Predicting Vaccine Hesitancy and Vaccine Sentiment using Topic Modeling and Evolutionary Optimization

Gokul S Krishnan[1,2⋆], Sowmya Kamath S[1] and Vijayan Sugumaran[3]

[1] Healthcare Analytics and Language Engineering (HALE) Lab, Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India
[2] Robert Bosch Centre for Data Science and Artificial Intelligence, Indian Institute of Technology Madras, India
[3] Department of Decision and Information Sciences, School of Business Administration, Oakland University, USA
gsk1692@gmail.com, sowmyakamath@nitk.edu.in, sugumara@oakland.edu

Abstract. The ongoing COVID-19 pandemic has posed serious threats to the world population, affecting over 219 countries with a staggering impact of over 162 million cases and 3.36 million casualties. With the availability of multiple vaccines across the globe, framing vaccination policies for effectively inoculating a country's population against such diseases is currently a crucial task for public health agencies. Social network users post their views and opinions on vaccines publicly and these posts can be put to good use in identifying vaccine hesitancy. In this paper, a vaccine hesitancy identification approach is proposed, built on novel text feature modeling based on evolutionary computation and topic modeling. The proposed approach was experimentally validated on two standard tweet datasets – the flu vaccine dataset and UK COVID-19 vaccine tweets. On the first dataset, the proposed approach outperformed the state-of-the-art in terms of standard metrics. The proposed model was also evaluated on the UKCOVID dataset and the results are presented in this paper, as our work is the first to benchmark a vaccine hesitancy model on this dataset.

Keywords: Evolutionary Computation · Machine Learning · Natural Language Processing · Population Health Analytics · Topic Modeling

1 Introduction

In the last few decades, the world has faced several epidemics and contagious viral diseases such as SARS, MERS, H1N1, Zika, Ebola etc, currently superceded by the "once-in-a-century pandemic", COVID-19 [17]. Organizations such as WHO and health governing bodies of most countries have a huge task of keeping their population healthy. During critical situations like pandemics like the ongoing

⋆ Work done as part of doctoral research work at HALE Lab, NITK Surathkal.

COVID-19 crisis, the world community has allocated huge amount of financial, research and human resources towards developing effective vaccines for managing disease outbreaks through structured vaccination of vulnerable population groups. Although the success of vaccines in disease control have been proven time and again, making them the obvious and successful measure for managing contagious disease outbreaks, it is quite unfortunate that a growing number of people deem it unnecessary, "against the natural order" and unsafe [4].

Public opinion on vaccinations can be diverse - e.g., majority of the population may be voluntarily ready to submit to the vaccination shots, whereas a significant number may be skeptical about it, despite strong recommendations from the medical community. Vaccine hesitancy is one of the most critical factors that affect effective vaccination policies, owing to the lack of confidence, disinclination or negative opinion towards a vaccine [9]. Vaccine hestinacy has resulted in reduced vaccination coverage and increased risk of epidemics and disease outbreaks that are often easily preventable via mass vaccination [4]. With the widespread adoption of Open Social Network (OSN) platforms such as Twitter and Facebook, such negative opinions can have a negative impact on the efforts of governments and public health organizations. Therefore, it is very important for public health governing national bodies to understand the prevalence of vaccine hesitancy in populations and public sentiment towards vaccination programmes. In normal cases, the public opinions are recorded through surveys and interactive programmes, which are not only difficult to organize and time-consuming, but also tend to under-represent all kinds of citizens, and hence may not be generalizable [6, 13].

Automated computational population health surveillance systems that can be modeled to identify vaccine hesitancy is a potentially advantageous solution to these challenges as mining OSN data can provide essential insights to health governing bodies for making informed and possibly better decisions. In this paper, we present a vaccine hesitancy prediction model that can effectively detect vaccine hesitancy in public based on OSN media posts. The proposed approach leverages the concepts of evolutionary computation (Particle Swarm Optimization (PSO)), topic modeling (Latent Dirichlet Allocation (LDA)) and neural networks like Convolutional Neural Networks (CNN) to achieve this objective. The key contributions of this work are as follows:

1. Design of a PSO based topic modeling approach that can dynamically determine the optimal number of latent topic clusters, for OSN data.
2. Design of a PSO-CNN wrapper for dynamically determining the optimal number of topics for LDA topic modeling and for effectively identifying any vaccine hesitancy.
3. Benchmarking the proposed vaccine hesitancy identification approach on open standard datasets.

