How to Cite:
Parimala, R., & Gomathi, K. (2022). Exploring brand perception in COVID vaccine tweets using TOPSA-NMF model. *International Journal of Health Sciences*, *6*(S3), 5422–5428. https://doi.org/10.53730/ijhs.v6nS3.7142

Exploring brand perception in COVID vaccine tweets using TOPSA-NMF mode

Dr. R. Parimala
Assistant Professor,CS Department, Periyar E.V.R. College, Tiruchirappalli-23, India
*Corresponding author email: rajamohanparimala@gmail.com

K. Gomathi
Research Scholar, CS Department, Periyar E.V.R. College, Tiruchirappalli-23, India, (Affiliated to Bharathidasan University)
Email: gomu2512@gmail.com

***Abstract*---**COVID-19 vaccines are primary biomedical preventive measures to combat the pandemic. Social media can have either a positive or negative impact on vaccine uptake, depending on their opinions about vaccines. An algorithm is proposed that can use a linear dimension reduction technique such as non-negative matrix factorization (NMF) to identify topics while using the sentiment lexicon to classify the tweet topics as positive, negative, and neutral. This study aims to determine public attitudes towards each topic of tweets about COVID-19 vaccines.

***Keywords*---**sentiment analysis, topics, page rank, non negative matrix factorization, COVID vaccine.

Introduction

Brand Sentiment Analysis is a way to determine general attitudes towards a brand, product or service. The easiest way to analyze brand sentiment is with media monitoring tools. People often express their voices on social media channel about what they feel about the trending theme in the society. Social media, especially Twitter and Facebook, are filled with different opinions, reviews and ratings on different topics, products, brands and people [8]. The results of the sentiment analysis can be very helpful for targeted marketing and designing new products that meet the customer's needs much more deeply.

The researchers used different topic modeling techniques, namely LSA (Latent Semantic Analysis), LDA (Latent Dirichlet Allocation), PLSA (Probabilistic Latent Semantic Indexing) and STM (Structural Topic Model). LSA works well for long

documents due to the small number of context vectors used to describe each document[1][2][3]. LDA performs poorly on short text. The NMF method is used to decompose the input matrix into two matrices of non-negative entries in a lower rank order. Several applications of NMF include: chemometrics, environmental metrics, pattern recognition, text mining and summarization, multimedia data analysis, DNA gene expression analysis, financial data analysis, and social network analysis [3]. Several algorithms have been developed to overcome the problems associated with the dimensionality reduction of data in classification, collaborative filtering and computational methods.

Social media websites like Twitter and Facebook generates a large amount of short texts every second that reveal a good knowledge of real-time information. Tweets are short pieces of text originally limited to 140 characters but more recently expanded to 280 characters; the short length poses a challenge for text analysis. Syed et al. (2017) reported that finding higher quality topics with greater coherence in models trained on larger documents [4]. Researchers have used social media datasets throughout the COVID-19 pandemic. Lyu et al. [5] used natural language processing techniques to analyze the topics and sentiment of Twitter conversations related to the COVID-19 vaccine using the Twitter chatter data set created by Georgia State University's Panacea Lab. In addition, a total of 22,587 tweets used the terms shot," "shots," "jab," "jabs," "jabbed," "vax," or "vaxxed." The words "shot" or "shots" were used in 9017 tweets. Extensive research is carried out to gain knowledge from it.

Short text is very sparse and ambiguous; therefore, it is a great challenge to find implicit topics in it. NMF can generate meaningful patterns, topics or aspects. Even the corpora have many words, and the words are ambiguous and sparse. NMF can reveal essential words and that these words significantly improve the performance of topic extraction in a text corpus. This work extracts the topic terms found in each topic to cover the topic distribution using the NMF method and to identify the polarity of the corpus from the extracted topics using sentiment lexicon. The paper is organized as follows, Section I explain about Introduction of Brand monitoring. Section II presents methodology of proposed model. Section III discusses the results. Finally concludes the study with future scope.

Methodology

This work presents a workflow model (Fig.1) of topic extraction (aspect detection) and sentiment detection of Tweets using NMF and Syuzhet Lexicon.

