Shot Or Not: Comparison of NLP Approaches for Vaccination Behaviour Detection

Aditya Joshi[1] **Xiang Dai**[1,2] **Sarvnaz Karimi**[1]
Ross Sparks[1] **Cécile Paris**[1] **C Raina MacIntyre**[3]
[1]CSIRO Data61, Sydney, Australia, [2]University of Sydney, Sydney, Australia
[3]The University of New South Wales, Sydney, Australia
{aditya.joshi,dai.dai,sarvnaz.karimi}@csiro.au
{ross.sparks,cecile.paris}@csiro.au, r.macintyre@unsw.edu.au

Abstract

Vaccination behaviour detection deals with predicting whether or not a person received/was about to receive a vaccine. We present our submission for vaccination behaviour detection shared task at the SMM4H workshop. Our findings are based on three prevalent text classification approaches: rule-based, statistical and deep learning-based. Our final submissions are: (1) an ensemble of statistical classifiers with task-specific features derived using lexicons, language processing tools and word embeddings; and, (2) a LSTM classifier with pre-trained language models.

1 Introduction

Public opinion about vaccines is diverse. Most people support vaccination, but some of these people do not receive vaccination. On the other hand, people who are vaccinated may also have concerns regarding the safety or efficacy of vaccines. In other words, a person's stance towards vaccines (referred to as '*vaccine hesitancy*') is distinct from whether or not they received a vaccine shot (referred to as '*vaccination behaviour*'). While automatic detection of vaccine hesitancy has been explored in the past, computational approaches to detect vaccination behaviour have been limited. Towards this, our paper deals with vaccination behaviour detection (SMM4H shared task #4). Vaccination behaviour and vaccine hesitancy can together help to understand penetration of vaccination programmes and the trust that communities place in large-scale vaccination programmes (Holt et al., 2016).

Vaccination behaviour detection is the task of predicting whether or not a given piece of text refers to a person receiving or intending to receive a vaccine. For example, the tweet '*I took the vaccine this morning, feeling great!*' is positive because the speaker reports having received the vaccine. On the contrary, '*Vaccines drastically reduce risks of infection*' is negative because the tweet describes vaccines but does not report a vaccine being administered.

Past work in vaccination behaviour detection uses n-grams as features of a statistical classifier (Skeppstedt et al., 2017; Huang et al., 2017). However, alternatives to n-grams have shown promise in several Natural Language Processing (NLP) tasks. Therefore, we compare three typical NLP approaches for vaccination behaviour detection: rule-based, statistical and deep learning techniques. Our submissions to the shared task use statistical and deep learning-based text classification. The systems are trained on a concatenation of the training and the validation set. The work reported in this paper ranked first among nine teams, as communicated by the shared task committee.

2 Approaches

In this section, we describe the three approaches that we employ for vaccination behaviour detection: Statistical, rule-based and deep learning-based.

2.1 Statistical Approach

Our statistical approach uses an ensemble of three classifiers: logistic regression, support vector machine with both using LIBLINEAR (Fan et al., 2008), and random forest using scikit-learn (Pedregosa et al., 2011). We use the following non-default parameters: (a) Positive misclassification cost is set to 3 in logistic regression; (b) 100 estimators in random forest. Majority voting is used to combine predictions from the classifiers, *i.e.*, a tweet must be predicted as positive by at least two classifiers for it to be predicted as positive by the ensemble.

The random forest classifier uses unigrams as features. The features for logistic regression and

Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, pages 43–47
Brussels, Belgium, October 31, 2018.

Feature	Description	Type
N-grams	Unigrams and bigrams in the tweet	Boolean
Special Characters	@ and # which indicate user mentions and hashtags, ! and ?	Boolean
POS	Number of words of each POS tag	Count
Negation	Presence of negation words	Count
Word Similarity	Maximum value of similarity of words in the tweet and words indicating administration/reception of a vaccine	Real
Sentence Vector	Average of word vectors of the words in the tweet	Real
Length	Number of characters and words	Count
Emotion	Number of words of each emotion category	Count

Table 1: Features of the statistical approach.

support vector machine are summarised in Table 1. These features are:

1. **Uni/Bigrams**: Boolean;
2. **Special Characters**: A feature each indicating four special characters ?, #, @, !
3. **POS**: Count of each POS tag using NLTK POS tagger (Bird and Loper, 2004). This feature follows the intuition that presence of certain POS tags such as verbs may serve as signals;
4. **Negation**: Presence of a negation word. This is to serve as a negation feature where, although the act of receiving a vaccine is mentioned, the negation word changes the output class;
5. **Word Similarity**: For each word, we obtain similarity with '*receive*, '*get*' and '*take*', and use the highest similarity as this feature. We use pre-trained embeddings from Mikolov et al. (2013). This is to allow presence of words related to the act of receiving to be used as a signal for prediction;
6. **Sentence Vector**: A sentence vector is computed as an average of word vectors using GloVe embeddings (Pennington et al., 2014);
7. **Length**: Number of characters and words;
8. **Emotion**: Word counts of each emotion category as given by SenticNet (Cambria et al., 2014).

