Russia’s Covid – 19 Vaccine: Social discussion and first emotions

Meena R (meenarajeswaran@gmail.com)
Prathyusha Institute of Technology and Management  https://orcid.org/0000-0002-7437-8159

Thulasi Bai V
KCG College of Technology

Research

Keywords: tweets, machine learning, LDA

DOI: https://doi.org/10.21203/rs.3.rs-95570/v1

License: ☺️ 📧 This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Social media chats related to healthcare are the prodigious basis of analysing the emotions of the people. Now, the Covid-19 vaccine is being the prevalent hope of nearly the entire mankind in the planet. Russia's first vaccine announcement kindled the various rays of emotions among the social media users which are shared as tweets. The tweet data is collected and analysed for the emotions and psychology of the users along with the topic of interest in their discussion. Using computational methods and algorithms such as machine learning and LDA, the social emotions are revealed and presented.

1. Introduction

The personal opinions of people collectively can be organized and mined for extracting the entity-based sentiment. Opinion mining drives the subjective analysis part on the texts which has expressions. Linguistic sentiment analysis can be done using the social blogging site such as Twitter from which emotions can be easily portrayed as sentiments. [1] The short tweets shared by the user pose several challenges in identifying the exact opinion and sentiment since it has a highly unstructured form of the text which contains emojis and sarcastic comments. [2 – 6] An outbreak of the covid-19 pandemic which spreads exponentially from china to many parts of the world creates the biggest impact on global health. [7 – 10] Since there is no prescribed drug to cure the corona virus infection, the World Health Organization suggests the standards for the caution and protection of public from the disease. [11] So, the prime hope for the people is the SARS-CoV-2 vaccine which gives the immunity against the virus. Vaccine development is a very tedious process which includes many stages of trail phases that can run through years. On a pandemic situation, where the world relies on the vaccine, researchers are developing the vaccine at a pandemic speed. [12] In a global race to the vaccine production, some of them has successfully reached the phase 2 and phase 3 of clinical trials. [13] Amidst the situation, Russia has launched its vaccine for Covid 19 named Sputnik V on August 11, 2020 and claimed to be first of its kind released for public use. [14] People around the world poured their views as tweets after the official declaration of the vaccine by Russia. The large scale emotion on the vaccine for pandemic is a great source to understand the sentiments of people and to uncover their topics of discussion along with it.

In this work, we address the sentiment analysis of the first vaccine launched for covid 19 pandemic.

We present the contributions of the paper as follows:

We have manually annotated the tweets containing the sentiment on the Russia’s covid – 19 vaccine. The data set consists of 4375 tweets on English language. Every tweet in the dataset is manually given either a positive or negative label by two annotators.

We did the analysis to reveal the sentiment of the tweets by using the popular VADER sentiment lexicon and evaluated the manual labelling using machine learning.
We have analysed the topics which were discussed along with the vaccine using Latent Dirichlet Allocation algorithm.

The paper is organized as follows: related works are listed in section II, Data collection and labelling in section III, sentiment analysis in section IV, topic mining in section V. The results are discussed in section VI followed by the conclusion in section VII.

2. Related Work

Social sentiment of the Covid-19 pandemic has been studied using the social media data such as tweets. Paper [15] collected tweets for a fourteen days period of time which uses nrc sentiment lexicon. They have presented that the majority of the tweets were of positive sentiment and also scaled the Plutchik’s emotions. Paper [16] perceives the topic modelling and the dominant emotions on the two weeks of twitter data on covid-19. Paper [17] extracts tweets on various hashtags such as “#COVID19”, #SARSCoV2”, “#CoronaVirus” over a week of time using twitteR API and the sentiments has been studied which revealed 65.6% of negative sentiment. Paper [18] collects nearly 63 million tweets and conducted the machine learning experiment using NLP on two cases which includes (i) ten topics based on LDA (ii) Sentiment analysis using CrystalFeel. Sentiments on the other attributes of covid – 19 such as “lockdown”, “stock market and economic impacts”, “work from home”, “news”, “masks”, “Reopening” were discussed in Papers [19 – 25]. Previous studies have analysed the sentiments on various aspects of Covid – 19 and its impacts. To the best of our knowledge, there is no separate work trailed on the global sentiment of Covid-19 vaccine release. Moreover, majority of the previous works uses automatic text annotation while our work annotated the tweets manually which reduces many pitfalls in the understanding of the emotions on tweets over the particular context.

