Sentiment Analysis of Customer Reviews

Abstract

The goal of this project is to develop a simple sentiment analysis application in Python that will be able to identify positive, negative, or neutral feedback/review provided by customers. This project's primary objective is to develop a method for evaluating the feelings of end users. This program will make it possible to identify the feelings that a customer is experiencing.

After a virtual environment has been created and it has been filled with virtual dependencies for a variety of Python libraries, it is necessary to create a code repository and download a dataset relating to customer review.

The dataset will then be cleaned, stemmed, and lemmatized as the following step in the preparation process, followed by vectorization:

1. Using a regular expression to convert uppercase letters to lowercase, eliminating URLs and special characters, web scraping, lemmatization, and stemming.

2. Implementing nltk's stop-word processing and tokenization.

3. Use Tf-IDF to vectorize texts once they have been cleaned. Tf-IDF/term frequency-inverse document frequency is good for vectorize sentiment analysis because it shows how important each word is in a set of documents. It not only focuses on the frequency of words present in the corpus but also provides the importance of the words.

4. N-gram range sets if the following characteristics of texts are going to be utilized:

- Unigrams or words (n-gram size = 1).

Bigrams or terms compounded by two words (n-gram size = 2). Trigrams or terms compounded by up to three words (n-gram size = 3).

ngram_rangetuple (min_n, max_n), default = (1, 1).

(1, 2) means unigrams and bigrams. (2, 2) means only bigrams. (1, 3) means unigrams, bigrams, and trigrams.

Vectorize cleaned tweets with Tf-IDF ngram_range(1,3).

Comparing accuracy of 3 model:

After analyzing the accuracy of the Linear, Bernoulli, and Logistic Regression algorithms, came to the conclusion is that the Logistic Regression model has the best accuracy and should be used.

The model is being trained, validated, and evaluated with the use of logistic regression. Separate the dataset into the Training data and the Testing data. 70% training data, 30% testing data. Python's pickle package is used to store the models.

Obstacle:

Because the dataset only comprised two emotions ("positive" and "negative"), a "pre-trained" version of Vader was used to discover a "neutral" feeling.

This model, called VADER (Valence Aware Dictionary and sEntiment Reasoner), was developed in 2014 and employs rule-based values that are adapted to feelings collected from social media. It has already been trained. It analyzes the words in a communication in order to determine not only the good and negative feelings evoked by the words, but also the intensity of those feelings. (Hutto, C.J., and Gilbert, E.E. both of the U.S. (2014).

When the value is larger than zero, a positive feeling is being conveyed. "Neutral sentiment" is the term used to describe anything that has no weight or significance. When the value is less than zero, a person is said to have a negative emotion.

Perform Sentiment Analysis:

The Streamlit API is used for the purpose of user input collecting due to the convenience and simplicity with which it. By loading the model and vectorizer, may make predictions about whether the sentiment analysis will be positive, negative, or neutral. Then display and report the result of the predictions.

Continues improvement.

It's likely that the model won't be quite as accurate in the beginning since the first dataset is so tiny because it only has 1000 customer reviews. The model is trained and verified with new data on an iterative basis. Periodically training automation using schedule as a Python library is used in order to create continuous improvements.

Conclusion:

The Artificial Intelligence (AI) provided by Natural Language Processing makes it feasible for the Customer review application to forecast the sentiment of customer reviews without requiring any assistance from a person. These predictions can be made regardless of whether the review was positive, negative, or neutral.

Literature:

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.

H.Wu and R. Luk and K. Wong and K. Kwok.(2008).Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems.

Hutto, C.J. & Gilbert, E.E. (2014).vaderSentiment. https://github.com/cjhutto/vaderSentiment

Scikit-learn 1.1.2. (2022). https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

What does Tf-idf means.(2021). http://www.tfidf.com/

Streamlit Community Cloud.(2021). https://streamlit.io/cloud

Web Article:

Sheel Saket.(2020). Count Vectorizer vs TFIDF Vectorizer | Natural Language Processing.

Topic Modeling and Sentiment Analysis with LDA and NMF on Moroccan Tweets. Retrieved from

https://link.springer.com/chapter/10.1007/978-3-030-66840-2_12

Anand Borad(2020). NLP text Pre-Processing: Text Vectorization. Retrieved from:

https://www.einfochips.com/blog/nlp-text-vectorization/#:~:text=There%20are%20three%20most%20used,separate%20article%20for%20word%20embedding