The rest of the paper is organized as follows. Section 2 provides an overview on existing related works in the domains of interest. Section 3 discusses in detail the system architecture of the proposed model. In Section 4, we present the

experimental results along with a discussion on the performance of the approach, followed by conclusion and potential scope for future work.

2 Related Work

The research community has shown significant interest in modeling OSN data for a wide variety of tasks. We discuss some relevant works in each of these categories, in the context of the chosen tasks. Computational techniques like Natural Language Processing (NLP) and ML have great potential in performing predictive analytics based tasks on OSN data. Several works in the areas of influenza or flu monitoring/detection [1, 2, 16], adverse drug event detection [3, 14], vaccine sentiment [6], vaccine behaviour / vaccine shot detection (whether vaccine shot was received or not) [6, 9], vaccine hesitancy/vaccine intent (whether vaccine is intended to be taken or not) [6], etc, have been proposed over the past decade. Huang et al. [6] presented a study that made use of several natural language classifiers to analyze Twitter users' behavior towards influenza vaccination. They performed prediction tasks such as vaccine relevance, vaccine shot detection, vaccine intent detection and vaccine sentiment.

Moslehi and Haeri [12] proposed a hybrid method based on PSO, Genetic Algorithm (GA) and gain ratio index to select optimal feature subsets. Gomez et al. [5] proposed a GA based evolutionary approach for learning some metarules which can help further optimize text classification. While these approaches showed capabilities of evolutionary compuation being applied towards text classification, these approaches fail to extract effective feature representations for a specific prediction task.

Li et al. [11] proposed an auxiliary word embedding based topic modeling approach for text classification. Steinskog et al. [15] proposed a topic modeling and pooling techniques based approach for aggregation of tweet texts. While these approaches showed the effectiveness of topic modeling in NLP tasks, other approaches (by Zhao et al. [18] and Ignatenko et al. [8]) were put forward by to determine the number of topic clusters, a known research problem in topic modeling techniques. Though these approaches could determine a certain number of topics for topic modeling, the choice is not based on the prediction task to be performed. In this paper, we propose the use of PSO, an evolutionary optimization algorithm, and ensemble it with a wrapper technique based on CNN to determine an optimal number of topic clusters for LDA topic modeling technique and use it to effectively model features for training a vaccine hesitancy identification model.

3 Proposed Approach

The overall workflow of the proposed approach is depicted in Fig. 1. We used standard datasets consisting of OSN data for the experiments. A preprocessing pipeline involving several basic NLP techniques were used to clean and preprocess the corpus. All special characters except white spaces were removed. Tokenization was performed on the corpus to break down the text into units

called tokens; stemming and lemmatization were applied to bring the words to root form and finally, stopping was also performed to filter out frequent unimportant words. The next set of processes that involve the feature modeling and prediction modeling are explained in subsequent sub sections.

