

PAPER • OPEN ACCESS

An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment analyzer

To cite this article: J Garay *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **482** 012043

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment analyzer

J Garay, R Yap and M J Sabellano

Department of Computer and Information Sciences, University of San Carlos,
Talamban, Cebu City, Cebu 6000, Philippines

lawrencetecho@gmail.com

Abstract. This study analyzes the insights and sentiments of the anti-vaccine movements in the social media. The data, comprising of tweets and excerpts, are pre-processed to omit noise and irrelevant data. They are clustered using the k-means clustering algorithm. Each word belonging to a cluster is processed by VADER sentiment analyzer. Prevalent sentiments per cluster label the mood associated in the cluster. The results suggest insights about vaccines such as: side effects, post-shot injuries, ineffectiveness, damage from ingredients, unvaccinated elite, reinforcement of the right to not vaccinate, toxic ingredients, big pharmaceuticals' profit maximization, links to autism, and health issues after getting vaccine shots. To evaluate k-means results, the silhouette score is determined to indicate how far a point is to other nearby clusters. The resulting average silhouette score of all points is 0.013540022 which indicates that the points are close to the decision boundaries.

1. Introduction

Curbing preventable diseases is a political, social, economic, and health-related endeavor. Many countries and institutions mandate vaccination so people most especially in public places are guarded against preventable diseases. [1] However, a significant number of parents in the United States show little confidence for vaccines. [2] In 1998, the rise in the number of people this movement has skyrocketed due to a faux scientific paper that linked vaccination to autism [3]. Together with the advancement of social media and the rise of technologies that make it easier to spread misinformation for its cause, the movement has garnered more followers and had been proliferating their causes in an even faster rate [4].

While in many respects, machine learning techniques could not extend as much as humans can in interpreting behavior and insights especially those involving heavy qualitative data, however, combining natural language processing and machine learning to help tackle social issues can be used as an advantage.

The purpose of this paper is: a) to demonstrate the use of the k-means clustering method in discovering insights from a corpus; b) to understand the sentiment in every cluster by using VADER sentiment analyzer to further examine the mood incorporated for every insight (cluster); and c) to gain useful insight and sentiment of the anti-vaccine movement.



2. Review of Related Literature

2.1. K-Means Clustering

A popular algorithm for clustering would be K-means clustering. [5], [6], [7] show k-means to have a relatively low time complexity ($O(nkd)$) which indicates faster performance. It also works well with large data set. Though it is quite sensitive to noisy data. With the pros in mind, the researchers concluded that this would be a good approach.

A study conducted by [8] shows the difference of speed between K-means and Fuzzy K-means. In their experiment where they cluster 39,000 tweets in order to compare the speed, it is shown that K-means outperforms Fuzzy K-means overall.

The researchers need a way to thoroughly analyze tweets to be able to make sense of anti-vaccine sentiments. A study analyzing disaster risk reduction (DRR) responses from the Philippines uses K-means clustering which allowed them to find out some keywords produced by the clusters and with the use of open coding, were able to interpret them. [9]

2.2. Sentiment Analysis

Sentiment Analysis has been gaining traction lately with the appearance of a large source of opinionated data through the web and its wide range of applicability in many fields of society. [10]

A study was conducted to find a way to use sentiment analysis in analyzing student evaluations of teaching (SET) for universities. They compared the sentiment of comments together with quantitative scores. They found that the polarity of the comments generally follows the trend of the meanings of the scores, making sentiment analysis a viable method for dealing with SETs. They used VADER as their sentiment analysis tool [11].

2.3. The Anti-Vaccine Movement

The anti-vaccination movement's rapid growth can be attributed to some very persuasive websites and information online. A study conducted to find out why these sites are persuasive was conducted. These sites contain a large amount of misinformation, such as vaccines are dangerous and can cause autism and brain damage. They use scientific evidence and anecdotes to support these claims, and they also commonly promoted the use of alternative medicine, homeopathy, and an organic lifestyle. These combined with many effective persuasion techniques to convince people to adapt their beliefs. [12]

3. Methodology

3.1. Data Gathering and Pre-processing

The data in this study refer to tweets and post excerpts from popular public Facebook anti-vaccination groups.

