Article

Enhancing Influenza Epidemics Forecasting Accuracy in China with Both Official and Unofficial Online News Articles, 2019–2020

Jingwei Li [1,2], Choon-Ling Sia [2], Zhuo Chen [3,4] and Wei Huang [5,*]

1   School of Management, Xi'an Jiaotong University, Xi'an 710049, China; li.jing.wei123@stu.xjtu.edu.cn
2   Department of Information Systems, City University of Hong Kong, Hong Kong 999077, China;
    iscl@cityu.edu.hk
3   College of Public Health, University of Georgia, Athens, GA 30602, USA; zchen1@uga.edu
4   School of Economics, University of Nottingham Ningbo China, Ningbo 315000, China
5   College of Business, Southern University of Science and Technology, Shenzhen 518000, China
*   Correspondence: huangw7@sustech.edu.cn

Abstract: Real-time online data sources have contributed to timely and accurate forecasting of influenza activities while also suffered from instability and linguistic noise. Few previous studies have focused on unofficial online news articles, which are abundant in their numbers, rich in information, and relatively low in noise. This study examined whether monitoring both official and unofficial online news articles can improve influenza activity forecasting accuracy during influenza outbreaks. Data were retrieved from a Chinese commercial online platform and the website of the Chinese National Influenza Center. We modeled weekly fractions of influenza-related online news articles and compared them against weekly influenza-like illness (ILI) rates using autoregression analyses. We retrieved 153,958,695 and 149,822,871 online news articles focusing on the south and north of mainland China separately from 6 October 2019 to 17 May 2020. Our model based on online news articles could significantly improve the forecasting accuracy, compared to other influenza surveillance models based on historical ILI rates ($p = 0.002$ in the south; $p = 0.000$ in the north) or adding microblog data as an exogenous input ($p = 0.029$ in the south; $p = 0.000$ in the north). Our finding also showed that influenza forecasting based on online news articles could be 1–2 weeks ahead of official ILI surveillance reports. The results revealed that monitoring online news articles could supplement traditional influenza surveillance systems, improve resource allocation, and offer models for surveillance of other emerging diseases.

Keywords: digital disease detection; seasonal influenza surveillance; online news; autoregressive exogenous model

1. Introduction

Influenza is a severe public health concern internationally. Every year, influenza is estimated to cause 1 billion cases, 3–5 million severe cases, and 290,000–650,000 influenza-related respiratory deaths worldwide [1,2] and cause an estimated 88,100 influenza-related excess respiratory deaths in China [3]. Through monitoring and evaluating influenza transmission, real-time influenza surveillance can contribute to early response and preparation for influenza epidemics [4]. The accuracy of real-time influenza activity forecasting strongly influences the effectiveness of the response and preparation [5]. The commonly used and well-established method to track influenza activity is through influenza-like illness (ILI) reporting from sentinel hospitals all over the country [4,6]. However, the official ILI reports suffer from a delay of one to three weeks due to processing ILI-related information from sentinel hospitals [5,7]. Traditional influenza epidemics forecasting methods based on historical ILI reports cannot capture the influenza activity in a timely manner and sometimes

Int. J. Environ. Res. Public Health **2021**, *18*, 6591

2 of 13

suffer from overfitting, especially during influenza outbreaks [5,8]. Thus, it is essential to seek real-time novel data sources to enhance influenza activity forecasting accuracy.

Public online data sources, including search queries, social media, and online news articles, accumulate a large volume of data every day and have shown a great potential to supplement traditional infection disease surveillance [9–11]. Specifically, public online data sources could improve the timelines and forecasting accuracy for infectious disease surveillance, compared with traditional official reporting results [12,13]. Public online data contains a large amount of unstructured data and a combination of methods has been applied to support the analysis. These methods include text mining, machine learning, and regression analysis [8,10,14]. Much influential infection disease surveillance has benefitted from using public online data. For example, Wilson and Brownstein (2009) used online articles to support timely surveillance of Listeria [15], He et al. (2018) made use of the Baidu search index to predict the incidence of HIV in China [16], and Liu et al. (2020) incorporated both search query data and online news articles data to provide real-time forecasting of COVID-19 outbreaks in different Chinese provinces [17].

Influenza is one of the most impactful infection diseases worldwide [1]. Digital influenza surveillance, which incorporates various kinds of real-time influenza-related information from the Internet, has been a valuable complement to the traditional influenza surveillance method based on historical ILI reports [6]. Previous studies made use of influenza-related online information, including the fraction of Google search queries [5,7], the number of tweets or microblogs [8,18], Wikipedia article views [19], Flu Near You [20], and self-reporting digital applications [6]. Among them, search queries and social media like Twitter received the most attention and accounted for 95.1% of internet-based health surveillance studies from a recent literature review [21]. However, methods based on Google search queries suffered an instability problem during the influenza outbreak [22], while methods based on tweets suffered from noise with linguistic errors and idiosyncratic style [8,23]. Online news articles are richer in information and written in a more formal language, compared with search queries and tweets. The number of online news articles has increased dramatically recently for the incorporation of new, emerging unofficial user-generated news into traditionally officially-certificated organization-generated news [24]. However, the value of online news articles in disease surveillance has not been fully studied. Previous studies only focused on the news articles that mention the "outbreak" of disease like Healthmap [25,26] or used the news articles from official news websites only [27,28].

In this study, we propose a novel approach to estimate influenza activity in both the south and north of mainland China by adding information extracted from both official and unofficial online news articles as a new exogenous input into the traditional auto-regression model. The provinces in south of mainland China include Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangdong, Guangxi, Hainan, Chongqing, Sichuan, Guizhou, and Yunnan, and provinces in the north of mainland China include Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia, Liaoning, Jilin, Heilongjiang, Shandong, Henan, Tibet, Shaanxi, Gansu, Qinghai, Ningxia, and Xinjiang. Our study used data from hundreds of millions of influenza-related online news articles in Chinese media. It tested whether incorporating online news article data can improve forecasting accuracy during influenza outbreaks. Our proposed model was compared with baseline models based on other online data sources.

The rest of this paper is organized as follows. First, we describe datasets used in this study, including influenza-related online news articles and influenza-related microblogs from a Chinese commercial online platform and influenza-like illness rates from the Chinese National Influenza Center (CNIC). A data mining method to extract influenza-related online news articles is also used in this part. We then present the descriptive analysis for the data we retrieved in the previous part. The fractions of influenza-related online news articles and microblogs are compared with the official ILI rates to characterize their temporal distribution during the seasonal influenza epidemics from 2019 to 2020. Further, autoregression analysis based on different online data sources is conducted to compare

the influenza epidemics forecasting accuracy between them over time. Moreover, paired t-tests are used to determine whether there is a statistically significant improvement in the forecasting accuracy comparing our proposed model to other baseline models.

2. Materials and Methods

2.1. Data Collection and Processing

Social media data used in this study, including online news articles and microblogs, are retrieved from the Sina Network Opinion Surveillance System (SNOSS) (https://www.yqt365.com, accessed on 15 October 2020). SNOSS is a commercially available online platform that continuously crawls all kinds of public online data in China. The platform crawls more than 140 million pieces of data every day, consisting of microblogs (Chinese Twitter), online news articles, announcements from Chinese Government websites, etc. Online news articles are retrieved from officially certified sources, including People's Daily, and unofficial sources, e.g., user-generated news in Wechat Official Account [24]. Data in SNOSS contains tags for time and location, representing when and where the data are posted. However, different from microblogs, where users usually post about events near themselves [29], online news articles can post news about incidents in other places. To extract online news articles that covered certain localities, we used an additional location filtering strategy. Specifically, we identified a news article covering a certain location when the place's name appeared to be the highest frequency of occurrence in the news article compared with other places' names. Our study used the 33 provinces' names to extract news articles covering the south and north of mainland China separately. The primary aim of SNOSS is to monitor real-time hot topics online, including breaking news and public concerns, which also shows a great potential to improve influenza surveillance through real-time online data [7].

2.1.1. Influenza-Related Online News Articles

We collected influenza-related online news articles from SNOSS from 6 October (week 40) 2019 to 17 May (week 20) 2020. We collected data from this timeframe based on World Health Organization's guidance. The guidance stated that temperate climate zones in the northern hemisphere (including mainland China) had clear seasonality and a well-defined flu season, which started from week 40 of one year to week 20 of the following year [30]. This well-defined flu season timeframe is used as the influenza surveillance period for other northern hemisphere temperate countries or organizations, including America [31], England [32], and the European Union [33], and this timeframe is also used by related influenza forecasting research [5,7].

We extracted online news articles that focus on influenza by using a combination of two sets of keywords [27]. One set of keywords includes 'flu' and 'influenza', and the other set contains keywords mostly related to different aspects of the flu. To identify keywords from online news articles that are most relevant to flu, we first used the bag-of-words method by treating all the online news articles with 'flu/influenza' as a document and identified 200 keywords with the highest frequency. We then used the TF-IDF method [34] to identify keywords that are low in frequency but represent a unique aspect of influenza. Developed by Salton and Buckley [35], the TF-IDF method is able to calculate the importance of a particular word in a document and provide more accurate representation and clustering results (see Supplementary Materials, Table S1, for detailed formula). Specifically, we built a term-document matrix based on 1000 randomly selected influenza-related articles and identified the top 100 terms (keywords) with the highest correlation with the word 'influenza' or 'flu'. Finally, we invited a medical doctor to assist in selecting exact influenza-related keywords from the pool of words identified in the previous two steps. The glossary of influenza (flu) terms from the CDC (https://www.cdc.gov/flu/about/glossary.htm, accessed on 20 October 2020) is also considered in the final selection process. Finally, we identified 91 keywords mostly related to different aspects of influenza from the text mining process (see Supplementary Materials, Table S2).

To further filter out the noise in online news articles, we adopted two information filtering methods, which exclude news articles with specific keywords and exclude similar news articles with an 85% repetition rate within 35 days. Following Kim and Ahn [27], we filtered out news articles containing the keyword 'avian influenza', which has a small number of human contact cases and has not observed sustained human-to-human transmission [36]. We also filtered out news articles containing the keyword 'novel coronavirus', which could refer to a new emerging coronavirus-related disease like SARS or COVID-19. Then, we filtered out the republished online news articles by excluding similar articles with an 85% repetition rate within 35 days. There is a tendency to republish hot news articles to attract public attention. Filtering out republished online news articles can reduce this noise, which has caused estimation instability in Google Flu Trend [22].

Finally, we collected the number of weekly influenza-related online news articles and overall crawled online news articles from the south and north of mainland China separately. The fraction of weekly influenza-related online news articles compared to overall crawled online news articles was computed for the south and north mainland China separately and used as an independent variable in the models.

2.1.2. Influenza-Related Microblogs

We also collected microblogs, which have been widely used in influenza surveillance with social media [37–39], for comparison. We followed the semantic filtering method proposed by Doan et al. [39]. We first extracted the microblogs containing at least one of 37 influenza-related symptom keywords or the keywords of 'flu' or 'influenza' (see Supplementary Materials, Table S3). We then filtered out the retweeted microblogs and microblogs with smiley emoticons, humor features, or URLs (see Supplementary Materials, Table S4). Finally, we collected the weekly number of influenza-related microblogs and overall crawled microblogs from the south and north of mainland China. We collected the data from online news articles and microblogs weekly and from the south and north mainland China separately to match the official influenza-like illness data, which is our baseline and will be introduced in the next section. The fraction of weekly influenza-related microblogs compared to overall crawled microblogs was then computed for the south and north mainland China separately and used as an independent variable in the models.

