Twitter Sentiment Analysis

Abstract

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

Design

This project has 5 major stages in its life cycle:

- Data Collection
- Data Processing
- Exploratory Analysis of the Data
- Data Modeling
- Interpreting the Data

This project will use of popular libraries such as Plotly for visualizations.

Pyspark would then be used to handle the big Data. Modelling would be done along with TF-IDF and Logistic Regression.

Data

It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment

Content

It contains the following 6 fields:

- target: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
- ids: The id of the tweet
- date: the date of the tweet
- flag: The query (*lyx*). If there is no query, then this value is NO_QUERY.
- user: the user that tweeted
- text: the text of the tweet

Algorithms

- EDA
- Logistic Regression

Tools

- Numpy and Pandas for data manipulation
- pyspark for modelling