Alerts from World Health Organization (WHO) and ProMed: A Prediction Task

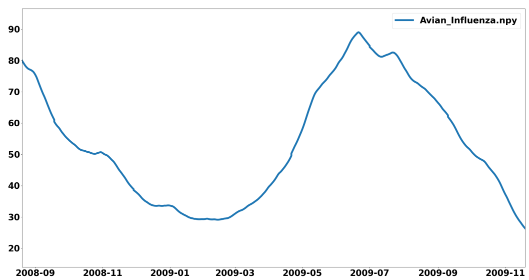

Figure 1: ProMed alerts on Avian Influenza.

Problem Statement: The World Health Organization (WHO) publishes about 100 alerts per year. We have a corpus of 2784 URLs from WHO [1], as well as 67,149 URLs from ProMed [2]. Both corpora span nearly 30 years starting in 1996. The corpora contain both structure data (disease, date, location) as well as unstructured data (natural language text). Figure 1 was computed based on structured data (disease & date) from ProMed. An alert on January 14, 2009, during a dip between two the peaks, asserted: *the number of cases seems to be diminishing.* The question is: how accurate are such predictions?

Our Approach: We will use standard methods in deep nets make forecasts. That is, based on the text in the alerts, can we predict how many alerts there will be over the next few weeks/months by disease and by location?

There is a considerable literature on deep nets. A popular prediction task involves the stock market [3, 4, 5]. One can model such prediction tasks like a game of poker, where people with inside information try not to reveal their cards, but a good poker player (and perhaps a machine) can pick up on subtle "tells." We suspect that predictions in WHO and ProMed alerts are more accurate during major outbreaks, and less accurate during dips when there is less evidence and more room for "spin." Simple models such as BERT [6] may be able to distinguish unsubstantiated spin with less evidence from more credible assessments with more evidence. We plan to start simple, by fine-tuning [7] BERT on training data with time-series labels (as illustrated in Figure 1).

An interesting challenge in computational linguistics is to summarize large collections of documents. Web search and summarization focus on a small number of relevant documents, whereas we are interested in trends that hold across many documents.

Reposting is an interesting challenge. Many alerts are similar to one another; [1, 8] updated 52 cases to 53 over 12 days. Reposting is common on social media, but less so for corpora used for training large language models (LLMs). Near duplicates are removed from web crawls [9], but on social media, reposting is a behavioral signal of importance. Reposting is problematic for classic methods such as Naive Bayes [10] that are dependent on assumptions such as independent and identically distributed (IID). It is said that deep nets are more robust to such concerns, though there is relatively little experience with data as repetitive as WHO/ProMed alerts. From a scientific perspective, it will be interesting to see if reposting is a blessing or a curse. Will deep nets find reposting to be a useful signal of importance (like social media), or confusing (like Naive Bayes)?

Evaluation: The task is to predict the time series over the next few weeks/months, based on the text. Stock market prediction systems use text from social media and/or earnings calls. We plan to use the alerts mentioned above, and start with a simple loss such as RMS error, comparing the number of observed alerts with predictions. We expect short-term forecasting to be easier than long-term forecasting. RMS errors will be reported by four variables: (1) forecasting window, (2) date, (3) disease and (4) location. As with stock market predictions, there are huge upsides, even if predictions are only slightly better than chance.

References

[1] https://www.who.int/emergencies/disease-outbreak-news/item/2009_01_14a-en.

[2] https://promedmail.org/.

[3] Z. Hu, W. Liu, J. Bian, X. Liu, and T.-Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3945120

[4] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Annual Meeting of the Association for Computational Linguistics*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:21665298

[5] R. Sawhney, S. Agarwal, A. Wadhwa, and R. R. Shah, "Deep attentive learning for stock movement prediction from social media text and company correlations," in *Conference on Empirical Methods in Natural Language Processing*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226262316

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[7] K. W. Church, Z. Chen, and Y. Ma, "Emerging trends: A gentle introduction to fine-tuning," *Natural Language Engineering*, vol. 27, no. 6, p. 763–778, 2021.

[8] https://www.who.int/emergencies/disease-outbreak-news/item/2009_01_26-en.

[9] A. Z. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1997, pp. 21–29.

[10] F. Mosteller and D. L. Wallace, "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, 1963.