**NYIT**

**Spring 2023**

# Sentiment Analysis of Movie Reviews Using Natural Language Processing and Machine Learning


**Name:** Yanting Wu, Gahyeon Back, Jabili Sandadi

**School ID#:** 1300990, 1307886, 1320015


**Course:** Machine Learning

**Course ID:** DTSC-710-M02

**Date:** 5/12/2023

## Table of Contents

# Introduction

The primary objective of this project was to conduct a sentiment analysis on movie reviews. The sentiment is either positive or negative, represented by '1' and '0', respectively. Two distinct sentiment analysis tools were employed: VADER (Valence Aware Dictionary and Sentiment Reasoner), a tool from the nltk library, and a RoBERTa-based model provided by the transformers library. Both tools were compared to evaluate their performance and precision in determining the sentiment of reviews. The purpose of this comparison was to understand the strengths and weaknesses of each tool and to gain a comprehensive understanding of their application in real-world sentiment analysis tasks.

# Methodology

## Dataset

The data used in this project was compiled from two popular movie review websites:

Rotten Tomatoes Movie Reviews: This dataset was primarily sourced from Rotten Tomatoes and consisted of critic reviews. The reviews were categorized as either 'Fresh' or 'Rotten', which were subsequently converted to '1' for positive sentiment and '0' for negative sentiment.

IMDB Movie Reviews: The second dataset was from IMDB, where the reviews were labelled as 'positive' or 'negative'. These labels were similarly converted to '1' for positive sentiment and '0' for negative sentiment.

## Data Preprocessing

The data was initially read into a pandas DataFrame to facilitate its manipulation and analysis. Each review was tokenized using the nltk library. Tokenization is the process of breaking down text into individual words or tokens. These tokens were then tagged with their respective Part-of-Speech (POS) tags, and named entities were identified. This step allowed for an understanding of the grammatical structure of the reviews and helped in the identification of key phrases and names.

## Baseline

The initial sentiment analysis was conducted using the VADER tool from the nltk library. VADER is a lexicon and rule-based sentiment analysis tool that is particularly sensitive to sentiments expressed in social media. It provides four sentiment metrics from these word ratings: positive, neutral, negative, and compound. The compound score, a metric that calculates the sum of all the lexicon ratings, standardizes the scores between -1 (most extreme negative) and +1 (most extreme positive). This served as the baseline for the sentiment analysis.

## Model Description and Implementation

To enhance the sentiment analysis results, a more sophisticated tool was used: a transformer-based model, specifically a RoBERTa-based model pre-trained on Twitter data for sentiment analysis. This model was loaded from the Hugging Face Model Hub.

The RoBERTa model was utilized to predict the sentiment of the review texts. The model returned three scores representing negative, neutral, and positive sentiments. These scores were then converted to probabilities using the softmax function to provide a more interpretable result. Each review's sentiment was predicted by comparing these scores and selecting the sentiment with the highest probability.

## Computational Requirements

This project was implemented using Python, with extensive use of libraries such as pandas for data manipulation, nltk for initial text processing and sentiment analysis, seaborn and matplotlib for data visualization, and transformers for implementing the RoBERTa model.

Due to the transformer model's computational intensity, it is recommended to run this code on a machine with sufficient memory and a powerful CPU, or even a GPU if available. The transformer models provided by Hugging Face are optimized for both CPU and GPU usage, enabling faster processing times.

## Results

The analysis of the movie reviews from the Rotten Tomatoes and IMDB datasets provided interesting insights into the performance of the Vader and RoBERTa sentiment analysis models.

In the Rotten Tomatoes dataset, we visualized the compound sentiment score as given by Vader for both review types. We found that the compound sentiment score was generally higher for positive reviews than for negative ones. Similar observations were made when using the RoBERTa model.

In the case of the IMDB dataset, a combined sentiment analysis approach was taken where the final sentiment prediction was made based on both Vader and RoBERTa outputs. The model's accuracy was calculated based on the proportion of reviews where the predicted sentiment matched the actual sentiment. This approach yielded a high accuracy rate.

In terms of individual sentiment scores (positive, negative, neutral), distributions of these scores were visualized for both Vader and RoBERTa outputs. It was interesting to see that the distributions varied between the two models, suggesting that they interpret and analyze sentiments differently.

The pair plots offered a comparison of the sentiment scores given by Vader and RoBERTa for both datasets, revealing interesting relationships and trends in the data.

The bar plot comparison of positive sentiment scores from both Vader and RoBERTa demonstrated that there can be significant differences in the sentiment analysis results from the two models.

## Conclusion

In conclusion, the study showed that both Vader and RoBERTa are robust sentiment analysis models, each with its unique strengths and analysis styles. However, the effectiveness of these models can depend on the nature and context of the data they are used on. Hence, it is crucial to understand the underlying data and the specific requirements of the task before choosing a sentiment analysis model. This project has demonstrated one approach to compare and contrast these models and should serve as a useful guide for similar tasks in the future.

# References


1. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkore