Project - Twitter US Airline Sentiment

Submission type
:
File Upload
Due Date
:
May 01, 8:25 AM
Total Score
:
60
Available from
:
Apr 15, 5:25 AM

Description

Hi Learners,

Welcome to the project on the module on NLP.

Background and Context:

Twitter posses 330 million monthly active users, which allows businesses to reach a broad population and connect with customers without intermediaries. On the other side, there's so much information that it's difficult for brands to quickly detect negative social mentions that could harm their business.

That's why sentiment analysis/classification, which involves monitoring emotions in conversations on social media platforms, has become a key strategy in social media marketing.

Listening to how customers feel about the product/services on Twitter allows companies to understand their audience, keep on top of what's being said about their brand, and their competitors, and discover new trends in the industry.

Data Description:

A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

Dataset:

The dataset has the following columns:

- tweet_id
- airline_sentiment
- airline_sentiment_confidence
- negativereason
- negativereason_confidence
- airline

- airline_sentiment_gold
- name
- negativereason_gold
- retweet_count
- text
- tweet_coord
- tweet_created
- tweet_location
- user_timezone

Objective:

To implement the techniques learnt as a part of the course.

Learning Outcomes:

- Basic understanding of text pre-processing.
- What to do after text pre-processing
- Bag of words
- Tf-idf
- Build the classification model.
- Evaluate the Model

Steps and tasks:

1. Import the libraries, load dataset, print shape of data, data description. (5 Marks)
2. Understand of data-columns: (5 Marks)
   a. Drop all other columns except "text" and "airline_sentiment".
   b. Check the shape of the data.
   c. Print the first 5 rows of data.
3. Text pre-processing: Data preparation. (16 Marks)

NOTE:- Each text pre-processing steps should be mentioned in the notebook separately.
   a. Html tag removal.
   b. Tokenization.
   c. Remove the numbers.
   d. Removal of Special Characters and Punctuations.
   e. Conversion to lowercase.
   f. Lemmatize or stemming.
   g. Join the words in the list to convert back to text string in the data frame. (So that each row contains the data in text format.)
   h. Print the first 5 rows of data after pre-processing.

4. Vectorization: (10 Marks)
   a. Use CountVectorizer.
   b. Use TfidfVectorizer.
5. Fit and evaluate the model using both types of vectorization. (6+6 Marks)
6. Summarize your understanding of the application of Various Pre-processing and Vectorization and performance of your model on this dataset. (8 Marks)

7.Overall notebook should have:(4 Marks)
   a. Well commented code
   b. Structure and flow