivariate correlation coefficient did not exceed 0.95, and the average MAPE was not lower than 53.94, indicating a broader issue with the model's predictive accuracy. Notably, the sentence embedding method outperformed the correlation-based approach, challenging our initial assumptions. The correlation-based method is designed to prioritise queries that exhibit a significant correlation with ILI rates. Hence, this was logically anticipated to be more predictive due to its direct reliance on explicit data relationships. In contrast, the sentence embedding approach, focusing solely on the semantic relevance of queries to flu-related terms, without any notion of feature and target variable relationships, yielded better results.

However, the relatively poor performance across all feature selection methods, raises a more critical concern around the sparsity of the selected features. The models consistently utilised a narrow subset of the available queries, retaining between 39 to 62 non-zero weights across various flu seasons and methods. This restrictive feature selection likely resulted in the omission of informative

Flu Season	Feature Selection	MAE	MAPE	r	λ_1	λ_2	NZWR
2014-2015	Correlation	1.73	26.23	0.8874	0.70	0.15	53/228
	Sentence Embedding	**1.47**	**24.17**	**0.9263**	0.27	0.06	62/400
	Hybrid	1.82	29.01	0.8732	0.68	0.15	51/330
2015-2016	Correlation	1.92	27.72	0.9172	0.59	0.13	57/238
	Sentence Embedding	**1.64**	26.24	**0.9406**	0.29	0.06	59/400
	Hybrid	1.67	**25.98**	0.9384	0.31	0.07	58/330
2016-2017	Correlation	2.56	62.47	0.9580	0.68	0.15	44/454
	Sentence Embedding	**2.30**	**57.50**	0.9565	0.32	0.07	60/400
	Hybrid	2.33	61.17	**0.9589**	0.63	0.14	47/371
2017-2018	Correlation	**2.69**	50.88	**0.9737**	0.70	0.15	39/470
	Sentence Embedding	2.81	**49.83**	0.9705	0.55	0.12	43/400
	Hybrid	2.90	54.27	0.9700	0.67	0.14	42/372
2018-2019	Correlation	3.56	115.55	**0.9549**	0.65	0.14	40/546
	Sentence Embedding	**3.48**	**111.97**	0.9492	0.46	0.10	45/400
	Hybrid	3.59	114.40	0.9464	0.48	0.10	47/394
Average	Correlation	2.49 (0.65)	56.57 (32.55)	0.9382 (0.03)	-	-	-
	Sentence Embedding	**2.34 (0.74)**	**53.94 (31.78)**	**0.9486 (0.01)**	-	-	-
	Hybrid	2.46 (0.71)	56.97 (31.83)	0.9374 (0.03)	-	-	-

Table 4.1: Nowcasting performance of the Elastic Net model for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during the each test flu season (from 2014-15 to 2018-19), for each feature selection method: correlation based, sentence embedding and hyrbid. The last three rows report the mean and standard deviation of the performances over the five test flu seasons. MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the nowcasts and reported ILI rates. The best results for each metric and flu season are shown in bold. The regularisation terms $\lambda_1 \in [0.0007, 7]$ and $\lambda_2 \in [0.00015, 1.5]$, corresponding to the regularisation strengths of the Lasso and Ridge regression respectively, where $\lambda_1 = \lambda \cdot \alpha$ and $\lambda_2 = \lambda \cdot \frac{(1-\alpha)}{2}$, and the resulting number of non-zero weights retained (NZWR) by the model are also presented. **The cross validation approach validates the λ hyperparamter on all flu seasons (k folds) in the training data.**

queries, constraining the model's learning capacity and thus its predictive performance. The impact of feature sparsity on model performance becomes more evident when comparing our findings to existing research, which achieved significantly superior performance when employing the Elastic Net, albeit on different test periods [1].

As discussed in chapter 3, both the trend in ILI rates and its relationship with the search queries varies across flu seasons, with significant differences being observed from the earlier to later seasons. Hence, this performance discrepancy could be attributed to our cross-validation strategy, validating hyperparameters across all flu seasons in the training data. This might have skewed the selection towards features effective in earlier seasons, but potentially less relevant for predicting more recent flu trends.

Cross Validating on the Last Three Flu Seasons in the Training Data

To address these shortcomings, we refined our cross-validation approach to focus on the three most recent flu seasons in the training data. This strategy is based on the rationale that recent flu seasons provide a more relevant basis for predicting future trends, thereby facilitating more relevant hyperparameter selection and enhancing the model's capacity to identify and utilise informative features for improved predictions.

With this revised cross-validation strategy, we replicated the model training and evaluation for each feature selection method. The outcomes of this adjusted approach are detailed in Table 4.2.

Flu Season	Feature Selection	MAE	MAPE	r	λ_1	λ_2	NZWR
2014-2015	Correlation	1.73	26.23	0.8874	0.70	0.15	53/228
	Sentence Embedding	1.43	25.61	0.9351	0.08	0.02	140/400
	Hybrid	**1.29**	**23.62**	**0.9447**	0.03	0.01	209/330
2015-2016	Correlation	**1.50**	**25.77**	**0.9481**	0.02	0.00	150/238
	Sentence Embedding	1.62	26.28	0.9423	0.24	0.05	65/400
	Hybrid	1.66	26.10	0.9402	0.26	0.06	59/330
2016-2017	Correlation	2.60	62.37	**0.9586**	0.61	0.13	43/454
	Sentence Embedding	2.44	57.58	0.9359	0.05	0.01	164/400
	Hybrid	**2.01**	**45.81**	0.9359	0.05	0.01	174/371
2017-2018	Correlation	2.69	50.88	0.9737	0.70	0.15	39/470
	Sentence Embedding	**2.21**	**31.16**	0.9685	0.00	0.00	394/400
	Hybrid	2.29	39.16	**0.9765**	0.00	0.00	369/372
2018-2019	Correlation	**1.50**	49.53	**0.9639**	0.04	0.01	202/546
	Sentence Embedding	2.10	**47.56**	0.9621	0.00	0.00	380/400
	Hybrid	1.96	47.82	0.9594	0.00	0.00	366/394
Average	Correlation	2.00 (0.53)	42.95 (14.55)	0.9463 (0.03)	-	-	-
	Sentence Embedding	1.96 (0.38)	37.64 (12.74)	0.9488 (0.01)	-	-	-
	Hybrid	**1.84 (0.34)**	**36.50 (9.96)**	**0.9514 (0.01)**	-	-	-

Table 4.2: Nowcasting performance of the Elastic Net model for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during the each test flu season (from 2014-15 to 2018-19), for each feature selection method: correlation based, sentence embedding and hyrbid. The last three rows report the mean and standard deviation of the performances over the five test flu seasons. MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the nowcasts and reported ILI rates. The best results for each metric and flu season are shown in bold. The regularisation terms $\lambda_1 \in [0.0007, 7]$ and $\lambda_2 \in [0.00015, 1.5]$, corresponding to the regularisation strengths of the Lasso and Ridge regression respectively, where $\lambda_1 = \lambda \cdot \alpha$ and $\lambda_2 = \lambda \cdot \frac{(1-\alpha)}{2}$, and the resulting number of non-zero weights retained (NZWR) by the model are also presented. **The cross-validation approach validates the λ hyperparameter on the last three flu seasons (three folds) in the training data.**

The refined strategy for cross-validation significantly improved the model's performance across all feature selection techniques. The hybrid method, in particular, showed substantial improvements, surpassing both the correlation and sentence embedding methods with an average Bivariate correlation of 0.95, MAE of 1.84, and MAPE of 36.50. These results not only mark a considerable advancement from our preliminary outcomes but also exceed the best MAE and Bivariate correlation performance reported in comparable models [1]. These results highlight the effectiveness of integrating direct data correlations with textual semantics for a more informative feature set in ILI rate prediction.

Despite these performance improvements, the correlation-based method lagged behind the sentence embedding approach in three out of the five flu seasons. This issue can again be attributed to the model's feature sparsity within these seasons, with the number of non-zero weights retained remaining notably low, < 100. In contrast, the sentence embedding and hybrid methods demonstrated a more consistent and denser selection of features, providing a robust basis for prediction and justifying their superior performance. However, in the remaining flu seasons where the correlation method exhibits comparable or greater feature densities, its effectiveness is highlighted through superior performance. This variability indicates the correlation model's effectiveness when sufficiently represented but also underscores the fluctuating feature densities across models and flu seasons. For example, in the 2015-16 season, feature density for the hybrid and sentence embedding models unexpectedly dropped before increasing again. This inconsistency still raises questions about the appropriateness of the cross-validation approach and thus the comparative predictive accuracy of the correlation-based method.

Flu Season	Feature Selection	MAE	MAPE	r	λ_1	λ_2	NZWR
2014-2015	Correlation	**1.27**	**21.46**	0.9443	0.01	0.00	184/228
	Sentence Embedding	1.41	25.59	0.9364	0.07	0.02	151/400
	Hybrid	1.29	23.62	**0.9447**	0.03	0.01	209/330
2015-2016	Correlation	1.50	25.77	0.9481	0.02	0.00	150/238
	Sentence Embedding	**1.43**	**23.37**	**0.9486**	0.07	0.02	151/400
	Hybrid	1.45	23.66	0.9460	0.04	0.01	190/330
2016-2017	Correlation	2.47	56.84	**0.9541**	0.04	0.01	166/454
	Sentence Embedding	2.44	57.58	0.9359	0.05	0.01	164/400
	Hybrid	**2.01**	**45.81**	0.9359	0.05	0.01	174/371
2017-2018	Correlation	**2.19**	46.47	**0.9773**	0.00	0.00	450/470
	Sentence Embedding	2.21	**31.16**	0.9685	0.00	0.00	394/400
	Hybrid	2.29	39.16	0.9765	0.00	0.00	369/372
2018-2019	Correlation	**1.50**	49.53	**0.9639**	0.04	0.01	202/546
	Sentence Embedding	2.10	**47.56**	0.9621	0.00	0.00	380/400
	Hybrid	1.96	47.82	0.9594	0.00	0.00	366/394
Average	Correlation	**1.79 (0.46)**	40.01 (13.88)	**0.9575 (0.01)**	-	-	-
	Sentence Embedding	1.92 (0.42)	37.05 (13.30)	0.9503 (0.01)	-	-	-
	Hybrid	1.80 (0.37)	**36.01 (10.50)**	0.9525 (0.01)	-	-	-

Table 4.3: Nowcasting performance of the Elastic Net model for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during the each test flu season (from 2014-15 to 2018-19), for each feature selection method: correlation based, sentence embedding and hyrbid. The last three rows report the mean and standard deviation of the performances over the five test flu seasons. MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the nowcasts and reported ILI rates. The best results for each metric and flu season are shown in bold. The regularisation terms $\lambda_1 \in [0.0007, 7]$ and $\lambda_2 \in [0.00015, 1.5]$, corresponding to the regularisation strengths of the Lasso and Ridge regression respectively, where $\lambda_1 = \lambda \cdot \alpha$ and $\lambda_2 = \lambda \cdot \frac{(1-\alpha)}{2}$, and the resulting number of non-zero weights retained (NZWR) by the model are also presented. **The cross-validation approach validates the λ hyperparameter on the last three flu seasons (three folds) in the training data, ensuring the best parameter chosen retains ≥ 150 features when trained on the entire training data.**

Cross Validating on the Last Three Flu Seasons in the Training Data with a Minimum Feature Density

In response to these observations, we implemented a minimum density threshold for the models. Following the same cross validation approach, we iterate through each hyperparameter λ value sorted by their MAPE performance, each time fitting the specific model setting on the entire training dataset. If the number of non zero weights retained is ≥ 150 we then take this hyperparameter as the best λ value, promoting a denser feature set for the model training. This experimental decision intended to address sparse feature selection issues without extensively delving into more complex cross-validation techniques, focusing instead on the best performance obtainable by the Elastic Net baseline model ahead of exploring improvements through more complex, neural networks.

The previous experiments were again repeated for each feature selection method, with this additional density criterion in place. This aimed to enforce a greater consistency in the feature density of models across flu seasons, refining predictive accuracy, specifically for the correlation based approach, which we expect to outperform the sentence embedding method. The outcomes of these modified experiments are detailed in table 4.3.

Upon introducing a minimum feature density threshold, our denser models demonstrated a notable improvement in performance across all feature selection methods. This adjustment was particu-

larly effective for the correlation-based approach, which saw a marked improvement, achieving an average MAE of 1.79 and an average Bivariate correlation of 0.96, outperforming both the sentence embedding and hybrid methods. This improvement underscores the effectiveness of the correlation approach in identifying informative features, when sufficiently represented, providing a strong basis for predictive modelling in ILI rate nowcasting. This finding aligns with our intuitive expectation that direct data relationships should offer more predictive value than solely relying on query similarities grounded on textual semantics.

Although the hybrid approach does not consistently outperform the correlation method, similar to previous research efforts [1], it achieves relatively similar performance, with a superior MAE in two (2015-16 and 2016-17) out of the five flu seasons and the best average MAPE of 36.01. These results suggest the hybrid model's predictions are more relatively accurate across flu seasons, with a smaller percentage deviation from the actual ILI rates. For most flu seasons where the correlation-based approach performs best, the hybrid method only slightly underperforms. However, when the hybrid method outperforms the correlation model, as seen in the 2016-17 flu season, its performance is significantly superior. This particular season showcases the adaptability of the hybrid model by placing significant predictive weight on broad-context queries like 'flu in adults' and 'winter flu', which capture a more holistic view of flu trends. These queries were not as influential in the predictions of the correlation model, reflecting the hybrid's ability to integrate broader public health concerns into its predictions. The correlation model, while focusing heavily on direct symptom queries like 'flu symptoms' and 'flu treatment', also attributed significant weights to queries such as 'neck warmer' and 'chilblains'. Though correlated with the flu season due to seasonal changes, these queries do not directly relate to influenza. This inclusion highlights a potential pitfall of relying solely on correlation data, where non-specific but seasonally correlated items might skew the model's effectiveness in purely medical contexts. This observation aligns with our earlier discussion on the potential drawbacks of the feature sets generated by the correlation-based method.

Interestingly, the sentence embedding approach led in performance during the 2015-16 flu season. Its success was likely due to its incorporation of a broad range of semantically relevant queries, placing greater weights on 'flu symptoms in adults' and 'flu in adults'. These queries are not only semantically relevant but also strongly correlated with the ILI rate, thus closely mirroring the performance of the hybrid method, which intentionally balances this dual focus. The slight edge in performance for the sentence embedding model may have stemmed from its emphasis on semantic relevance, avoiding the undue influence of seasonally correlated but non flu related queries, such as 'neck warmer' that hampered the hyrbid and in particular correlation methods.

All models generally struggled to predict the ILI rates for the 2016-17 and 2017-18 flu seasons, with the latter presenting a particular challenge due to an unexpected peak in ILI rates during the winter (Figure 3.1). These challenges underscore potential limitations of the Elastic Net model, as performances were similar and poor despite the application of various feature selection methods. During the 2017-18 season, the correlation method demonstrated relatively better performance, which could be attributed to its consistent focus on direct clinical indicators of ILI such as 'fever and cough', 'chest infection', and 'winter vomiting bug.' These symptom-based searches tend to remain reliable predictors, unaffected by contextual changes. In contrast, the hybrid model often gave greater weight to queries more responsive to public concerns, such as 'swine flu NHS', 'swine flu jab', and 'swine flu news'. These terms were not indicative of ILI cases that season, suggesting a misalignment of the hybrid model's predictive focus. The continued reliance of the correlation model on proven clinical symptoms allowed it to maintain its superiority in the subsequent final season.

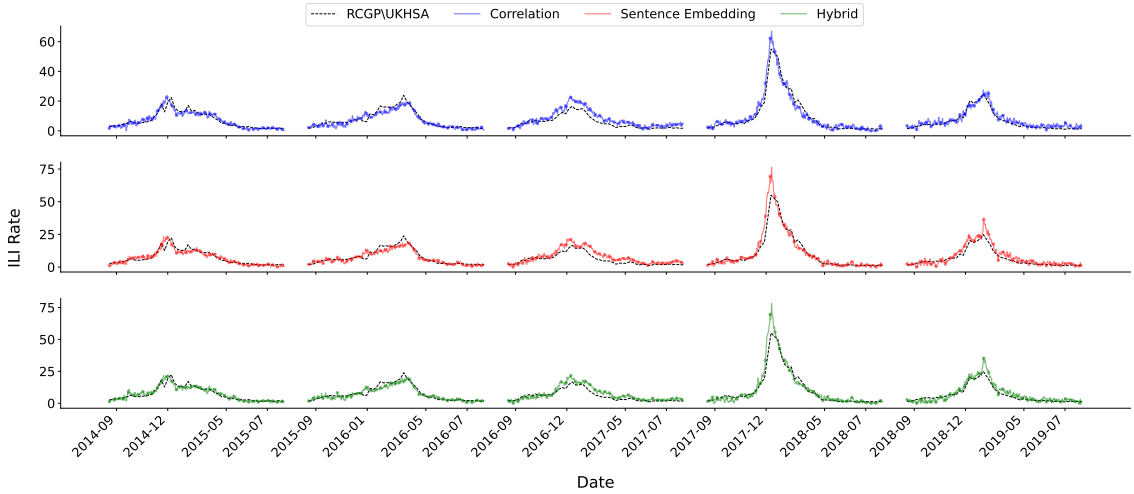

Figure 4.1: Comparative plot of the nowcasting predictions made by the Elastic Net model for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during each test flu season (from 2014-15 to 2018-19), for the different feature selection methods: correlation, sentence embedding and hybrid.