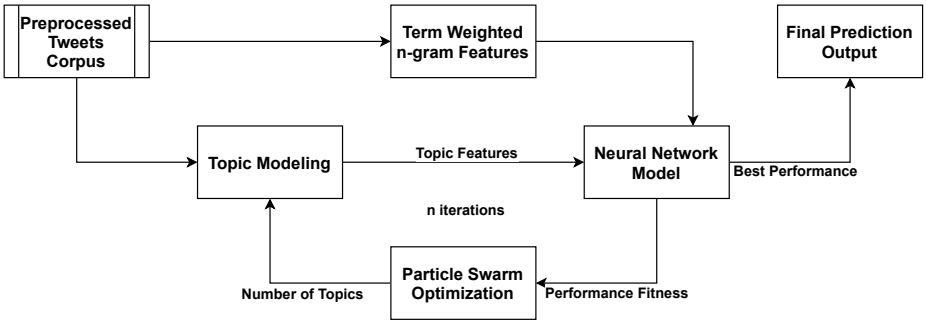

Fig. 1: Workflow of Proposed Vaccine Hesitancy Identification Approach

Term Weighted n-gram Feature Generation. The preprocessed tokens obtained from the tweets corpus were modeled into a vector representation using a Term Frequency - Inverse Document Frequency (TF-IDF) vectorizer to create term weighted n-gram features. TF-IDF, a statistical measure that signifies the weightage or importance of words within a document, is often considered as textual features in text mining based prediction modeling. In the proposed approach, TF-IDF weights for n-grams (i.e., n = 1, 2, 3) were extracted and the top 2000 weights were considered to be Feature Set 1 (hereafter referred to as FS1).

Latent Dirichlet Allocation. Topic modeling is an approach that clusters documents into a set of topics that most represent the documents in an unsupervised manner. Latent Dirichlet Allocation (LDA) is a popular probabilistic topic modeling approach that assigns a given set of documents to topic clusters. LDA theorizes that for a set of words appearing in a given document, it belongs to a certain number of topics with certain probabilities. In this work, LDA is applied to the preprocessed tweets corpus, and vectorized probabilities of each topic for a document are considered as features for the proposed prediction model, i.e., Feature Set 2 is hereafter referred to as FS2.

Similar to unsupervised clustering techniques, determining the number of topics while performing the LDA topic modeling approach is a challenging and critical task. Determining the optimal number of LDA topic clusters also pertains to the process of deriving the optimal number of features in the topic feature vector generated by the topic model. The solution subspace to search for the optimal number of topics is quite large and therefore, evolutionary optimization approaches are an apt choice. We adopted the usage of PSO algorithm for this and the adaptation is explained in detail next.

PSO based LDA Topic Modeling. We utilize PSO, an evolutionary optimization algorithm, for dynamically determining the optimal number of topics for various prediction tasks. Towards this objective, a wrapper, named PSO-CNN, is proposed. The feature sets – FS1 and FS2 are combined and fed into a PSO-CNN wrapper, in which the neural network model was adopted from the popular TextCNN model [10] for effective text classification. The performance of the TextCNN model in terms of F-score was considered as the fitness performance for the proposed PSO based topic modeling approach.

Initially, a swarm of particles, along with particle positions were initialized as a set of number of topic clusters for the LDA model. For each position, say i, the best classification performance of the TextCNN model in terms of F-score is considered as the $local\_best_i$ score, and the same of the entire swarm is considered as the $global\_best$ score. The new next positions, x_{i+1}, and next velocities, v_{i+1}, of the initalized particles are calculated and updated based on PSO equations (Eq. 1 and 2), where, $c1$ and $c2$ are constants, whose values were empirically found to be 0.5 and 0.2 respectively and $r1$ and $r2$ are random real numbers.

$$v_{i+1} = w * v_i + c_1 * r_1 * (local\_best_i - x_i) + c_2 * r_2 * (global\_best - x_i) \quad (1)$$

$$x_{i+1} = x_i + v_{i+1} \quad (2)$$

In our work, a set of eight particles were used and number of iterations was set to 50. The position at which the best performance was observed, i.e., the position of $global\_best$, was considered to be the optimal number of topic clusters for LDA topic modeling for the task of vaccine hesitancy identification.

The TextCNN model adopted in the proposed PSO-CNN wrapper model consists of three 1D convolution layers with 512 filters which were of sizes 5,6 and 7 respectively. The number of nodes in the input layer indicates the optimal number of topics as determined by the PSO based topic modeling approach. Further, 1D Maxpool layers and a Rectified Linear Unit (ReLU) activation function were used along with each convolution layer, followed by concatenating and flattening layers. Additionally, a 50% dropout was also introduced to reduce chances of overfitting. Finally, the output layer consisted of a sigmoid activation function. The optimizer used for training was rmsprop and the loss function used was binary cross-entropy. The performance for vaccine hesitancy prediction task was extensively tested, the details of which are presented in Section 4.

4 Experimental Results and Discussion

The performance of the proposed vaccine hesitancy identification approach was benchmarked on two standard tweet datasets, created specifically for vaccine hesitancy tasks. The performance was measured using standard classification metrics – precision, recall and F-score. The proposed model was also benchmarked against state-of-the-art approach for one dataset.

Datasets. We used two datasets to benchmark the performance of the proposed vaccine hesitancy idetification model. First, the flu vaccine dataset (FVD) provided by Huang et al. [6] was used for this prediction task which consists of

around 10,000 tweets related to influenza vaccine. It is to be noted that, the irrelevant tweets labelled in the dataset and any rows with missing labels were dropped, after which a total of 9,513 instances were available for the analysis. Second, the UK COVID-19 Vaccine tweets dataset (hereafter referred as UKCOVID) collected and released by Hussain et al. [7] was also used for the experiments. The dataset originally consists of 40,268 tweets from the United Kingdom with respect to the context of vaccination for COVID-19 pandemic. The original dataset released by the authors consisted of only tweet IDs as per policy of Twitter. However, only 24,309 tweets could be retrieved due to issues such as deleted tweets or private accounts. The vaccine hesitancy is indicated as sentiment labels for tweets in three categories – positive, negative and neutral. The characteristics of the two datasets are as shown in Table 1a and 1b.

Table 1: Dataset Statistics

(a) Dataset Statistics of FVD

Feature	**Frequency**
Unique tweets	9,513
Users	9,334
Words	1,54,204
Intend/Receive *(Positive)*	3,148
Hesitancy *(Negative)*	6,365

(b) Dataset Statistics of UKCOVID

Feature	**Frequency**
Unique tweets	24,309
Words	63,863
Positive	10,230
Negative	8,245
Neutral	5,834

Results. When applied on FVD dataset, the PSO-CNN wrapper based topic modeling technique determined the optimal number of LDA topic clusters to be 634. Along with 2,000 top n-gram features, total number of textual features came to 2,634. The performance of the proposed approach was compared to that of the state-of-the-art approach by Huang et al. [6]. Similar to their approach, the performance of the proposed approach was also measured after the 5-fold cross validation. The comparison of performance is as shown in Table 2, from which, it can be observed that the proposed approach outperformed Huang et al's approach in terms of Recall and F-score by 5% and 2% respectively. Higher values of recall and F-score indicate that the proposed approach was able to reduce the number of False Negatives (FNs).

The proposed approach was also applied on the UKCOVID dataset, and the optimal number of LDA topic clusters were determined to be 600. Using 2,000 top n-gram features, total number of textual features used were about 2,600. The performances in terms of precision, recall and F-score has been benchamrked for this dataset and the results are shown in Table 2. The dataset is quite recent and therefore, no other works have benchmarked any performance on this dataset yet due to which we did not perform any comparison. The classification performance in terms of Recall and F-score shows that there is scope for improvement. This is due to the misclassification of true neutral sentiment as either positive or negative. This is one of the limitations of the current model, towards which we plan to design techniques as part of future work.

Table 2: Flu Vaccine Hesitancy: Performance of Proposed Approach

Approach	**Dataset**	**Precision**	**Recall**	**F-Score**
Huang et al. [6]	FVD	0.84	0.80	0.82
LDA+PSO+TextCNN *(Proposed)*	FVD	0.84	0.84	0.84
LDA+PSO+TextCNN *(Proposed)*	UKCOVID	0.75	0.51	0.60

Discussion. From Table 2, it can be observed that the proposed vaccine hesitancy identification approach outperforms the existing approach by Huang et al. Huang et al. [6] by 2% in terms of F-score. As the proposed approach is an entirely text-dependent model, it is able to 'understand' the natural language text and figure out the vaccine hesitancy sentiment of the user, which demonstrates its suitability for quantifying vaccine hesitancy sentiment. The performance benchmarking on the UKCOVID dataset not only ensures future research promotion, but also highlights an approach that can be put to use in the current real world scenario of the COVID-19 pandemic.

