Using a Recurrent Neural Network Model for Classification of Tweets Conveyed Influenza-related Information

Chen-Kai Wang[1], Onkar Singh[2,3] Zhao-Li Tang[1] and **Hong-Jie Dai[4,5*]**

[1]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C.
[2]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.
[3]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan, R.O.C.
[4]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, R.O.C.
[5]Interdisciplinary Program of Green and Information Technology, National Taitung University, Taitung, Taiwan, R.O.C.

Abstract

Traditional disease surveillance systems depend on outpatient reporting and virological test results released by hospitals. These data have valid and accurate information about emerging outbreaks but it's often not timely. In recent years the exponential growth of users getting connected to social media provides immense knowledge about epidemics by sharing related information. Social media can now flag more immediate concerns related to outbreaks in real time. In this paper we apply the long short-term memory recurrent neural network (RNN) architecture to classify tweets conveyed influenza-related information and compare its performance with baseline algorithms including support vector machine (SVM), decision tree, naive Bayes, simple logistics, and naive Bayes multinomial. The developed RNN model achieved an F-score of 0.845 on the MedWeb task test set, which outperforms the F-score of SVM without applying the synthetic minority oversampling technique by 0.08. The F-score of the RNN model is within 1% of the highest score achieved by SVM with oversampling technique.

1 Introduction

With the popularity of WWW, the use of data mining techniques to analyze the big data generated by users provides a feasible way for identification and exploration of health-related information. For example, Ginsberg et al. (2009) utilized search query logs of Google to develop models for influenza surveillance. With the recent increased use of social media platforms, users can communicate each other by updating their status led to wide sharing of personal information timely. The information spreads over these platforms is now a valuable information resource for building social sensors to develop real time event detection systems for detecting events like earthquakes (Sakaki, Okazaki, & Matsuo, 2010) and abuse of medications (Sarker, O'Connor, et al., 2016).

A review conducted by Charles-Smith et al. (2015) demonstrated evidence that the use of social media data can provide real-time surveillance of health issues, speed up outbreak management, identify target populations necessary to support and even improve public health and intervention outcomes. Facebook, microblogs, blogs, and discussion forums are examples of such media. Among them, Twitter, the leading micro-blogging platform, has become the primary data sources for digital disease surveillance and outbreak management. The

* Corresponding author

Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017), pages 33–38,
Taipei, Taiwan, November 27, 2017.

platform has been used for creating first hand reports of adverse drug events (Bian, Topaloglu, & Yu, 2012). Shared tasks (Aramaki, Wakamiya, Morita, Kano, & Ohkuma, 2017; SARKER, NIKFARJAM, & GONZALEZ, 2016) and hackathon style competitions (Adam, Jonnagaddala, Chughtai, & Macintyre, 2017) for digital disease detection or biosurveillance are also emerging. The rationale behind social media-based surveillance systems is based on the assumption that target events occur in the real world will immediately reflect on social media. Therefore, systems that aggregate and determine the degree of related information from social media can monitor or even forecast the current or future outbreak events.

Iso, Wakamiya, and Aramaki (2016) have demonstrated words such as "fever" present clues for upcoming influenza outbreaks. They concluded that an approximately 16-day time lag exists between the frequency of the word "fever" mentioned in tweets and the number of influenza patients announced by the infectious disease surveillance center in Japan. Although their results are promising, the use of word-level information was noisy and impedes precise influenza surveillance. Take the following two tweets described in the work of Aramaki, Maskawa, and Morita (2011) as an example.

"*Headache? You might have flu.*"
"The World Health Organization reports the avian influenza, or bird flu, epidemic has spread to nine Asian countries in the past few weeks."

Although the above two tweets include mentions of "flu", apparently they do not indicate any influenza patient has presented nearby. Therefore it is required to develop classifiers to categorize diseases/symptoms related to influenza in order to have an accurate influenza forecasting model. With this in mind, we consider to develop classifiers for diseases/symptoms related to influenza. We formulate the task as a classification problem and employ several baseline algorithms and recurrent neural networks (RNNs) to develop our models. The performance of the developed models are evaluated on a corpus annotated with eight disease/symptoms including influenza, cold, hay fever, diarrhea, headache, cough, fever and runny nose.

2 Method

2.1 Preprocessing

In the preprocessing step, we normalize all web links and usernames into "@URL" and "@REF" respectively. The part-of-speech tagger developed by Gimpel et al. (2011) is then used to tokenize tweets followed by removing the hashtag symbol "#" from its attached keywords or topics. Finally, we followed the numeric normalization procedure proposed by (Dai, Touray, Wang, Jonnagaddala, & Syed-Abdul, 2016) to normalize all numeral parts in each token into "1".

