Alexandria University
Institute of Graduate Studies and Research
Department of Information Technology
Master program in Information Technology

Python Sentiment Analysis

Project with NLTK and IMDB Dataset of movies reviews and translation

Course title: data analytics

Group 7

Students' names:	Id
1- Khaled Mohamed Fathallah	1404-3-101
2- Tamer Amer Zohdy	1404-3-086
3- Tarek Mohamed Ahamed	1404-3-031
4- Fatma Ebrahim Eldosoky	1404-3-103

Introduction:

This report aims to provide comprehensive analysis of public opinion on social media regarding movies. social media provides all stakeholders in the film industry with valuable information. Data was collected using advanced sentiment analysis tool python to understand people's overall sentiment towards the topic.

Data:

1- Data collection (data source):data was collected form http://www.kaggle.com
2- data definition's: type of data is text and data structure are text file.it is tweets about opinions about a group of movies and these opinions vary between positive, negative and neutral reviews

Steps of coding

First step importing important libraries like: pandas, NumPy, so, and time.

After that we read the csv file of the used data set

1- We import necessary libraries including NLTK for text preprocessing tasks.

2- We define a function preprocessing text to perform the preprocessing steps described above.

3- Inside this function, we sequentially apply each preprocessing step to the input text.

4- Finally, we apply this preprocessing function to each review in the dataset and store the preprocessed text in a new column called 'Preprocessed _Reviews'.

5-After running this code, we'll have preprocessed text data ready for sentiment analysis.

After that we apply the word scheme to visualize the frequency of the most common words in the data set

After that we made the 3d word cloud of the most common words in the data set

After that we apply:

1- We import the Sentiment Intensity Analyzer from NLTK's Vader module.

2- We initialize the sentiment analyzer.

3- We define a function get _ sentiment _score to compute the sentiment score (compound score) for each preprocessed review using VADER.

4- We apply this function to each preprocessed review and store the sentiment scores in a new column called 'Sentiment _ Score'.

5- We define another function classify _ sentiment to classify the sentiment as positive, negative, or neutral based on the compound score.

6- We apply this classification function to each sentiment score and store the sentiment labels in a new column called 'Sentiment'.

7- Now, each review in the dataset will have a sentiment score and a sentiment label assigned to it.

Now it’s time to train our model

This step involves training a model using historical sentiment data to forecast future sentiment trends related to your chosen topic.

A basic approach using a simple machine learning model:

1- Feature Extraction: Extract features from the preprocessed text data. This could include word frequency counts, TF-IDF scores, or word embeddings.

2- Split Data: Split the dataset into training and testing sets.

3- Model Training: Train a machine learning model (such as logistic regression, random forest, or support vector machine) using the training data and their corresponding sentiment labels.

4- Model Evaluation: Evaluate the trained model's performance using the testing data. Common evaluation metrics for sentiment analysis include accuracy, precision, recall, and F1-score.

5- Predictive Analysis: Once you have a trained model, you can use it to predict future sentiment trends by feeding it with new text data.

After that:

1- We use TF-IDF (Term Frequency-Inverse Document Frequency) to extract features from the preprocessed text data.

2- We split the dataset into training and testing sets.

3- We train a logistic regression model using the training data.

4- We evaluate the trained model's performance using accuracy and classification report on the testing data.

5- Finally, we predict sentiment for new data using the trained model.

The next step is evaluation and reporting

1-WE need to assess the performance of our sentiment analysis model and document our findings. Here's how we can proceed:

2- Model Performance Evaluation: We'll evaluate the performance of the sentiment analysis model by using appropriate metrics such as accuracy, precision, recall, and F1-score.

3-Sentiment Distribution Analysis: We'll analyze the distribution of sentiment labels in the dataset to understand the overall sentiment trends.

4- Insights and Findings: We'll provide insights into the sentiment analysis results, discussing any patterns or trends observed and highlighting key findings.

5- Implications: We'll discuss the implications of the sentiment analysis findings for the chosen topic or area of interest, considering how they can be applied in real-world scenarios.

Evaluate the performance of the sentiment analysis model using metrics such as accuracy, precision, recall, and F1-score. It also provides a detailed classification report.

Next, we will analyze the distribution of sentiment analysis by this code

Insights and Implications

- Model Performance Insights:
- The sentiment analysis model achieved an accuracy of X%, indicating its ability to classify sentiment accurately.
- Precision, recall, and F1-score provide additional insights into the model's performance across different sentiment classes.

Sentiment Distribution Analysis:

- Positive sentiment accounts for X% of the dataset, followed by negative sentiment (X%) and neutral sentiment (X%).
- Understanding the distribution of sentiment labels provides context for interpreting the sentiment analysis results.

Implications and Results:

- The sentiment analysis findings can be leveraged to gain insights into customer opinions and preferences.

- Businesses can use sentiment analysis to identify areas for improvement, tailor marketing strategies, and enhance customer satisfaction.
- Monitoring sentiment trends over time enables businesses to adapt to changing consumer sentiments and market dynamics.

Overall, the sentiment analysis provides valuable insights that can inform decision-making processes and drive business strategies.

In this section, we provide insights into the sentiment analysis results and discuss the implications for real-world applications.

Conclusion:

Sentiment analysis of social media tweets is a power tool that can be used to better understand the world around us. these technologies continue to develop. they will become even more important for both individuals and organizations