2.1.3. Influenza-Like Illness Rates in China

Following the period of the online data sources we gathered, we collected weekly Influenza-like illness (ILI) rates for both the south and north of mainland China separately from the CNIC from 6 October 2019 to 17 May 2020. Based on guidelines from CNIC (http://ivdc.chinacdc.cn/cnic/, accessed on 15 October 2020), ILI case counts are defined as having a temperature of 38 °C or greater and having symptoms of cough or sore throat. Sentinel hospitals upload ILI case counts and total physician visits data to CNIC's system on Monday of the following week. The CNIC would release the aggregated data at the end of next week. Due to the vast difference in climate and space, the CNIC reports the ILI rates for the south and north of mainland China separately. The data from the CNIC only includes seasonal influenza, while avian flu (H7N9) and swine flu (H1N1) are not included.

2.2. Statistical Analysis
2.2.1. Descriptive Analysis

We summarized the number of influenza-related and overall data of online news articles and microblogs. The percentages of official and unofficial online news articles were also presented. We then compared the fraction of online news articles and microblogs with the ILI data to depict the data's tendency and show their potential to provide earlier warnings for influenza outbreaks.

2.2.2. Model Formulation

Similar to previous influenza surveillance research [5,6,8] that incorporates data sources into official ILI rates, we used a model in the form of logistic auto-regression with exogenous inputs [40]. Let $y_t = logit(p_t)$ be the logit-transformed CNIC's ILI rates p_t at week t and x_t be the log-transformed fraction of influenza-related online news articles at week t. Our proposed model incorporated the online news articles as a new exogenous input into the auto-regression model:

$$y_t = \sum_{i=1}^{Lag_{ILI}} a_i y_{t-i} + \sum_{j=0}^{Lag_{News}} b_j x_{t-j} + c + \varepsilon_t, \ \varepsilon_t \ \sim \ N\left(0, \ \sigma^2\right) \tag{1}$$

where b_j quantifies the contribution from the fraction of online news articles at time $t - j$ to predict CNIC's ILI rates at time t, in addition to the historical CNIC's ILI rates' contribution. In this model, c is a constant term, ε_t is a sequence of independent random variables, and Lag_{ILI} and Lag_{News} are the lagged indicators. Lag_{ILI} and Lag_{News} are unknown and need to be estimated to achieve the highest prediction accuracy. Following previous research settings [5,6], we considered three weeks' time lag and varied Lag_{ILI} from 1 to 3 and Lag_{News} from 0 to 3. We then denoted our model as AR(Lag_{ILI}) + News(Lag_{News}), in which Lag_{ILI} and Lag_{News} were the optimal values to maximize prediction accuracy.

2.2.3. Parameters Estimation and Baseline Models Comparison

Retrospective estimations of weekly ILI rates for the south and north of mainland China were produced through our proposed model, from 6 October 2019 to 17 May 2020. By calculating a collection of accuracy measures, we compared our proposed model's estimates with the actual value, i.e., the CNIC's weekly reported ILI rates for the south and north of mainland China, published with a delay of one to two weeks. The accuracy measures include the root-mean-squared error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), correlation with forecasting target, and correlation of increment with forecasting target (accuracy indexes are described in Appendix A). The same accuracy indexes were also calculated for other baseline models we built for comparison. The baseline models include:

1.　Autoregression model based on official ILI rates only [22,41], denoted as AR(lag_{ILI1}),

$$y_t = \sum_{i=1}^{Lag_{ILI1}} a_i y_{t-i} + c + \varepsilon_t, \ \varepsilon_t \ \sim \ N\left(0, \ \sigma^2\right) \tag{2}$$

2.　Autoregression model based on official ILI rates and fraction of flu-related microblogs [8], denoted as AR(lag_{ILI2}) + Mblog(lag_{Mblog2}),

$$y_t = \sum_{i=1}^{Lag_{ILI2}} a_i y_{t-i} + \sum_{h=0}^{Lag_{Mblog2}} d_h z_{t-h} + c + \varepsilon_t, \ \varepsilon_t \ \sim \ N\left(0, \ \sigma^2\right) \tag{3}$$

3.　Autoregression model based on official ILI rates, fraction of flu-related online news articles, and fraction of present week's flu-related microblogs [26,41], denoted as AR(lag_{ILI3}) + News(lag_{News3}) + Mblog(0),

$$y_t = \sum_{i=1}^{Lag_{ILI3}} a_i y_{t-i} + \sum_{j=0}^{Lag_{News3}} b_j x_{t-j} + d_t z_t + c + \varepsilon_t, \ \varepsilon_t \ \sim \ N\left(0, \ \sigma^2\right) \tag{4}$$

where z_t is the log-transformed fraction of influenza-related microblogs at week t.

All the time lags represent the number of previous weeks' data that we considered in the model. Mblog(0) means that only the present week's microblog was added to our proposed model, as suggested in [25,38]. lag_{ILI1}, lag_{ILI2}, and lag_{ILI3} vary from 1 to 3, while lag_{Mblog2} and lag_{News3} range from 0 to 3. The values of these different time lags were determined by choosing the value leading to the lowest RMSE for each model [8,42] (see Supplementary Materials, Table S5).

We used the repeated k-fold cross-validation approach [43] to validate our proposed model and all the baseline models for fair comparisons. In k-fold cross-validation, the dataset is divided into k (approximately) equally sized subsets. In each interaction, one subset is used as the validation data, and the remaining (k − 1) subsets are used as training data. The process is repeated k times until each subset is treated as validation data once. We then repeated the cross-validation procedure multiple times by splitting the data into k different subsets. In this study, we set k = 10 and repeated the cross-validation procedure ten times, as suggested in [26]. We tested the hypothesis at a significance level of 0.05 and conducted all the analyses with the R version 4.0.2 statistical software packages DAAG [44] version 1.24 and caret [45] version 6.0-86.

3. Results

Overall, we retrieved 153,958,695 online news articles focusing on the south of mainland China and 149,822,871 online news articles focusing on the north of mainland China, from 6 October 2019 to 17 May 2020, from the SNOSS. In the south, 10,851,588 online news articles were from official sources, which accounted for 7.0% of the overall online news articles. In the north, 10,556,304 online news articles were from officially certificated sources, accounting for 7.0% of the overall online news articles. SNOSS also averagely traced 54,572,133 microblogs from the south and 72,182,715 microblogs from the north. The weekly time series of CNIC's ILI rates, the fraction of influenza-related online news articles, and the fraction of influenza-related microblogs are displayed in Figures 1 and 2. In both the south and north of mainland China, the highest official ILI rate occurred on 5 February 2020, while the highest fractions for influenza-related online news articles and microblogs occurred on 26 January 2020, which was one week earlier compared to the official report.
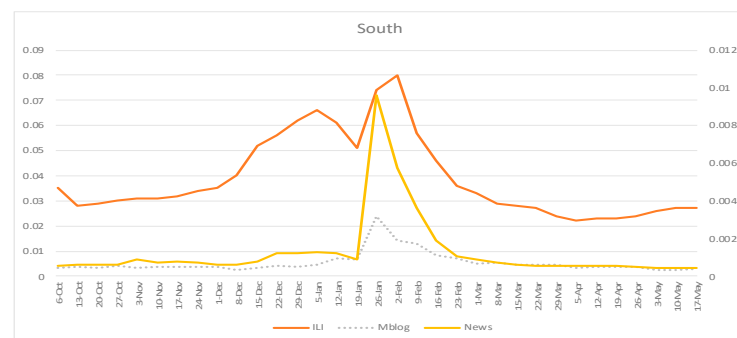

Figure 1. Weekly time series of ILI rates from the Chinese National Influenza Center (CNIC), fraction of influenza-related online news articles, and fraction of influenza-related microblogs: South of mainland China, 6 October 2019–17 May 2020. ILI = influenza-like illness rates from CNIC; Mblog = fraction of influenza-related microblogs; News = fraction of influenza-related online news articles.
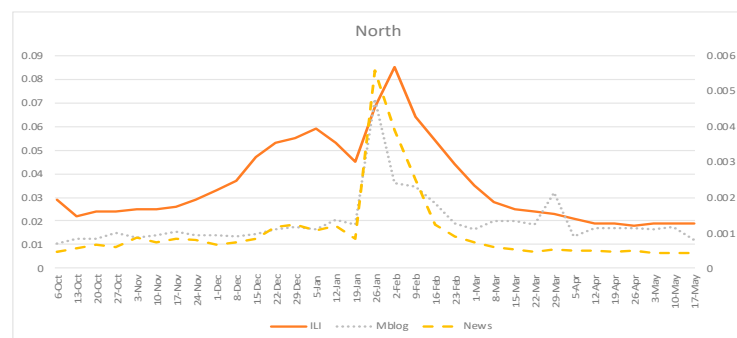

Figure 2. Weekly time series of CNIC's ILI rates, fraction of influenza-related online news articles, and fraction of influenza-related microblogs: north of mainland China, 6 October 2019–17 May 2020. ILI = influenza-like illness rates from CNIC; Mblog = fraction of influenza-related microblogs; News = fraction of influenza-related online news articles.

In Tables 1 and 2, the optimal lags for different data sources, leading to the lowest RMSE for each model, are shown in the parentheses. Boldface highlights the best performance for each index in south and north mainland China. The paired *t*-test results, comparing our proposed model with baseline models, are shown in the last column. The *t* value is present in the last column with the related *p*-value in the parentheses below it. *** means significant at 1% level, ** means significant at 5% level, and * means significant at 10% level.

Table 1. Comparison of different models for the forecasting of influenza epidemics: south of mainland China.

Model (lag)	RMSE	R^2	MAE	Correlation	Correlation of Increment	*t* (sig.)
AR(1) + News(2)	**0.087**	**0.938**	**0.072**	0.980	0.824	
AR(2)	0.126	0.918	0.098	0.903	0.286	−3.164 *** (0.002)
AR(1) + Mblog(2)	0.150	0.872	0.113	0.946	0.517	−2.212 ** (0.029)
AR(1) + News(2) + Mblog(0)	0.107	0.905	0.083	**0.985**	**0.839**	−1.047 (0.297)

RMSE = root-mean-squared error; R^2 = coefficient of determination; MAE = mean absolute error. Values in bold highlight the best performance for each index in the south of mainland China, *** means significant at 1% level, and ** means significant at 5% level.

Table 2. Comparison of different models for the forecasting of influenza epidemics: north of mainland China.

Model (lag)	RMSE	R^2	MAE	Correlation	Correlation of Increment	*t* (sig.)
AR(1) + News(2)	**0.119**	**0.943**	**0.099**	0.978	**0.834**	
AR(2)	0.151	0.898	0.122	0.910	0.317	−4.196 *** (0.000)
AR(1) + Mblog(2)	0.142	0.908	0.116	0.967	0.703	−5.721 *** (0.000)
AR(1) + News(2) + Mblog(0)	0.129	0.922	0.107	**0.980**	0.820	−1.878 * (0.063)

RMSE = root-mean-squared error; R^2 = coefficient of determination; MAE = mean absolute error. Values in bold highlight the best performance for each index in the north of mainland China, *** means significant at 1% level, and * means significant at 10% level.