3. Method

DATA COLLECTION AND LABELLING

We have collected the tweets using the Twitter API which is related to the Russia’s vaccine on Covid -19 using the hashtags “#russia”, “#russiavaccine”, “vaccine”.

In this work, we have collected the tweets using the Twitter API on August 11, 2020 the day which Russia claimed to have the vaccine for Corona virus.

The initial labelling is done manually for all the tweets in two types such as positive or negative. The data contain a variety of sarcastic tweets such as “This tree needs a vaccine, clearly.” express the sense of dissatisfaction of the people regarding the vaccine release. The tweet which mentions “Russia has a COVID vaccine. Polonium is so versatile.” generally would be labelled positive since there is no negative lexicons present. But it express negative emotion to the context of vaccine. The other tweet mentioned, “This appears to be a 20ml shot of real vodka” which gives the sarcastic comment on the vaccine. Labelling such tweets on the particular context either as positive or negative is crucial and often mislead
the sentiments of the overall text. So we have done manual annotation of the tweets as the part of our work.

4. Sentiment Analysis

Valence Aware Dictionary and sEntiment Reasoner is a specially developed to read the social media text sentiments which contains more of emojis and short texts[26]. Vader lexicon is the proven to be the outperformer in analysing the social media text snippets. The polarity scores of the tweets were calculated using the Intensity Analyzer of vader lexicon. The scores were on two major classes such as positive, negative. Followed by that, the compounded score of each tweet is calculated and tabulated. After computing the score if the sentiment is assigned as,

Pos if Cs>=0

Neg if Cs<0

Table 1. Sample scores on polarity of tweets and its classification based on sentiment

| Tweet                                                                 | Polarity using vader lexicon | Score | Class of Tweet |
|----------------------------------------------------------------------|------------------------------|-------|----------------|
| Putin claims first coronavirus vaccine without providing evidence    | {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0} | 0     | pos            |
| Putin added that one of his daughters had already taken it; he said she had a slightly higher temperature after each dose, but that: "Now she feels well." | {'neg': 0.0, 'neu': 0.911, 'pos': 0.089, 'compound': 0.3919} | 0.39   | pos            |
| The new Cold War will be centered around #vaccines . It is the new atomic energy and the new oil. | {'neg': 0.134, 'neu': 0.793, 'pos': 0.072, 'compound': -0.4215} | -0.42  | neg            |

Table 1: Vader Scores for tweets

The sentence level sentiment analysis is carried out in the tweets using a two part process of labelling them. The tweets are investigated using the vader lexicon is compared with the manual labels using the accuracy calculation. Accuracy, Precision and Recall [27] is calculated using the following equation:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]
Recall = \( \frac{TP}{TP + FN} \)

TP – True Positive
TN – True Negative
FP – False Positive
FN – False Negative

The precision, recall and f1 scores has also been calculated.

5. Topic Mining

To identify the topics discussed along with the vaccine, we have used Latent Dirichlet Allocation for topic allocation. The topic model follows the Markov approach where the next state of the model depends on the current state [28]. Initially the state is set as random. Repeatedly the best topic for the word is selected.

Consider document D has multiple number of topics \( T_N \). The LDA algorithm has two steps (i) Generative process (ii) Topic distribution.

(i) Choose \( N \sim \text{Poisson}(\xi) \)
(ii) Choose \( \theta \sim \text{Dir}(\alpha) \)
(iii) For each of the N words \( w_n \):
   a. Choose a topic \( z_n \sim \text{Multinomial} \)
   b. Choose a word \( w_n \) from \( p(w_n \mid z_n, \beta) \)

\[ [\beta]_{k \times V} \beta_j = p(w_j = 1 \mid z = 1) \]

Dirichlet Distribution:

A k – dimensional Dirichlet random variable \( q \) can take values in the (k-1) simplex. The probability density is,

\[
(1) \quad p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}
\]

\[
(2) \quad p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta)
\]

\[
(3) \quad p(w | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d^k \theta
\]

\[
p(D | \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d^k \theta_d
\]
More concentratedly, we use the topic modelling algorithm to cluster the set of topics based on the word occurrences in the tweet. The algorithm calculates the probability of words which binds to a particular topic from the mixture of topics.