2.2 Network Architecture

The network architecture used in this study is a recurrent model consisting of an embedding layer, a bi-directional RNN layer followed by a dense layer to compute the posterior probabilities for each disease or symptom. Figure 1 illustrates the architecture.

2.3 Embedding Layer

The pre-trained word vectors for Twitter generated by GloVe (Pennington, Socher, & Manning, 2014) was used to initialize the embedding layer. The pre-trained 200-dimensional vectors were trained on two billion tweets which can be downloaded from the project website[2]. For word cannot be found in the pre-trained vectors, we initialized it with values closed to zero.

2.4 Recurrent Neural Network and Dense Layer

RNNs have proven to be a very powerful model in many natural language tasks (Mesnil, He, Deng, & Bengio, 2013; Tomá, Martin, Luká, Jan, & Sanjeev, 2010). This work used the long short term memory (LSTM), which is a special kind of RNN capable of learning long-term dependencies, to implement the RNN layer. As traditional RNNs, LSTM networks have the form of a chain of repeating cells of its neural network. LSTM uses gates to control the information flow inside its network. In Figure 1, assume that the first cell in the forward-LSTM is the tth token. The cell uses Equation 1 to output a value by considering the value h_{t-1} generated by the previous cell and the current input x_t.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

[2] https://nlp.stanford.edu/projects/glove/

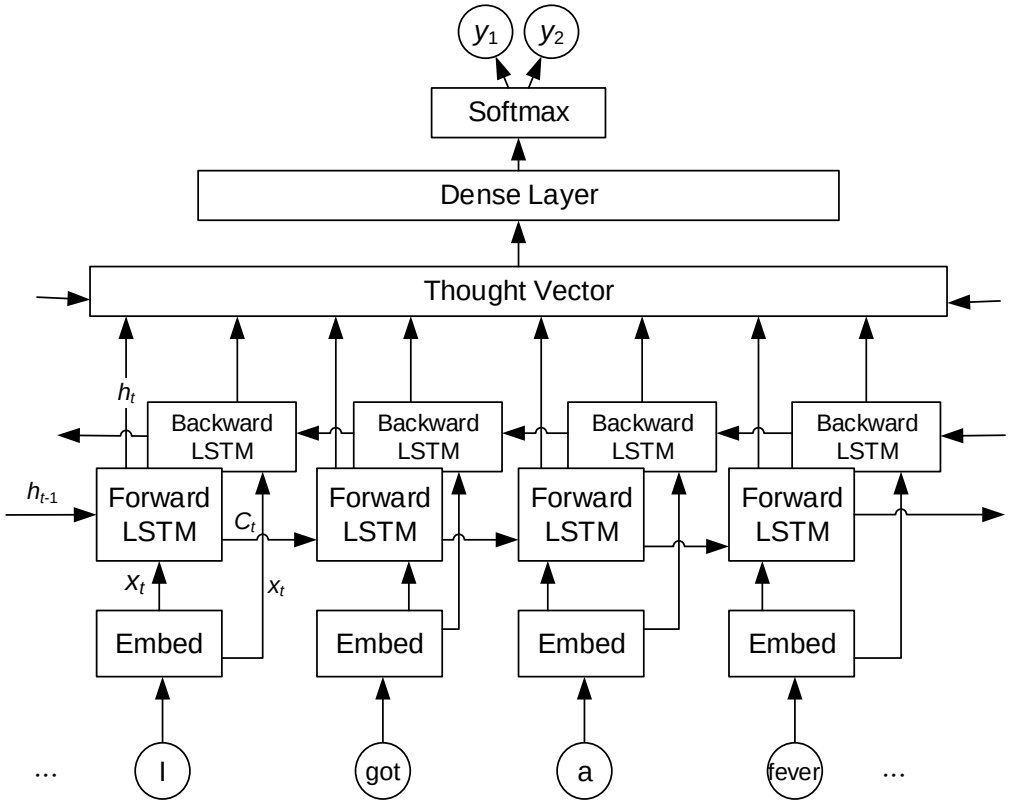

Figure 1: Network architecture developed in this work.

The cell then produces its two outputs C_t and h_t by using equation 4 and equation 6 respectively.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

In our network architecture, we duplicate the first recurrent layer to create a second recurrent layer so that there are now two layers side-by-side. The second layer is denoted as the Backward LSTM in Figure 1. For the backward layer, the input sequence is provided as a reversed copy of the input sequence.

Finally, we concatenate the last frame from the forward recursion, and the first frame from the backward recursion to build the thought vector (Gibson). The vector is then be the input of the dense layer with a dimension of 150 for soft max classification.