Tables 1 and 2 summarize the accuracy indexes for all the forecasting methods in the south and north of mainland China separately, from 6 October 2019 to 17 May 2020. To test whether our proposed model can significantly improve the forecasting accuracy compared to other baseline models, we conducted paired *t*-tests on the RMSE of different models [46]. The paired *t*-test results are also shown in Tables 1 and 2. Plots comparing different models' forecasting results for south and north mainland China are shown in Figures 3 and 4.

Results in Tables 1 and 2 and Figures 3 and 4 show that our proposed model incorporating online news articles data (AR + News) could improve the forecasting accuracy in terms of all accuracy indexes in both the south and north of mainland China. This result outperformed the baseline models using other data sources, including using previous official ILI data (AR) or incorporating microblog data (AR + Mblog). When our proposed model incorporates the present week's microblog data (AR + News + Mblog), the accuracy declined in terms of RMSE, R^2, and MAE, compared to the proposed model (AR + News). In terms of correlations and incremental correlations in the south and north of mainland China, AR + News (*r* = 0.98, incremental *r* = 0.824 in the south; *r* = 0.978, incremental *r* = 0.834 in the north) has similar performance to AR + News + Mblog (*r* = 0.985, incremental *r* = 0.839 in the south; *r* = 0.98, incremental *r* = 0.82 in the north). Close inspection in the south suggests that adding microblogs as exogenous input (AR + Mblog) can decrease the

Int. J. Environ. Res. Public Health **2021**, *18*, 6591

8 of 13

forecasting accuracy compared to the traditional autoregression model (AR) in terms of RMSE, R2, and MAE. Referring to RMSE, the accuracy decrease was statistically significant ($t = -1.762$, $p = 0.081$).
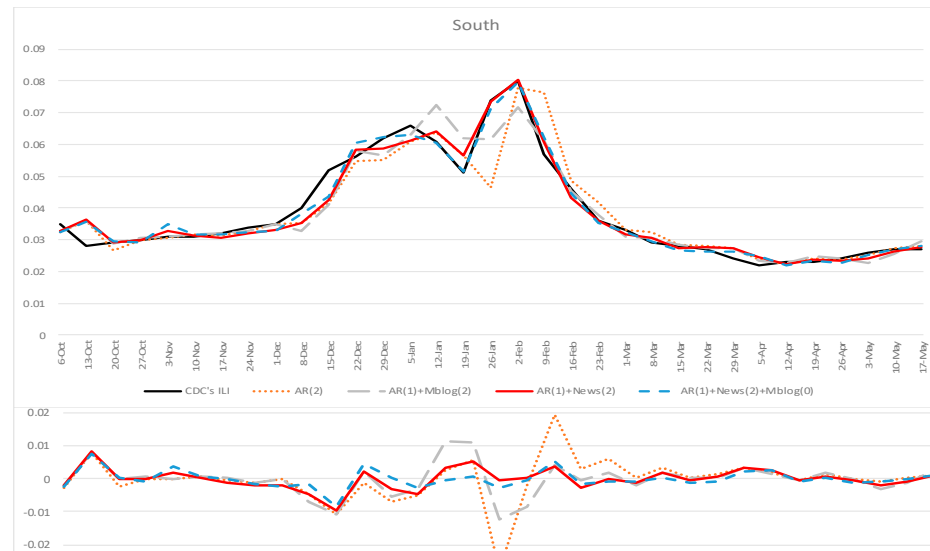

Figure 3. Forecasting results for the south of mainland China, 6 October 2019–17 May 2020. The weekly estimations of AR(2) (dotted orange), AR(1) + Mblog(2) (long dashed gray), AR(1) + News(2) (red), and AR(1) + News(2) + Mblog(1) (dashed blue) models were compared against the CNIC's ILI rates (black). Bottom: The estimation error, defined as the estimated value minus the CNIC's ILI rates.
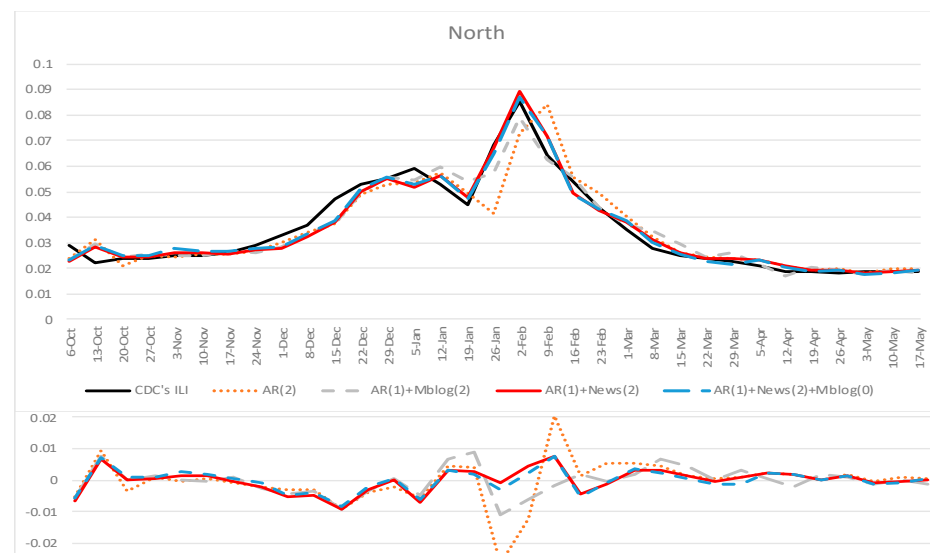

Figure 4. Forecasting results for the north of mainland China, 6 October 2019–17 May 2020. The weekly estimations of AR(2) (dotted orange), AR(1) + Mblog(2) (long dashed gray), AR(1) + News(2) (red), and AR(1) + News(2) + Mblog(1) (dashed blue) models were compared against the CNIC's ILI rates (black). Bottom: The estimation error, defined as the estimated value minus the CNIC's ILI rates.

