Predicting Influenza Dynamics with Neural Networks Using Signals from Social Media

Ellyn Ayton, Western Washington University
Advisor: Svitlana Volkova, PhD Pacific Northwest National Laboratory

Motivation: Every year there are 500,000 deaths worldwide attributed to influenza [7]. The Center for Disease Control and Prevention (CDC) reports weekly on the level of influenza-like illness (ILI) seen year round in hospitals and doctor visits. These values are used to monitor the spread and impact of the influenza, however by the time the ILI data is released, the information is already 1-2 weeks old and is frequently inaccurate until revisions are made [8]. To overcome this, we propose making use of large amounts of social media data, such as Twitter to be a secondary source of information in order to predict current and future ILI proportions — the total number of people seeking medical attention with ILI symptoms. In previous related work, flu forecasting has been accomplished through the use of basic linear autoregressive models, linear autoregression exogenous models, SVM regressions, logistic regression classifiers, SIR models, and more [1, 2, 5, 6]. The addition of social media features to several of these models such as the linear autoregressive model, has improved the model's performances over ILI data alone [3, 4, 8]. Our work is geared toward applying these data sources to more powerful machine learning models. Having this predictive power can aid health officials to properly prepare for and respond to yearly flu outbreaks.

Approach: By integrating the information that people tweet about e.g., topics, syntax, style and their communication behavior e.g., hashtags, mentions, we built predictive models for ILI and confirmed influenza activity across different geographical locations in the U.S. We experiment and evaluate the predictive power of a variety of features and machine learning models e.g., Support Vector Machines with radial basis function or linear kernels, AdaBoost with Decision Trees [10]. We are the first to evaluate the predictive power of neural networks — Long Short Term Memory (LSTM) for ILI nowcasting and forecasting [9]. An LSTM is a special type of recurrent neural network (RNN) that is capable of preserving information and learning long-term dependencies in data, which traditional RNNs struggle with. For this specific reason, we chose LSTMs to model our data over the course of several weeks.

Results: We found that LSTMs achieve the best performance regardless of which text representations are included e.g., embeddings vs. raw tweets. Of our nine features extracted from Twitter, AdaBoost models learned from unigrams, hashtags, and word embeddings consistently outperform all other features. Using up to four weeks of past data, our models are capable of accurately predicting ILI proportions for the current week and predicting ILI values for up to the next two weeks. We have found that a model tailored to a specific location shows a greater performance than a general model encompassing all regions. In our

future work, we will apply our LSTM model to 25 additional locations, and combine our ILI and social media data into one predictive LSTM model.

References

1. Broniatowski, David A., Michael J. Paul, and Mark Dredze. "National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic." *PloS one* 8.12 (2013): e83672.
2. Santillana, Mauricio, et al. "Combining search, social media, and traditional data sources to improve influenza surveillance." *PLoS Comput Biol* 11.10 (2015): e1004513.
3. Paul, Michael J., et al. "SOCIAL MEDIA MINING FOR PUBLIC HEALTH MONITORING AND SURVEILLANCE." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. Vol. 21. 2016.
4. Smith, Michael C., et al. "Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter." (2015).
5. Riley, Pete, et al. "Early Characterization of the Severity and Transmissibility of Pandemic Influenza Using Clinical Episode Data from Multiple Populations." *PLoS Comput Biol* 11.9 (2015): e1004392.
6. Shaman, Jeffrey, and Alicia Karspeck. "Forecasting seasonal outbreaks of influenza." *Proceedings of the National Academy of Sciences* 109.50 (2012): 20425-20430.
7. WHO (2009) *Influenza (Seasonal), Fact Sheet Number 211.* Available at http://www.who.int/mediacentre/factsheets/fs211/en/index.html. Accessed August, 2016.
8. Paul, Michael J., Mark Dredze, and David Broniatowski. "Twitter improves influenza forecasting." *PLOS Currents Outbreaks* (2014).
9. Chollet, Francois. *Keras* (2015) github.com/fchollet/keras.
10. Scikit-learn: Machine Learning in Python, Pedregosa *et al.,* JMLR 12, pp. 2825-2830, 2011. http://scikit-learn.org/stable/index.html