ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter60892020.pdf
https://doi.org/10.30534/ijeter/2020/60892020

Sentiment Analysis of Social Media Data in Vaccination

N. H. Abd Rahim[1], S. H. Mohd Rafie[2]
[1] Faculty Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia, noorhafhizah@umt.edu.my
[2] Faculty Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia, hajarrafie98@gmail.com

ABSTRACT

Sentiment analysis helps doctors and researchers to identify the cause of vaccine hesitancy among societies by distinguishing between opinions expressed through social media. As time goes by, various views related to vaccination have been discussed through social media platforms such as Twitter, Facebook, and others. The real question is how we interpret the opinion given and how vaccine resistance turns out to be an issue in society. The Support Vector Machine (SVM) classifier has been used to classify the sentiment in this study. A few keywords related to vaccination are used to extract the Twitter data. The data is crawled from the 30th of October 2019 until the 30th of March 2020. As our data provides three categories of polarity, the classifier needs to train more than two classes, which brings to the use of two types of kernels to handle our data. The kernels are polynomial and RBF kernels. Experiment results indicate RBF kernel has higher accuracy compared to the polynomial kernel.

Key words: Sentiment Analysis, Machine Learning, Vaccination, Twitter, Social Media

1. INTRODUCTION

The immune system in humans is indispensable for longevity in the human body. They try to keep track of supposed microorganisms or pathogens to get rid of them, but sometimes they stray from their usual means. Therefore, in 1796, the first vaccine was introduced to fight smallpox to help the immune system to get on track detecting the pathogens [1]. As the years passed by, numerous vaccines were developed to fight serious illness; for example, Hepatitis B, Rotavirus, Diphtheria and Human papillomavirus (HPV).

As the first child of a family is born, the parent's minds will be wondering whether vaccination is vital for their precious child. Although doctors have advised them about how vaccination can help to prevent many health problems, some of them are still reluctant to get the vaccine for their children due to the possible side effects and religious beliefs. Hence, this situation is considered as 'vaccine hesitancy' [2].

The rejection of some parts of the world towards vaccination has become a long term concern for doctors, especially when the number of cases for polio increases. Through social media, specifically Twitter, sentiments are sent by account users as 'tweets' to express their opinions regarding this matter. Twitter is a social platform that contributes to the Web 2.0 where it provides interaction between users in the form of short messages that usually contains from personal thoughts to public statements [3].

Figure 1 shows the example of a typical tweet on Twitter that is written by a user via personal computer. Twitter audience is represented by users from various countries which makes it possible to collect data in any languages. Furthermore, a wide range of people uses Twitter from ordinary users to celebrities, influencers, politicians and even country presidents. It allows the possibility to collect data from users with different backgrounds [3].

In this paper, we develop a model that uses SVM classifier to classify the polarity of sentiments. Two kernels have been used in the classifier; namely polynomial and RBF kernels. The rest of the paper is organised as follows. Section 2 describes the previous works related to sentiment analysis of social media data that have been done. Section 3 mentions the research framework, and Section 4 discusses experiments, results, and discussion. In the final section, it concludes the research.

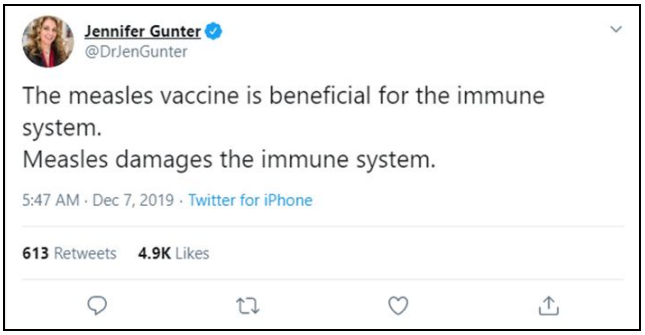

Figure 1: Example of a Twitter Post

2. RELATED WORKS

During the emerging of Web 2.0, the usage of Web as a social platform for generating content grew with the introduction of microblogs such as Facebook, Twitter and others [4]. This phenomenon encourages people to share information among themselves on any topics they need. They also have the freedom to publish their opinions and voice them out. These opinions are often used by marketers to identify the feedback of their product in the community. However, with the rapid growth of blogs and posts, it is impossible to keep up with the information to monitor. Thus, the implementation of sentiment analysis is much required. The factor behind the idea of sentiment analysis is because of the rise of machine learning in processing data retrieved. Sentiment analysis gives the concept of using intelligence towards data to produce broader types of information from a single data.

