Opinion Mining Project

1 Sentiment Analysis Approaches

Sentiment analysis is a crucial task in opinion mining, focusing on determining the sentiment expressed in a given text. The approaches to sentiment analysis can generally be categorized into **rule-based methods** and **machine learning-based methods**.

1. Rule-Based Approach

- Relies on predefined linguistic rules and sentiment lexicons.
- Words and phrases are assigned sentiment scores (e.g., "excellent" = +2, "terrible" = -2).
- Overall sentiment is calculated by combining these scores, often adjusted for factors like negation (e.g., "not good") and intensity (e.g., "very bad").
- Popular tools for rule-based sentiment analysis include **VADER** and **TextBlob**.

2. Machine Learning-Based Approaches

Traditional Machine Learning: In this approach, text is converted into numerical features using techniques like Bag-of-Words (BoW) or TF-IDF, and machine learning models are trained to classify sentiment. Common algorithms used include:

- Naive Bayes
- Support Vector Machines (SVM)
- Logistic Regression
- Linear Regression

This approach requires extensive feature engineering but often provides good results for structured datasets.

3. Handling Linguistic Challenges

Sentiment analysis models need to address several linguistic complexities, such as:

- **Negation:** Phrases like "not good" or "wasn't terrible" can invert sentiment. Both rule-based and machine learning models use heuristics or training data to handle negation.
- **Sarcasm:** Sarcasm often expresses the opposite of the literal meaning (e.g., "I love waiting for hours!"), making it difficult for models to detect.
- **Idiomatic Expressions:** Expressions like "kick the bucket" or "break the ice" carry meanings that differ from the literal translation, which can be challenging for both rule-based and traditional machine learning approaches.

Summary Table

Approach	Description
Rule-Based	Uses predefined lexicons and linguistic rules to calculate sentiment scores. Simple, interpretable, but less flexible.
Traditional Machine Learning	Converts text into numerical features (TF-IDF, BoW) and uses classifiers such as SVM, Naive Bayes, or Logistic Regression. Requires feature engineering but performs well on structured data.

2 Steps of Opinion Mining

Opinion mining, also known as sentiment analysis, involves a series of steps to extract and classify sentiments from text data. Here are the key steps involved:

1. **Data Collection:** In this step, I started by testing a small dataset generated from AI to understand the different sentiment cases and get familiar with the data.
2. **Text Preprocessing:** After collecting the dataset, I started with cleaning and preparing the text data before feeding it into any model. Here's the exact sequence I followed:

- **Lowercasing:** First, I converted all the text to lowercase. This helped avoid treating the same word differently just because of letter casing — for example, "Happy" and "happy" would be counted as the same.
- **Punctuation and Symbol Removal:** Then, I removed all punctuation marks and symbols like commas, exclamation marks, hashtags, etc. This made the text cleaner and easier to process.
- **Stopword Removal:** After that, I removed common stopwords like "is", "the", and "in" because they don't carry meaningful sentiment and would just add noise to the model.
- **Emoji Removal:** Since the data included emojis, I used a Python library to strip them out. Emojis can be useful in some cases, but for my model they weren't necessary and could affect the tokenization.
- **Tokenization:** I tokenized the text — which means I broke each sentence down into individual words or tokens. This step was crucial for feeding the text into the machine learning pipeline.
- **Lemmatization:** I reduced each word to its base form using POS tagging and the WordNet lemmatizer to improve text consistency.