To assess the statistical significance of the improved prediction power, we conducted a paired *t*-test on RMSE to compare our proposed model's relative efficiency with other baseline models. The results showed that our proposed model (AR + News) could significantly improve the forecasting accuracy, compared with other baseline models. Specifically,

our proposed model outperformed the traditional influenza surveillance model based on historical official ILI rates (AR, $t = -3.164$, $p = 0.002$ in the south; $t = -4.196$, $p = 0.000$ in the north) and also outperformed the forecasting model incorporating microblog data as an exogenous input (AR + Mblog, $t = -2.212$, $p = 0.029$ in the south; $t = -5.721$, $p = 0.000$ in the north). When adding microblog data into our proposed model, the forecasting accuracy in RMSE decreased in both the south and north of mainland China. In the south, the accuracy decrease was not statistically significant ($t = -1.047$, $p = 0.297$), while in the north, the accuracy decrease was statistically significant at the level of 0.1 ($t = -1.878$, $p = 0.063$).

4. Discussion

Our proposed model innovatively incorporated online news articles as a new data source for influenza activity forecasting in both the south and north of mainland China. The results showed that online news articles could significantly improve the forecasting accuracy of influenza activity during outbreaks, compared to other online data sources. This accuracy improvement by mining online news articles could support the public health agency in making better preparations for influenza outbreaks, including improving resource allocation efficiency. This implication is even more important in a geographically vast and densely populated country like China.

Traditional influenza surveillance data from the CNIC is based on ILI reports from sentinel hospitals all over the country, which suffers from a 1–2 week delay for the processing and aggregating all the sentinel hospitals' reports [5,7]. Online news articles, consisting of both user-generated news and organization-generated news [24], can help capture real-time influenza activity. Our results showed that the fraction of influenza-related online news articles peaked one week earlier than the official reports, and the two-week time lag for online news articles led to the highest prediction accuracy in the auto-regression model. This implies that online news articles, as a new online data source, could produce estimation results 1–2 weeks ahead of official ILI surveillance reports.

According to our analysis, about 93.0% of online news articles are from unofficial user-generated news, allowing users to post events happening around themselves in time. This character could help to capture a real-time snapshot of influenza-related activities. A large part of user-generated news comes from Wechat Official Account in China, where users can post news articles freely to attract clicks and generate income through advertisements. Similar characters have been found in other worldwide platforms like LinkedIn Articles, Medium, Instant Articles [47], etc., which shows that online news articles have the potential to enhance forecasting accuracy on influenza epidemics all over the world.

Other online data sources, such as search queries, Twitter, or Microblogs, have shown great potential to supplement influenza detection. However, they also suffered from instability and linguistic noise [8,22], especially during influenza-like illness outbreaks [5,22]. Our study's data is from 6 October 2019 to 17 May 2020, during which a pandemic of COVID-19 also broke out in January 2020 in China. COVID-19 has many similar symptoms compared to influenza, causing additional noise in online data sources. In our study, the forecasting accuracy decreased in the south of mainland China after incorporating microblog data into historical ILI rates. The decrease may be because of the noise produced by COVID-19. Meanwhile, online news articles contribute to a significant improvement in forecasting accuracy in both the south and north of mainland China. This proves that online news articles have a great potential to enhance influenza epidemics forecasting accuracy and robustness, especially during influenza-like illness outbreaks.

Our study also proposed a keyword-based method to extract influenza-related online news articles. This method framework could also help future research monitoring other diseases based on online news articles.

Several limitations and future research directions should be noted. First, only one flu season's data was used in our analysis, and COVID-19 happening during this period could have had a confounding factor during the analysis. We could not extract additional data from the SNOSS, which only provides data within the most recent one and a half

years. Even though the results from the south and north of mainland China both supported our arguments, future studies may test the methods in different years' flu seasons for robustness.

Second, this study used the keywords-based method to extract influenza-related online news articles. We used this method because we wanted to explore the whole dataset of online news articles from the SNOSS. The original raw data could not be downloaded from SNOSS, and many other advanced machine learning methods could not be performed. Future research can try to test the model on actual raw data of official and unofficial online news articles and make use of other machine learning methods. More accurate data filtering processes and deeper content analyses can then be conducted, e.g., using classification algorithms to improve accuracy in filtering results [38] or using topic modeling algorithms to extract richer information from online news articles [48].

Finally, we did not compare our model with methods based on search query data, such as Google Flu Trends [7], in our research. Google Flu Trends stopped providing service on 9 August 2015. Even though people can still get flu-related Google search query data, Google Flu Trends has been shown to suffer from instability during influenza outbreaks [22]. However, future research could still add a comparison of search query data, if possible, to show a clearer picture of the difference between online news articles and other online data sources.

We would also like to discuss the public health implications of our findings in developing countries, including China. About 940 million internet users, 67.0% of the overall population, read and generate information online in China [49]. Among these users, 99.2% use a mobile phone to access the Internet and 30.4% are from rural areas. This character allows internet-based sources such as online news articles to monitor influenza activity in rural areas, where there are no sentinel hospitals to monitor influenza-related activities. Innovative approaches are also needed to monitor vulnerable subgroups who do not even have access to the Internet. Compared with other internet-based sources, online news articles contain richer information, helping extract more detailed disease characteristics and contribute to better preparation. As more and more online news articles are generated, it would be possible to detect various disease outbreaks in more specific areas and subgroups.

5. Conclusions

This study showed that unofficial online news articles accounted for the major part of present online news articles. Online news articles from both official and unofficial sources may serve as an innovative and effective exogenous input to supplement the traditional influenza surveillance model based on historical ILI data. Our proposed autoregression model based on online news articles significantly improved influenza forecasting accuracy during outbreaks, compared to the autoregression model based on historical ILI rates or additional microblogs. Our findings also revealed that online news articles could produce forecasting results 1–2 weeks ahead of official ILI surveillance reports.