For the task studied in this work, we created one RNN model for each disease/symptom. Therefore, in total of eight RNN models were constructed for the eight types of diseases/symptoms.

2.5 Baseline Systems

Owing to the evaluation of the MedWeb task is ongoing, for the purpose of performance comparison, we implemented several baseline algorithms with Weka (Holmes, Donkin, & Witten, 1994) including C4.5 decision tree, simple logistic, support vector machine(SVM) trained with sequential minimal optimization, and Naïve Bayes Multinomial. The features used by the baseline were *n*-gram features with TF-IDF as the weighing function. Based on the tokens generated by the preprocessing step, we generate lowercased unigrams, bigrams and trigrams and filtered out stop words by using the list developed by McCallum (1996) along with a custom stop word list created by analyzing the training set. Finally, the Snowball stemmer (Porter, 2001) is used to perform stemming.

Symptom/Disease	Positive	Negative
Influenza	112	1808
Diarrhea	189	1731
Hay fever	201	1719
Cough	237	1683
Headache	254	1666
Cold	284	1636
Fever	355	1565
Runny nose	417	1503
Total	2049	13311

Table 1: Statistics of the training set.

3 Results

3.1 Dataset

The dataset released by NTCIR13-MedWeb task was used in this study (Kato, Kishida, Kando, Saka, & Sanderson, 2017). The dataset contains annotations indicating whether a Twitter user or someone around the user has symptoms or diseases like influenza, cold, hay fever, diarrhea, headache, cough, fever or runny nose at that point in time. Each tweet in the dataset was assigned with one of the following two labels. The label "p" (positive) is given if a tweet is determined as having symptom while the label "n" (negative) if it is not determined as a symptom/disease.

The training set consists of 1,920 tweets. Table 1 shows the statistics of the training set. As one can see that the dataset suffered the class imbalance problem.

3.2 Evaluation Metrics

We used the standard precision, recall and F-measure to evaluate the developed methods. We considered both the micro- and macro-averaged F-measure (Sokolova & Lapalme, 2009). A micro-F-score is generated by pooling all true positives, false positives and false negatives and calculate the F-score from that. A macro-average, on the other hand, is obtained by calculating the F-score for each class, and then averaging those F-scores to get a single number.

	P	R	F
SVM	0.869	0.880	0.875
C4.5	0.849	0.861	0.855
Naïve Bayes (NB)	0.262	0.966	0.413
Simple Logistic	0.822	0.924	0.870
NB Multinomial	0.666	0.874	0.756
SVM-SMOTE	0.867	0.882	0.875

Table 2: Baseline algorithm performance on the training set (micro-averaged).

	P	R	F
SVM	0.865	0.874	0.869
C4.5	0.841	0.851	0.845
Naïve Bayes (NB)	0.265	0.967	0.406
Simple Logistic	0.817	0.915	0.863
NB Multinomial	0.662	0.874	0.751
SVM-SMOTE	0.861	0.876	0.869

Table 3: Baseline algorithm performance on the training set (macro-averaged).

3.3 Results of the Baseline System on the Training Set

Table 2 and 3 show the two fold cross validation results on the English corpus of the MedWeb task. The baseline classifier with the best overall F-measure is SVM, which achieves the highest F-scores in the categories of "Cold" and "Influenza". The detail per-category performance of SVM is shown in Table 4.

Consider the imbalance observed in the training set, we try to increase the weight of examples when the classifiers makes errors on false positives. However the F-scores of all baseline classifiers didn't be improved. We also applied the synthetic minority oversampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to create new instances for training SVM. The result is indicated by SVM-SMOTE in Table 2 and 3. We can observe that the precision and the recall is improved in micro-average and macro-average respectively. The overall F-score is also slightly improved.

	P	R	F
Influenza	0.746	0.759	0.752
Diarrhea	0.849	0.894	0.871
Hay fever	0.848	0.915	0.880
Cough	0.982	0.911	0.945
Headache	0.887	0.929	0.908
Cold	0.982	0.911	0.945
Fever	0.837	0.865	0.850
Runny nose	0.864	0.882	0.873

Table 4: Performance of the best performed baseline model on the training set.

Configuration	P	R	F
SVM	0.734	0.809	0.770
SVM-SMOTE	0.807	0.911	0.856
RNN	0.836	0.854	0.845

Table 5: Performance on the test set (micro-averaged).

3.4 Results on the Test Set

Table 5 and 6 show the performance of the two best performed baselines and the proposed RNN model on the test set of the MedWeb task. The developed RNN model can achieve an F-score of 0.845, which outperforms the F-score of the SVM model by 0.0754 and is closed to that of the best performed baseline SVM-SMOTE. Comparing SVM-SMOTE with the proposed RNN model, RNN has better precision while lower recall. We can also observe that the precision and recall of the developed RNN model has similar scores while SVM-based models have better recall.

