Sentimental Analysis of youtube video comments

By

Rohan Chhetry , Sujan Sharma , Nimesh Gautam & Asim Upreti
[MINSKY]

Agendas:

- Introduction To The Projects
- Feasibility Study
- Problem Statement
- System design
- Possible target
- Performance Metrics
- Estimated Project timeline

Introduction

Sentiment analysis is a machine learning techniques which can be used to determine the sensibility behind the texts, i.e. tweets, movie reviews, youtube comments, any incoming message, etc. For this very first project of the fellowship we decided to perform the analysis on youtube comments.
First we have to choose the youtube video and then extract the comments of that video.

Feasibility Study

Technical : Keeping in mind the time given and simplicity of the project chosen we feel as our project is feasible.

Economically: No cost will be encountered as we will be using all open source softwares.

Timescale : We have given a rough estimate for the time we may need for the project.

Problem Statement

In this project, we try to implement sentimental analysis on the comments of the song on youtube that helps to overcome the challenges of identifying the sentiments of the various types of comments.

System Design

Data source for the problem

We are planning to scrape the comments of the youtube video and create our own dataset , the labeling of the dataset may be done by the python package textBlob which generates polarity for sentiment of the comment and we can classify it for our need as -1(negative) , 0(neutral) or 1(positive).

Block diagram

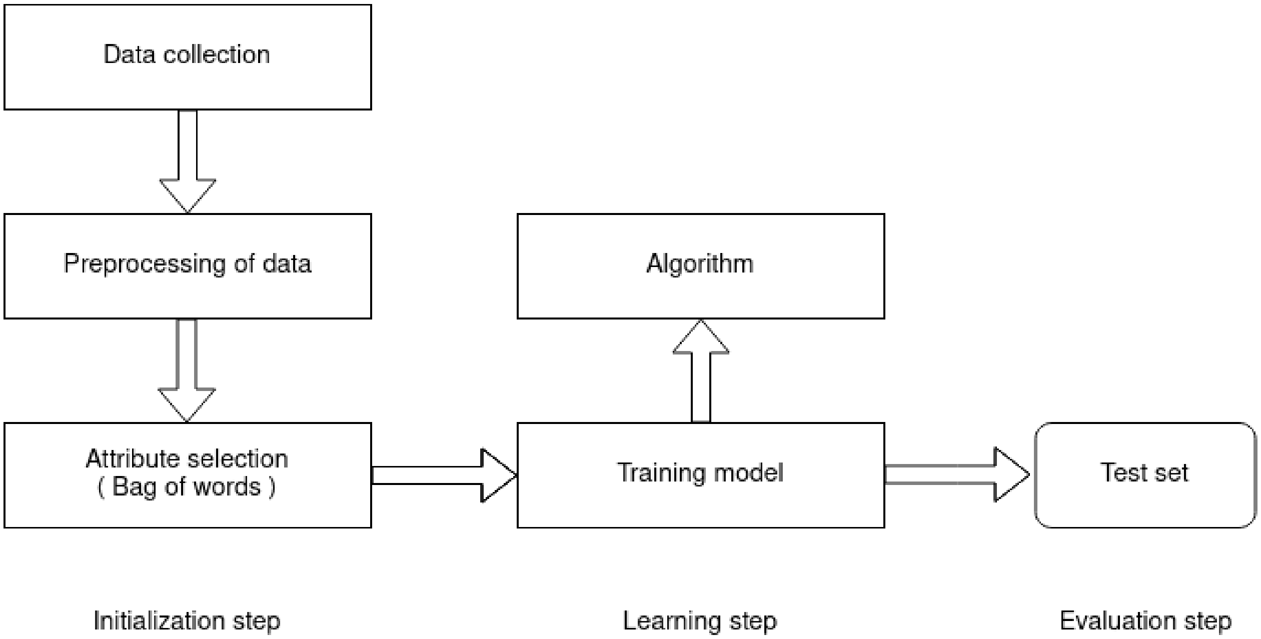

Fig: General block diagram

We are planning to use the NLTK package for initial processing of data. After producing the bag of words , and splitting the data into training and testing sets which will be done by sklearn package, the next step that we will have to perform is the selection of an appropriate machine learning algorithm for classification.

Selection of Machine learning Algorithm

Logistic Regression

It's a predictive modeling algorithm for the classification where there is a labeled dataset with the categorical target variable. It falls into the category of the supervised machine learning algorithm. It will help in predicting the probability of outcomes i.e. binary classification or multi-classification. There are various classification examples which can be done using this algorithm, where they are classified i.e. positive (1) , negative (-1) and neutral (0), which is a multinomial classification problem.

Naive Bayes Classifier

Naive Bayes classification is also a form of supervised learning. It is one of the most straightforward and fast classification algorithms & very well suited for large volumes of data. When used for textual data analysis, such as Natural Language Processing, the Naive Bayes classification yields good results.

One of these algorithm will be used and their corresponding accuracy will be compared.

Possible target

We can expect the model to perform well with new youtube comment data on that video.

Performance Metrics

We have planned on using the confusion matrix (F1-score) as it seems to be the best matric when dealing with unbalanced data. We will also be calculating the precision and recall metrics. As our chosen dataset is of a youtube song we may see a trend of comments being more positive.

Project Timeline Estimate

2 days - scrape data
7 days - clean, label and visualize data
5 days - train model
5 days - evaluate performance
7 days - optimize model
2 days - re evaluate performance

Total: 4 weeks