Author Contributions: Conceptualization, J.L., C.-L.S. and W.H.; methodology, J.L., C.-L.S., Z.C. and W.H.; software, J.L.; validation, J.L. and W.H.; formal analysis, J.L.; investigation, J.L., C.-L.S., Z.C. and W.H.; resources, J.L. and W.H.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L., C.-L.S., Z.C. and W.H.; supervision, J.L., C.-L.S. and W.H.; project administration, J.L. and W.H.; funding acquisition, C.-L.S. and W.H. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Accuracy Indexes. The RMSE, R2, and MAE of the estimator $\hat{p}$ to the target ILI rate p are defined, respectively, as RMSE $= \left[(1/n) \sum_{t=1}^{n} (\hat{p}_t - p_t)^2 \right]^{1/2}$, R2 $= 1 - \sum_{t=1}^{n} (\hat{p}_t - p_t)^2 / \sum_{t=1}^{n} \left[p_t - (1/n) \sum_{t=1}^{n} p_t \right]^2$, MAE $= (1/n) \sum_{t=1}^{n} |\hat{p}_t - p_t|$. The correlation of estimator $\hat{p}$ to the target ILI rate p is their sample correlation coefficient. The correlation of increment between $\hat{p}$ and p is defined as Corr. of increment $(\hat{p}_t, p_t) = \text{Corr}(\hat{p}_t - \hat{p}_{t-1}, p_t - p_{t-1})$.

References

1. World Health Organization. *Global Influenza Strategy 2019–2030*; World Health Organization: Geneva, Switzerland, 2019.
2. Iuliano, A.D.; Roguski, K.M.; Chang, H.H.; Muscatello, D.J.; Palekar, R.; Tempia, S.; Cohen, C.; Gran, J.M.; Schanzer, D.; Cowling, B.J. Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet* **2018**, *391*, 1285–1300. [CrossRef]
3. Li, L.; Liu, Y.; Wu, P.; Peng, Z.; Wang, X.; Chen, T.; Wong, J.Y.; Yang, J.; Bond, H.S.; Wang, L.; et al. Influenza-associated excess respiratory mortality in China, 2010–2015: A population-based study. *Lancet Public Health* **2019**, *4*, e473–e481. [CrossRef]
4. Yang, X.; Liu, D.; Wei, K.; Liu, X.; Meng, L.; Yu, D.; Li, H.; Li, B.; He, J.; Hu, W. Comparing the similarity and difference of three influenza surveillance systems in China. *Sci. Rep.* **2018**, *8*, 2840. [CrossRef] [PubMed]
5. Yang, S.; Santillana, M.; Kou, S.C. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 14473–14478. [CrossRef] [PubMed]
6. Hswen, Y.; Brownstein, J.S.; Liu, J.; Hawkins, J.B. Use of a digital health application for influenza surveillance in China. *Am. J. Public Health* **2017**, *107*, 1130–1136. [CrossRef]
7. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 1012–1014. [CrossRef] [PubMed]
8. Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; Liu, B. Twitter Improves Seasonal Influenza Prediction. In Proceedings of the Healthinf, Algarve, Portugal, 1–4 February 2012; SciTePress: Setúbal, Portugal, 2012; pp. 61–70.
9. Nsoesie, E.O.; Brownstein, J.S. Computational approaches to influenza surveillance: Beyond timeliness. *Cell Host Microbe* **2015**, *17*, 275–278. [CrossRef]
10. Gupta, A.; Katarya, R. Social media based surveillance systems for healthcare using machine learning: A systematic review. *J. Biomed. Inform.* **2020**, *108*, 103500. [CrossRef]
11. Rees, E.; Ng, V.; Gachon, P.; Mawudeku, A.; McKenney, D.; Pedlar, J.; Yemshanov, D.; Parmely, J.; Knox, J. Early detection and prediction of infectious disease outbreaks. *CCDR* **2019**, *45*, 5. [CrossRef]
12. Yan, S.; Chughtai, A.; Macintyre, C. Utility and potential of rapid epidemic intelligence from internet-based sources. *Int. J. Infect. Dis.* **2017**, *63*, 77–87. [CrossRef]
13. Bernardo, T.M.; Rajic, A.; Young, I.; Robiadek, K.; Pham, M.T.; Funk, J.A. Scoping review on search queries and social media for disease surveillance: A chronology of innovation. *J. Med. Internet Res.* **2013**, *15*, e147. [CrossRef]
14. Allam, Z.; Dey, G.; Jones, D.S. Artificial intelligence (AI) provided early detection of the coronavirus (COVID-19) in China and will influence future Urban health policy internationally. *AI* **2020**, *1*, 156–165. [CrossRef]
15. Wilson, K.; Brownstein, J.S. Early detection of disease outbreaks using the Internet. *CMAJ* **2009**, *180*, 829–831. [CrossRef]
16. He, G.; Chen, Y.; Chen, B.; Wang, H.; Shen, L.; Liu, L.; Suolang, D.; Zhang, B.; Ju, G.; Zhang, L.; et al. Using the Baidu search index to predict the incidence of HIV/AIDS in China. *Sci. Rep.* **2018**, *8*, 9038. [CrossRef] [PubMed]
17. Liu, D.; Clemente, L.; Poirier, C.; Ding, X.; Chinazzi, M.; Davis, J.; Vespignani, A.; Santillana, M. Real-time forecasting of the COVID-19 outbreak in Chinese provinces: Machine learning approach using novel digital data and estimates from mechanistic models. *J. Med. Internet Res.* **2020**, *22*, e20285. [CrossRef] [PubMed]