3.5 Discussion

Because the organizers of the MedWeb task have not released the gold annotations for the test set, we cannot conduct in-depth error analysis on the test set. Herein, we list some important key terms for each symptoms/diseases based on the results of the training set in Table 7. The list is generated by using the tree structures of C4.5 to prioritize the important terms for each symptoms/diseases. In addition to the terms directly related to the corresponding symptoms/diseases, we can see some interesting terms like “dog” for runny nose and “Nepali” for diarrhea.

4 Conclusion

In this paper, we presented a neural network architecture based on a bi-directional RNNs that can classify tweets conveying influenza-related information. We study the performance of this architecture and compare it to the best performing baseline algorithm on the test set of the MedWeb task. Using th e micro-F1 measure, the developed RNN model outperforms SVM by 0.087 and is within 1% of the highest score achieved by SVM with oversampling technique. In the future, we will continue to improve the performance of our model and conduct in depth error analysis regarding to the different symptoms/diseases.

Configuration	P	R	F
SVM	0.733	0.835	0.770
SVM-SMOTE	0.796	0.918	0.849
RNN	0.818	0.844	0.828

Table 6: Performance on the test set (macro-averaged).

Type	Terms
Cold	cold, fever
Cough	coughing, cough, phlegm
Hayfever	because of, allergies, spring, pollen,
Headache	headache, head hurt
Influenza	flu, vaccinate
Runny-nose	nose, dog
Fever	temperature
Diarrhea	stomach, Nepali

Table 7: Key terms observed on the training set.

References

Adam, D., Jonnagaddala, J., Chughtai, A. A., & Macintyre, C. R. (2017). *ZikaHack 2016: A digital disease detection competitio*. Paper presented at the Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017, Taipei, Taiwan.

Aramaki, E., Maskawa, S., & Morita, M. (2011). *Twitter catches the flu: detecting influenza epidemics using Twitter*. Paper presented at the Proceedings of the conference on empirical methods in natural language processing.

Aramaki, E., Wakamiya, S., Morita, M., Kano, Y., & Ohkuma, T. (2017). *Overview of the NTCIR-13: MedWeb task*. Paper presented at the Proceeding of the NTCIR-13 Conference, Tokyo, Japan.

Bian, J., Topaloglu, U., & Yu, F. (2012). *Towards large-scale twitter mining for drug-related adverse events.* Paper presented at the Proceedings of the 2012 international workshop on Smart health and wellbeing.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H. Y., Olsen, J. M., . . . Corley, C. D. (2015). Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. *PLoS ONE, 10*(10), e0139701. doi:10.1371/journal.pone.0139701

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic

minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Dai, H.-J., Touray, M., Wang, C.-K., Jonnagaddala, J., & Syed-Abdul, S. (2016). Feature Engineering for Recognizing Adverse Drug Reactions from Twitter Posts. *Information*.

Gibson, C. N., Adam. Thought Vectors, Deep Learning & the Future of AI - Deeplearning4j: Open-source, distributed deep learning for the JVM.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2011). *Part-of-speech tagging for Twitter: annotation, features, and experiments*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, Portland, Oregon.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012-1014.

Holmes, G., Donkin, A., & Witten, I. H. (1994). *Weka: A machine learning workbench.* Paper presented at the Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on.

Iso, H., Wakamiya, S., & Aramaki, E. (2016, December 11-17). *Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction.* Paper presented at the Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan.

Kato, M. P., Kishida, K., Kando, N., Saka, T., & Sanderson, M. (2017). *Report on NTCIR-12: The Twelfth Round of NII Testbeds and Community for Information Access Research.* Paper presented at the ACM SIGIR Forum.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. *http://www.cs.cmu.edu/~mccallum/bow*.

Mesnil, G., He, X., Deng, L., & Bengio, Y. (2013). *Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding.* Paper presented at the INTERSPEECH.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 12*, 1532-1543.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* Paper presented at the Proceedings of the 19th international conference on World wide web.

SARKER, A., NIKFARJAM, A., & GONZALEZ, G. (2016). *SOCIAL MEDIA MINING SHARED TASK WORKSHOP*. Paper presented at the Pacific Symposium on Biocomputing 2016.

Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety, 39*(3), 231-240. doi:10.1007/s40264-015-0379-4

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Tomá, M., Martin, K., Luká, B., Jan, È., & Sanjeev, K. (2010). *Recurrent neural network based language model*. Paper presented at the Proceedings of the 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan.