Paper Review 01

Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm

Md. Alahi Almin Tansen (190321028)

Department of Computer Science and Engineering, European University of Bangladesh

CSE – 459: Internet of Things

Date: 26-03-2023

Purpose of This Research

The purpose of this research was to analyze the sentiments of general people towards the COVID-19 vaccines Pfizer, Moderna, and AstraZeneca. The researchers extracted tweets from Twitter using a Twitter API authentication token and processed them using Natural Language Processing (NLP). The processed data was then classified using a supervised KNN classification algorithm into three classes: positive, negative, and neutral. These classes refer to the sentiment of the general people whose tweets were extracted for analysis. The results of the analysis showed the percentage of positive, negative, and neutral sentiment towards each of the three vaccines.

Previous Research Gap

There are four previous research mentioned here

- Jia Xue et al. conducted a study that involved extracting data from Twitter using multiple hashtags and performing sentiment analysis on this data using the LDA machine learning algorithm the study found that many tweets related to COVID-19 express fear as a prevalent emotional sentiment.
- One study mentioned where twitter data was extracted manually by data crawling using Twitter API access token with “Vaccine” and “COVID-19” as keywords. The sentiment analysis was conducted using the Naïve Bayes algorithm, and it was found that the majority of tweets related to COVID-19 and vaccines expressed negative sentiment.
- Another study referenced where Twitter data related the keyword "COVID". The data was then preprocessed using NLP techniques. Finally, sentiment classification was conducted on the preprocessed data using RNN.
- Additionally, another research mentioned where extracted raw tweets from Twitter using keywords, used NLP preprocessing, and then conducted topic modeling with an unsupervised LDA algorithm. To determine tweet sentiment, they implemented the valence aware dictionary and sentiment reasoner (VADER).

In this paper the researcher conducted a study on public sentiments regarding COVID-19 vaccines using Twitter data and NLP techniques with a supervised KNN classification algorithm. Tweets related to Pfizer, Moderna, and AstraZeneca are preprocessed, and polarity and subjectivity are determined before being classified by the KNN algorithm into three categories: positive, negative, or neutral.

Proposed System

This paper proposes a system for analyzing Twitter tweets about the COVID-19 vaccines from Pfizer, Moderna, and AstraZeneca. The system uses Natural Language Processing (NLP) and a supervised machine learning classification algorithm to determine the sentiment of each tweet as positive or negative. To implement the system, tweet data is fetched from Twitter using Tweepy library, saved in CSV file format, and preprocessed using NLP techniques like tokenization, normalization, and lemmatization. After preprocessing, polarity and subjectivity are calculated, and a supervised KNN classifier is used to classify the polarity data. Finally, the classified data is visualized and compared.

Architecture

There is no specific architecture used in this research. The researcher used Natural Language Processing (NLP) including text conversion to lower case, stop word removal, fixing misspelled words, replacing emojis with plain English, removing special characters/URLs/HTML tags, tokenization, normalization, and lemmatization. Object identification is the final step of preprocessing where each data column is checked if it is blank and set to value 0 or 1 accordingly in a new identification column. And K-nearest neighbor classification algorithm is used. It involves loading the dataset, selecting the value of K, calculating the distance between each data point using Euclidean distance, sorting the data point according to the calculated distance, selecting the top K rows, assigning the data point based on the most frequent class, and ending the algorithm.

Experimental Procedure

Researcher used different visualization techniques to analyze the processed tweet data related to Pfizer, Moderna, and AstraZeneca vaccines. Word cloud and bar diagrams are used for visualizing the most commonly occurring words and polarity/subjectivity scores respectively. Scatter plots are also used for a better understanding of frequency distribution. Additionally, maximum average polarity scores for 10 tweets are calculated and visualized separately for each vaccine. Then converting the data into polarity scores. The KNN classification algorithm is used to classify the polarity scores into three categories: positive, negative, and neutral.

To calculate the distance of the data point in the KNN algorithm Euclidean distance is calculated using,

$$d\,(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

The resulting classification for Pfizer, Moderna, and AstraZeneca vaccines is shown in table indicating positive, negative, and neutral sentiments in the tweet data. Also, visualization of the sentiment analysis results, highlighting that people have generally less positive sentiment towards AstraZeneca vaccine compared to Pfizer and Moderna vaccines, with higher negative sentiment as well.

Future Plan

The results using NLP preprocessing and KNN classification algorithm show that Pfizer had a positive sentiment rate of 47.29%, Moderna had a positive sentiment rate of 46.16%, while AstraZeneca had a positive sentiment rate of 40.08%. Therefore, people have a higher positive sentiment towards Pfizer and Moderna vaccines compared to AstraZeneca. These findings can help authorities to provide people with the vaccine they trust, which may lead to peaceful control of the pandemic. More research can be done towards sentimental analysis to determine newly invented vaccines and current vaccine performance.

References

1. Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm by FMJM Shamrat, Sovon Chakraborty, MM Imran, Jannatun Naeem Muna, Md Masum Billah, Protiva Das, OM Rahman