18. Wang, Y.F.; Xu, K.; Kang, Y.; Wang, H.Y.; Wang, F.; Avram, A. Regional Influenza Prediction with Sampling Twitter Data and PDE Model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 678. [CrossRef] [PubMed]

19. Hickmann, K.S.; Fairchild, G.; Priedhorsky, R.; Generous, N.; Hyman, J.M.; Deshpande, A.; Del Valle, S.Y. Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput. Biol.* **2015**, *11*, e1004239. [CrossRef]

20. Smolinski, M.S.; Crawley, A.W.; Baltrusaitis, K.; Chunara, R.; Olsen, J.M.; Wojcik, O.; Santillana, M.; Nguyen, A.; Brownstein, J.S. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. *Am. J. Public Health* **2015**, *105*, 2124–2130. [CrossRef]

21. Barros, J.M.; Duggan, J.; Rebholz-Schuhmann, D. The application of internet-based sources for public health surveillance (infoveillance): Systematic review. *J. Med. Internet Res.* **2020**, *22*, e13680. [CrossRef]

22. Lazer, D.; Kennedy, R.; King, G.; Vespignani, A. The parable of Google Flu: Traps in big data analysis. *Science* **2014**, *343*, 1203–1205. [CrossRef]

23. Derczynski, L.; Ritter, A.; Clark, S.; Bontcheva, K. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference Recent Advances in Natural Language Processing Ranlp 2013, Hissar, Bulgaria, 9–11 September 2013; pp. 198–206.

24. Gu, J.; Wu, Y.; Xu, Y. Linguistic Feature and Temporal Pattern of User-Generated News: Evidence from an Online News Portal in China. In Proceedings of the PACIS 2018, Yokohama, Japan, 26–30 June 2018; p. 19.

25. Ghosh, S.; Chakraborty, P.; Nsoesie, E.O.; Cohn, E.; Mekaru, S.R.; Brownstein, J.S.; Ramakrishnan, N. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *Sci. Rep.* **2017**, *7*, 40841. [CrossRef]

26. McGough, S.F.; Brownstein, J.S.; Hawkins, J.B.; Santillana, M. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0005295. [CrossRef] [PubMed]

27. Kim, J.; Ahn, I. Weekly ILI patient ratio change prediction using news articles with support vector machine. *BMC Bioinform.* **2019**, *20*, 259. [CrossRef] [PubMed]

28. Liu, N.; Chen, Z.; Bao, G. Role of media coverage in mitigating COVID-19 transmission: Evidence from China. *Technol. Forecast. Soc. Chang.* **2020**, *163*, 120435. [CrossRef]

29. Lamb, A.; Paul, M.; Dredze, M. Separating fact from fear: Tracking flu infections on twitter. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 789–795.

30. World Health Organization. *Global Epidemiological Surveillance Standards for Influenza*; World Health Organization: Geneva, Switzerland, 2013.

31. Centers for Disease Control and Prevention. The Flu Season. Available online: https://www.cdc.gov/flu/about/season/flu-season.htm (accessed on 29 May 2021).

32. Public Health England. Annual Flu Reports. Available online: https://www.gov.uk/government/statistics/annual-flu-reports (accessed on 29 May 2021).

33. European Centre for Disease Prevention and Control. Indicators of Influenza Activity. Available online: https://www.ecdc.europa.eu/en/seasonal-influenza/surveillance-and-disease-data/facts-indicators (accessed on 29 May 2021).

34. Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, Piscataway, NJ, USA, 3–8 December 2003; pp. 133–142.

35. Salton, G.; Buckley, C. *Term Weighting Approaches in Automatic Text Retrieval*; Cornell University: Ithaca, NY, USA, 1987.

36. The Writing Committee of the World Health Organization (WHO) Consultation on Human Influenza A/H5. Avian influenza A (H5N1) infection in humans. *N. Engl. J. Med.* **2005**, *353*, 1374–1385. [CrossRef] [PubMed]

37. Zhang, Y.; Yakob, L.; Bonsall, M.B.; Hu, W. Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data. *Sci. Rep.* **2019**, *9*, 3262. [CrossRef]

38. Broniatowski, D.A.; Paul, M.J.; Dredze, M. National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. *PLoS ONE* **2013**, *8*, e83672. [CrossRef] [PubMed]

39. Doan, S.; Ohno-Machado, L.; Collier, N. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In Proceedings of the 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, La Jolla, CA, USA, 27–28 September 2012; pp. 62–71.

40. Lennart, L. *System Identification: Theory for the User*; PTR Prentice Hall: Upper Saddle River, NJ, USA, 1999; Volume 28, pp. 1–14.

41. Paul, M.J.; Dredze, M.; Broniatowski, D. Twitter improves influenza forecasting. *PLoS Curr.* **2014**, *6*. [CrossRef]

42. Zou, H.; Yang, Y. Combining time series models for forecasting. *Int. J. Forecast.* **2004**, *20*, 69–84. [CrossRef]

43. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.

44. Maindonald, J.H.; Braun, W.J.; Braun, M.W.J. Package 'DAAG'. Data Analysis and Graphics Data and Functions. 2015. Available online: https://cran.r-project.org/package=DAAG (accessed on 17 June 2021).

45. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]

46. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural. Comput.* **1998**, *10*, 1895–1923. [CrossRef] [PubMed]

47. Kreft, J. Instant Articles (Facebook): The Impact of Trust and Relations Among the Partners Pursuing the Strategy of Coopetition. In *Eurasian Business Perspectives*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 243–253.
48. Zhang, Y.; Ibaraki, M.; Schwartz, F.W. Disease surveillance using online news: Dengue and Zika in tropical countries. *J. Biomed. Inform.* **2020**, *102*, 103374. [CrossRef] [PubMed]
49. CNNIC. *The 46th China Statistical Report on Internet Development*; CNNIC: Beijing, China, 2020.