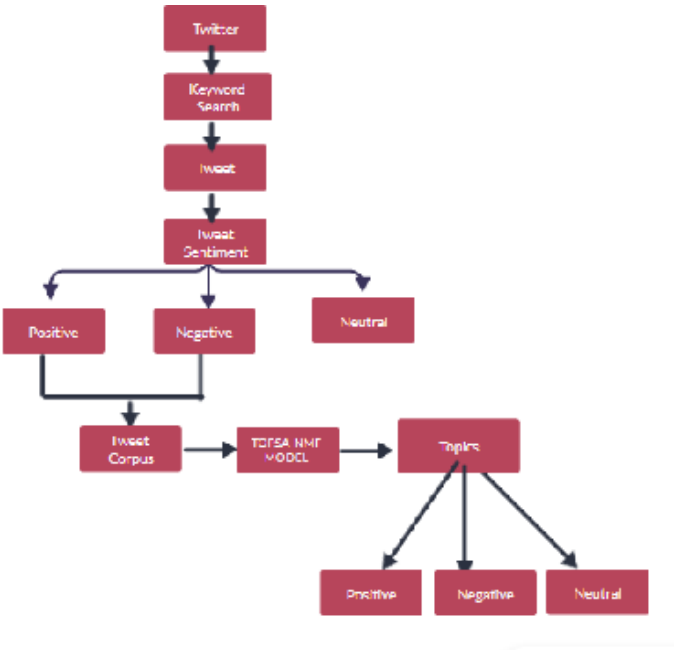

Fig 1. Systematic workflow of proposed work

NMF Model

Given Tweet word matrix A, NMF decomposition find mixture matrix (M) and basis matrix (B) such that $A_{mxn} \approx M_{mxr} B_{rxn}$, where all elements of A, M, and B are strictly non-negative [3]. Mathematically,

$$\underset{M,B \geq 0}{\text{Min}} \left\| A - MB \right\|$$

M and B are updated using the multiplication update rule [16]

$$A = M \otimes \frac{AB^T}{MBB^T}$$

$$A = B \otimes \frac{M^T A}{M^T MB}$$

TOPSA -NMF Model

The algorithm for the TOPSA-NMF model is as follows.

Algorithm TOPSA-NMF

- Read Tweet Corpus
- Perform preprocessing
- Create TWM using Wf-iTf
- Remove sparse words
- Construct Word Graph

- Call Pagerank algorithm
- Select Top words from TWM
- Iterate NMF on TWM over Number of topics
- Evaluate Cophenetic Score for each topic
- Obtain optimum topic size, M and B
- Return top-20 words of each topic using M.
- Apply Sentiment Lexicon for each topic word.
- Record the attitude of each topic and corpus

Results

Brand reputation is the real-time, continuous process that monitors the reputation of one or more Covid-19 vaccine brands in the text corpus. The proposed algorithm TOPSA-NMF implemented in R programming language [7]. The corpus used in these experiments was collected from Twitter during the period from 13.04.2022 to 15.04.2022, which allows tweets to be collected based on a keyword #covidvaccine. 8546 English tweets were collected and processed. There are 4121 positive tweets, 939 neutral tweets, and 3486 negative tweets in corpus. Neutral sentiment tweets are removed from the corpus since it denotes the lack of sentiment. 7607 tweets are considered for this study. Tweets were pre-processed, tokenized and converted to lowercase; Removed emoticons, URLs and HTML tags, as well as common English stop words. Tweet Word Matrix (TWM) is constructed whose elements are word Frequency-Inverse Tweet Frequency which is also a bag-of-words model. The TWM consists of 7607 tweets and 10133 words. The sparse words are removed from TWM. The resultant TWM is of size 7607 x 287. Google Page rank algorithm is applied to select the most influential words [9]. Word Graph (Word by Word) is constructed. Page rank algorithm applied on Word Graph to select the significant words. The top most 100 words are selected. The NMF model is trained on 7607 x 100 tweets with various topic size. A key challenge in applying topic modeling is selecting an appropriate number of topics k. Cophenetic correlation coefficient, extensively used to select the optimal topic size k. The value of k is chosen where the magnitude of the cophenetic correlation coefficient begins to fall [7]. Experiments are performed for 5 different NMF runs, varying the number of topics from 2 to 12 and choosing the model with the optimal number of topics. The topics and its cophenetic measures are depicted in Fig.2. It is observed that the large decrease in stability for k=4 compared to lower topic number.

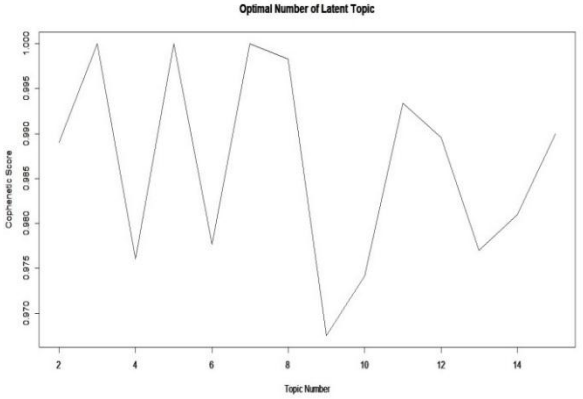

Fig 2. Topic and its Cophenetic Measures

Four topics were extracted from TWM. A sentiment lexicon is a list of words, with each word assigned a positive, negative or neutral value that reflects its sentiment polarity and strength [6]. Sentiment lexicon and its polarity of topics are summarized in Table 1. Topics are Topic1 (Vaccine Effectiveness), Topic2 (Protection), Topic3 (Child Imunization) and Topic4 (Elderly Vaccine). From the table, it is inferred that a public attitude towards Topic1 and Topic2 are positive whereas Topic3 and Topic 4 are neutral and negative. Overall the corpus is positive. The finding illustrates that attitude of covid vaccine brand play in leading better future decisions.