We have followed the following algorithm to analyse the topics using LDA.

Algorithm:

Step 1: Create C β Corpus where,

\[ C \equiv D_1, D_2, D_3, ..., D_n \]

Step 2: Create G β ToLowerCase(C)

Step 3: Gn ß Pre-process G

Step 3.1: Tokenize G using NLTK

\[ T = \{VBP, VB, VBG, NN, RB, JJ\} \]

Step 3.2: Remove Punctuationsm URLs

Step 3.3: Remove Stop words in English

Step 4: LDA Topic Modelling

Step 4.1: Structuring Input G_n

Step 4.2: Matrix Representation

Step 4.3: Extract Features F

Step 4.4: Set Components and Iteration

Step 4.5: Print Topics

We conduct the work in two parts : Pre processing tweets and applying the LDA Gensim algorithm. We have considered each tweet as a document D and assigned the set of tweets to a Corpus C where \( D_1, D_2, D_3, ..., D_n \). The corpus is converted to lower case for efficient cleaning and stored in G. The cleaning of tweets follows tokenization process includes removing punctuations and URLs present in each document. Words belong to the following four categories (Verbs – VBP; VB; VBG, Noun - NN, Adverb - RB and Adjective - JJ) are being tokenized after stop word removal. The documents in the form of word set has to be converted into a structured input for LDA. We use CountVectorizer for feature extraction which assigns an integer to each term in the corpus and convert them into vector. The parameter tuning of LDA algorithm is essential to get accurate results. The dimensionality (number of topics) ‘k’ is
assigned to ‘2’ since the context of the tweets are converging to a small topic called as Covid vaccine. The number of iterations is set to 200 for evaluation.

6. Discussion And Results

In this section, we discussed the results derived in the analysis. Firstly, the tweets which are manually annotated has been segregated into two classes such as positive and negative and the graph is plotted to show the percentage of tweets in each class. The graph depicts that, nearly 90% of them are positive and 10% are negative tweets.

Secondly, the labelled tweet classes are compared with the sentiment lexicon (VADER) and the performance is tabulated. The overall accuracy achieved is 74.2%. The precision, recall, f1-score is been tabulated for both the classes positive and negative.

|                | Precision | Recall | F1-score | Support |
|----------------|-----------|--------|----------|---------|
| Neg            | 0.14      | 0.35   | 0.20     | 416     |
| Pos            | 0.92      | 0.78   | 0.84     | 3959    |
| Micro Avg      | 0.74      | 0.74   | 0.74     | 4375    |
| Macro Avg      | 0.53      | 0.57   | 0.52     | 4375    |
| Weighted Avg   | 0.87      | 0.74   | 0.78     | 4375    |

Table 2. Performance of the labelling of tweets

Finally, the discussions on the topic Covid vaccine is retrieved. The Latent Dirichlet Allocation algorithm extracts the highly influencing topics over the tweets. The high scaled probability topics contain the words such as, Safety of the vaccine” and “Trump, Election and Putin”

7. Conclusion And Future Work

In this work, we collect a sample set of tweets to understand the sentiment of Covid 19 vaccine proposed by Russia using machine learning. The topic modelling is carried out with LDA. In future work, this can be extended with advanced sentiment analysis models with specific features and comparative analysis can be done with several other discussions on vaccines. Topic modelling can be extended to reveal detailed analysis.

Abbreviations

LDA-Latent Dirichlet Allocation

VADER-Valence Aware Dictionary and sEntiment Reasoner
Declarations

Availability of data and materials

Data has been collected by the authors themselves using the twitter API.

Competing interests

The authors declare that they have no competing interests

Funding

No Funding

Authors' contributions

Author 1 has collected, annotated, analysed and interpreted the data.

Author 2 has annotated the data and has checked the accuracy for the analysis.

Acknowledgements

Not applicable

References

[1] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREC. Vol. 10. No. 2010. 2010.

[2] Novak, Petra Kralj, et al. "Sentiment of emojis." PloS one 10.12 (2015).

[3] Wolny, Wieslaw. "Sentiment analysis of Twitter data using emoticons and emoji ideograms." Studia Ekonomiczne 296 (2016): 163-171.

[4] Maynard, Diana G., and Mark A. Greenwood. "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis." LREC 2014 Proceedings. ELRA, 2014.

[5] Bharti, Santosh Kumar, et al. "Sarcastic sentiment detection in tweets streamed in real time: a big data approach." Digital Communications and Networks 2.3 (2016): 108-121.

[6] Archana, R., and S. Chitrakala. "Explicit sarcasm handling in emotion level computation of tweets-A big data approach." 2017 2nd International Conference on Computing and Communications Technologies (ICCCT). IEEE, 2017.

[7] Song, Fengxiang, et al. "Emerging 2019 novel coronavirus (2019-nCoV) pneumonia." Radiology 295.1 (2020): 210-217.
[8] Novel, Coronavirus Pneumonia Emergency Response Epidemiology. "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China." Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi 41.2 (2020): 145.

[9] World Health Organization. "Novel Coronavirus (2019-nCoV): situation report, 3." (2020).

[10] Wang, Chen, et al. "A novel coronavirus outbreak of global health concern." The Lancet 395.10223 (2020): 470-473.

[11] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses

[12] Nicole Lurie, M.D., M.S.P.H., Melanie Saville, M.D., Richard Hatchett, M.D., and Jane Halton, A.O., P.S.M, Developing Covid-19 Vaccines at Pandemic Speed

[13] https://www.ecdc.europa.eu/en/covid-19/latest-evidence/vaccines-and-treatment

[14] https://www.livemint.com/news/world/russia-develops-world-s-first-covid-19-vaccine-putin-s-daughter-gets-vaccinated-11597135192189.html

[15] Ashish Kumar, MBBS, Safi U Khan, MD, Ankur Kalra, MD, FACP, FACC, FSCAI, COVID-19 pandemic: a sentiment analysis, European Heart Journal, , ehaa597, https://doi.org/10.1093/eurheartj/ehaa597

[16] Medford, Richard J., et al. "An" Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak." medRxiv (2020).

[17] Chukwusa, Emeka, Halle Johnson, and Wei Gao. "An exploratory analysis of public opinion and sentiments towards COVID-19 pandemic using Twitter data." (2020).

[18] Gupta, Raj Kumar, Ajay Vishwanath, and Yinping Yang. "COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes." arXiv preprint arXiv:2007.06954 (2020).

[19] Dubey, Akash Dutt, and Shreya Tripathi. "Analysing the sentiments towards work-from-home experience during covid-19 pandemic." Journal of Innovation Management 8.1 (2020).

[20] Buckman, Shelby R., et al. "News Sentiment in the Time of COVID-19." FRBSF Economic Letter 8 (2020): 1-05.

[21] Sanders, Abraham, et al. "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse." medRxiv (2020).

[22] Kraemer-Eis, Helmut, et al. The market sentiment in European private equity and venture capital: Impact of COVID-19. No. 2020/64. EIF Working Paper, 2020.
[23] Duan, Yuejiao, Lanbiao Liu, and Zhuo Wang. "COVID-19 Sentiment and Chinese Stock Market: Official Media News and Sina Weibo." Available at SSRN 3639123 (2020).

[24] Barkur, Gopalkrishna, and Giridhar B. Kamath Vibha. "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India." Asian journal of psychiatry (2020).

[25] Samuel, Jim, et al. "Feeling Positive About Reopening? New Normal Scenarios from COVID-19 Reopen Sentiment Analytics." medRxiv (2020).

[26] C.J. Hutto, Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text"

[27] Goutte C., Gaussier E. (2005) A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Losada D.E., Fernández-Luna J.M. (eds) Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science, vol 3408. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25

[28] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3 (2003) 993-1022

**Figures**

![Word Cloud representation of tweets](image-url)

**Figure 1**

Word Cloud representation of tweets
M – Number of Documents
N – Number of words
α - Document topic density
β - Topic Word density
Θ - Topic distribution
Z_{ij} – Topic in i^{th} document for word j
W_{ij} – word

Figure 2

LDA Graphical Model
Figure 3

Graph showing the percentage of tweets in two classes
Figure 4

Topics with high probability words