Fake News Detection about COVID-19 Vaccine on Social Media

اكتشاف الأخبار المزيفة بخصوص مطعوم كوفيد 19 ـعلى منصات التواصل الاجتماعي

By
Noura Zeiad Al-telawei
Supervisor:
Dr. Mohammad Al-azzeh
Full Professor
Co-Supervisor:
Dr. Fadi Almasalha
Associate Professor

Master Thesis
This thesis proposal was submitted in Partial Fulfilment of the Requirements for the Master's Degree in Computer Science
Deanship of Scientific Research and Graduate Studies
Applied Science Private University
Amman-Jordan

Abstract

This thesis aims to detect fake news of the COVID-19 vaccine on the Twitter platform based on Valence Aware Dictionary and Sentiment Reasoning (VADER) value, percentage of positive words in the tweet aimed at encouraging vaccination, and percentage of negative words in the tweet aimed at frightening against vaccination. The datasets are taken from tweets on the Twitter social media platform, containing 4000 tweets not labeled as fake news or not fake news about COVID-19 vaccines, which are used by the entire world on large scale. We processed tweets using NLP techniques such as (text normalization, tokenization, and stemming) into a form that is predictable and analyzable. We applied our rule to label the tweets as fake news or not fake news based on VADER sentiment analysis value, positive words percentage, and negative words percentage based on the data published by the world health organization website and many authorized websites about the Covid-19 vaccines. The VADER sentiment analysis value, positive and negative words percentage in tweets has been entered into machine learning classification algorithms (Random Forest, Support Vector Machine, Naive Bayes, and Logistic Regression) to predict the Covid-19 vaccine fake news. Using our solution as an API service reduces and stops the spread of false rumors about the Covid-19 vaccine and increases the demand for Covid-19 vaccines to limit the spread of the virus Covid-19. For results, the Random Forest model achieved the best accuracy of 91% than other models, while the Support Vector Machine model was achieved an accuracy of 83%, the Logistic Regression classifier achieved an accuracy of 55%. Finally, the Naïve Bayes classifier achieved a low accuracy of 15%.