Visualize the Influenza: An Inference Model and Opinion Mining

Aonan Guan, Binyan Xiong, Bo Su, Pranjal Gupta, Qiming Sun
Georgia Institute of Technology
Atlanta, GA 30332, USA
{aguan,bxiong37,bqu7,pranjalg96,qiming}@gatech.edu

1 Introduction

1.1 Motivation

Influenza is a common topic in our society and it brings us lots of troubles every year, especially during the seasonal flu outbreak period. Our project aims at generating several views using different visualization methods to help people understand more about flu in our daily life. This project mainly contributes in two ways: 1) visualize the past influenza-like activities and propose a model for predictions; and 2) dive into the public opinion, basically is, how people treat the flu, via analysis on twitter posts. In this way, we can deliver a comprehensive and multi-dimensional knowledge related to influenza. For individuals, it is useful for us to get prepared for the flu season. More importantly, it could help officials make prevention plans and collect public opinions.

1.2 Problem Definition

This project contains three parts: visualize the past flu activities, generate a model to predict future flu trends, and conduct opinion mining on the social media posts related to flu.

We use the Influenza-Like Illness (ILI) Activity Level as an indicator to visualize the flu situation in past years. The ILI level data from 2008 to March 2020 is visualized on a weekly basis in our project. The inference model is trained on the ILI level on Week 40 to Week 52 (which is considered as the flu season) and the predictions are made for the flu season.

For the opinion mining part, over 2 million tweets are retrieved from the Internet and we conduct word-vector analysis as well as sentiment analysis on the dataset. The result includes the public's feeling during the flu season and the flu-related trendings at different times.

2 Literature Review

2.1 Research on ILI activity

In the past, a multitude of approaches have been used for the Forecasting of Influenza.[17] uses Support Vector Machine (SVM) and AdaBoost for predicting the ILI dynamics. They also experimented with multiple neural network architectures that rely on LSTM layers that combine ILI historical data and social media signals through a merged layer. [19] employs a multi-channel LSTM neural network that can draw information from different types of inputs. To improve accuracy, they add an Attention mechanism. [16] took a more data-driven approach, by applying both deep learning methods and incorporating environmental and spatio-temporal factors such as humidity, temperature, precipitation and sun exposure, to improve the performance of their influenza forecasting models.

2.2 Opinion Mining

2.2.1 Words Clustering Clustering is the process of grouping similar entities together as one important technique in unsupervised learning. Some novel approaches to cluster words include Word2Vec from Google [11] and FastText from Facebook[4]. In 2015, [6] presents a approach on conducting dimension reduction on the word vectors, which enables a bird's-eye view of the text as well as helping understand the opinion. In 2019, another group has proposed an approach of leveraging graph technique to visualize the word vectors [7].

2.2.2 Sentiment Analysis Starting from 2010, people have been doing sentiment analysis on social networks especially on twitter [15] [1] [9]. Some current researches have proved the value of text mining to monitor and analyze the flu trend. The communications in social media are at a relatively low cost and thus is a good tool to track people's response to health issues as well as identify the potential communities which may need further intervention [13]. For instance, one approach

combined sentiment analysis and TF-IDF weighting to filter the tweets and track the flu outbreaks [2]. Besides, sentiment analysis can also give us the knowledge of how people perceive a disease in different areas, communities, and at different periods [5][3]. To conduct sentiment analysis, emotion analysis tools, such as NRC emotion lexicon [12] and Stanford CoreNLP [10], are specially designed to detect the sentiments contained in the posts.

3 Proposed Methods

3.1 Intuition

Normally, to make accurate predictions, historical data needs to be integrated with other environmental and spatio-temporal factors, such as humidity, temperature, precipitation, sun exposure, etc. We attempt to make reasonable predictions using only the historical data, thus avoiding the need for complex, forecasting methods.

We dig deeper into the social media posts to find out emotional patterns that people feel about flu at different periods in a year and compare it with the existing data. Thus our results can give us a view of citizen's attention and sentiments to flu, which shows the public opinions during the flu season.