Vaccination has come to a point where hesitancy has become a concern to the doctors and medical researchers. Vaccine hesitancy happens when parents doubt deciding to vaccinate their children. A wide variety of variables was discussed, but it came to a point where [5] emphasised that there was insufficient evidence to help doctors and researchers find a solution to this reluctance. When there are no standardised tool to measure vaccine hesitancy that can be used, they developed a series of survey questions to improve the measurement of hesitancy. However, this action takes a massive amount of time and labour work to achieve such data. With sentiment analysis, we could achieve broad ways of obtaining data to analyse the polarity of the opinion made by community.

As sentiment analysis was introduced during the emerge of Web 2.0, it is important to stress that sentiment analysis is applicable to be applied towards the data from social media. [6] is an example of sentiment analysis study which uses Twitter data. As can be seen in Table 1, a few approaches have been used to perform sentiment analysis for vaccination. A paper by [7] suggested the usage of Facebook in identifying the landscape of anti-vaccination sentiment on Facebook. With 197 user accounts observed, they obtained the visualisation of the network representing Facebook profiles which discussed about vaccine which provides insight of the communication between the pro- and anti- vaccine users. However, they were not able to retrieve information from private account users. The reason behind this is because of the ethical procedures that are needed to follow when mining data from social medias [8]. Another researcher based on Table 1, [9] uses Youtube as their social media platform to research identifying the expression of pro- and anti-vaccine sentiment by using Natural Language Processing tool called cleanNLP. It identifies 275 videos through its likeability by immunisation advocacy category and ranking of word frequencies which was able to identify the trends that grew among anti-vaccine users. But, using Youtube lacks the precision to identify perspectives from various regions. On the other hand, using Twitter can pinpoint the geographic detail of a Tweet or a user to ensure better comprehensive coverage of content. Hence, to fill in the research, we initiate the sentiment analysis of vaccination through Twitter. Although Twitter can be the source of sentiments for a study, it always depends on the methodology applied to the data.

The next research is implemented by [10], where they developed two approaches towards classifying the stance towards vaccination focusing on Twitter users. With a sample size of 95,566 tweets, the study stated that they have used Multinomial Naïve Bayes and SVM algorithm approach in analysing the samples to come to a conclusion on which approach is the best for future use of gathering opinions using data. With a clear aim in improving performance, [10] achieved an outcome where using SVM classifiers gives out the best performance in differentiating unnecessary tweets and polarity changes. This study gives a point that using SVM algorithm would give a slight difference and a better performance to study the opinions of people in vaccination. Unfortunately, the machine was able only to be trained to handle a small amount of data.

Another study conducted by [11] expressed an opinion which influences among pro-vaccine and anti-vaccine users in Twitter plays an essential role in analysing the sentiment of vaccination. By having a sample size of 669,136 tweets, they identified the communicative patterns between both types of users and how they influenced each other by using linear SVM algorithm to have a better structured data. From the data they obtained, they found out that anti-vaccine Twitter users are a growing and enclosed community, which makes it difficult for health organisations to get useful information. However, they stated that emojis were not taken into consideration during the study when these characters are also essential for identifying sarcasm in Tweets. Emojis are essential as people constantly use them in expressing opinions especially in identifying if they are a part of the pro-vaccine or anti-vaccine users. It might be ambiguous, but it is important to identify where the emojis are taken account from [12].

[13] suggested in their research that the opinion of a Twitter user can drift based on a specific event that they experience. They used 112,397 tweets from a period of the 1st of September 2016 until the 30th of June 2017 studying the possibilities of people having opinions on multiple events happened in that period of time. To their observation for 10 months they gathered, it was found that events happening regarding to vaccination can change the amount of Twitter users having interest on vaccine related news. Each event left an impact on Twitter users to leave their own opinion to it whether the opinion is positive or negative in terms of polarity. It gives an idea in doing this research to identify specific dates in recent years which have events contributing to vaccination. However, this approach would be a concern when it comes to dealing with multiple concept drifts especially when the scope of the research is wide.

Table 1: Comparison of Some Existing Works Related to Vaccine Sentiment

Study	**Methods Applied**	**Source of Sentiments**	**Appli-cation Fields**	**Measurement Metrics**	**Benefits**	**Limitations**
[7]	Quantitative coding, descriptive analysis, social network analysis, visualisation of network and in-depth qualitative assessment.	Social Media (Facebook) (197 user accounts)	Health	-Frequency of anti-vaccination post by category -Visualisation of the network representing Facebook profiles discussing vaccine topics.	-Provides insight of the communicati on between the pro and anti-vaccine users through visualisation of network.	-Cannot retrieve the information from private account users. -Cannot ensure the authenticity of information.
[9]	Natural Language Processing using cleanNLP	Social Media (YouTube) (275 videos using Google's automatic caption track)	Health	-Video likeability by immunization advocacy category - Ranking of word frequencies by advocacy category	-Able to identify the trends that grew among anti-vaccine users.	-Lacks precision to identify perspectives from various regions
[10]	Multinomial Naive Bayes and SVM algorithm	Social Media (Twitter) 95,566 Tweets	Health	-Machine learning performance on predicting the label of tweets with negative stance.	-Outperforms the general-purp ose sentiment analysis tools.	-Machine only trained on a small amount of data.
[11]	Sentiment Analysis: Machine Learning Approach (Linear SVM algorithm)	Social Media (Twitter) 669,136 Tweets	Health	-Distributions of each opinions in every structural community by network visualisation, percentage and frequency	-Can identify the interaction between anti and pro vaccine users.	-Emoji are not taken into considerations
[13]	Sentiment Analysis: Machine Learning Approach (SVM algorithm)	Social Media (Twitter) 112,397 Tweets	Health	-Distribution of opinion polarity by events and month.	-Ability to relate the pattern of vaccination sentiment with events.	-Concern on dealing with multiple concept drifts

3. METHODOLOGY

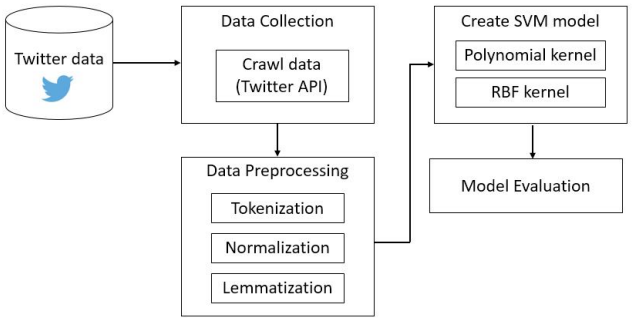

Figure 2: The Framework of Analyzing Vaccine Sentiment Through Twitter

In this study, it is essential to classify the opinions which are expressed by Twitter users to identify who are pro-vaccination and anti-vaccination community. Figure 2 shows the research framework for analysing the sentiment.

The first stage in the framework starts with collecting data from Twitter. Data collection is performed by using a library in Python called Tweepy. Tweepy is a library which is used to access the Twitter API. In any cases, it can be used to create simple automation and also Twitter bots. As Twitter is a platform where users can find tweets based on keywords, it is applicable to use it through Tweepy. The most important part in accessing Tweepy library is to have access and consumer keys where it can be found when we registered at a Twitter Developer for using Twitter API. Tweepy accepts parameters and provide Twitter account's data as a return.

A few keywords related to vaccination are used to extract the data from Twitter referring to Table 2 from the 23rd of November 2019 until 15th of May 2020. The data is then manipulated in spreadsheets using another library called Panda where each data is divided into separate fields. Panda is used to manipulate the data from Twitter into a proper table. The gathered tweets are numbered 100,000 according to benchmark.

The next following stage is data preprocessing. In this phase, there are three processes involved; namely tokenization, normalization, and lemmatization.

Table 2: Keywords Used in Crawling Data from Twitter

Keywords
vaccine , vaccination , polio vaccine, polio, measles, mumps, rubella, flu shots, hpv vaccine, hpv

3.1 Tokenization

Tokenization is used to separate sequence of strings from the Tweets into pieces such as words, phrases, and symbols as stated in Table III. This step is important in separating the Twitter handler of accounts and links from the Tweet. After the separation, the Twitter handler will be removed.

3.2 Normalization

Stop words, punctuations, uppercase letters are removed during normalization. Stop words are commonly used words ("the", "a", "an"," in"). Looking at the tweets, the possibility of handling hashtags is big therefore explains why normalization is indeed important in this procedure. Referring to Table III most punctuations and stop words are removed, resulting in a few words left for later lemmatization.

3.3 Lemmatization

Lemmatization is where the words are generated to their root forms. Unlike stemming, lemmatization preserves the word's POS (part-of-speech) without dividing the suffixes [14]. For example, the word 'caring' changes to 'care' through lemmatization unlike stemming which turns to 'car'. Looking through Table 3 most words like 'jabbed' and 'means' from previous step are changed into their root words.

After the data preprocessing stage, then the Support Vector Machine (SVM) model is created. In this study, we use the algorithm to perform the classification of data. According to [10], one of the advantages of using SVM is that it provides a solution for an extensive optimization problem, hence decreases the generalization error of a classifier. To classify the data, there are three classes which are positive, negative and neutral class. The modules involved in the process are Natural Language Tool Kit (NLTK) and Text Blob.

Furthermore, SVM (Support Vector Machine) classification algorithm provides one of the highest accuracy performance among the other classifiers used in previous research [15]. In performing SVM classification method, it requires drawing a hyperplane on the data to separate the data into classes. This is helped by using a mathematical function which is kernel. There are various types of kernels that can be used in sentiment analysis which are linear, sigmoid, Radial Basis Function (RBF) also known as Gaussian, non-linear and polynomial that are the most popular ones used. In this study, two types of kernels are tested which are polynomial and RBF to find out which kernel is able to reach the highest accuracy to be applied solely in the system. Both are able to classify more than two classes unlike linear and sigmoid SVM [15].

Table 3: Text Representation Through Preprocessing

Step	Features
Before Tokenization	@DrAmalinaBakri @KKMPutrajaya @DrDzul @SyedSaddiq I'm not antivaxx but if tumor's cell used to create polio vaccine means live potential cancer is jabbed into human body, is it safe ? science need to prove it.
After Tokenization	'I', '", 'm', 'not', 'antivaxx', 'but', 'if', 'tumor', '", 's', 'cell', 'used', 'to', 'create', 'polio', 'vaccine', 'means', 'live', 'potential', 'cancer', 'is', 'jabbed', 'into', 'human', 'body', ',', 'is', 'it', 'safe', '?', 'science', 'need', 'to', 'prove', 'it', '.'
Normalization	'antivaxx', 'tumor', 'cell', 'used', 'create', 'polio', 'vaccine', 'means', 'live', 'potential', 'cancer', 'jabbed', 'human', 'body', 'safe', 'science', 'need', 'prove'
Lemmatization	'antivaxx', 'tumor', 'cell', 'use', 'create', 'polio', 'vaccine', 'mean', 'live', 'potential', 'cancer', 'jab', 'human', 'body', 'safe', 'science', 'need', 'prove'

At final stage, the model is evaluated on its performance. Each kernel is compared with four parameters; namely precision, recall, f1-score, and accuracy.

4. RESULT AND DISCUSSION

By using Tweepy to access the Twitter API, a total of 105,965 tweets are collected from 23rd of November 2019 until 15th of May 2020. However, about 28.5% of Twitter are removed as they are not suitable to be used in this study. Textblob allows the tweets to be categorized into three categories which are positive, negative and neutral. These categories are distinguished based on their polarity of the tweets. Figure 3 shows the percentages of tweets categorized in the following categories. The negative tweets are deduced to come from users to are among the anti-vaccine community, while the positive tweets came from pro-vaccines. About 20.03% of the Twitter users who shared their opinion on vaccination are negative.

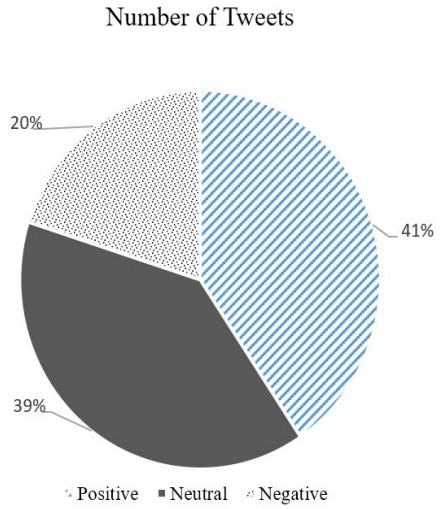

Figure 3: Percentage of Tweets based on Category

Table 4: Performances of SVM model with different kernels.

Type of Kernel	Category	Precision	Recall	F1-score	Accuracy
Polynomial Kernel	Positive	0.79	1.00	0.88	87%
	Neutral	0.99	0.76	0.86	
	Negative	1.00	0.75	0.86	
RBF Kernel	Positive	0.88	0.97	0.92	91%
	Neutral	0.94	0.88	0.91	
	Negative	0.95	0.82	0.88	

Then, the next process is model evaluation. Among the data, 30% of the data is allocated as test data. The balance is treated as training data. Table 4 shows the performance of the classifier using both types of kernels.

Based on Table 4, using RBF Kernel has a higher accuracy in classifying the data than Polynomial Kernel. RBF kernel is known to be used in optical remote sensing data for their high accuracy. In some cases, Polynomial kernel tends to have higher accuracy on other types of data depending on the improvement of the normalization of data [16]. The implementation of RBF kernel in SVM classifier outstands the other.

5. CONCLUSION

Many kinds of research have been focusing on analyzing the opinion of people about vaccination. The existence of social media made it easier to gain the sentiment data from pro- and anti- vaccine community. This study aids in providing solution on sentiment analysis of about 100,000 tweets accumulated using Textblob and a Support Vector Machine (SVM) Classifier. Using RBF kernel brings more accuracy on

data as it refines the hyperplane of the classification graph. However, due to weakness in pre-processing and filtering the data, there is a tendency that the accuracy would become much lower. In future, the result of the accuracy could be improved using Spider Monkey Optimization (SMO) algorithm as suggested by [17].

REFERENCES

1. E. Andreano, U. D'Oro, R. Rappuoli, & O. Finco. **Vaccine evolution and its application to fight modern threats,** *Frontiers in immunology*, Vol.10, pp.1-5, 2019.
2. Bedford, H., Attwell, K., Danchin, M., Marshall, H., Corben, P., & Leask, J. **Vaccine hesitancy, refusal and access barriers: The need for clarity in terminology.** *Vaccine.* pp.1-3, 2017.
3. A. Pak & P. Paroubek. **Twitter as a corpus for sentiment analysis and opinion mining.** *LREc*, Vol. 10, No. 2010. pp. 1320-1326, 2010.
4. B. Pang & L. Lillian. **Opinion mining and sentiment analysis.** *Computational Linguist*, Vol. 35 Ed-2 .pp. 311-312, 2009.
5. D. A. Salmon, M.Z. Dudley, J.M. Glanz & S.B. **Vaccine hesitancy: causes, consequences, and a call to action. Vaccine**, Vol. 33, pp. 66-71, 2015.
6. K.H. Saglani & N.J. Janwe. **Machine Learning Based Sentiment Analysis on Twitter Data.** *International Journal of Emerging Trends in Engineering Research (IJETER).* Vol. 8. No. 8, pp4413-4419, 2020.
7. B. L. Hoffman, E. M. Felter, & K.H. Chu. **It's not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook.** *Vaccine*, Vol.16, pp. 2216-2223, 2019.
8. R.F. Hunter, A. Gough, N. O'Kane, G. McKeown, A. Fitzpatrick & T. Walker. **Ethical issues in social media research for public health.** *Am J Public Health*, Vol. 108. pp. 343-348, 2018.
9. N. Yiannakoulias, C.E. Slavik, &. M. Chase, M. **Expressions of pro-and anti-vaccine sentiment on YouTube.** *Vaccine*, Vol. 37 Ed-15, pp.2057-2064, 2019.
10. F. Kunneman, M. Lambooij, A. Wong, A. V. D. Bosch, & L. Mollema, **Monitoring stance towards vaccination in Twitter messages.** *arXiv preprint.* pp.1-16, 2019.
11. X. Yuan & A. T. **Examining Online Vaccination Discussion and Communities in Twitter** in *Proceedings of the 9th International Conference on Social Media and Society*. 2018, pp. 197-206.
12. G. Guibon, M. Ochs, & P. Bellot. **From Emoji Usage to Categorical Emoji Prediction** in *19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2018).* 2018.
13. E. D'Andrea, P. Ducange, A. Bechini, A. Renda, & F. Marcelloni. **Monitoring the public opinion about the vaccination topic from tweets analysis.** *Expert Systems with Applications*, Vol.116. pp. 209-226. 2019.
14. E. Cambria. **Affective computing and sentiment analysis.** *IEEE Intelligent Systems*, Vol. 31 Ed-2. pp. 102-107, 2016.
15. S. Suthaharan. **Support vector machine** in *Machine learning models and algorithms for big data classification*. Springer US. 2016, pp. 207-235.
16. C. Obimbo & E. Nyakundi. **Comparison of SVMs with Radial-Basis function & Polynomial Kernels** in *Classification of Categories in Intrusion Detection*. 2019.
17. P. Suryachandra & P.V.S. Reddy. **A Novel Hybrid Machine Learning Approach to Classify the Sentiment Value of Natural Language Processing in Big Data**. International Journal of Emerging Trends in Engineering Research (IJETER). Vol. 8. No. 8, pp4460-4465, 2020.