The combination of classifiers, misclassification costs and features has been experimentally validated.

2.2 Rule-based Approach

Since vaccination behaviour detection may appear to be only about detecting administration of a vaccine, we implement a naïve method to detect vaccination behaviour. Our rule-based approach looks for words indicating '*receive*' (without negation) to predict vaccination behaviour as follows:

1. If a tweet contains one among the words 'give', 'take', 'taking', 'gave', 'giving', 'get', 'getting', 'took', 'receive' or 'received' and no negation word, predict the tweet as positive.
2. Else, predict the tweet as negative.

2.3 Deep Learning-based Approach

We experiment with five typical deep learning-based models:

1. **Sentence Vector**: 200 dimensions; Logistic Regression. (*SV*)
2. **Dense Neural Network**: 64 dimensions, 1 inter. layer + 5 epochs (DNN)
3. **BiLSTM**: GloVe840B + 3 epochs + 50 lstm units + 0.25 dropout (*BiLSTM*)
4. **CNN**: GloVe840B + 5 epochs + 50 filters + 2 filter length + 0.75 dropout (CNN)
5. **LSTM-LM**: We pre-train a *language model* on the training dataset with a 3-layer LSTM. We then build a softmax layer on top of this pretrained LSTM, and fine-tune the neural network with supervision (Howard and Ruder, 2018).

All models are implemented using TensorFlow (Abadi et al., 2016). The parameters are experimentally determined.

Approach	F-Score	Accuracy
Skeppstedt et al. (2017)	76.84	87.01
Huang et al. (2017)	77.64	87.65
Statistical	80.75	**88.97**
Rule-based	40.48	64.91
SV	77.87	87.39
DNN	78.74	87.66
BiLSTM	78.30	87.30
CNN	78.40	87.60
LSTM-LM	**80.87**	88.94

Table 2: 10-fold cross-validation results (%) on the training dataset.

3 Experimental Setup

The shared task provided three labeled datasets of tweets for evaluation: a training dataset (5751 tweets of which 1692 are positive), a validation dataset (1215 tweets of which 306 are positive) and a test dataset (161 tweets, labels undisclosed).

We re-implement two past works as baselines (Skeppstedt et al., 2017; Huang et al., 2017). The two baselines use n-grams as features of statistical classifier.

4 Results

We present our results in six parts. We first describe the performances on the training, validation and test sets. Then, to understand the components contributing to the performance, we perform additional evaluation: (a) impact of the size of the training set on the performance; (b) impact of data source from which language model is trained in case of the deep learning approach; and (b) impact of the features on the performance of the statistical approach. Finally, we present an analysis of errors made by our system.

4.1 Performance on Training Set

The performance of our methods using 10-fold cross-validation is shown in Table 2. The performance of the re-implementation of baselines are comparable to the original papers. The low values in case of the rule-based approach highlight that vaccination behaviour detection is not a trivial task of detecting words that indicate administration of a vaccine. The best F-scores are achieved by the statistical approach (80.75%) and LSTM-LM (80.87%). This is an improvement of 3-4% over the baseline.

Approach	F-Score	Accuracy
Statistical	86.06	85.71
LSTM-LM	88.74	89.44

Table 3: Performance (%) on the test dataset.

	Statistical	LSTM-LM
20%	73.59	77.69
40%	75.17	78.58
60%	79.26	78.95
80%	80.54	79.52
100%	81.56	80.43

Table 4: F-scores (%) of the two best-performing approaches for varying size of the training set.

4.2 Performance on Validation and Test Sets

The statistical approach achieves an average F-score of 81.56%, while LSTM-LM achieves 80.43% on the validation set. Similarly, the performance of our methods on the test dataset is in Table 3. We obtain a F-score of 86.06% with the statistical approach and 88.74% with the LSTM-LM on the test set of 161 instances.

4.3 Impact of Size of the Training Set

To analyse the impact of the training set size on the resultant performance, we show the F-scores for the two best approaches for varying sizes of the training set in Table 4. '20%' indicates that 20% of the training set was used to train the system while the validation set was used for evaluation. We observe that when training on a small size of labeled data, LSTM-LM performs much better than statistical model. This shows the benefit of transfer learning that it can utilize knowledge learned from unlabeled data to train a model with a small number of labeled instances.