3.1.1. Data Cleaning. The mined data were cleaned by removing "stopwords" – non-important words (e.g. "to"), non-alphanumeric characters, words with lengths less than 4, and non-English words were removed from the data

3.1.2. Data Pre-Processing. The data were converted into vectors using the term-TF-IDF vectorization method.

3.2. TF-IDF Vectorization Method

Term frequency times inverse document frequency (TF-IDF) is a numerical statistic method used to determine how important a word is to a corpus. This method compares the frequency of the word in a document versus its frequency in the whole aggregate of documents (corpus). The study uses scikit's TF-IDF vectorizer to achieve vectorization.

3.3. K-Means Clustering Algorithm

The use of k-means clustering is to create k number of groups and to designate each word (vector) in the corpus a corresponding group. This method works by iteratively designating a data to one of the k groups based on a common mean vector. The clusters are formed as follows:

1. Set k
2. Designate random k locations; this will serve as the centroids.
3. For each data, do the following until there is no longer a significant shift after measuring the new centroids (shift is significant if it's above a threshold):
 - i) Find a centroid whose mean has the least squared Euclidean distance with the current data
 - ii) Find the new centroid within the space in this centroid by calculating the mean of all points in the space.

3.4. Validation Methods

3.4.1. Silhouette Score. The silhouette score describes how near a point is to its own cluster and how far the point is from other clusters. The study uses this to measure the correctness of the assignment of the points/vectors in their respective clusters. This is calculated using the mean intra-cluster distance i and the mean nearest-cluster distance j for each word (vector) in the data. In practice, a score nearing 1 indicates more distance from other clusters. A score nearing -1 indicates that the point is assigned in the wrong cluster. Nearing 0 means that the point is close to the decision boundary. The formula is as follows:

$$\frac{i - j}{\max(i, j)} \quad (1)$$

3.4.2. Picking the Optimal Number of Clusters. The number of clusters that yields a silhouette score closest to 1 is the optimal number of clusters. Table 1 shows the number of clusters and their corresponding silhouette score. In the result, having 10 clusters has the silhouette score closest to 1. This means having 10 clusters yield clusters farther from each other with the least likely occurrence of overlap relative to other k number of clusters. The silhouette score started to decrease after the 10th cluster. Implied from this, any k cluster beyond 10 will yield a silhouette score fatter from 1 which indicates that weaker clusters are formed since points become nearer the decision boundary. The 10th cluster however, given 0.013540022 is the most optimal silhouette score relative to other k.

Table 1. Resulting silhouette score after K iterations of the corpus

Number of Clusters	Silhouette Score
2	0.005830464
3	0.007237809
4	0.006643489
5	0.009101259
6	0.010489809
7	0.012620094
8	0.012190529
9	0.012351144
10	0.013540022
11	0.012131123
12	0.011417081

3.5. VADER Sentiment Analyzer

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based sentiment analyzer specifically trained sentiments from social media. This API is used in the study by determining the corresponding sentiment for every word. The analyzer works by passing to it a word and it returns the sentiment associated with the word. This method is used in the study by determining the prevalent sentiment in every cluster. Every word in the a is processed by VADER sentiment analyzer and the cluster is then labelled a sentiment by identifying the most prevalent sentiment.

4. Results and Discussions

All results below are the results of k-means clustering algorithm with a set k of 10 to optimize resulting clusters after having tested various k and producing different silhouette scores.

Table 2. Clusters generated by K-means with the closest words it contains and the prevalent sentiment from all words in the cluster as generated by VADER analyzer.

Cluster	Nearest frequent keywords in the cluster	Sentiment
1 – Possible side effects of vaccines	kill, cripple, death, injection, disorder, neurological	negative
2 – Experience of vaccine injury	parents, children, kids, injured, sick, doctors	neutral
3 – A call to question whether vaccines really work	unnecessary, effectivity, expire, herd, immunity, gardasil	neutral
4 – Chronic damage caused by toxic vaccines	damage, brain, toxic, learn, cause, ones	neutral
5 – Emphasis on the elite who choose not to vaccinate	children, unvaccinated, elite, rich, doctors, unlikely	neutral
6 – Expression of their right to choose	child, choice, right, parent, decision, choose	neutral
7 – “toxic” Ingredients present, research	research, children, ingredients, wanting, kids, toxicity	neutral
8 – Capitalism of vaccine products	kids, sick, pediatricians, profit, pharmaceutical, big	negative
9 – Autism being linked to vaccines	autism, amish, cause, children, relation, link	neutral
10 – children started getting health issues after vaccine shots	health, cause, issues, daughter, shots, jab	neutral

4.1. Cluster 1: Possible side effects of vaccines (kill, cripple, death, injection, disorder, neurological; sentiment: negative)

The discouragement of people who fall into this cluster relies heavily on the prevalent fear-mongering in social media. The social media space has a subset of people who use fear to discourage people opting to vaccination. Although vaccines in fact may have side effects, side effects in this cluster are exaggerated to evoke fear within social media (kill, cripple, death). The negative sentiment supports such claim. It can be implied as though advocacies and campaigns should focus more on debunking the claimed side effects.

4.2. Cluster 2: Experience of vaccine injury (parents, children, kids, injured, sick, doctors; sentiment: neutral)

Campaigns and advocacies should start to raise awareness as to the improbability of vaccine injuries from happening.

4.3. Cluster 3: A question on the effectivity of vaccines (unnecessary, effectivity, expire, herd, immunity, Gardasil; sentiment: neutral)

People who fall into this cluster are skeptical of vaccine's effectivity; resulting to claims that vaccines are unnecessary. It becomes apparent that there is a need to create a discourse about the scientifically-

tested benefits of vaccines. The neutral sentiment manifests that such skepticism does not come with other negative notions about vaccines.

4.4. Cluster 4: Chronic damage caused by toxic vaccines (damage, brain, toxic, learn, cause, ones; sentiment: neutral)

There is a prevalent claim in this cluster that vaccines could cause brain damage. Another way to interpret this result is to say that people in this cluster encourage to learn as to how toxicity in vaccines could cause brain damage. Either interpretation is necessitates debunking such claim.

4.5. Cluster 5: Emphasis on the rich, elite few whose choice is to opt out from vaccination (children, unvaccinated, rich, elite, doctors, unlikely; sentiment: neutral)

This cluster allows us to understand how the decisions of the rich and elite being unlikely to opt from vaccination has an impact on the middle class' decision whether or not they should vaccinate. Their influence does have a significant impact on the decisions of people who fall in this cluster.

4.6. Cluster 6: Expression of their right to choose (child, choice, right, parent, decision, choose; sentiment: neutral)

Vaccinations have become a requirement for different social and economic activities in many countries. It is vague to interpret the reason why parents might not vaccinate their child. It is apparent in this cluster however, that parents are reinforcement the value of choice and the ability to decide for themselves to opt to vaccination or not.

4.7. Cluster 7: Toxic ingredients present in vaccines (research, children, ingredients, wanting, kids, toxicity; sentiment: neutral)

In this cluster, research is introduced and is associated with other keywords like toxicity and ingredients. A way to interpret this cluster is that there may be "research" about the toxicity of the ingredients found in the vaccines that parents believe. This cluster allows us to understand that people in this cluster are not motivated by fear but are people who believe in findings in "research", although there has been no research so far that suggests with confidence about the significant harms of the ingredients present in vaccines.