5 Conclusion and Future Work

In this paper, a novel approach leveraging topic modeling and evolutionary optimization for predicting vaccine hesitancy and identifying negative sentiments towards vaccination using OSN data has been proposed. The proposed approach is built on effective usage of the PSO algorithm and LDA topic modeling approach, along with CNN based prediction model to identify vaccine hesitancy in tweets by users. Experimental validation revealed that the proposed approach outperformed state-of-the-art approaches on the FVD dataset. In addition, the performance of the same on the newly released COVID-19 based UKCOVID dataset was also benchmarked. The purely natural language text dependent model proved to be effective in identifying vaccine hesitancy and can be considered as a tool to identify population sentiment towards COVID-19 vaccines in the current pandemic scenario. As part of future work, we plan to explore further improvements for effective identification of neutral sentiment towards vaccine policies. We further intend to benchmark the proposed approach on more diverse datasets. Moreover, we also plan to explore the use of other topic modeling approaches and also word embedding approaches as part of textual feature modeling.

References

[1] Alshammari, S.M., Nielsen, R.D.: Less is More: With a 280-character limit, Twitter Provides a Valuable Source for Detecting Self-reported Flu Cases. In: Proceedings of the 2018 International Conference on Computing and Big Data. pp. 1–6. ACM (2018)

[2] Byrd, K., Mansurov, A., Baysal, O.: Mining Twitter data for influenza detection and surveillance. In: Proceedings of the International Workshop on Software Engineering in Healthcare Systems. pp. 43–49. ACM (2016)

[3] Cocos, A., Fiks, A.G., Masino, A.J.: Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. JAMIA 24(4), 813–821 (2017)
[4] Dubé, E., Laberge, C., Guay, M., Bramadat, P., Roy, R., Bettinger, J.A.: Vaccine hesitancy: an overview. Human vaccines & immunotherapeutics 9(8), 1763–1773 (2013)
[5] Gomez, J.C., Hoskens, S., Moens, M.F.: Evolutionary learning of meta-rules for text classification. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. pp. 131–132 (2017)
[6] Huang, X., Smith, M.C., Paul, M.J., et al.: Examining patterns of influenza vaccination in social media. In: Workshops at 31st AAAI Conference on Artificial Intelligence (2017)
[7] Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., et al.: Artificial intelligence-enabled analysis of uk and us public attitudes on facebook and twitter towards covid-19 vaccinations (2020)
[8] Ignatenko, V., Koltcov, S., Staab, S., Boukhers, Z.: Fractal approach for determining the optimal number of topics in the field of topic modeling. In: Journal of Physics: Conference Series. vol. 1163. IOP Publishing (2019)
[9] Joshi, A., Dai, X., Karimi, S., Sparks, R., Paris, C., MacIntyre, C.R.: Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In: Proceedings of the 2018 EMNLP Workshop. pp. 43–47 (2018)
[10] Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
[11] Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: ACM SIGIR conference on Research and Development in Information Retrieval. pp. 165–174 (2016)
[12] Moslehi, F., Haeri, A.: An evolutionary computation-based approach for feature selection. Journal of Ambient Intelligence and Humanized Computing pp. 1–13 (2019)
[13] Parker, A.M., Vardavas, R., Marcum, C.S., Gidengil, C.A.: Conscious consideration of herd immunity in influenza vaccination decisions. American journal of preventive medicine 45(1), 118–121 (2013)
[14] Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: a review. Journal of biomedical informatics 54, 202–212 (2015)
[15] Steinskog, A., Therkelsen, J., Gambäck, B.: Twitter topic modeling by tweet aggregation. In: Proceedings of the 21st nordic conference on computational linguistics. pp. 77–86 (2017)
[16] Wakamiya, S., Kawai, Y., Aramaki, E.: Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. JMIR public health and surveillance 4(3), e65 (2018)
[17] Yang, H., Ma, J.: How an epidemic outbreak impacts happiness: Factors that worsen (vs. protect) emotional well-being during the coronavirus pandemic. Psychiatry research 289, 113045 (2020)
[18] Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W.: A heuristic approach to determine an appropriate number of topics in topic modeling. In: BMC bioinformatics. vol. 16, p. S8. Springer (2015)