4.4 Impact of Language Model Source in LSTM-LM

A pre-trained language model represents knowledge learned from source data that is applied to a classifier. To understand if the domain of this source data has an impact on the performance of the resultant classifier, we compare how effective different domains are for vaccination behaviour detection. We compare three datasets in Table 6. The SMM4H dataset is the training dataset for the task while WikiText-103 (Merity et al., 2016)

Feature	ΔF-score (%)
POS	1.16
Special characters	0.97
Negation	0.66
Word similarity	0.15
Sentence vector	0.20
Length	0.39
Emotion	0.33

Table 5: Degradation in F-scores (%) of the statistical approach when each of the features is removed.

Source data	# of tokens	F-score
WikiText-103	101M	80.84 (± 0.37)
IMDB	17M	81.15 (± 0.83)
SMM4H	884K	80.43 (± 0.67)

Table 6: F-scores (%) of the LSTM-LM when language model is pretrained on different source data.

and IMDB (Maas et al., 2011) are datasets from wikipedia and a movie review corpus respectively. The latter are significantly larger than the SMM4H dataset. However, they only result in a marginally higher performance.

4.5 Impact of Features in the Statistical Approach

To understand how the features contribute to the statistical approach, we conduct ablation tests. The degradation in F-score when each of the features is removed is in Table 5. The positive values in all fields validate the value of the proposed features. The highest degradation is observed in case of POS-based features.

4.6 Error Analysis

We analyse incorrectly predicted instances from the validation set. About 50% of errors have first or second person pronouns. Nearly 44% of false

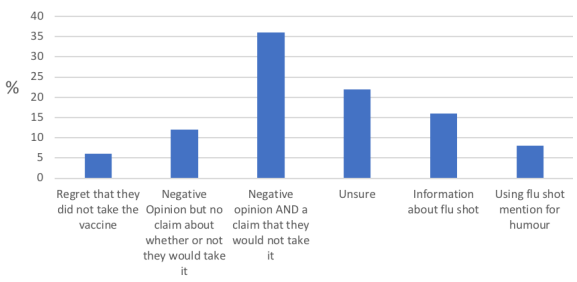

Figure 1: Sources of errors in false positives.

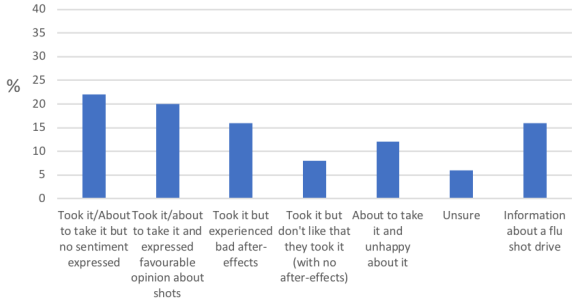

Figure 2: Sources of errors in false negatives.

negatives have negative sentiment about flu shots because of actual or expected, unpleasant side-effects. The ratio of false negatives to false positives is 1.40. An analysis of 50 random false positives and 50 random false negatives are shown in Figures 1 and 2 respectively. The label '*Unsure*' indicates that the error could not be assigned to any of the other categories. Some incorrectly classified instances for the different error sources are:

- Negative opinion but no claim whether they would take it, as in the case of '*Getting a flu vaccine after reading this article is crazy!*'.
- Mentions of taking a flu shot without expressing sentiment, such as '*Flu shots for hubby and daughter... check.*'.
- Took it or about to take it and expressed favourable opinion about shots, as in the case of the tweet '*We're headed to the @Brigham-Womens flu shot clinic! Getting vaccinated is good for you and your community.*'.

5 Conclusions

We evaluate three text classification approaches for the task of vaccination behaviour detection. The rule-based approach considers simple presence of words, the statistical approach uses an ensemble of classifiers and task-specific features while the deep learning approaches employ five neural models. On comparing the three approaches, we observe that an ensemble of statistical classifiers using task-specific features and a deep learning model using pre-trained language model and LSTM classifier obtain comparable performance for vaccination behaviour detection. Our findings in the error analysis which show that vaccine hesitancy often conflicts with vaccination behaviour detection, will be helpful for future work.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, Savannah, GA.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *ACL*, page 31, Barcelona, Spain.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI*, Quebec, Canada.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

D Holt, Fredric Bouder, C Elemuwa, G Gaedicke, A Khamesipour, B Kisler, S Kochhar, R Kutalek, W Maurer, P Obermeier, et al. 2016. The importance of the patient voice in vaccination and vaccine safetyare we listening? *Clinical Microbiology and Infection*, 22:146–153.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Xiaolei Huang, Michael C Smith, Michael J Paul, Dmytro Ryzhkov, Sandra C Quinn, David A Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *AAAI Joint Workshop on Health Intelligence*, pages 542–546, San Francisco, CA.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR abs/1609.07843*.

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 26*, pages 3111–3119.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, Doha, Qatar.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2017. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media*, pages 1–8, Taipei, Taiwan.