Notably, both the correlation and hybrid approaches exceeded the best Elastic Net performance metrics reported in [1], although they predicted three flu seasons compared to our five. The high performance achieved by the correlation-based approach seems difficult to beat, suggesting that for this specific task, ensuring model density may negate the need for combining correlation and textual semantics in feature selection, at least within the constraints of the Elastic Net model. This outcome highlights the potential limitations of the Elastic Net model regarding the hyperparameter validation process but also further points to the robustness of the hybrid-based method when model density is not ensured. Additionally, despite the strong performances of the correlation-based model in specific flu seasons when feature density is ensured, the hybrid method's consistent reliability, demonstrated by its lower standard deviation in MAE and smaller relative error, points to its deeper utility, especially in complex or atypical flu seasons, in adapting to variable ILI trends. This consistency and gap in performance over the correlation model has shown to increase when employing non-linear regressive models, specifically a Gaussian Process (GP) [1]. Hence, for our subsequent experiments using non-linear neural networks, we look to take this hybrid feature selection forward in an effort to benefit from its dual consideration of correlation and textual semantic relationships for improved performance.

The predictive performances of each feature selection approach can be further analysed by plotting the predicted ILI rates against the actual observed rates, as shown in Figure 4.1 This visual representation can be complemented with a deeper periodic error analysis, highlighting specific periods within each season where predictions closely align with actual trends and where discrepancies occur.

We observe that although the overarching trends are generally captured by all models, there are distinct deviations during specific periods. For each flu season, the models reasonably capture both the onset and the tail ends of the ILI rate peaks, though there are occasional slight over and under predictions. For example, during the tail phase of the 2016-17 season, from January 31 to August 31, all models consistently overestimate, with the sentence embedding model registering an MAE of 2.62, correlation at 2.46, and the hybrid model at 2.05. This indicates that the hybrid

model's predictions align more closely with the actual rates during this period.

This trend of over-prediction continues in the tail end of the 2018-19 season, from February 19 to August 31, where this time the correlation model provides the most accurate predictions with an MAE of 1.59, lower than the hybrid's 1.65 and the sentence embedding's 1.83. During the onset of this season, from September 1 to December 28, the correlation and sentence embedding methods effectively capture the trend, whereas the hybrid method underestimates, reflected by its MAE of 1.80 versus 1.27 for the correlation and 1.63 for sentence embedding.

A similar pattern of variable performance is observed in the 2017-18 season. Despite less fluctuation in the predictions across all models, suggesting improved performance, the reality is more complex. Near the peak periods, both before and after, the models yield higher error rates than in other seasons. Specifically, the sentence embedding model more accurately captures the tail period of the 2017-18 season with an MAE of 1.23, better than the correlation's 1.93 and the hybrid's 1.36. For the onset period, the correlation model outperforms the others with an MAE of 1.60, followed by the sentence embedding at 1.91 and the hybrid at 2.04. These varied performances demonstrate the contrasting capabilities of the models in accurately capturing these phases across different flu seasons.

Overall, these periods are mostly well-predicted, as shown by their general capture of the ILI rate trend and relatively low MAEs. However, the most challenging and critically evaluated predictions occur during the peak periods of flu activity. Figure 4.1 shows that the discrepancies between the models' predictions and the actual data are more significant during these times. In the 2014-15 flu season, from December 2, 2014, to January 20, 2015, each model's predictions closely mirror the actual peak, albeit slightly premature, with correlation recording an MAE of 2.53, sentence embedding at 2.79, and hybrid at 2.55. However, this alignment diminishes in the following 2015-16 season, as all models underestimate the peak intensity, resulting in higher MAEs. Specifically, the sentence embedding records an MAE of 2.82, correlation at 2.58, and hybrid at 2.61. Additionally, minor fluctuations in ILI rates, particularly around February in both 2015 and 2016, are again not accurately captured by any model. This recurring pattern of the same type of misprediction across all feature selection methods suggests potential limitations in the Elastic Net model and its ability to adapt to rapid increases and short-term variances in ILI rates.

These similarities among models are again present in the 2016-17 flu season, with a common trend of overestimation across all models, aligning with the poorer performance metrics listed in Table 4.3. However, during the peak period, from December 2, 2016, to January 30, 2017, the hybrid model gains an apparent advantage, being the closest to the ground truth, recording an MAE of 3.07, followed by the sentence embedding at 3.50 and correlation at 4.55. In subsequent seasons, 2017-18 and 2018-19, there is a noticeable shift; predictions from both the hybrid and sentence embedding models significantly overshoot the peak ILI rates, whereas the correlation-based method provides more accurate nowcasts closer to the ground truth. Notably, during the peak period of 2018-19, from December 29, 2018, to February 18, 2019, the correlation method achieves an MAE of 1.70, significantly more precise than the sentence embedding's 4.20 and hybrid's 3.43. These observations generally align with the aggregated performance of these models over the flu season as reported in Table 4.3.

Across flu seasons, the graphs reveal overall similarities in the trends predicted by each model, often showing instances where all models either overestimate, underestimate, or accurately predict ILI rates during specific periods. Although there are instances where they diverge, with some models slightly outperforming others, no single model consistently excels within a given flu season or across different flu seasons. This observation is supported by the results presented in Table

4.3, which show comparable performance across the methods. All models perform relatively well in particular seasons and struggle in others, while the average performances of the Hybrid and Correlation methods are very close. This consistency underscores the potential efficacy of these models in achieving a robust baseline predictive performance that generally captures seasonal trends. However, it also raises questions about the scope for further improvements. Specifically, it suggests a limitation in the Elastic Net in better capturing the trends in ILI rate that have proven to be difficult for all feature selection methods.

Additionally, a notable property of the Elastic Net predictions is their erratic nature, presenting sharp ebbs and flows that starkly contrast the smoother ground truth ILI rates. This irregularity underscores a shortfall in the Elastic Net's capacity to incorporate time series dynamics, which leads to these volatile, point-to-point predictions.

These findings highlight the need to explore more complex models to better capture trends throughout the course of flu seasons and enhance predictive accuracy. As we conclude the discussion on linear models, we turn our attention to deep learning models in an effort to achieved this improved accuracy for nowcasting ILI rates and mitigate the erratic tendencies displayed by the Elastic Net model.

Chapter 5

Non-linear Regression Models For Predicting Influenza-Like Illness Rates

This chapter details the experiments conducted and the corresponding analysis of results regarding the Neural Network models. We continue with the task of nowcasting, by first exploring a minimal Feed Forward Neural Network (MFFNN) model that outperforms the predictive performance of the best Elastic Net model. We then explore deeper FFNN architectures, including the integration of historical lagged features to achieve competitive predictive performance, while still maintaining model simplicity.

Following this we shift our attention to the task of Forecasting, where we evaluate the predictive performance of our best nowcasting architecture when forecasting the ILI rate at different horizons γ. We compare and assess this performance against both a baseline persistence model and the MFFNN.

5.1 Nowcasting

5.1.1 Minimal Feed Forward Neural Network

After training various Elastic Net models, our next objective is to train a simple Feed Forward Neural Network (FFNN) with the fewest parameters possible. This aims to demonstrate the efficacy of the neural network for the prediction task by showcasing a vanilla FFNN with minimal parameters that can outperform the best Elastic Net baseline, introducing a new baseline performance for the prediction task. This step serves as a foundation for our subsequent experiments, where we plan to gradually increase the complexity of our Minimal FFNN (MFFNN) to further improve predictive accuracy. Through this, we aim to emphasise the effectiveness of these deep learning models over traditional approaches for the prediction tasks, whilst providing a competitive performance benchmark upon which further complex deep learning model architectures can be validated against.

From our Elastic Net experiments, we found that the hybrid feature selection method was the most effective and robust. It consistently performed well, regardless of whether feature density was enforced. Therefore, we have decided to utilise the hybrid feature selection method for our neural network models. This choice aims to maximise performance while ensuring a consistent comparison to the best Elastic Net model trained on the same set of features, particularly the one with ensured feature density by maintaining ≥ 150 non-zero weights during validation.

Methodology

The MFFNN was trained and evaluated on the five flu seasons, following a methodology similar to that of the Elastic Net model. Each flu season prediction involved dividing our dataset into training (01-09-2009 to 31-08-y) and test (01-09-y to 31-08-y+1) data before applying the hybrid selection method to derive our $4\tilde{0}0$ most relevant queries. Subsequently, we standardised both the training and test data with respect to the training data, this time utilising min-max scaling as it has been proven to be a more effective normalisation technique for training Neural Network models in these prediction tasks [3], [30].

To construct the architecture of our MFFNN, we incrementally increased the number of parameters in the model until we found an architecture that could outperform the average performance across flu seasons of the best Elastic Net model. We began with a single hidden dense layer and progressively added more layers and units. Our input layer had $4\tilde{0}0$ units, matching the number of features, while the output layer consisted of one unit representing the predicted ILI rate for that day.

Each input was flattened and then passed into the hidden dense layer with a predetermined number of units, applying a ReLu activation function (max(0, x)) to its outputs. This processed data was then fed into the output layer to generate predictions. After initial experimentation, the MFFNN model was trained using an initial learning rate of 0.001 and the Adam optimiser, with a mini-batch size of 14. This low initial learning rate was chosen to prevent overfitting of the training data for flu season predictions allowing us to train till convergence, for a fixed number of epochs, specifically 200.

Due to the random initialisation of weights when training a neural network, we sought to mitigate the potential impact of lucky performances and tuning our models to a specific random seed and weight initialisations. Therefore, we opted to train our models on 10 fixed seeds, averaging their predictions and performance metrics for each flu season. This approach aimed to increase the robustness of our proposed model, ensuring it achieved consistent performance across 10 random initialisations of its weights.

The architecture with the minimum number of parameters found to surpass the performance of the baseline consisted of a single hidden dense layer with just 4 units, resulting in $(4\tilde{0}0 \times 4) + (4 \times 1) = 1\tilde{6}04$ parameters. Figure 5.1 displays a computation graph, which further illustrates the architecture of this model. Utilising the notation defined in the graph's caption, the formal list of equations demonstrating how the data processes through the network layers is as follows:

$$\mathbf{z} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]} \tag{5.1}$$

$$\mathbf{z}^{[a]} = \sigma(\mathbf{z}) \tag{5.2}$$

$$\hat{y} = \mathbf{w}^{[2]} \cdot \mathbf{z}^{[a]} + b^{[2]} \tag{5.3}$$

Results

The results of the MFFNN model's predictions averaged across the 10 seeds can be seen against those of the best Elastic Net in Table 5.1.

This showcases the performance improvements in nowcasting predictions made by the MFFNN compared to the best Elastic Net model trained with the same set of features. Specifically, the MFFNN achieves an improvement of 10.56% in MAE, 13.86% in MAPE, and 2.1% in Bivariate

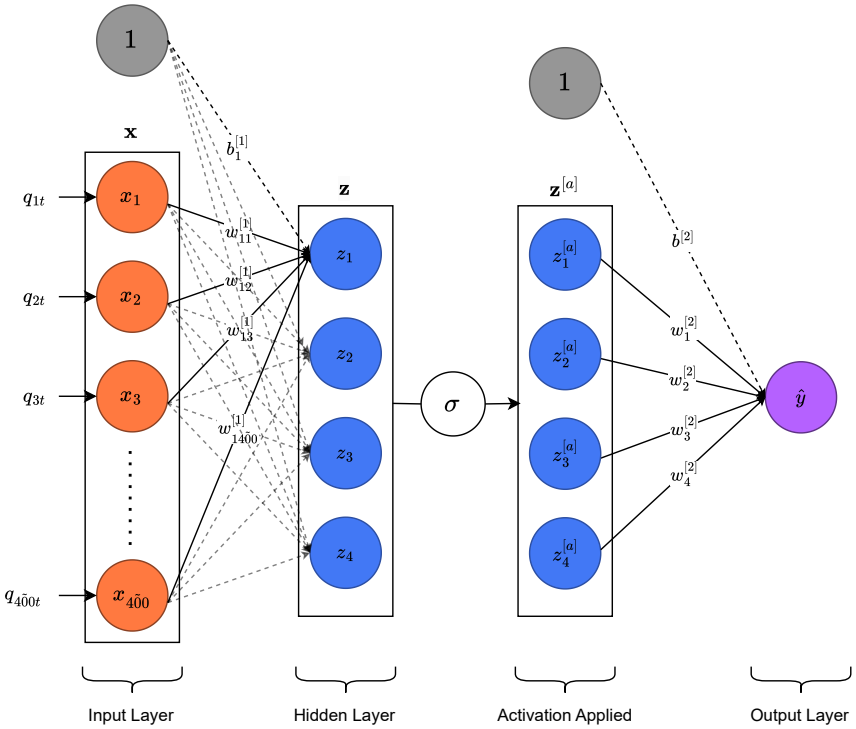

Figure 5.1: Computation graph of our MFFNN model. The input layer, $\mathbf{x} \in \mathbb{R}^{4\tilde{0}0}$, represents the normalised frequencies of each web search query q_i on day t, denoted as q_{it}. The hidden layer, $\mathbf{z} \in \mathbb{R}^4$, calculates the weighted sum of the inputs plus a bias vector $\mathbf{b}^{[1]} \in \mathbb{R}^4$, with components $b_i^{[1]}$ corresponding to each hidden unit. The weight matrix for this hidden layer, $\mathbf{W}^{[1]} \in \mathbb{R}^{4\times 4\tilde{0}0} = [\mathbf{w}_1^{[1]}, \mathbf{w}_2^{[1]}, \mathbf{w}_3^{[1]}, \mathbf{w}_4^{[1]}]^{\mathrm{T}}$, where each $\mathbf{w}_i^{[1]} = [w_{i1}^{[1]}, w_{i2}^{[1]}, \ldots, w_{i4\tilde{0}0}^{[1]}]$ links the input features to the corresponding i^{th} hidden unit. Upon application of the ReLu activation function σ, the activated hidden layer $\mathbf{z}^{[a]} \in \mathbb{R}^4$ is obtained. The output layer prediction $\hat{y} \in \mathbb{R}$ is obtained via the weight vector $\mathbf{w}^{[2]} \in \mathbb{R}^4 = [w_1^{[2]}, w_2^{[2]}, w_3^{[2]}, w_4^{[2]}]$ and bias $b^{[2]} \in \mathbb{R}$.

Correlation (r), across the five flu seasons. The MFFNN demonstrates superior MAE in four out of the five flu seasons, with a slight underperformance only in the 2016-17 season. It consistently delivers superior MAPE and Bivariate correlation values for all seasons, emphasising the consistent performance gains of the neural network model.

Through observing the standard deviation of these performance metrics, we can gauge the variability in the MFFNN model's performance across different random seeds. The model exhibits a good stability for the first two flu seasons, achieving standard deviations of 0.03 and 0.04, respectively. However, this consistency slightly diminishes in subsequent seasons, with standard deviations exceeding 0.1, indicating a marginal decrease in predictability. This is most pronounced in the 2018-19 season, with a significantly high standard deviation of 0.97. This season appears to have been particularly challenging for the MFFNN, suggesting a volatility in the model's performance across seeds. Despite these minor inconsistencies and some questions regarding the robustness of the MFFNN model, we observe an overall tangible improvement in predictive performance over the Elastic Net baseline across flu seasons.