Table 1
Topic and its sentiment polarity of Tweet Corpus using TOPSA-NMF algorithm

No	Topic 1	Polarity	Topic 2	Polarity	Topic 3	Polarity	Topic 4	Polarity
1	dose	0	children	0	mandat	0	approv	0
2	book	0	year	0	school	0	valneva	0
3	least	0	pfizer	0	student	0	use	0
4	new	0.8	recommend	0.5	coronavirus	0	regul	0
5	one	0	age	0	least	0	first	0
6	avail	0	booster	0	state	0	moderna	0
7	now	0	mrna	0	requir	0	news	0
8	amp	0	shot	-0.4	kid	0	get	0
9	receiv	0	biontech	0	mani	0	new	0.8
10	protect	0.75	ano	0	parent	0	covid	0
11	peopl	0	moderna	0	news	0	coronavirus	0
12	age	0	today	0	fda	0	jab	-0.6
13	first	0	effects	0	health	0	vaccin	0
14	studi	0	old	0	covid	0	old	0
15	spread	0	peopl	0	vaccin	0	virus	-0.5
16	booster	0	announc	0	make	0	fda	0
17	effect	0	kid	0	mask	0	author	0
18	safe	0.75	studi	0	still	0	just	0
19	death	-0.75	data	0	announc	0	shot	-0.4

20	news	0	immun	0	just	0	requir	0

A word cloud representations of the topic sentiments are shown in Figure 3. Positive and negative topic words are highlighted with blue stone and red colors to show its polarity. Public sentiment on COVID-19 vaccines varied significantly over time will help public health policymakers to make desired decisions.

Fig. 3. A word cloud of Topic wise sentiment of Covid vaccine corpus

Conclusion

In recent years, COVID-19 is the greatest human challenge with all nations governments and researchers trying to reduce the mortality rate of this widespread disease. A Public debate of COVID-19 vaccines on Twitter has been largely driven by major events over COVID-19 vaccines. Brand reputation makes it possible to find out what people think about brands, where people express those opinions and what expresses sentiment. NMF is an unsupervised method that can perform dimension reduction and topic extraction simultaneously. The proposed NMF method map tweets to a Topic. Learning meaningful topics from brand reputation in tweet is considered to be a challenge due to ambiguous contextual information in it. Each tweet consists of different words and each topic can be associated with a few words. The goal behind the proposed work was achieved to find the sentiment of the topics to which the tweets belong based on the words in the topic. This study examined topic modeling of tweets related to COVID-19 vaccines and found 4 topics and opinions of tweets about vaccines to monitor vaccine brands. For future work, we propose more effective topic extraction methods to provide more efficient and relevant information using sentiment analysis. Our future goal will be to improve the challenges of current model and explore more brand monitoring model.

Acknowledgment

We would like to acknowledge R core Team Shell and other contributors for developing and maintaining the R Packages which have been used in the preparation of this proposed work.

References

1. Blei DM. *Probabilistic topic models.* Communications of the ACM. 2012;55(4):77–84. Available from:. https://doi.org/10.1145/2133806.2133826

2. Korshunova I, Xiong H, Fedoryszak M, Theis L, *Discriminative topic modeling with logistic LDA* [Internet]. arXiv [stat.ML], 2019. Available from: http://arxiv.org/abs/1909. 01436
3. Mifrah S. Topic modeling coherence: *A comparative study between LDA and NMF models using COVID'19 corpus.* International Journal of Advanced Trends in Computer Science and Engineering,2020;9(4):5756–61. Available from: http://dx.doi.org/10.30534 /ijatcse/2020/231942020
4. Syed S, Spruit M. *Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation.* In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE; 2017.
5. Carvalho PR, Munita CS, Lapolli AL. *Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficien*t. Brazilian Journal of Radiation Sciences. 2019;7(2A). Available from: http://dx.doi.org/10.15392/bjrs.v7i2a.668
6. Garcia K, Berton L. *Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA*. Applied Soft Computing [Internet]. 2021;101(107057):107057. Available from: http://dx.doi.org/10.1016/j.asoc.2020.107057
7. Feinerer I, Hornik K, Meyer D. *Text Mining Infrastructure in R*. Journal of Statistical Software. 2008;25(5):1–54.
8. Özdemïr Akcan E, Sütütemïz N. *Analysis of brand perceptions of covid-19 vaccines with sentiment analysis on social media.* Pamukkale University Journal Social Science Institute[Internet]. 2021; Available from: http://dx.doi.org/10.30794/pausbed.994700
9. Rousseau C, Saint-Aubin Y. *Google and the page rank algorithm*. In: Mathematics and Technology. New York, NY: Springer New York; 2008. 1–24.