Also, our visualization will include the official statistical data, the prediction results, and the opinion-mining outcomes, which links people's thoughts with ILI activities.

3.2 ILI Activity visualization

The dataset we use comes from the CDC website, under the ILI activity section. This dataset contains the weekly ILI activity for each state in the US, over the years 2008-2020, and continues to be updated. The ILI activity has 11 levels, ranging from 'Level 0' to 'Level 10', with 'Level 10' being the highest. We use this data for our visualization and inference models.

For the visualization part, ILI activity from 2008 to 2020 is used in this project and we applied D3 choropleth to show the spread of flu of different years. To switch between different years and weeks in a particular year, dropbox and slider bar are used respectively. In order to show the ILI activity of state at a particular time, the tooltip feature is added to the choropleth.

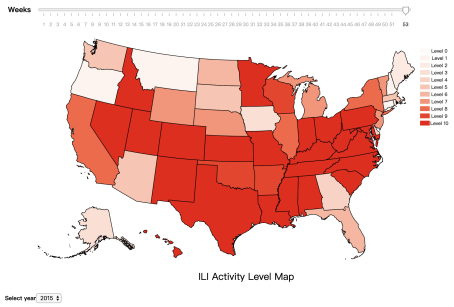

Figure 1: ILI Activity Level

3.3 ILI Activity Inference

For the Inference part, we use Recurrent Neural Networks (RNNs) together with Long Short Term Memory (LSTM) units. These LSTM units can capture long (or short) range sequence dependencies in the input and are ideal for our problem of forecasting ILI activity. We propose an approach, where each state is modeled independently. We also use ILI activity as our only feature. Although it could be argued that a unified state model together with additional information, such as weather and population of the state might do better, we feel the data insufficiency does not allow this integration, due to the increased network size and parameters that need to be learned.

In addition, we only make predictions for the flu season (weeks 40 to 52) in each season. There are two main reasons for this. The first is that making predictions during this season carries a lot more weight than predictions during the rest of the year. The second is that this allows our RNN to focus on learning the trends and seasonality pattern of the flu season, rather than irrelevant patterns that may occur, if all the 52 weeks were to be used. For our training, we use data from seasons 2008-2009 to 2017-2018, and then evaluate our performance on the 2018-2019 season.

Even though our ILI activities are categorical, we model them as continuous. To do this, we simply integer encode these activities. For example, 'Level 7' is encoded as 7. We then normalize these to a 0-1 range. We previously experimented with a categorical model, where we one-hot encoded the activities. But, the results were not good enough. We believe that this was because the one-hot encoded vectors added features, which our model could not effectively learn, due to insufficient data. We also believe that treating the activity as continuous and performing regression on it had an added advantage.

Regression ensures that our predicted activity levels are at least close to the true activity levels. If we treat it as categorical, this would not be the case, which might be another reason for the poor performance of our categorical model.

3.4 Opinion Mining on Flu-Related Tweets

3.4.1 *Data Crawling*

To retrieve the online tweets, a crawler based on TweetScraper[1] and Scrapy[2] is developed. Our crawler uses 'flu' and 'influenza' as the keyword and gets tweets by querying Twitter's search operations. All related tweets posted from January 1st, 2019 to February 29, 2020 were downloaded and each tweet is stored in a JSON file. A total of 2,689,583 flu-related tweets were crawled (with 1,715,986 tweets in 2019 and 973,597 in 2020). To improve the efficiency of our following analysis process, all tweets were imported to the MongoDB database. Considering in year 2020, a majority of the tweets containing 'flu' are related to the COVID-19 and the novel coronavirus topic, the tweets posted in 2019 and 2020 are managed separately.

3.4.2 *Word Vectors and text mining*

After the tweets were crawled,we used two word-embedded models to train the dataset based on gensim[3] and generated the word vector so that we could conduct exploratory data analysis. We used Wrod2Vec [11] and FastText [4] to generate a 100 dimensions vector for each word with the sliding window size as 5 in the sentence. The generated word vectors represent the word's semantically meaning where semantically similar words have similar vectors and closer distance in the geometric space. The data was also cleaned by removing the common words and the stop words for a model tunning. The comparison and evaluation results of these models were discussed in the section Four. Then we completed the visualization based on the model we got before and generated the interactive HTML page below.

Vec2Graph [18] was used to help generate the words' neighbors as a node-edge graph. Here we set influenza as the gateway word, and recursively query the closest

[1]https://github.com/jonbakerfish/TweetScraper
[2]http://scrapy.org/
[3]https://github.com/RaRe-Technologies/gensim

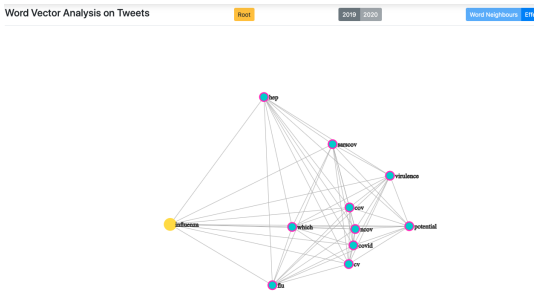

Figure 2: Node-edge words' neighbour Graph

words in each iteration to compute a node-edge graph. Here we set a threshold distance value for drawing the edge and connecting nodes and limit the graph size to terminate the graph's growth. D3.js was used for drawing node-edge and enabled the hyperlink for neighbor nodes calling neighbors recursively. To perform a better visual effect, we applied a gamma correlation to map the words' correlation interval from [0,0.8] to [0,0.4] and keep the total interval form as [0,1] by adding a power function. Here we selected 4.1 as the power so that the nodes will not be huddled together.

Besides, ScatterText [8] was used to help generate the comparison document visualization. In this case, we generated the following interactive plot charts, which enabled us to query the words from the input box and presented the related tweets comparatively in Figure 3.

The chart in Figure 4 was generated to present the

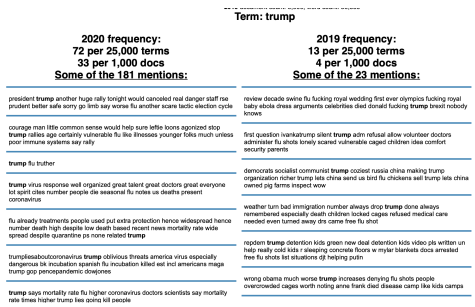

Figure 3: Query Example

statistical effect size [14] based on the corpus we have crawled. To calculated the effect size of each word, we used the TF/IDF liked method to calculate the term frequency in two of the documents (2019, 2020 respectively). We then calculated the standard error to perform the Cohen's d type effect sizes.

Here, $d = \frac{\mu_1 - \mu_2}{s}$, in which $s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

To comparatively present the 2019 and 2020's words'

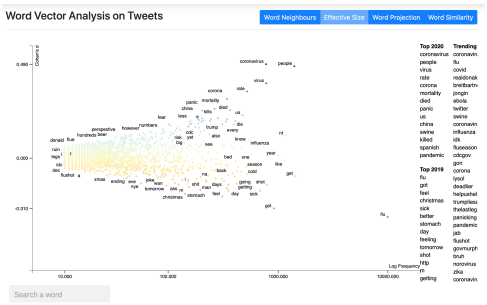

Figure 4: Cohen's D effect size plot

usage in the same plot chart. We used axis x to show the log frequency of each word use, and the axis y as the Cohen's d value to present 2019 and 2020's corpus, in which the positive value represented the corpus in 2020 and negative value represented the corpus in 2019.

The chart in Figure 5 is a word projection chart. In this

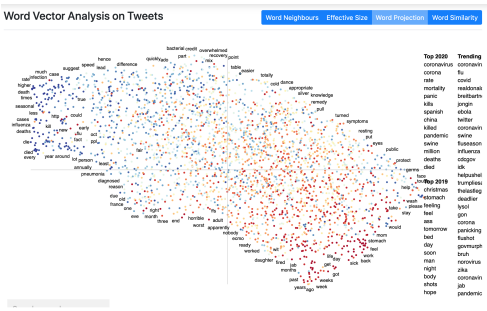

Figure 5: Words vector mapped to 2D projection

one, we simply conducted a dimension reduction on the models we have trained before, for the year 2019 and the year 2020(plotted as red and blue respectively). Here, we used the T-distributed Stochastic Neighbor Embedding method(TSNE) to perform the reduction procedures on the 100 dimensions vectors we generated. Then we projected them to a 2D vector space and placed them based on the 2D coordinates.

This chart presents us with a holistic and heuristic view of the people's word usage on twitter when they talking about influenza. The chart in Figure 6 we generated here was the word similarity plot chart when querying flu as the root word among the tweets in 2019 and 2020. Here the axis x is the usage in the year 2019, the axis y is the usage in the year 2020. Hence the flu was located on the top right corner which performs both large amounts in both 2019 and 2020. Then the other similarity words were scattered on the canvas based on the frequency and semantic distance.

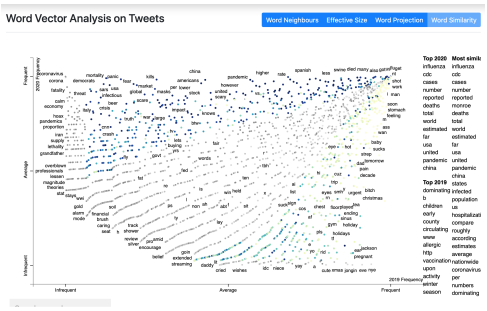

Figure 6: Similarity plot respond to "influenza"

3.4.3 Sentiment Analysis

We used StanfordCoreNLP to complete sentiment analysis. Comparing to other sentiment analysis package, the performance of StanfordCoreNLP is outstanding. The model used is a RNN based on grammatical structures and trained by a well-labeled dataset - Standford Sentiment Treebank.

The Java program has computed all the crawled tweets' text. The average score of each sentence in one tweet text has been adopted. 1 - 5 represents a very negative, negative, neutral, positive, and very positive. There are 0.713GB of data which consists of 1,715,986 in entire 2019 and 97,359 in the first two months of 2020.

As far as a current analyzed result, the overall attitude towards influenza is shown in the table below. The pre-

Year	2019	2020
Sentiment Score	1.4536126657875	1.4014825327134

Table 1: Sentiment score of the tweets

liminary result shows the flu-related tweets in 2020 are relatively negative than tweets in 2019. We then produced weekly or monthly analysis to dig out presentable data visualization. Another approach we have done is filtering the tweet and show the representative tweets which could be selected based on the number of retweets time. Note that we do no take a whole tweet, a complete sentence will be appropriate enough.

4 Experiments and Evaluation

4.1 Evaluate the Inference model

To train our model, we set up the data so that every week, we predict the ILI activity for the next week. We would then assume that the true value for the next

week is given, and continue the process, to get a predicted value for all the weeks in the flu season of 2018-2019. This is called the one-step walk forward validation model and mimics a real-world scenario where we would forecast the next week activity based on the current activity, and continue this process as new data becomes available.

We then design the architecture of the model. We choose the number of units in the LSTM layer of our network as 4. In our experiments, this number seemed to do well, and increasing it further did not improve the performance by much. Finally, we have a single hidden unit as the output. We choose the loss function as 'Mean squared error' to penalize large differences between prediction and true values. We also used the reliable 'Adaptive Moment Estimation, or Adam' optimizer. We set the number of epochs as 100, and used a batch size of 1(stochastic descent).

After we train our model for a particular state, we evaluate it, using Root mean squared error (RMSE). We calculate the RMSE between the predicted values and the true activity levels. We do this for both the training and the test set. After this, we plot the true activity levels as well as predicted values (both training and test set) against the flu weeks over all seasons, to get a visual idea of our predictions. We then round our predictions, and inversely encode the integer values back into the categorical values.

As an example, for the state of Georgia, we got an

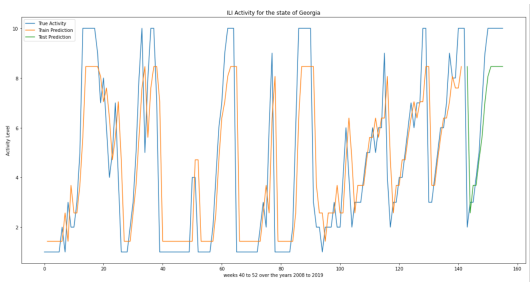

Figure 7: True and Predicted Activity levels for the state of Georgia

RMSE of 1.90 for the training set and 2.22 for the test set. This means a difference of about two activity levels on average, which is pretty good. The plot obtained for the state of Georgia is shown in Fig. 7. From this figure, we can see that our model is able to learn the trends and seasonality patterns pretty effectively. It only seems to not predict extreme values very well. This could perhaps be fixed with more training data. Similarly, the RMSE for the state of Wyoming was 1.65 for training and 2.27 for the test.

We then repeat this process for all states and store the final categorical values for each state. We then use these to visualize our predictions, using the choropleth map. We also calculate the average RMSE over all states, which comes out to be around 1.702 for training and 1.971 for test. The predictions on our choropleth map are shown in Fig. 8. This is interactive, and the user can cycle across the predictions for all 13 weeks using the sliding bar. There is also a tooltip functionality, with the state name and the ILI activity for the state.

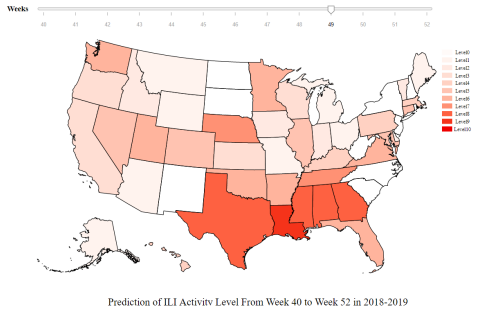

Figure 8: Predictions of ILI activities for the 2018-2019 season over all states.

4.2 Evaluation of Text Mining

From the text mining perspective, we are supposed to implement an intuitionistic visualization based on the tweets data in 2019 and 2020 to reveal the people's opinion under the influenza topic and help present a comparative presentation along with the CDC's quantitative data visualization.

Here we completed our approach by training multiple models with different strategies. Then we completed a model evaluation based on similarity and analogy tasks.The evaluation results are shown below. We then analyzed the tweets in 2019 with the Word2Vev model, the ones in 2020 with FastText and completed the visualization by using the D3.js library.

The node-edge words' neighbor graph in Figure 2 shows that in 2020 the result was highly affected by COVID-19 since there exist 7 words among 10 related to coronavirus connected closely with influenza, on the contrary, the results in 2019 did not show abnormal. Besides, since

Similarity Task	2019	2020
Word2Vec	0.6762688	0.48389888
Word2Vec(cleaned)	0.6019994	0.4278322
FastText	0.71620524	0.7815212
FastText(cleaned)	0.60729575	0.66706836

Table 2: Accuracy results of the similarity task for all models.

Analogy Task	2019	2020
Word2Vec	True	True
Word2Vec(cleaned)	True	False
FastText	False	False
FastText(cleaned)	False	False

Table 3: Validation results of the analogy task for all models.

the data crawled in 2020 was stopped by February, it is believed that the COVID-19 has been dominating the flu-related topic on the internet at least in February 2020.

The effect size plot chart based on term frequency in Figure 4 reveals two appreciable outlier plot in 2019 and 2020, which was coronavirus and flu. There are also some words that reflect people's consideration during 2020 including "panic", "US", "China", "Spanish", "pandemic", "Ebola".

The words' projection plot chart in Figure 5 presents a holistic view of the words' use. In this case, the words used in 2019 (red), and 2020 (blue) hold a distinct boundary. Since the data used to present visualization was consistently querying influenza. It is not supposed to hold this distribution, while this abnormal plot could imply that the flu that happened in 2019 should not be the same type as it is in 2020, for this abnormal data distribution. The words' similarity plot chart also shows consistency with this hypothesis.

4.3 Evaluation on the Sentiment Analysis

We conducted sentiment analysis on more than two million tweets. Overall, the sentiment score is low since we crawled the data according to flu, influenza, and related words. It does make sense that the sentiment score is near to 1 that is very negative. We try the simple way which computes the average of each month and week, then presents them with the line chart to show the trends.

The figure below is the monthly average sentiment

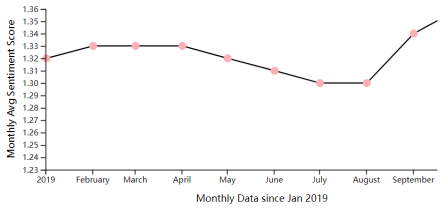

Figure 9: "Monthly Avg Sentiment Score"

score which suggests the outbreak of flu spread the sadness on the Internet. As we can see the curve goes straight down at the end part of the line chart. There is a peak valley from July to August. It is weird since the summer is not the flu season because of the high temperature. After diving into the database, it shows that the data in July and August is unbalanced which may lead to bias.

The figure above presents a similar idea of the monthly

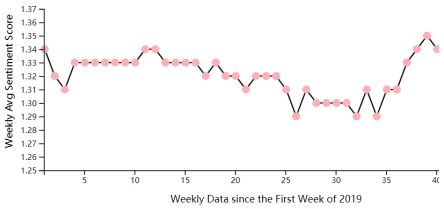

Figure 10: "Weekly Avg Sentiment Score"

line chart which gives more tiny fluctuation in detail.

5 Conclusions and discussion

Overall, our team has completed the flu-related data visualization, influenza trend prediction with a machine learning model and opinion mining on tweets data crawled with the flu-like keyword. From cover to cover, the above experiment results in plenty of work to gather flu-related data, clean out dirty data, finally visualize the fine-grained data. Our analysis shows that the public opinions of influenza are affected by COVID-19 significantly. The opinion mining we have done with more than two million Tweets data shows people's emotional fluctuation. Also, the model we generated using RNNs together with LSTM does show certain accuracy.

6 Distribution of Work

All team members have contributed similar amount of effort.

References

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. 30–38.

[2] Ali Alessa and Miad Faezipour. 2018. Tweet classification using sentiment analysis features and TF-IDF weighting for improved flu trend detection. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 174–186.

[3] Ali Alessa, Miad Faezipour, and Zakhriya Alhassan. 2018. Text classification of flu-related tweets using fasttext with sentiment and keyword features. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 366–367.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[5] Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri. 2015. Using twitter data and sentiment analysis to study diseases dynamics. In *International Conference on Information Technology in Bio-and Medical Informatics*. Springer, 16–24.

[6] Hendrik Heuer. 2016. Text comparison using word vector representations and dimensionality reduction. *arXiv preprint arXiv:1607.00534* (2016).

[7] Nadezda Katricheva, Alyaxey Yaskevich, Anastasiya Lisitsina, Tamara Zhordaniya, Andrey Kutuzov, and Elizaveta Kuzmenko. 2019. Vec2graph: a Python library for visualizing word embeddings as graphs. In *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 190–198.

[8] Jason S. Kessler. 2017. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. (2017).

[9] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!. In *Fifth International AAAI conference on weblogs and social media*.

[10] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[12] Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada* (2013).

[13] S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research* 15, 4 (2013), e85.

[14] Shinichi Nakagawa and Innes C Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews* 82, 4 (2007), 591–605.

[15] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.. In *LREc*, Vol. 10. 1320–1326.

[16] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony Maida, and Stephen Nichols. 2017. A novel data-driven model for real-time influenza forecasting. bioRxiv. (2017).

[17] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one* 12, 12 (2017).

[18] Tamara Zhordaniya, Andrey Kutuzov, and Elizaveta Kuzmenko. 2020. Vec2graph: A python library for visualizing word embeddings as graphs. In *Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers*, Vol. 1086. Springer, 190.

[19] Xianglei Zhu, Bofeng Fu, Yaodong Yang, Yu Ma, Jianye Hao, Siqi Chen, Shuang Liu, Tiegang Li, Sen Liu, Weiming Guo, et al. 2019. Attention-based recurrent neural network for influenza epidemic prediction. *BMC bioinformatics* 20, 18 (2019), 1–10.