We can further explore the performances of these models by examining their nowcasting predictions for each flu season, as shown in Figure 5.2. The MFFNN features a slightly smoother prediction

Flu Season	MAE		MAPE		r	
	Elastic Net	MFFNN	Elastic Net	MFFNN	Elastic Net	MFFNN
2014-2015	1.29	**1.24** (0.03)	23.62	**22.62** (2.35)	0.9477	**0.9506**
2015-2016	1.45	**1.25** (0.04)	23.66	**19.38** (2.11)	0.9460	**0.9622**
2016-2017	**2.01**	2.08 (0.35)	45.81	**45.12** (11.23)	0.9359	**0.9587**
2017-2018	2.29	**1.72** (0.20)	39.16	**23.73** (6.46)	0.9765	**0.9849**
2018-2019	1.96	**1.78** (0.97)	47.82	**44.26** (25.36)	0.9594	**0.9667**
Average	1.80	**1.61** (0.19)	36.01	**31.02** (5.14)	0.9525	**0.9646**

Table 5.1: Nowcasting performance of the best Elastic Net baseline - the hyrbid based feature selection with a minimum threshold of ≥ 150 non zero weights, against the Minimal Feed Forward Neural Network (MFFNN) for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during the each test flu season (from 2014-15 to 2018-19). The last row shows the performances averaged over the five test flu seasons. MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the nowcasts and reported ILI rates. For the MFFNN, for each flu season, and the average results across flu seasons, we also report the standard deviation of the MAE and MAPE across the 10 fixed seeds. The variation in the bivariate correlation reported across seeds is < 0.01, and thus is omitted from the table. The best results for each metric and flu season are shown in bold.

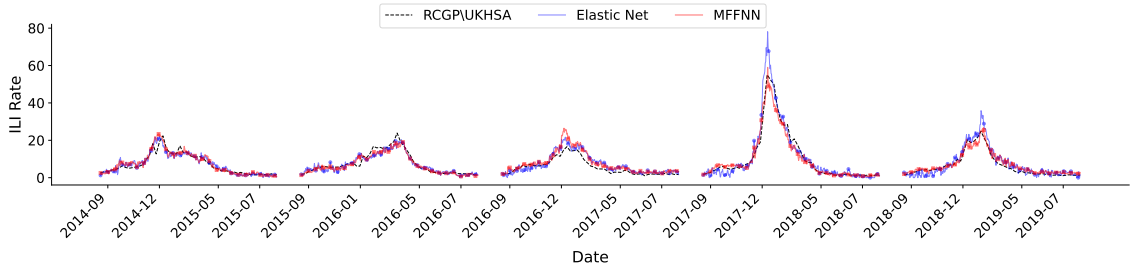

Figure 5.2: Comparative plot of the nowcasting predictions made by the best Elastic Net baseline - the hybrid based feature selection with a minimum threshold of ≥ 150 non zero weights and the Mininmal Feed Forward Neural Network (MFFNN), for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during each test flu season (from 2014-15 to 2018-19).

curve, suggesting it might better capture the underlying flu trends with reduced daily fluctuations. This characteristic is particularly evident in the early flu seasons. Although both models exhibit similar overall trends, the MFFNN's smoother predictions are more closely aligned with the actual ILI rates throughout the onset, peak, and tail phases, in contrast to the Elastic Net, which exhibits greater fluctuations. For instance, in the 2015-16 flu season, the MFFNN exhibits lower MAEs of 1.05, 2.50, and 0.69 in the onset (01 September to 30 December), peak (31 December to 24 March), and tail (25 March to 31 August) periods respectively, compared to the Elastic Net which recorded MAEs of 1.45, 2.61, and 0.82 in the same periods.

The differences between the models become more pronounced in later flu seasons. During the 2016-17 season, the MFFNN accurately reflects the onset and tail of the ILI peak, adhering closely to the actual data with an onset MAE of 0.83, while the Elastic Net is somewhat slower to respond, recording an MAE of 1.24. However, despite both models overestimating the peak from 02 December 2016 to 30 January 2017, the Elastic Net comes distinctly closer to the ground truth with an MAE of 3.07 compared to the MFFNN's 4.31, thus resulting in its overall superior performance for this season (Table 5.1).

Despite this periodic lapse, the remaining seasons again showcase the consistent improvements of the MFFNN over the Elastic Net. During the 2017-18 flu season, the Elastic Net model tended

to overestimate the peak rate of ILI, particularly from December 31, 2017, to February 8, 2018, with an MAE of 7.71. This issue is mitigated in the case of the MFFNN, with its ILI rate peak predictions much closer to the actual rates, albeit with a slight overestimation, recording a significantly lower MAE of 4.74 during the same period. This demonstrates the efficacy of the MFFNN in better capturing areas of the ILI rate trend that proved difficult for the Elastic Net model across all feature selection approaches (Figure 4.1). This superiority can again be seen in the last flu season, from September 1, 2018, to December 28, 2018, where the MFFNN more accurately nowcasts the ILI rates far better than the underpredictions of the Elastic Net, recording an MAE of 0.72 compared to the Elastic Net's 1.80. Again in the tail phase, while both models demonstrate similar trends in their slight overpredictions, the smoother curve of the MFFNN achieves a lower MAE of 1.57, compared to the 1.65 of the Elastic Net.

In summary, the combination of visual predictions and performance metrics underscores the MFFNN's ability to consistently better capture the ILI rate trend, providing smoother and more precise predictions across flu seasons and throughout their seasonal periods. This demonstrates the great potential of neural network models for the prediction task, with such a minimal architecture significantly outperforming the best Elastic Net model, effectively establishing a new baseline for the nowcasting task. This invites exploration for further performance improvements through the adoption of deeper FFNN architectures, guiding us towards our upcoming experiments, with the aim of achieving a competitive nowcasting performance, while still preserving the simplicity of using an FFNN.

5.1.2 Deeper Feed Forward Neural Network

Methodology

Following the training of our Minimal Feed Forward Neural Network (MFFNN), we aimed to improve our nowcasting performance by exploring deeper FFNN architectures. After some initial experimentation we propose a slightly deeper model to that of the MFFNN, with the same input and output layers, however with 2 hidden dense layers, each with 25 units. This resulted in a configuration with approximately $(4\tilde{0}0 \times 25) + (25 \times 25) + (25 \times 1) = 10\tilde{6}50$ parameters.

We also introduced historical lagged features into our model's input. The idea was to capture the time series data's inherent patterns by using past observations as part of the input. By adding these lagged features, we aimed to provide the network with a richer context for making future predictions. This approach also helps smooth out the predictions by leveraging information on how the trends have evolved over time, rather than relying solely on daily data.

Incorporating x lagged features means adding the normalised frequencies of each query from previous days $t-1, t-2, \ldots, t-x-1$, in addition to the frequencies on day t, resulting in x frequencies for each query. Consequently, the model's input size expanded from $4\tilde{0}0$ features to $4\tilde{0}0 \times x$. We experimented with increasing the number of lagged features in multiples of 7, aiming to capture weekly to monthly search trends. While we anticipated improvements from expanding the input, we recognised there would be limits to the benefits of adding more features without adjusting the FFNN's architecture, especially considering the challenge of managing the increased complexity and the FFNN's inherent limitations in capturing temporal dependencies. Nevertheless, our main aim was to gain performance improvements by integrating historical lagged features, establishing a competitive performance for the prediction task.

Figure 5.3 displays the computation graph of these more complex model architectures, incorporating x lagged features. Utilising the notation defined in the graph's caption, the formal list of

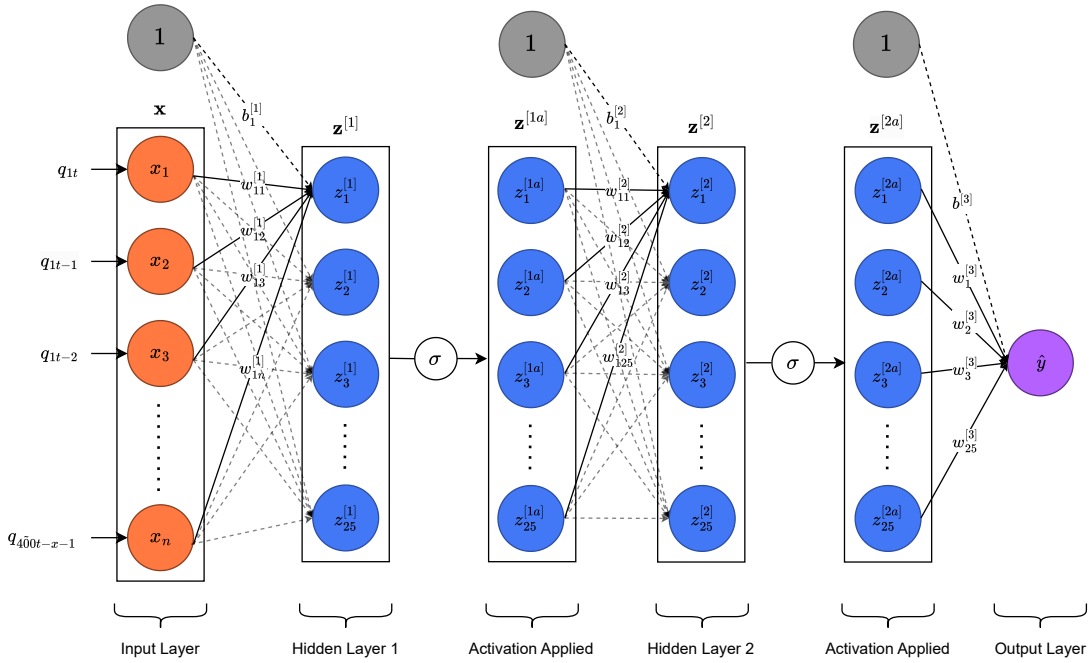

Figure 5.3: Computation graph of the deeper FFNN model incorporating x lagged features. The input layer $\mathbf{x} \in \mathbb{R}^n$, where $n = 4\tilde{0}0 \times x$, represents the normalised frequencies of each web search query q_i across the current and past $x-1$ days, denoted as $q_{it}, q_{it-1}, \ldots, q_{it-x-1}$. The first hidden layer $\mathbf{z}^{[1]} \in \mathbb{R}^{25}$ computes the weighted sum of these inputs plus a bias vector $\mathbf{b}^{[1]} \in \mathbb{R}^{25}$, with components $b_i^{[1]}$ for each hidden unit. The weight matrix for this hidden layer $\mathbf{W}^{[1]} \in \mathbb{R}^{25\times n} = [\mathbf{w}_1^{[1]}, \mathbf{w}_2^{[1]}, \ldots, \mathbf{w}_{25}^{[1]}]^{\mathrm{T}}$, where each $\mathbf{w}_i^{[1]} = [w_{i1}^{[1]}, w_{i2}^{[1]}, \ldots, w_{in}^{[1]}]$, links the input features to the corresponding i^{th} hidden unit. The activated hidden layer $\mathbf{z}^{[1a]} \in \mathbb{R}^{25}$, obtained after applying the ReLu activation function σ, is fed to the second hidden layer $\mathbf{z}^{[2]} \in \mathbb{R}^{25}$, which further includes a bias vector $\mathbf{b}^{[2]} \in \mathbb{R}^{25}$. The weight matrix $\mathbf{W}^{[2]} \in \mathbb{R}^{25\times 25}$ and bias vector $\mathbf{b}^{[2]}$ for the second hidden layer are detailed in the same manner. After applying the ReLu activation σ on the second hidden layer $\mathbf{z}^{[2]}$, the output layer prediction $\hat{y} \in \mathbb{R}$ is obtained via the weight vector $\mathbf{w}^{[3]} \in \mathbb{R}^{25} = [w_1^{[3]}, w_2^{[3]}, w_3^{[3]}, w_4^{[3]}]$ and the bias $b^{[3]} \in \mathbb{R}$.

equations demonstrating how the data processes through the network layers is as follows:

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]} \tag{5.4}$$

$$\mathbf{z}^{[1a]} = \sigma(\mathbf{z}^{[1]}) \tag{5.5}$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{z}^{[1a]} + \mathbf{b}^{[2]} \tag{5.6}$$

$$\mathbf{z}^{[2a]} = \sigma(\mathbf{z}^{[2]}) \tag{5.7}$$

$$\hat{y} = \mathbf{w}^{[3]} \cdot \mathbf{z}^{[2a]} + b^{[3]} \tag{5.8}$$

For our Deeper FFNN (DFFNN), we adapted our training methodology to account for the model's increased complexity. We noticed that with more parameters, the optimisation steps during training showed more fluctuation, suggesting that our initial learning rate might be too high. Therefore, we continued using the Adam optimiser with a mini batch size of 14 but lowered the learning rate to 0.0001. Without signs of overfitting, we trained the model for 200 epochs until convergence, consistent with our approach for the MFFNN. Again, we trained our models across 10 fixed seeds, averaging their predictions and performance metrics for each flu season, to ensure

Metric	MFFNN	DFFNN				
Lagged Features	0	0	7	14	21	28
MAE	1.61 (0.19)	1.56 (0.04)	1.43 (0.11)	**1.25** (**0.07**)	1.28 (0.08)	1.37 (0.12)
MAPE	31.02 (5.14)	29.09 (1.15)	25.63 (2.70)	**22.34** (**1.87**)	23.24 (1.48)	23.69 (2.67)
r	0.9646	0.9669	0.9678	0.9713	**0.9723**	0.9701

Table 5.2: Nowcasting performance of the Minimal Feed Forward Neural Network (MFFNN) against the Deeper Feed Forward Neural Network (DFFNN) with varying numbers of historical lagged features passed into the model's input (0, 7, 14, 21, 28), for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) averaged across the 5 test flu seasons (from 2014-15 to 2018-19). MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the nowcasts and reported ILI rates. We also report the standard deviation of the MAE and MAPE across the 10 fixed seeds. The variation in the bivariate correlation reported across seeds is < 0.01 for both models, and thus is omitted from the table. The best results for each metric are shown in bold.

robustness and consistency in our findings.

Results

Table 5.2 presents the predictive performance of the DFFNN utilising various numbers of historical lagged features, compared to our baseline MFFNN. The performance metrics are averaged across the five flu seasons.

Initially, it's apparent that the DFFNN, even without incorporating any lagged features, outperforms our MFFNN across all performance metrics. This advantage is incrementally enhanced as we introduce more historical lagged features, peaking at 14 lagged features. At this point, we achieve our optimal model performance within this architecture, obtaining an average MAE of 1.25, MAPE of 22.34, and a Bivariate correlation of 0.97, marking improvements of 22.36%, 27.98%, and 0.62% over the MFFNN, respectively. Additionally, we note an enhancement in the stability and robustness of these performances compared to the MFFNN. This suggests that our deeper model with 14 lagged features is generally more consistent in its predictions across different seeds and weight initialisations. However, adding more than 14 lagged features leads to a decline in predictive performance, indicating that the model struggles to learn effectively with the added complexity and parameter count in the current training setup.

Nonetheless, the key takeaway is that we have demonstrated a relatively straightforward FFNN model, with only two dense layers, that achieves commendable nowcasting performance, surpassing the performance metrics reported in other studies [1], including the preliminary results in the Flu Detector paper [2].

Table 5.3 further illustrates the predictive performance of our best performing DFFNN - incorporating 14 historical lagged features, against our newly established baseline, the MFFNN, across each flu season.

The table of results reinforces the DFFNN-14's enhanced performance compared to the MFFNN, with the DFFNN-14 outperforming the MFFNN in four out of five flu seasons. In particular, the 2016-17 flu season presents the most significant improvement in performance, with the DFFNN-14 achieving 46.6% and 47.6% better MAE and MAPE respectively. Furthermore, the significantly greater stability across seeds in the 2018-19 season underscores the robustness of the improved model. While the 2015-16 season presents an anomaly with the MFFNN slightly bettering the DFFNN-14 in MAE by 1.6%, the DFFNN-14 still attains a smaller relative percentage error and

Flu Season	MAE		MAPE		r	
	MFFNN	DFFNN-14	MFFNN	DFFNN-14	MFFNN	DFFNN-14
2014-2015	1.24 (0.03)	**1.06** (**0.03**)	22.62 (2.35)	**18.06** (**1.37**)	0.9506	**0.9615**
2015-2016	**1.25** (**0.04**)	1.27 (0.18)	19.78 (2.105)	**16.77** (**2.53**)	0.9622	**0.9711**
2016-2017	2.08 (0.35)	**1.11** (**0.18**)	45.12 (11.23)	**23.63** (**4.06**)	0.9587	**0.9654**
2017-2018	1.72 (0.20)	**1.56** (**0.11**)	**23.73** (**6.46**)	23.77 (5.46)	0.9849	**0.9883**
2018-2019	1.78 (0.97)	**1.25** (**0.16**)	44.26 (25.36)	**29.48** (**5.42**)	0.9667	**0.9701**
Average	1.61 (0.19)	**1.25** (**0.07**)	31.02 (5.14)	**22.34** (**1.87**)	0.9646	**0.9713**

Table 5.3: Nowcasting performance of the Minimal Feed Forward Neural Network (MFFNN) against the best Deeper Feed Forward Neural Network (DFFNN) - using 14 historical lagged features, for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during the each test flu season (from 2014-15 to 2018-19). The last row shows the performances averaged over the five test flu seasons. MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the nowcasts and reported ILI rates. For each flu season, and the average results across flu seasons, we also report the standard deviation of the MAE and MAPE across the 10 fixed seeds. The variation in the bivariate correlation reported across seeds is < 0.01 for both models, and thus is omitted from the table. The best results for each metric and flu season are shown in bold.

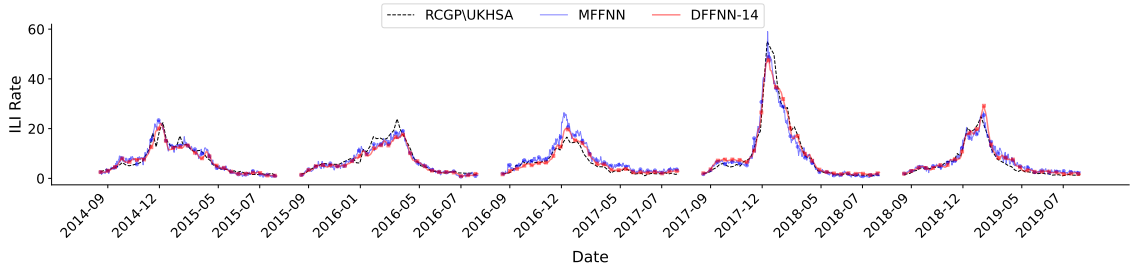

Figure 5.4: Comparative plot of the nowcasting predictions made by the Minimal Feed Forward Neural Network (MFFNN) against the best Deeper Feed Forward Neural Network (DFFNN) - using 14 historical lagged features, for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during each test flu season (from 2014-15 to 2018-19).

a lower deviation. However, it's notable that the MFFNN outperforms the DFFNN-14 in the 2017-18 flu season in MAPE. Nonetheless, the DFFNN-14 sets a high standard of performance across flu seasons, as evident in its average metric improvements. The toughest challenge for the DFFNN-14, similar to the Elastic Net model, occurs in the 2017-18 flu season, likely due to the distinctive peak of this season, which has proven more challenging to predict compared to other seasons.

The prediction plots for each flu season, as illustrated in Figure 5.4, highlight the significant refinement in prediction smoothness achieved with the incorporation of historical lagged features in the DFFNN-14. This model's predictions closely mirror the smooth trends of the ground truth, particularly enhancing its accuracy over the MFFNN, as observed in the flu seasons of 2014-15 and 2018-19. Although both models present similar trends, the DFFNN-14's smoother curve aligns more precisely with the actual rates, leading to a higher predictive accuracy throughout seasonal periods. For instance, in the 2014-15 flu season, the DFFNN-14 exhibits lower MAEs of 1.03 in the onset (01 September to 01 December), and 2.10 during the peak (02 December to 20 January), compared to the MFFNN's MAEs of 1.35 and 2.50 in these respective periods.

The 2016-17 flu season displays more variability between the models' predictions, with distinct periods where the DFFNN-14's superiority is evident. During the onset (01 September to 01

December) and peak (02 December to 30 January) periods, the DFFNN-14's predictions align significantly closer with the actual ILI rates, with MAEs of 0.70 and 2.00 respectively, despite a substantial overprediction in the latter. However, this is still markedly lower than the MFFNN's MAEs of 0.83 and 4.31 for the same periods, which overpredicts the peak period even further. This corresponds to the marked performance differential in the results table.

While the differences in the prediction trends in the remaining two seasons aren't as distinct, reflecting the closer performance metrics, there still remain some interesting insights into deficiencies of the DFFNN-14. Despite the DFFNN-14 significantly better capturing the peak in 2016-17, this is not uniform across seasons. In particular, the deeper model slightly underestimates the peaks (MAE 3.11) compared to the baseline, which aligns closer to the ground truth (MAE 2.50) in the 2015-16 season. This results in a marginal performance edge as detailed in Table 5.3. A similar trend can be observed during the peak of 2017-18, from 31 December to 08 February, where the MFFNN again better captures the sharp rise in ILI rate with an MAE of 4.74, albeit slightly overpredicting, compared to the underpredictions made by DFFNN-14, resulting in an MAE of 5.13. The DFFNN's overall superior performance in this season is instead attributed to its greater accuracy in the tail period. These observations unveil difficulties in the lagged feature model, with it not demonstrating a consistent superior performance across given seasons but struggling to capture peak periods in comparison to the baseline, possibly due to its increased complexity.

Despite these slight deficiencies, overall, the DFFNN-14 significantly outperforms the baseline model. It demonstrates significant enhancements in predictive accuracy across flu seasons and the specific utility of incorporating historical lagged features into the model's input, providing greater context on recent search behaviour and thus ILI rates. This establishes competitive performance, exceeding the outcomes of related studies that utilise more conventional machine learning techniques to estimate ILI rates in England [1], [2]. This underscores the capabilities of deep learning approaches in the predictive task and sets a benchmark for evaluating more advanced deep learning models, as suggested in other location studies [12], [17]. The analysis not only outlines the efficiency of the DFFNN-14 but also indicates specific limitations of the model in capturing certain peak periods of ILI rates which can be potentially better captured by these more complex architectures, such as recurrent models equipped with temporal context. These comparisons can determine if the incremental benefits, if any, provided by more complex models warrant the increased complexity and training challenges, especially when the DFFNN-14 model, with its relative simplicity, already attains a high level of nowcasting accuracy.

5.2 Forecasting

Following the training of our DFFNN-14 model for nowcasting ILI rates, we turn our attention to its forecasting capabilities. The goal remains consistent with our earlier efforts: to leverage neural networks for competitive forecasting results and to establish a point of comparison for evaluating more advanced deep learning architectures prevalent in forecasting tasks [3], [28]–[30]. We scrutinise the DFFNN-14's accuracy in forecasting over different horizons of 7, 14, 21, and 28 days ahead, shifting the prediction target from the ILI rate on day t to that on day $t + \gamma$. The performance of these forecasts will be compared with our previously defined baseline, the MFFNN, and a traditional baseline forecasting model known as the persistence method.

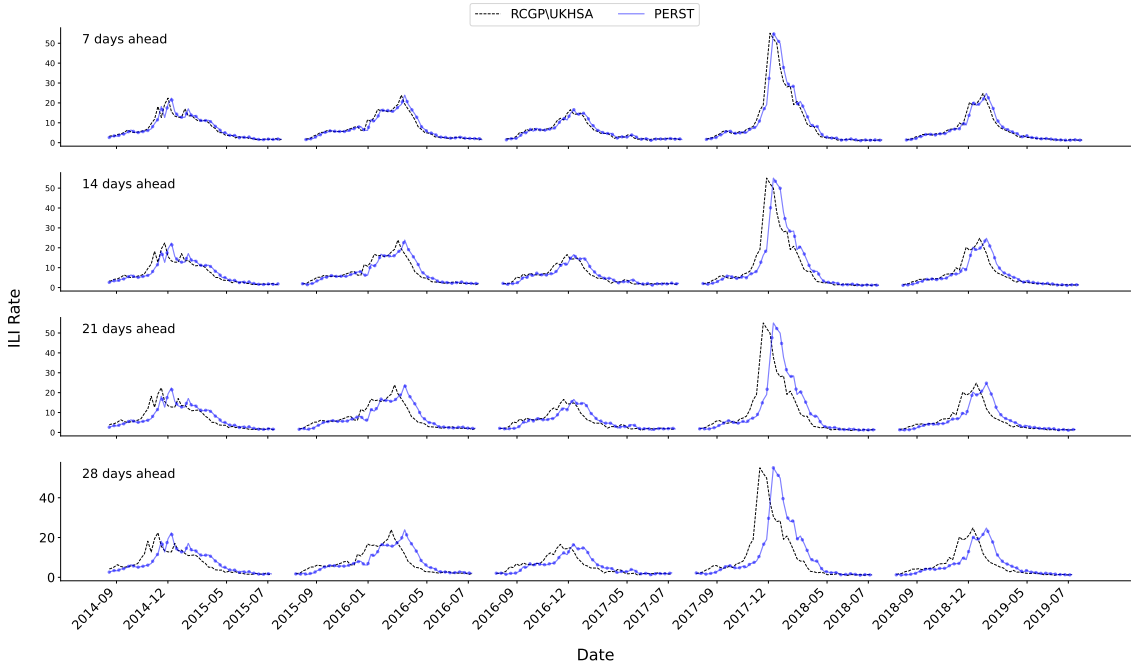

Figure 5.5: Forecasting predictions made by the baseline Peristence model (PERST) for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during each test flu season (from 2014-15 to 2018-19), for four different forecasting horizons γ - 7, 14, 21 and 28.

5.2.1 Persistence Model

The persistence model is a basic yet commonly employed benchmark in time series forecasting. It assumes the future values of a variable will be the same as the most recent values. For instance, if the ILI rate on day t is 20, the persistence model would forecast that the ILI rate on day $t+\gamma$ will also be 20. Effectively, the forecasted ILI rate trend from the persistence model would appear as a temporal shift of the actual data, where the predicted value on day t mirrors the real value on day $t-\gamma$. This is further illustrated in Figure 5.5, where we plot the predictions of the persistence model against the ground truth ILI rate for each flu season and forecasting horizon.

The model is predicated on the stability of the system in the short term. While simple and less precise over longer forecasting durations or in highly variable conditions, it offers surprisingly accurate short-term forecasts, especially when compared with our models that do not include the autoregressive components of ILI rates. It thus serves as a simple point of reference to assess the ability of our forecasting methodologies.

5.2.2 Forecasting Performance

Methodology

We adopted the same training and evaluation methodology for our models as we did in the now-casting task. The only difference however arised in firstly shifting the ground truth values for each sample in our dataset such that we were predicting the ILI rate on day $t+\gamma$ as opposed to day t. We ensured to retain the same set of features our model trained using the same hybrid feature selection method, where the feature filtering was still applied on the correlations of query frequencies on day t and the ILI rate on day t - as opposed to the correlations with the forecasted

γ	Flu Season	MAE			MAPE			r		
		PERST	MFFNN	DFFNN-14	PERST	MFFNN	DFFNN-14	PERST	MFFNN	DFFNN-14
7	2014-2015	**1.06**	1.31	1.27	**13.92**	23.11	19.30	0.9483	0.9439	**0.9513**
	2015-2016	**0.96**	1.54	1.83	**12.71**	21.66	23.93	**0.9681**	0.9399	0.9393
	2016-2017	**0.73**	2.46	1.21	**14.79**	56.99	26.74	**0.9697**	0.9392	0.9437
	2017-2018	**2.13**	2.27	2.28	**18.62**	25.62	29.69	0.9551	**0.9828**	0.9852
	2018-2019	**0.95**	1.92	1.49	**13.93**	46.09	38.96	**0.9664**	0.9501	0.9478
	Average	**1.17**	1.90	1.62	**14.79**	34.7	27.73	**0.9615**	0.9512	0.95
14	2014-2015	1.63	1.75	**1.60**	**21.70**	25.96	21.91	0.8900	0.9039	**0.9252**
	2015-2016	**1.70**	1.85	2.15	**22.86**	23.38	25.19	0.9076	0.9187	**0.9274**
	2016-2017	**1.24**	2.38	1.80	**24.2**	55.49	35.13	**0.9126**	0.9098	0.8720
	2017-2018	4.04	3.10	**3.02**	33.23	**28.42**	32.77	0.8542	**0.9414**	0.9313
	2018-2019	**1.70**	2.33	2.06	**24.42**	55.21	47.32	0.8945	0.9176	**0.9235**
	Average	**2.06**	2.28	2.13	**25.28**	37.69	32.47	0.8918	**0.9182**	0.9159
21	2014-2015	2.15	**2.03**	2.15	30.11	27.23	**24.74**	0.8309	0.8530	**0.8803**
	2015-2016	2.29	**2.29**	2.70	31.74	**25.38**	28.77	0.8306	**0.8949**	0.8885
	2016-2017	**1.70**	3.17	2.26	**32.04**	70.03	44.06	0.8402	**0.8416**	0.8205
	2017-2018	5.82	3.45	**3.32**	47.67	**29.25**	32.9	0.7297	0.8757	**0.8850**
	2018-2019	2.42	2.65	**2.03**	**34.96**	64.46	43.47	0.8032	0.8723	**0.8989**
	Average	2.88	2.72	**2.49**	35.30	43.27	**34.79**	0.8069	0.8675	**0.8746**
28	2014-2015	2.55	**2.34**	2.39	37.79	30.05	**27.54**	0.7553	0.8086	**0.8569**
	2015-2016	2.82	**2.62**	2.81	40.93	**25.38**	28.88	0.7372	**0.8468**	0.8287
	2016-2017	**2.17**	3.49	2.66	**40.57**	76.52	52.77	0.7454	**0.7924**	0.7847
	2017-2018	7.31	4.04	**3.96**	62.20	**30.73**	39.75	0.6020	0.7995	**0.8374**
	2018-2019	3.11	2.56	**1.94**	46.46	59.09	**41.14**	0.6952	0.8694	**0.9110**
	Average	3.59	3.01	**2.75**	45.59	44.75	**38.02**	0.7070	0.8223	**0.8437**

Table 5.4: Forecasting performance of the baseline Persistence model (PERST), Minimal Feed Forward Neural Network (MFFNN) and the best Deeper Feed Forward Neural Network (DFFNN) - using 14 historical lagged features, for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during the each test flu season, for four different forecasting horizons γ - 7, 14, 21 and 28. The last row for each forecasting horizon shows the performances averaged over the five test flu seasons (from 2014-15 to 2018-19). MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the forecasts and reported ILI rates. The best results for each metric and flu season are shown in bold.

targets.

Results

Table 5.4 presents the predictive performances of the DFFNN-14, MFFNN and persistence models for each forecasting horizon averaged across the the 10 fixed seeds. Figure 5.6 provides an alternative visual of the forecasting performance metrics of the different models when averaged over the five flu seasons.

Beginning with the 7-day forecast horizon, we observe the persistence model typically surpassing both neural network models. This is largely attributed to the short-term nature of the forecast, where recent trends are likely to continue. Despite this, the DFFNN-14 demonstrates its potential by outperforming the MFFNN in three of the five flu seasons, particularly noting a significant decrease in MAE during the 2016-17 season. This performance gap, also consistent across longer forecast horizons, mirrors our findings from the nowcasting phase, highlighting the baseline model's difficulties with accurately capturing the ILI rate trend in this specific season.

Expanding the forecast to a 14-day horizon reveals the diminishing accuracy of the persistence model, suggesting its predictions become less reflective of the actual trends. In this context, the DFFNN-14 begins to demonstrate its strengths, outperforming the persistence model in two of the five seasons, notably in the 2017-18 season where it better captures unexpected peaks within the data (Figure 3.1). Additionally the correlation of the persistence model's forecasts with the ground truth ILI rates significantly decreases, achieving the best correlation in only one of the flu seasons.

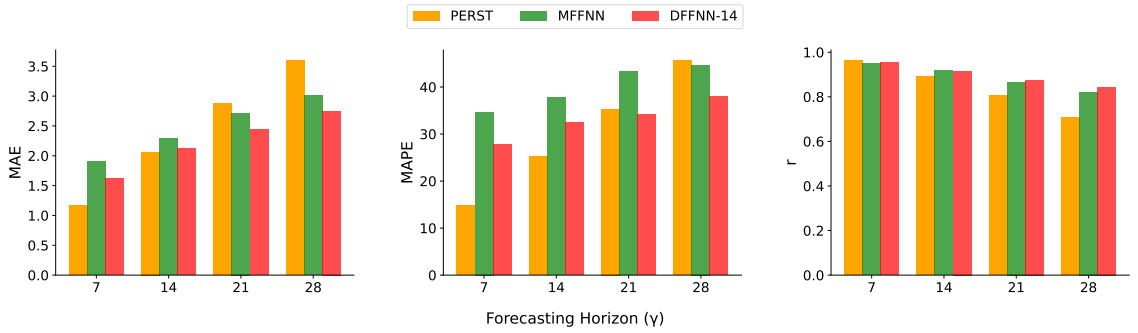

Figure 5.6: Comparative plot of the forecasting performance metrics of the Persistence model (PERST), Minimal Feed Forward Neural Network (MFFNN) and the best Deeper Feed Forward Neural Network (DFFNN) - using 14 historical lagged features, for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth), for four different forecasting horizons γ - 7, 14, 21 and 28. MAE is the mean absolute error, MAPE is the mean absolute percentage error and r is the Bivariate correlation between the forecasts and reported ILI rates, averaged across the 5 test flu seasons (from 2014-15 to 2018-19).

The persistence model's limitations become markedly apparent as we extend the forecasting period to 21 and then 28 days. The method's reliance on the continuation of recent trends fails to capture the evolving nature of ILI rates, making its predictions increasingly divergent from actual trends. In contrast, the DFFNN-14 showcases its robustness in these extended forecasts, achieving notable improvements in MAE, MAPE, and Bivariate correlation of 23.4%, 16.6%, and 15.5% respectively, across all seasons when predicting 28 days ahead. This superior performance over both the MFFNN and persistence model, highlights the DFFNN-14's proficiency in navigating the complexities of flu trends over prolonged horizons.

While the DFFNN-14 generally presents as the superior forecasting solution for longer durations, particularly in later seasons characterised by intricate patterns, it's interesting to note the persistence model's resilience in shorter-term forecasts. Moreover, the MFFNN demonstrates an unexpected efficiency in earlier seasons, consistently outperforming the DFFNN-14, albeit with smaller margins. These observations, reminiscent of the nowcasting phase where the MFFNN displayed a stronger performance in the 2015-16 season, reveal the nuanced effectiveness of each model under various conditions.

The model predictions presented in Figures 5.5 and 5.7 reinforce our previous inferences, offering clear visual evidence of the trends observed in performance metrics. For shorter forecasting horizons, the persistence model aligns more closely with actual ILI rates, adeptly mirroring true trends with its prediction shifts. In contrast, the neural network models exhibit their superiority over extended forecasting horizons, particularly during flu seasons characterised by significant fluctuations, such as that of 2017-18. Although the neural network models tend to underpredict the peak ILI rates in this season, they nonetheless more effectively capture the overall trend compared to the now larger, less accurate shifts of the persistence model. For the 28-day forecasting horizon in particular, the persistence model recorded MAEs of 2.86, 27.07, and 6.08 across the onset, peak, and tail periods respectively. In comparison, the MFFNN-14 demonstrated slightly lower MAEs of 2.52, 23.92, and 1.13, while the DFFNN-14 showed further improved performance with MAEs of 2.26, 20.78, and 1.46 for the same periods. This substantiates the neural networks' advantage over the persistence model in capturing the ILI rate trend throughout more variable flu seasons in

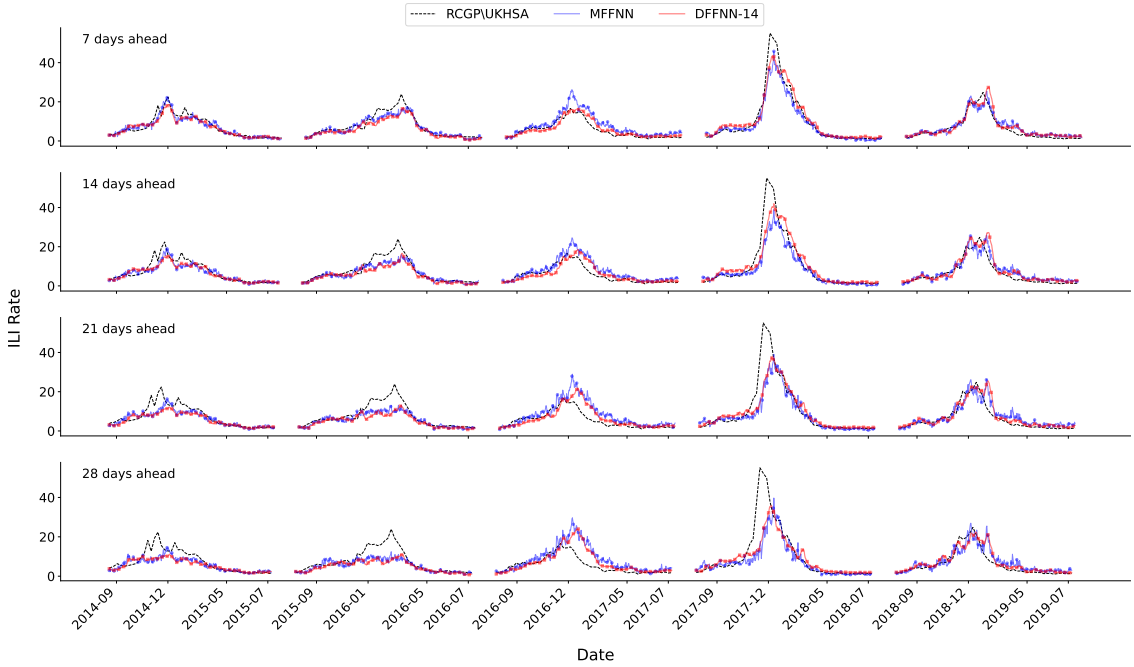

Figure 5.7: Comparative plot of the forecasting predictions made by the Minimal Feed Forward Neural Network (MFFNN) and the best Deeper Feed Forward Neural Network (DFFNN) - using 14 historical lagged features, for the estimation of ILI rates in England (RCGP/UKHSA ILI rates denote the ground truth) during each test flu season (from 2014-15 to 2018-19), for four different forecasting horizons γ - 7, 14, 21 and 28.

longer forecasting horizons.

Conversely, during the more stable ILI rate trend of the 2016-17 flu season, the persistence model's straightforward shift of ground truth data appears more effective, outperforming the neural network models that tend to overpredict the trend. This emphasises the persistence model's potential in conditions with less variability, where a simple shift can reliably anticipate future values, even over longer forecasting horizons. Delving deeper into the periodic errors of the models in this flu season, with a 28-day forecasting horizon, we can observe that while the neural network models edge the accuracy in capturing the onset and peak periods, the overall superior performance of the persistence model is attributed to its significantly better MAE during the tail period, achieving 1.95 compared to the much higher 3.28 and 4.46 of the DFFNN-14 and MFFNN, respectively.

The peak periods provide additional insight into our models' forecasting abilities. Although neural network models exhibit superior performance over longer forecasting horizons, this consistency does not extend across the entire duration of flu seasons. Specifically, during the early seasons of 2014-15 and 2015-16, the neural network models significantly underpredict the peak flu period, with a considerably flat forecasted trend, resulting in higher MAEs than those achieved by the persistence model. For instance, in the 2015-16 flu season with a 28-day forecasting horizon, the persistence model achieved a MAE of 4.18, markedly lower than the 6.82 and 8.05 reported by the MFFNN and DFFNN-14 respectively. The overall superior performance of the neural network models is instead attributed to their considerably lower MAEs in capturing the onset and tail periods. During the same season, the MFFNN and DFFNN-14 recorded MAEs of 1.43 and 1.05 respectively for the onset, and 1.32 and 1.29 for the tail periods, in stark contrast to the persistence model's 1.39 and 6.66. This notable disparity in non-peak periods compensates for

their peak period shortcomings, affirming their overall effectiveness during these seasons in longer forecasting horizons.

Continuing our analysis on the longer forecasting horizons, we observe notable differences between the MFFNN and DFFNN-14 which highlight the contrasting prediction performance between the models. The MFFNN's forecasts are closer to the actual rates during the earlier flu seasons of 2014-15 and 2015-16, a trend that aligns with previously identified patterns (Table 5.4). In particular, when forecasting 21 days ahead in the 2014-15 flu season, the MFFNN recorded MAEs of 5.84 during the peak and 1.95 during the tail, whereas the DFFNN-14 exhibited slightly higher MAEs of 6.59 and 2.23 for these respective periods. Conversely, the DFFNN-14's predictions align more closely with the ground truth in later seasons, particularly in the 2018-19 flu season, where it demonstrated MAEs of 0.93 and 2.08 during the onset and peak periods, compared to the MFFNN's 1.03 and 3.68. Considering the similarities in their predicted trends, the DFFNN-14's smoother prediction curves, a trait carried over from the nowcasting task, significantly contributes to its improved accuracy. This visual smoothness underscores the efficacy of incorporating historical features when forecasting over extended periods, through its enhanced capability to accurately follow and capture the underlying trends in ILI rates.

One clear observation across all models is the diminishing ability to capture the ILI rate trend as the forecasting horizon extends. This trend is logical, as predicting further ahead in time becomes increasingly challenging. Notably, as the forecasting horizon extends, the neural network models exhibit an increasing tendency to over predict the peak ILI rates during more stable periods, such as in the 2016-17 flu season, and under predict during peak periods in remaining seasons. Additionally, it becomes evident that the differentiation in the prediction trends of the neural network models decreases for longer forecasting horizons. For instance, during shorter forecasting periods in the 2016-17 flu season, the differences in MAE between the DFFNN-14 and the MFFNN are more pronounced with the DFFNN-14 having a peak MAE of 1.67 compared to the MFFNN's 3.85. However, in later forecasting horizons of 28 days ahead, in the same flu season, both models exhibit closer MAEs of 2.67 (MFFNN) and 1.97 (DFFNN-14), with the DFFNN-14 increasingly growing closer to the overprediction trend of the MFFNN. This demonstrates a convergence in their predictive performances.

This suggests a potential limitation of the DFFNN-14 in distinguishing itself from our baseline MFFNN model, possibly reaching the limits of improvement with the vanilla FFNN architecture despite the inclusion of lagged features. This is further supported by the DFFNN's struggle to enhance nowcasting performance with the addition of more lagged features, which led to a downturn in nowcasting results. These findings provide grounds for the further exploration we have threaded throughout our study, in exploring more complex network architectures, like recurrent models, which now have a competitive performance benchmark in the DFFNN-14 to be evaluated against. These models can be assessed not only for their ability to improve forecasting performance across various horizons but also for their resilience against the tendency to over or under predict as the forecasting horizon extends, and instead maintaining a trend that stays closer to the ground truth.

Overall, the forecasting results of our DFFNN-14 still show promise. Although it exhibits variability in performance, sometimes struggling with the simpler, earlier seasons, it generally outperforms both the persistence and MFFNN models in longer-term predictions, underscoring its robustness as a forecasting model. This competes with the forecasting accuracy reported in related studies for England [3], which employed simpler FFNN models as well as more complex LSTM models with uncertainty. Even though our study encompasses five flu seasons, compared to their four, the DFFNN-14's on par-to superior performance on each forecasting horizon relative to their

best models signifies that we have achieved a highly competitive performance for the prediction task. This contributes a benchmark for evaluating the performance and necessity of more complex models, which are typically more challenging to train, against our simpler performant DFFNN-14.

Chapter 6

Conclusion

6.1 Summary

In summary, this project aimed to explore the predictive potential of deep learning models for nowcasting and forecasting ILI rates in England, with a focus on developing relatively simple yet effective Feed Forward Neural Network (FFNN) architectures. Our objective was twofold: to achieve competitive prediction performance, surpassing traditional machine learning methods, and to provide a robust benchmark for evaluating the performance and necessity of more complex models discussed in related studies [3], [12], [17], [29].

We began our investigation by training a baseline Elastic Net model for the nowcasting task, exploring various feature selection methods. Our findings favoured the hybrid approach, which aligned with expectations from previous studies [1], effectively addressing the first two goals outlined in the Introduction chapter and laying a solid foundation for subsequent analyses.

Building upon this groundwork, our next objective was to enhance predictive performance for the nowcasting task by training a minimal FFNN (MFFNN) with a single hidden fully dense layer containing 4 units. This simplistic deep learning model surpassed the performance of the Elastic Net baseline across flu seasons, thereby introducing a new baseline for the prediction task and demonstrating the potential of deep learning models. This achievement provided a basis for progressing to our next goal of exploring deeper architectures for competitive performance.

Subsequently, we further improved our predictive models by training a deeper FFNN (DFFNN), incorporating 14 historical lagged features, which yielded significant advancements in accuracy. These results not only exceeded performances reported in similar studies [1], [2] but also provided a competitive benchmark for further evaluations, thus fulfilling the fourth goal.

Finally, we evaluated the forecasting capabilities of our DFFNN model against the MFFNN and a persistence model across various forecasting horizons ($\gamma = 7, 14, 21$, and 28). Our results highlighted the superior performance of the DFFNN, particularly with extended forecasting horizons, outperforming both baseline models and metrics reported in comparable studies. This includes those that explored more complex neural network architectures such as the LSTM, thereby achieving the fifth and final goal of our study.

In conclusion, this project successfully achieved its objectives of attaining highly competitive performance for both prediction tasks using neural network models. The simplicity of our FFNN architectures not only demonstrated their effectiveness but also provides a clear benchmark for evaluating more complex models, ensuring that any performance gains are substantial, justifying

their increased training difficulty and complexity.

6.2 Evaluation

The results presented in our study align closely with the initial goals we set out to achieve, providing a consistent narrative that demonstrates the potential of deep learning models, particularly simpler Feed Forward Neural Network (FFNN) architectures, in the prediction tasks. We achieved competitive performances across both nowcasting and forecasting, reflecting the effectiveness of our approaches. Although we analysed some deficiencies in our deeper FFNN architecture (DFFNN), particularly its inferior forecasting performance on earlier flu seasons compared to our minimal FFNN (MFFNN), overall, the performance of the DFFNN across flu seasons and longer forecasting horizons demonstrated its efficacy and robustness.

However, our project was not without its challenges. As discussed, during the training of the baseline Elastic Net model, we encountered difficulties in refining our cross-validation strategy to obtain intuitive and consistent results. Despite this challenge, it also underscored the limitations of the baseline model and the complexity involved in achieving the performance levels we eventually attained. This was further reinforced by the ease with which we surpassed this performance with neural networks, as showcased by our minimal FFNN architecture.

Another significant challenge we encountered was the lengthy training times of our models, particularly with the resources available. We initially faced limitations with GPU caps on Google Colab, prompting us to transition to UCL lab machines and a time-shared machine named Blaze. Although these machines offered better GPU capabilities, the issue of long training times persisted, albeit to a lesser extent with Blaze due to its uninterrupted training time.

Given more time and resources, we could have further investigated the performance of more complex models such as recurrent models like the LSTM or GRU. This would have added further depth to our experiments regarding the comparison of these more difficult models to train against the high performance of our simpler FFNN architectures. These more complex models require significantly more settings to configure and parameters to learn, thus necessitating a validation of parameters and architectures to achieve reasonable performance. This would have been significantly more GPU-intensive and time-consuming, which was constrained by the resources and time available to us.

Additionally, a key challenge we encountered was the nature of user-generated content (UGC), particularly web search queries, which served as our main predictor for ILI rates. Despite employing various feature selection techniques, this approach is limited to the occurrence of search queries without contextual backgrounds to the purpose of these specific searches. However, despite this limitation, numerous studies, including ours, have demonstrated the high prediction performance achievable through training machine learning models on such data, providing valuable insights into ILI prediction.

6.3 Future Work

There are several aspects of our study that can be extended and further explored in the domain of nowcasting and forecasting ILI rates in England.

One avenue we have touched upon is the incorporation of autoregressive components into our models, through the inclusion of historical ILI rates as additional features, alongside web search

queries. Integrating the predicted ILI rates of our model for given days can provide additional context, potentially improving performance in both nowcasting and forecasting tasks. This could especially aid in forecasting, where our proposed DFFNN architecture struggled with earlier forecasting horizons due to a lack of awareness of ILI rate trends, unlike the persistence model which benefited from short-term shifts in the ground truth. By providing this context, we may achieve further performance gains for our relatively simpler models.

Another aspect to investigate is the potential performance gains from refining the feature set. In our study, we retained the same feature set employed by the elastic net to demonstrate direct performance gains obtained when training on the same set of optimal features. However, neural networks can effectively learn from a greater number of features, providing additional context and potentially improving predictive performance. In the case of forecasting, this could involve modifying the feature selection approach to focus on correlations with forecasted targets as opposed to those of nowcasting or even a combination of both.

As discussed previously, an additional extension to our work could involve training more complex deep neural network architectures such as recurrent models like LSTMs or GRUs, which are adept at modelling time series data. This would allow us to evaluate the performances of these models and assess whether their increased difficulty and complexity are justified by performance improvements over our simpler DFFNN architecture.

When training these more complex models, we mentioned the necessity to validate the increased number of settings present. This invites further exploration into different validation strategies to determine the best generalisation settings for model parameters. Exploring strategies such as validation on the most recent flu season data or segments from multiple recent flu seasons could provide interesting insights into obtaining optimal parameters for improved model performance.

Lastly, we could enhance our understanding of the forecasting abilities of our models by modifying the nature of the prediction task. Instead of predicting a single day at different forecasting horizons, we could assess the models' capabilities by predicting multiple outputs and sequences of future values. For instance, rather than predicting day t+7 in isolation, we could predict a sequence of days from t to t+6, providing a more comprehensive assessment on ILI rate forecasting predictions.

Bibliography

[1] R. P. V. Lampos E. Yom-Tov and I. J. Cox, "Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance," *Proceedings of the 26th International Conference on World Wide Web*, pp. 695–704, 2017. DOI: 10.1145/3038912.3052622.

[2] V. Lampos, "Flu Detector: Estimating influenza-like illness rates from online user-generated content," *ArXiv*, vol. abs/1612.03494, 2016. DOI: 10.48550/arXiv.1612.03494.

[3] I. J. C. M. Morris P. Hayes and V. Lampos, "Estimating the Uncertainty of Neural Network Forecasts for Influenza Prevalence Using Web Search Activity," *arXiv*, vol. abs/2105.12433, 2021. DOI: 10.48550/arXiv.2105.12433.

[4] H. P. H. Kwak C. Lee and S. Moon, "What Is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 591–600. DOI: 10.1145/1772690.1772751.

[5] H. Choi and H. Varian, "Predicting the Present with Google Trends," *Economic Record*, vol. 88, pp. 2–9, 2012. DOI: 10.1111/j.1475-4932.2012.00809.x.

[6] P. R. M. Cha H. Kwak and Y. Ahn, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007, pp. 1–14. DOI: 10.1145/1298306.1298309.

[7] M. G. M. D. Choudhury and S. Counts, "Predicting Depression via Social Media," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2013, pp. 128–137. DOI: 10.1609/icwsm.v7i1.14432.

[8] W. H. H. Johnson M. Wagner and W. Chapman, "Analysis of Web Access Logs for Surveillance of Influenza," *Studies in Health Technology and Informatics*, vol. 107 Pt 2, pp. 1202–1206, 2004. DOI: 10.3233/978-1-60750-949-3-1202.

[9] D. M. P. P. Polgreen Y. Chen and F. Nelson, "Using internet searches for influenza surveillance," *Clinical Infectious Diseases*, vol. 47, pp. 1443–1448, 2008. DOI: 10.1086/593098.

[10] R. S. P. J. Ginsberg M. H. Mohebbi and L. Brammer, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009. DOI: 10.1038/nature07634.

[11] I. J. C. V. Lampos and N. Cristianini, "Advances in nowcasting influenza-like illness rates using search query logs," *Scientific Reports*, vol. 5, p. 12 760, 2015. DOI: 10.1038/srep12760.

[12] Y. C. Y. He Y. Zhao and H.-Y. Yuan, "Nowcasting influenza-like illness (ILI) via a deep learning approach using google search data: An empirical study on Taiwan ILI," *International Journal of Intelligent Systems*, vol. 37, pp. 2648–2674, 2021. DOI: 10.1002/int.22788.

[13] A. F. S. Cook C. Conrad and M. Mohebbi, "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic," *PLoS ONE*, vol. 6, 2011. DOI: 10.1371/journal.pone.0023610.

[14] M. P. D. Olson K. Konty and C. Viboud, "Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales," *PLoS Computational Biology*, vol. 9, 2013. DOI: 10.1371/journal.pcbi.1003256.

[15] G. K. D. Lazer R. Kennedy and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, vol. 343, pp. 1203–1205, 2014. DOI: 10.1126/science.1248506.

[16] T. Preis and H. S. Moat, "Adaptive nowcasting of influenza outbreaks using Google searches," *Royal Society Open Science*, vol. 1, 2014. DOI: 10.1098/rsos.140095.

[17] P. Jiang, "Nowcasting Influenza Using Google Flu Trend and Deep Learning Model," *Advances in Economics, Business and Management Research*, vol. Proceedings of the 2020 2nd International Conference on Economic Management and Cultural Industry (ICEMCI 2020), pp. 2352–5428, 2020. DOI: 10.2991/aebmr.k.201128.079.

[18] M. D. A. Lamb M. J. Paul and R. Pebody, "Separating Fact from Fear: Tracking Flu Infections on Twitter," *Proceedings of National Assessment Accreditation Council*, vol. 13, pp. 789–795, 2013.

[19] M. D. D. A. Broniatowski M. J. Paul and I. J. Cox, "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic," *PLoS ONE*, vol. 8, 2013. DOI: 10.1371/journal.pone.0083672.

[20] P. G. S. A. Tumasjan T. Sprenger and I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2010, pp. 178–185. DOI: 10.1609/icwsm.v4i1.14009.

[21] B. R. R. B. T. O'Connor R. Balasubramanyan and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2010, pp. 122–129. DOI: 10.1609/icwsm.v4i1.14031.

[22] X.-J. Z. J. Bollen H. Mao and L. Brilliant, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, pp. 1–8, 2010. DOI: 10.1016/j.jocs.2010.12.007.

[23] S. L. S. Goel J. M. Hofman and D. M. Pennock, "Predicting consumer behavior with Web search," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 17 486–17 490, 2010. DOI: 10.1073/pnas.1005962107.

[24] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," in *2010 2nd International Workshop on Cognitive Information Processing*, 2010, pp. 411–416. DOI: 10.1109/CIP.2010.5604088.

[25] S. Kandula and J. Shaman, "Reappraising the utility of Google Flu Trends," *PLoS Computational Biology*, vol. 15, 2019. DOI: 10.1371/journal.pcbi.1007258.

[26] D. H. S. Kandula and J. Shaman, "Subregional Nowcasts of Seasonal Influenza Using Search Trends," *Journal of Medical Internet Research*, vol. 19, 2017. DOI: 10.2196/jmir.7486.

[27] Y. G. A. F. Dugas M. Jalalpour and S. Levin, "Influenza Forecasting with Google Flu Trends," *PLoS One*, vol. 8, 2013. DOI: 10.1371/journal.pone.0056176.

[28] L. L. R. R. Q. Xu Y. Gel and K. Nezafati, "Forecasting influenza in Hong Kong with Google search queries and statistical model fusion," *PLoS ONE*, vol. 12, 2017. DOI: 10.1371/journal.pone.0176690.

[29] K.-L. T. Y. Liu G. Feng and S. Sun, "Forecasting influenza epidemics in Hong Kong using Google search queries data: A new integrated approach," *Expert Systems with Applications*, vol. 185, 2021. DOI: 10.1016/j.eswa.2021.115604.

[30] I. J. C. M. Morris P. Hayes and V. Lampos, "Neural network models for influenza forecasting with associated uncertainty using Web search activity trends," *PLoS Computational Biology*, vol. 8, 2023. DOI: 10.1371/journal.pcbi.1011392.

[31] F. Galton, "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 15, pp. 246–263, 1886.

[32] X. Guyon and J. Yao, "On the Underfitting and Overfitting Sets of Models Chosen by Order Selection Criteria," *Journal of Multivariate Analysis*, vol. 70, no. 2, pp. 221–249, 1999. DOI: 10.1006/jmva.1999.1828.

[33] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[34] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, vol. 12, pp. 69–82, 1970. DOI: 10.1080/00401706.1970.10488635.

[35] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x.

[36] L. P.-P. M.-C. Popescu V. E. Balas and N. E. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

[37] S. K. S. S. R. Dubey and B. B. Chaudhuri, "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark," *Neurocomputing*, vol. 461, pp. 657–673, 2021. DOI: 10.48550/arXiv.2109.14545.

[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, vol. 9, pp. 249–256, 2010.

[39] S. R. K. He X. Zhang and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.

[40] S. Ruder, "An overview of gradient descent optimization algorithms," *ArXiv*, vol. abs/1609.04747, 2016. DOI: 10.48550/arXiv.1609.04747.

[41] C. S. B. L. E Lansbury and J. S. Nguyen-Van-Tam, "Influenza in long-term care facilities," *Influenza Other Respir Viruses*, pp. 356–366, 2017. DOI: 10.1111/irv.12464.

Appendices

Appendix A

Software and Tools

The following section outlines and motivates the various software, programming languages, frameworks and libraries employed for our study. The choices made were guided by the specific requirements of the prediction task and the advantages offered by each tool.

A.1 Programming Language

Python was the programming language of choice for this experimentation, due to its versatility and extensive support in the machine learning community. It provides a ecosystem of libraries and frameworks for tasks ranging from data preprocessing to complex machine learning model training.

A.2 Machine Learning Libraries

A.2.1 Scikit-Learn

Scikit-Learn was the primary library employed for training linear machine learning models. It is a widely-used, user-friendly machine learning library offering a comprehensive set of tools for data preprocessing, feature engineering, and model evaluation.

A.2.2 Pytorch

For more advanced deep learning models, PyTorch was employed. Pytorch is known for its dynamic computation graph and flexibility, facilitating the training of complex neural networks. Its popularity and seamless integration with GPU acceleration were key factors for its use.

A.2.3 Matplotlib

Matplotlib is a data visualisation library in Python. This was utilised for creating insightful visualisations to aid in the analysis of experimental results.

A.2.4 SentenceTransformers

As discussed, one of the feature selection methods we aimed to explore involved capturing the textual semantics of web search queries to discern their relevance to ILI rates. This approach is facilitated by employing a pre-trained sentence transformer model from the SentenceTransformers

Python library. This is a framework designed for generating state-of-the-art sentence, text, and image embeddings, providing support for over 100 languages. The embeddings are compared via a measure of similarity, such as cosine similarity, to identify sentences with similar meanings. The framework is based on PyTorch and Transformers and provides access to a vast collection of models pre-tuned for a variety of tasks.

For our specific task, we selected the all-MiniLM-L12-v2 model, a fine-tuned variant of microsoft /MiniLM-L12-H384-uncased. This model has been utilised in a wide range of studies and was extensively trained across a range of datasets, including healthcare-related terminology.

The fine-tuning of this model was conducted on a corpus that combined multiple datasets, exceeding 1 billion sentence pairs, to ensure a diverse exposure to various sentence types and contexts. Among these datasets were TriviaQA, WikiHow, and the SNLI corpus, offering a comprehensive coverage from trivia questions to complex natural language inference tasks.

The training regime for the model involved 100,000 steps on a TPU v3-8, with a batch size of 1024 and a sequence length limited to 128 tokens. The optimisation was carried out using the AdamW optimiser, a variant of Adam explained earlier, with a learning rate of 2e-5, following a learning rate warm-up period of the first 500 steps. This training process, aimed at leveraging the contrastive learning objective to compute cosine similarity among sentence pairs.

A.3 Development Environment

For the entire development of the project, encompassing data preprocessing, feature selection, model training, and evaluation, we employed Jupyter Notebooks due to their simplicity and user-friendly interface for setup and execution. A combination of Google Colab and UCL lab and Blaze machines were selected as the development environment for their capability to accelerate model training, specifically for complex neural network architectures. Their provision of access to Graphics Processing Units (GPUs) made them a valuable asset for handling computationally intensive model training and significantly reduced the time required for experimentation. Blaze was eventually resorted to for more efficient model training which were incapable of running on Colab due to their GPU usage limits and the UCL lab machines which were reset periodically.

Appendix B

Data Analysis Figures

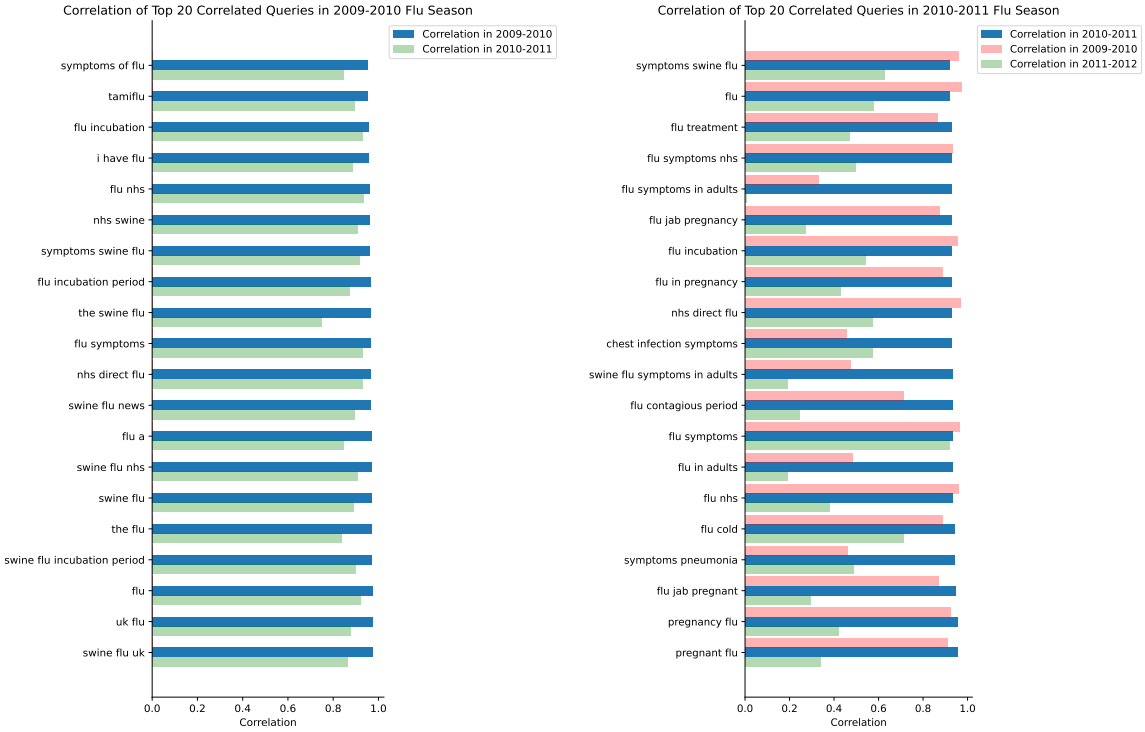

Figure B.1: Correlations of the top 20 correlated web search queries for flu seasons 2009-10 and 20010-11. Correlation values of these queries reported for $y - y + 1$ and adjacent flu seasons. The graph highlights fluctuations in the top 20 correlated queries across flu seasons, focusing particularly on correlations between adjacent seasons.

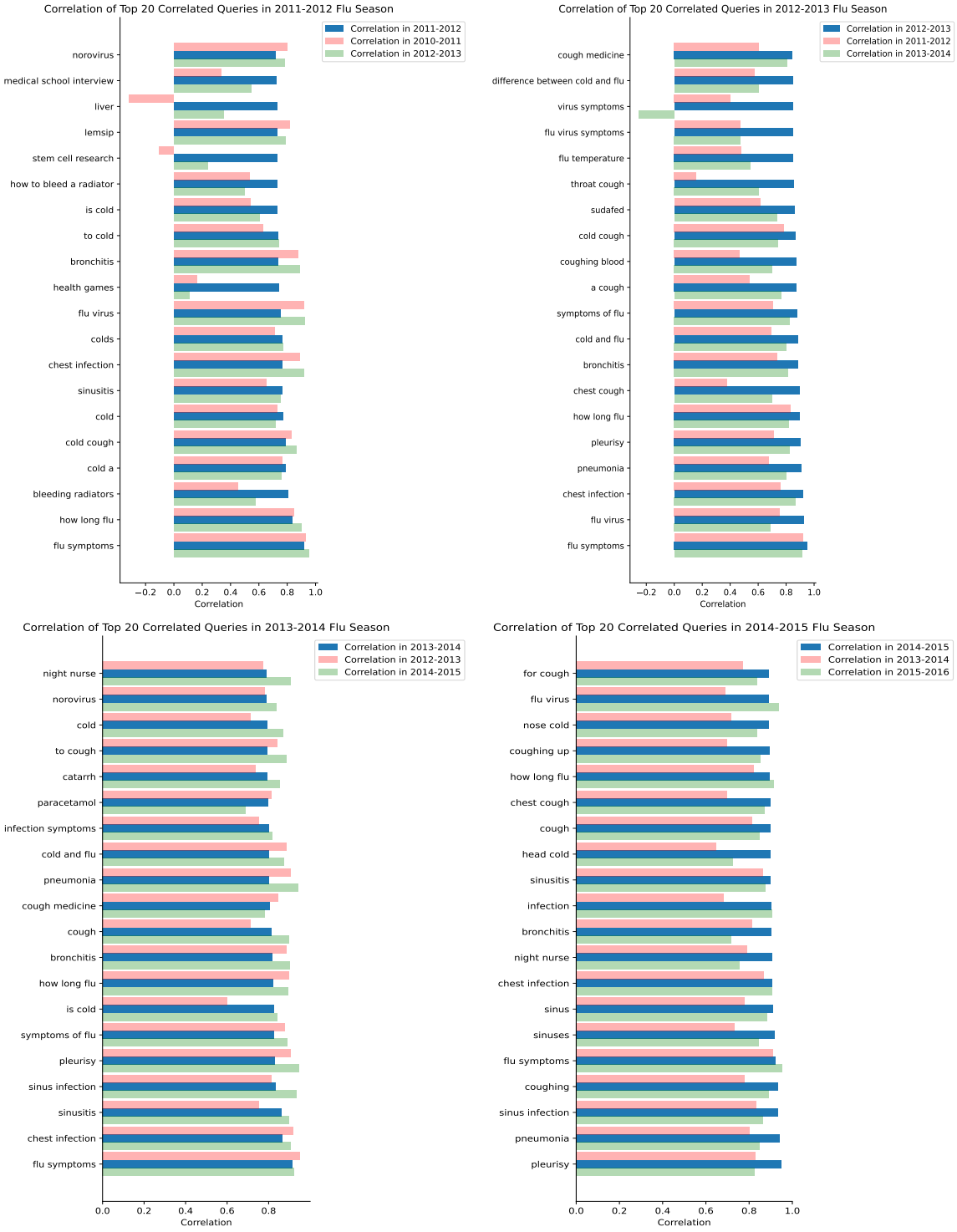

Figure B.2: Correlations of the top 20 correlated web search queries in each flu season from 2011-12 to 20014-15. Correlation values of these queries reported for $y - y + 1$ and adjacent flu seasons. The graph highlights fluctuations in the top 20 correlated queries across flu seasons, focusing particularly on correlations between adjacent seasons.