4.8. Cluster 8: Capitalism of vaccine products (kids, sick, pharmaceutical, profit, big; sentiment: negative)

People who fall into this cluster are skeptical of the vaccines produced by pharmaceutical companies. This finding could be interpreted in a way that people have negative perception/sentiment as to how reliable vaccines are. The association of "profit" in this cluster may imply that people's perception of vaccine as a necessity only benefits pharmaceutical companies as it may be thought of as a marketing ploy.

4.9. Cluster 9: Autism linked to vaccines (autism, amish, cause, children, relation, link; sentiment: neutral)

This cluster reinforces the existence of the already prominent idea with in antivaxxers that vaccines cause autism in children.

4.10. Cluster 10: Vaccines caused other health-related issues in children (health, cause, issues, daughter, shots, jab; sentiment: neutral)

People in this cluster claim that vaccine shots caused health issues to their children.

5. Conclusion

The resulting average silhouette score computed for all points is 0.013540022. Any silhouette scores higher than zero indicates that the data were fit to the clusters they were assigned to.

Majority of the sentiments were neutral. Implied from this, the subjects were more cautious with the reasons provided per cluster than they are negative about it.

The study manifests that k-means clustering can be used as an approach to clustering a corpus of qualitative data such as tweets and posts from social media. Furthermore, the study shows that quantitative methods in machine learning, particularly k-means algorithm, can be used to analyze qualitative data such as opinions.

The methods in this study can be replicated in many other practical domains. The same method of clustering data and finding sentiment for each formed cluster can be applied to social, political, and business domains with data such as reviews and feedbacks. This opens possibilities of new types of case studies that require interpreting feedback by extracting features and understanding the sentiment associated with it.

We recommend the use of a stronger, context-sensitive machine learning algorithm in determining the associated emotion of a word. The used model — that VADER utilizes to determine the emotion — is trained with general, non-context-sensitive statements and topics from their own massive dataset. While massive datasets create a more neutral, less-biased, and less context-sensitive output, in a world political and social issues, context can be a strong factor for determining emotion within a text.

Acknowledgements

We wish to acknowledge the University of San Carlos and its Department of Computer and Information Sciences for the research opportunity.

References

- [1] Walkinshaw E 2011 Mandatory vaccinations: The international landscape *Canadian Medical Assoc. J.* **183(16)** pp 1167-8
- [2] Kennedy A, LaVail K, Nowak G, Basket M and Landry S 2011 Confidence about vaccines in the United States: understanding parents' perceptions *Health Affairs* **33(6)** pp 1151-9
- [3] Olpiński M 2012 Anti-vaccination movement and parental refusals of immunization of children in USA *Pediatrics Polska* **87(4)** pp 381-5
- [4] Stien R 2017 The golden age of anti-vaccine conspiracies *Germs* **7(4)** pp168-70
- [5] Xu D and Tian Y 2015 A comprehensive survey of clustering algorithms *Annals of Data Science* **2** pp165-93
- [6] Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya A, Fofou S and Bouras A 2014 A survey of clustering algorithms for big data: Taxonomy and empirical analysis *IEEE transactions on emerging topics in computing* **2** pp 267-79
- [7] Jain E and Jain S 2014 Using Mahout for clustering similar Twitter users: Performance evaluation of k-means and its comparison with fuzzy k-means *Int. Conf. on Computer and Communication Technology (ICCCCT)* pp 29-33
- [8] Go Bui S M, Gorro K, Aquino G A and Sabellano M J 2017 An analysis of DRR suggestions using K-means clustering *Proc. of the 2017 Int. Conf. on Information Technology* pp 76-80
- [9] Liu B 2012 Sentiment analysis and opinion mining *Synthesis Lectures on Human Language Technologies* vol **5** no **1** pp 1-67
- [10] Newman H and Joyner D 2018 Sentiment analysis of student evaluations of teaching *Int. Conf. on Artificial Intelligence in Education* pp 246-50
- [11] Moran M B, Lucas M, Everhart K, Morgan A and Prickett E 2016 What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment *J. of Communication in Healthcare* **9** pp 151-63
- [12] Kata A 2012 Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement *Vaccine* **30** pp 3778-89