World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:02, 2023

Open Science Index, Computer and Information Engineering Vol:17, No:02, 2023 publications.waset.org/abstracts/153206.pdf

Interpretation of the Russia-Ukraine 2022 War via N-Gram Analysis

Authors : Elcin Timur Cakmak, Ayse Oguzlar

Abstract : This study presents the results of the tweets sent by Twitter users on social media about the Russia-Ukraine war by bigram and trigram methods. On February 24, 2022, Russian President Vladimir Putin declared a military operation against Ukraine, and all eyes were turned to this war. Many people living in Russia and Ukraine reacted to this war and protested and also expressed their deep concern about this war as they felt the safety of their families and their futures were at stake. Most people, especially those living in Russia and Ukraine, express their views on the war in different ways. The most popular way to do this is through social media. Many people prefer to convey their feelings using Twitter, one of the most frequently used social media tools. Since the beginning of the war, it is seen that there have been thousands of tweets about the war from many countries of the world on Twitter. These tweets accumulated in data sources are extracted using various codes for analysis through Twitter API and analysed by Python programming language. The aim of the study is to find the word sequences in these tweets by the n-gram method, which is known for its widespread use in computational linguistics and natural language processing. The tweet language used in the study is English. The data set consists of the data obtained from Twitter between February 24, 2022, and April 24, 2022. The tweets obtained from Twitter using the #ukraine, #russia, #war, #putin, #zelensky hashtags together were captured as raw data, and the remaining tweets were included in the analysis stage after they were cleaned through the preprocessing stage. In the data analysis part, the sentiments are found to present what people send as a message about the war on Twitter. Regarding this, negative messages make up the majority of all the tweets as a ratio of %63,6. Furthermore, the most frequently used bigram and trigram word groups are found. Regarding the results, the most frequently used word groups are “he, is”, “I, do”, “I, am” for bigrams. Also, the most frequently used word groups are “I, do, not”, “I, am, not”, “I, can, not” for trigrams. In the machine learning phase, the accuracy of classifications is measured by Classification and Regression Trees (CART) and Naïve Bayes (NB) algorithms. The algorithms are used separately for bigrams and trigrams. We gained the highest accuracy and F-measure values by the NB algorithm and the highest precision and recall values by the CART algorithm for bigrams. On the other hand, the highest values for accuracy, precision, and F-measure values are achieved by the CART algorithm, and the highest value for the recall is gained by NB for trigrams.

Keywords : classification algorithms, machine learning, sentiment analysis, Twitter
Conference Title : ICBSML 2023 : International Conference on Bayesian Statistics for Machine Learning
Conference Location : Lisbon, Portugal
Conference Dates : February 06-07, 2023