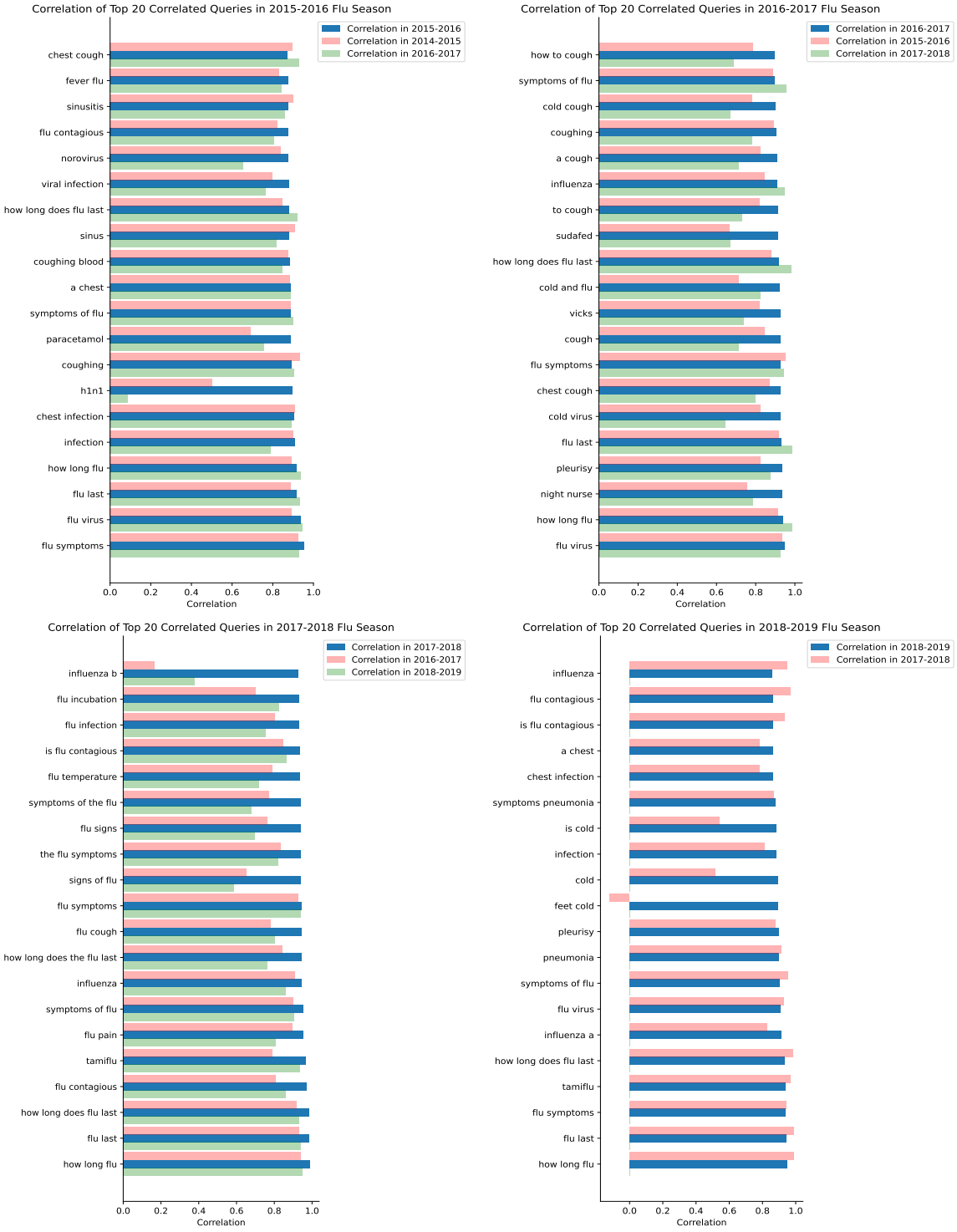

Figure B.3: Correlations of the top 20 correlated web search queries in each flu season from 2015-16 to 20018-19. Correlation values of these queries reported for $y - y + 1$ and adjacent flu seasons. The graph highlights fluctuations in the top 20 correlated queries across flu seasons, focusing particularly on correlations between adjacent seasons.

Appendix C

Project Plan

Project Title:
Nowcasting and Forecasting Influenza-Like Illness (ILI) Rates Using Web Search Queries

C.1 Aims and Objectives

Aims:

- Develop accurate and reliable predictive models for forecasting and nowcasting ILI rates in the UK.
- Contribute to the field of epidemiology by leveraging web search query data for disease rate predictions.

Objectives:

1. Literature Review:
    - Conduct a comprehensive review of existing literature, covering:
        - ILI rate nowcasting, exploring various models and feature selection methods.
        - The application of machine learning in epidemiology, its motivations, and practical applications.
        - The use of Neural Networks in Nowcasting and forecasting methods in general machine learning prediction tasks.
2. Data Pre-processing:
    - Download and preprocess the ILI and web search query datasets to ensure data quality and consistency for machine learning analysis.
        - Interpolate daily ILI rates.
        - Take the intersection of the datasets and drop zero feature (query) columns
        - Clean and smooth the normalised web search query frequencies.
3. Data Exploration:
    - Investigate the relationships between ILI rates and web search query frequencies, focusing on:.

- Identifying seasonal trends in ILI rates.
- Assessing the correlation of individual queries with ILI rates.
- Analysing how query-ILI correlations vary during flu seasons and different periods.
- Cluster queries by similarity and study their correlation with ILI rates

4. Baseline Linear Model for Nowcasting:
   - Download and preprocess the ILI and web search query datasets to ensure data quality and consistency for machine learning analysis.
     - Conduct feature selection using various approaches, such as BERT sentence models, correlation-based and hybrid methods.
     - Train Elastic Net models for nowcasting on each flu season and compare their performance.
5. Neural Network Architectures for Nowcasting:
   - Train different Neural Network architectures, including FFNN and LSTM, to predict ILI rates.
6. Neural Network Architectures for Forecasting:
   - Extend the neural network modelling to forecasting ILI rates, following the approach used for nowcasting.
7. Evaluation and Comparison:
   - Assess model performance, presenting results in tables and graphs.
   - Conduct a detailed comparative analysis between developed models and related work, offering insights into factors influencing model performance.

C.2 Expected Outcomes and Deliverables

- Literature Survey: Provide a comprehensive literature survey summarising relevant research in ILI and flu predictions, epidemiology, public health motivation, and machine learning techniques for nowcasting and forecasting.
- Extensive Data Analysis: Present a thorough analysis of data relationships, including query correlations, seasonality trends in ILI rates, and their variations over time.
- Detailed Experiments: Document a series of comprehensive experiments demonstrating the application of diverse machine learning models and approaches to ILI rate prediction.
- Results Presentation: Explain model performances clearly through tables and graphs for easy interpretation.
- Analysis: Offer an in-depth comparative analysis of experimental findings, providing insights into model behaviour, factors influencing performance, and a comparative assessment against existing research in the field.

C.3 Work Plan

- Project start to end of October (4 weeks):
  - Begin with Objective 1.
  - Complete Objective 2.
  - Make significant progress on Objective 4.
- November to mid-December (6 weeks):
  - Finalise Objective 4.
  - Complete Objective 3.
  - Start on Objective 5.
- Mid-December to mid-February (8 weeks):
  - Complete Objective 5.
  - Complete Objective 6.
- Mid-February to end of March (6 weeks):
  - Complete the project report.

C.4 Ethics Review

This project does not require an ethics review.

Appendix D

Interim Report

Initial/Current Project Title:
Nowcasting and Forecasting Influenza-Like Illness (ILI) Rates Using Web Search Queries

Supervisor Name: Vasileios Lampos

D.1 Progress So Far

The project began with an extensive review of literature demonstrating the application of Online User-Generated Content (UGC) across various research fields such as politics, finance, commerce, and health. The focus then shifted towards how UGC is utilised in epidemiological surveillance, starting with a broad examination of UGC methods in public health, particularly in epidemiology. This included a detailed analysis of the first real-time system using UGC for flu surveillance, Google Flu Trends (GFT). The literature review progressed to scrutinise a series of papers evaluating GFT's performance and adaptability. Subsequently, the review concentrated on research employing web search queries for influenza surveillance. Noteworthy among these was the UK-based web query influenza surveillance study, 'Flu Detector'. Additionally, papers like 'Advances in Nowcasting Influenza-Like Illness Rates Using Search Query Logs' and 'Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance' were reviewed. The methodologies in these papers informed the initial experiments of our study.

Following the literature review, I proceeded to address the primary task, beginning with activities related to the dataset. My dataset comprised weekly Influenza-Like Illness (ILI) rates and daily normalised frequencies of web search queries. The initial phase involved preprocessing the dataset to address discrepancies in data granularity, achieved by interpolating weekly rates into daily values. This interpolation assumed each weekly ILI rate to represent the ILI rate for the corresponding Thursday, with linear interpolation applied to the remaining days of the week. The datasets were then aligned by synchronising dates across both sets. As my study aimed to predict individual flu seasons using data from previous seasons, all query features with zero normalised frequencies were excluded for the initial training period of each flu season. A 14-day window was subsequently used to smooth the normalised query frequencies in the data.

I then commenced an initial statistical analysis of the data. This entailed plotting daily ILI rates across various flu seasons to visually identify trends and seasonal fluctuations, as depicted in Figure 7.1. Figure 7.2 further illustrated the daily mean ILI rates across all flu seasons, presenting

an average trend of ILI rates throughout the year. The correlation between query features and ILI rates was also examined, particularly how this correlation varied across different seasons. This analysis involved identifying the top 20 correlated queries for each season and assessing their correlation in subsequent seasons. Numerical values were assigned to each query, reflecting their frequency in the top 20 lists across seasons. A subset of the results from this analysis is presented in Figure 7.3.

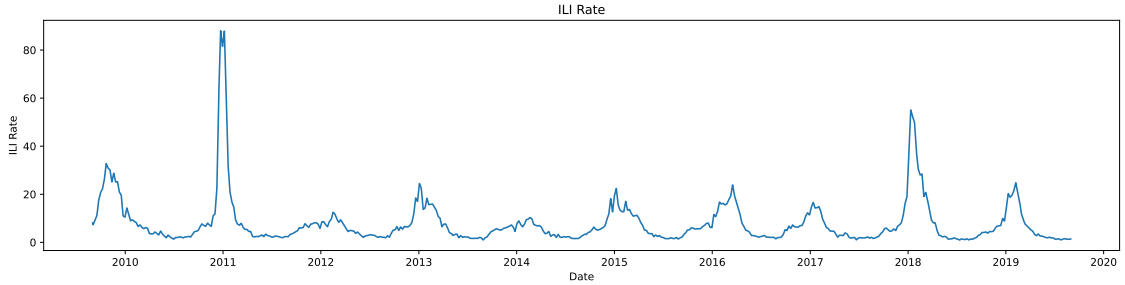

Figure D.1: ILI Rate Across Flu Seasons

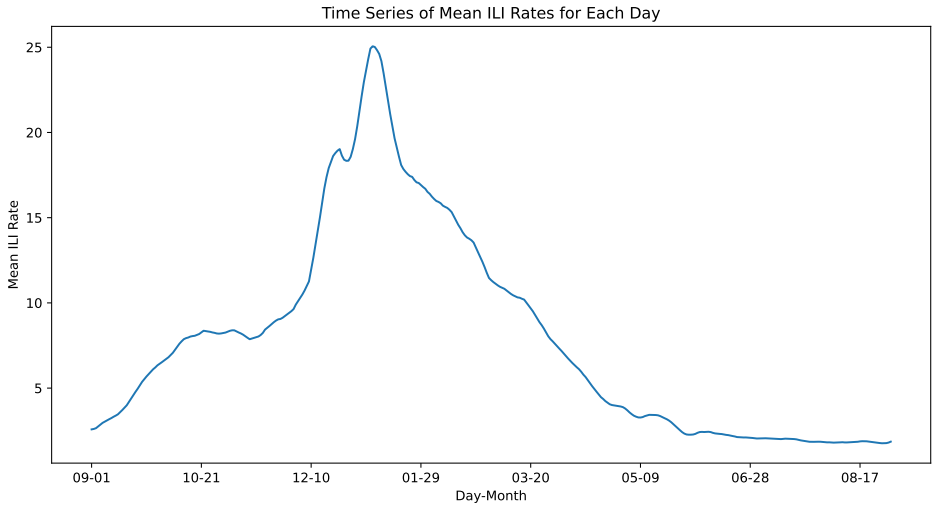

Figure D.2: Mean Daily ILI Rate Across Flu Seasons

Following the initial exploratory phase, I began training a baseline Elastic Net model for now-casting Influenza-Like Illness (ILI) rates.

The Elastic Net Formula is as follows:

$$\frac{1}{2n}|\mathbf{Y} - \mathbf{X}\mathbf{W}|_2^2 + \lambda_1|\mathbf{W}|_1 + \frac{\lambda_2}{2}|\mathbf{W}|_2^2$$

Predictions were made for five distinct flu seasons, from 2014-15 to 2018-19. For each season Y-Y+1, the test data consisted of information from 01-09-Y to 31-08-Y+1. Training data included all previous seasons from 01-09-2009 to 31-08-Y. Both training and test data were standardised using z-scoring relative to the training data. The regularisers λ_1 and λ_2, were cross-validated using a manual 3-fold cross-validation approach. This involved dividing the training data into K folds, each representing data for a specific flu season, and validating over the three most recent seasons. The hyperparameter grid was designed to explore a broad spectrum of sparsity and density within

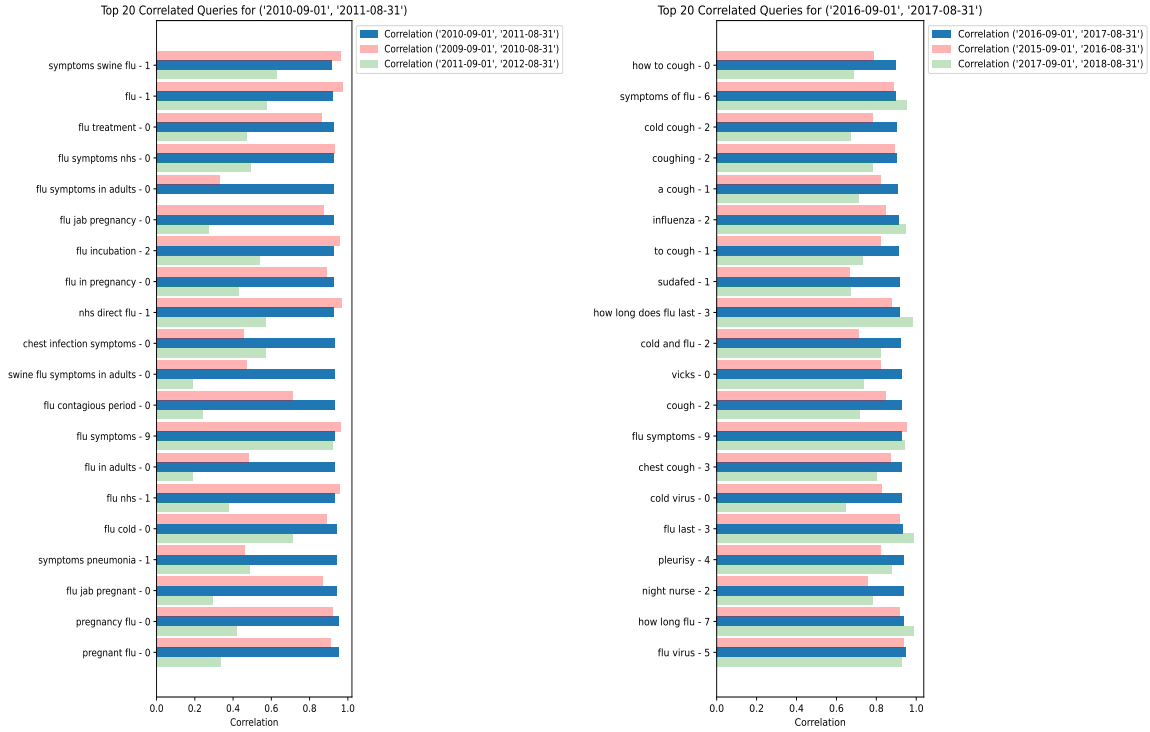

Figure D.3: Top 20 Correlated Queries By Flu Season

the feature set, with the best hyperparameter selected based on achieving the lowest average MAPE across all validation folds.

Given the extensive feature space, I explored various feature selection methods. Initially, a pre-trained BERT Sentence model was employed to analyse textual semantics during feature selection. After obtaining sentence embeddings for each query, cosine similarities with three base queries ('flu', 'flu NHS', and 'influenza') were calculated. This process aided in selecting a diverse set of base queries indicative of flu, which were then used in a final BERT Sentence model with aggregated cosine similarity scores. In addition, a correlation-based feature selection method was employed, focusing on queries with a Pearson correlation ≥ 0.5 with ILI rates over the past five seasons, as inspired by our earlier statistical analysis. A hybrid approach combined both methods, applying a correlation-based selection with a threshold of 0.3 to the top 1000 cosine similar queries from the BERT model. The Elastic Net model was then trained for each flu season using queries obtained from each feature selection approach. The MAE, MAPE, and Pearson Correlation were recorded, along with the number of non-zero weights selected by the cross-validated models on the test data for each flu season prediction.

The initial training results of the Elastic Net model, as shown in Table 7.1, displayed suboptimal prediction performance. The best performance, with an average MAE of 1.9608, was surprisingly achieved by the BERT Sentence model, outperforming the correlation model. Notable sparsity was observed in the non-zero weights selected across all methods.

To address these issues of sparsity and underperformance, I modified the hyperparameter grid to ensure a minimum selection of 150 features, with results presented in Table 7.2. A denser selection of features significantly improved the performance of all models, particularly the Hybrid and Correlation-based methods. The Hybrid method's performance exceeded that of the BERT Sentence model and closely matched the Correlation-based model. The best performance was

Model	Year	MAE	Pearson Correlation	MAPE	λ_1	λ_2	Min Non-Zero Weights	Max Non-Zero Weights	Non-Zero Weights Chosen
BERT Sentence	2014-2015	1.4253	0.9351	25.6088	0.0770	0.0165	29.0000	389.6667	140
BERT Sentence	2015-2016	1.6236	0.9423	26.2775	0.2380	0.0510	22.3333	390.6667	65
BERT Sentence	2016-2017	2.4381	0.9359	57.5788	0.0532	0.0114	17.6667	392.6667	164
BERT Sentence	2017-2018	2.2136	0.9685	31.1564	0.0007	0.0002	13.0000	390.3333	394
BERT Sentence	2018-2019	2.1034	0.9620	47.5623	0.0014	0.0003	13.6667	391.3333	380
BERT Sentence	Average	1.9608	0.9488	37.6368					
Correlation	2014-2015	1.7263	0.8874	26.2307	0.7000	0.1500	29.3333	225.0000	53
Correlation	2015-2016	1.5003	0.9481	25.7649	0.0189	0.0040	22.6667	231.0000	150
Correlation	2016-2017	2.5994	0.9586	62.3692	0.6090	0.1305	16.3333	442.0000	43
Correlation	2017-2018	2.6944	0.9737	50.8753	0.7000	0.1500	11.6667	456.3333	39
Correlation	2018-2019	1.4993	0.9639	49.5317	0.0378	0.0081	13.6667	533.6667	202
Correlation	Average	2.0039	0.9463	42.9544					
Hybrid	2014-2015	1.4081	0.9337	23.4002	0.0273	0.0059	29.3333	208.6667	125
Hybrid	2015-2016	1.6527	0.9412	28.2896	0.0210	0.0045	22.6667	204.3333	137
Hybrid	2016-2017	2.3386	0.9561	59.3719	0.5950	0.1275	16.3333	248.6667	39
Hybrid	2017-2018	2.8131	0.9698	53.6509	0.7000	0.1500	11.6667	263.3333	36
Hybrid	2018-2019	2.6818	0.9419	83.2603	0.0700	0.0150	13.6667	313.0000	107
Hybrid	Average	2.1789	0.9485	49.5946					

Table D.1: Elastic Net Performance With Different Feature Selection Methods

achieved by the Correlation model with an average MAE of 1.7878. However, the Hybrid model exhibited a lower standard deviation in its MAE across the flu seasons while also achieving the lowest average MAPE.

Model	Year	MAE	Pearson Correlation	MAPE	λ_1	λ_2	Min Non-Zero Weights	Max Non-Zero Weights	Non-Zero Weights Chosen
BERT Sentence	2014-2015	1.4138	0.9363	25.5932	0.0693	0.0149	33.0000	391.3333	148
BERT Sentence	2014-2015	1.4253	0.9351	25.6088	0.0770	0.0165	29.0000	389.6667	140.0000
BERT Sentence	2015-2016	1.4392	0.9484	23.4459	0.0700	0.0150	22.3333	390.6667	149.0000
BERT Sentence	2016-2017	2.4381	0.9359	57.5788	0.0532	0.0114	17.6667	392.6667	164.0000
BERT Sentence	2017-2018	2.2136	0.9685	31.1564	0.0007	0.0001	13.0000	390.3333	394.0000
BERT Sentence	2018-2019	2.1034	0.9621	47.5623	0.0014	0.0003	13.6667	391.3333	380.0000
BERT Sentence	Average	1.9239 +- 0.4157	0.9500	37.0704					
Correlation	2014-2015	1.2695	0.9443	21.4624	0.0060	0.0013	29.3333	225.0000	184.0
Correlation	2015-2016	1.5003	0.9481	25.7649	0.0189	0.0040	22.6667	231.0000	150.0
Correlation	2016-2017	2.4838	0.9576	56.9895	0.0546	0.0117	16.3333	442.0000	142.0
Correlation	2017-2018	2.1862	0.9773	46.4719	0.0007	0.0001	11.6667	456.3333	450.0
Correlation	2018-2019	1.4993	0.9639	49.5317	0.0378	0.0081	13.6667	533.6667	202.0
Correlation	Average	1.7878 +- 0.4643	0.9583	40.0441					
Hybrid	2014-2015	1.2933	0.9447	23.6200	0.0336	0.0072	29.3333	388.0000	209.0
Hybrid	2015-2016	1.4460	0.9460	23.6640	0.0364	0.0078	22.6667	392.3333	190.0
Hybrid	2016-2017	2.013	0.9359	45.8076	0.0497	0.0106	17.6667	388.6667	174.0
Hybrid	2017-2018	2.2870	0.9765	39.1566	0.0021	0.0004	13.0000	389.6667	369.0
Hybrid	2018-2019	1.9559	0.9594	47.8232	0.0021	0.0004	13.6667	389.3333	366.0
Hybrid	Average	1.7990 +- 0.3712	0.9525	36.0143					

Table D.2: Elastic Net Performance With Different Feature Selection Methods ($\geq$ 150 non-zero weights enforced in hyperparamter cross validation)

Subsequently, I advanced to training various neural network architectures, utilising the hybrid feature selection method for the same prediction task. Instead of normalising the data via Z-scoring as before, I opted for the min-max normalisation technique. Initially, I started with a feed-forward neural network, incorporating early stopping based on a validation set comprising the most recent flu season in the training data. I progressively increased the network's complexity by adding additional layers and adjusting the number of nodes in each layer. Parameters such as learning rates and early stopping patience were varied, while the batch size was consistently maintained at 14. The results obtained using an architecture with five hidden layers—comprising 200, 100, 50, 25, and 12 nodes respectively—are presented in Table 7.3. The optimal performance, with an average MAE of 1.354, was achieved using a learning rate of 0.001 and a patience setting of 20.

Year	MAE	Pearson Correlation	MAPE	Epochs
2014-2015	1.2596	0.9393	18.5853	158.0000
2015-2016	1.0016	0.9598	15.1724	57.0000
2016-2017	0.9604	0.9681	19.0539	61.0000
2017-2018	2.4450	0.9772	46.7831	54.0000
2018-2019	1.1036	0.9618	19.5058	56.0000
Average	1.3540494 +- 0.5551583	0.9612	23.8201	

Table D.3: Feed Forward Neural Network Performance

Additionally, I explored the use of lagged features, expanding the feature set by approximately 400 times 'k', where 'k' represents the number of lagged features (i.e., the daily normalised fre-

quency of queries up to day t-k-1). For instance, with two lagged features, the dataset included features for day 't' and day 't-1'. Following the same methodology of experimenting with various layer depths and node sizes, as well as different numbers of lagged features such as 7, 14, 21, and 28, the best performance recorded was a marginally improved average MAE of 1.339. This was attained with an architecture of six hidden layers, containing 650, 300, 150, 75, 32, and 16 nodes respectively, a learning rate of 0.0001, and a patience setting of 50. These results are detailed in Table 7.4.

Year	MAE	Pearson Correlation	MAPE	Epochs
2014-2015	1.1696	0.9561	20.1643	70.0000
2015-2016	1.2311	0.9613	18.2686	134.0000
2016-2017	0.8903	0.9613	18.0695	80.0000
2017-2018	1.7823	0.9873	33.3988	75.0000
2018-2019	1.6203	0.9329	25.6652	102.0000
Average	1.3387047 +- 0.32165855	0.9598	23.1133	

Table D.4: Feed Forward Neural Network With 7 Lagged Features Performance

Graphs depicting the ILI rate predictions for each flu season using both the baseline Elastic Net and the 7-lagged FFNN models are illustrated in Figures 7.4-7.8.

Figure D.4: Elastic Net vs 7-Lagged FFNN Flu Season Predictions - 2014-2015

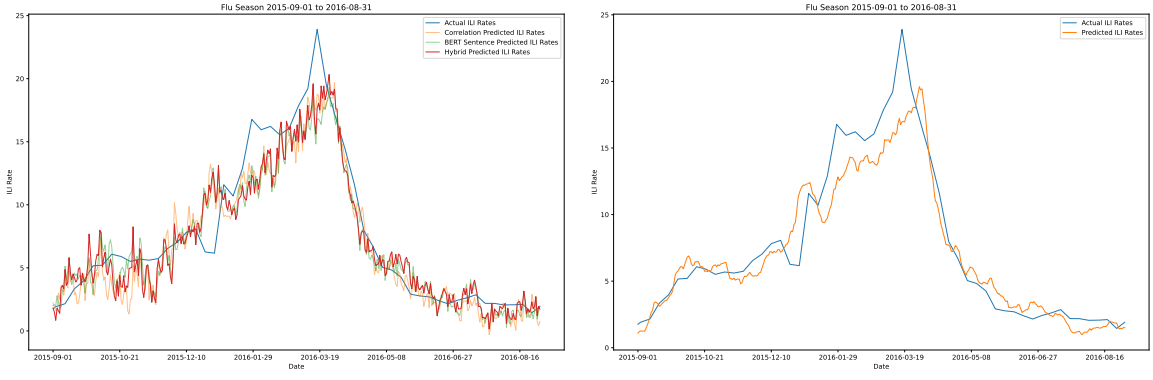

Figure D.5: Elastic Net vs 7-Lagged FFNN Flu Season Predictions - 2015-2016

Currently, my research is focused on investigating multi-output FFNN models for nowcasting ILI rates. This involves training the neural network to predict ILI rates for a week (from day t to t+6) using query frequencies from day t, with the test error being evaluated based on the

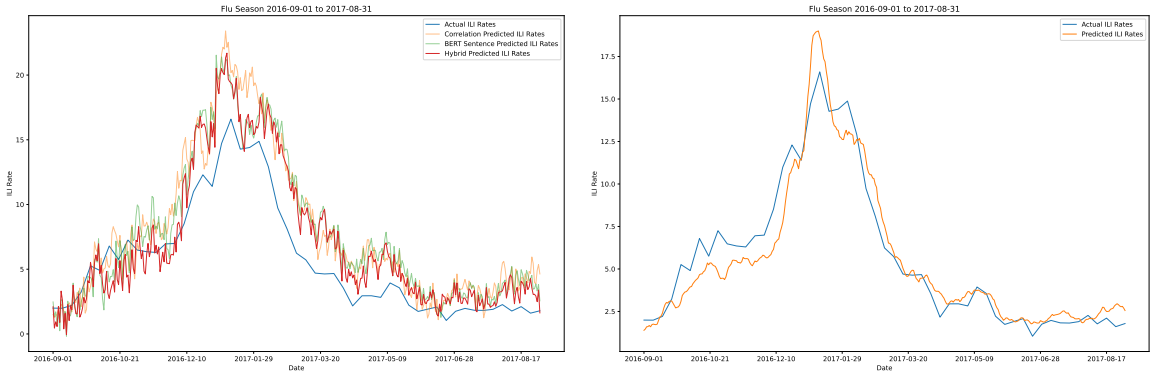

Figure D.6: Elastic Net vs 7-Lagged FFNN Flu Season Predictions - 2016-2017

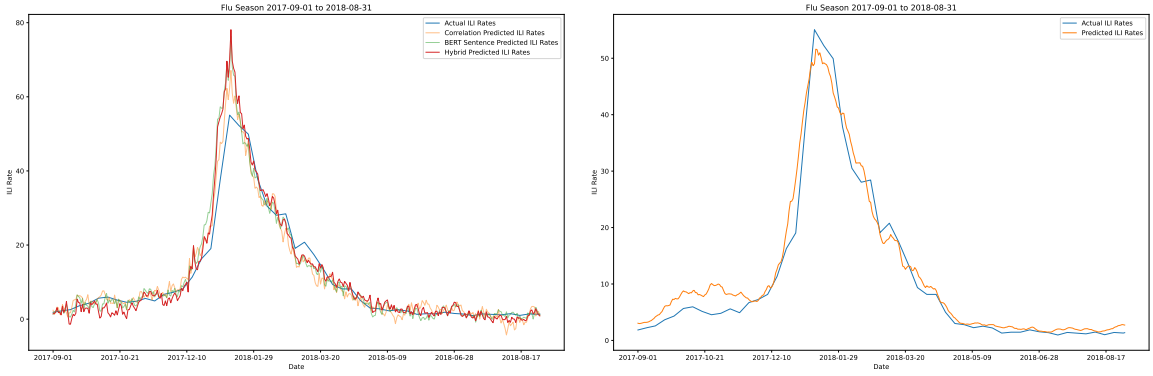

Figure D.7: Elastic Net vs 7-Lagged FFNN Flu Season Predictions - 2017-2018

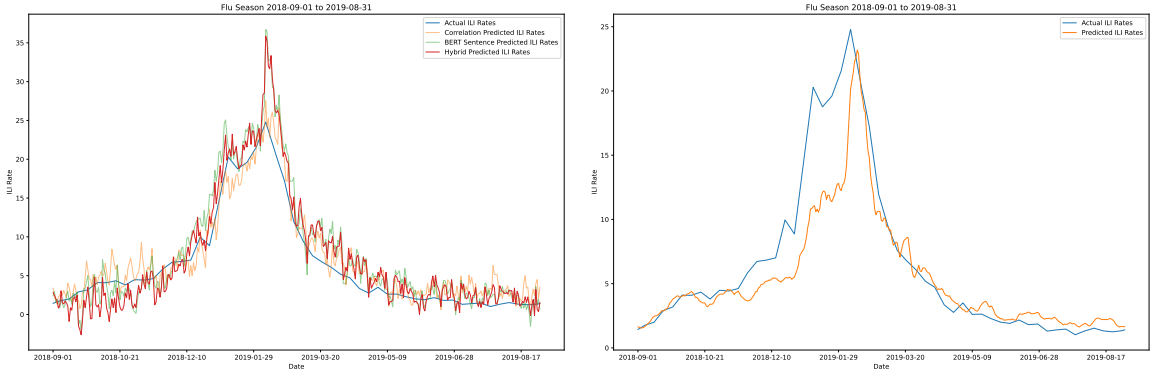

Figure D.8: Elastic Net vs 7-Lagged FFNN Flu Season Predictions - 2018-2019

predictions for day t. This exploration includes variations of hindcasting and forecasting, such as predicting the preceding six days alongside or in place of the following six days. In addition, I have also started exploring LSTM architectures for nowcasting ILI rates.

D.2 Remaining Work

The ongoing experiments are centred around training multi-output and LSTM models for nowcasting purposes. Should time permit, I also plan to explore forecasting ILI rates. This phase will

be followed by comprehensive documentation of the experiments, encompassing both analysis and presentation of the results.

D.3 Work Plan

- Project start to end of October (4 weeks):
  - Begin with Objective 1. *Completed*
  - Complete Objective 2. *Completed*
  - Make significant progress on Objective 4. *Completed*
- November to mid-December (6 weeks):
  - Finalise Objective 4. *Completed*
  - Complete Objective 3. *Completed*
  - Start on Objective 5. *Completed*
- Mid-December to mid-February (8 weeks):
  - Complete Objective 5. *In Progress*
  - Complete Objective 6. *Not Started*
- Mid-February to end of March (6 weeks):
  - Complete the project report. *Not Started*

Appendix E

Code

The source code for this project can be found at the following Github Repository.