**SENTIMENT ANALYSIS OF HENRY HARVIN**

**Data collection:**

**Twitter:** [CODE](https://github.com/him2506/Data_scraper)

Got nearly 200 tweets from Twitter using twint and searching for “Henry Harvin Education”.

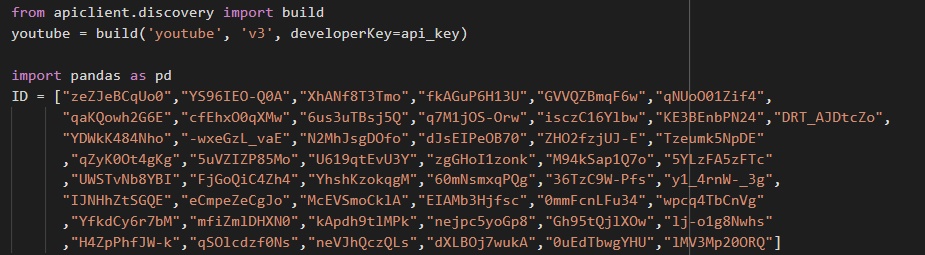
**Google comments:**

Got the reviews on google.Got around 500 reviews.

**Youtube comments:**

Used a scrapping algorithm along with youtube API to get comments from the youtube videos.

\*we had to manually enter all the youtube video IDs since we had many videos with disabled comments and there isn’t a way to find whether a video has disabled comments with youtube API.



**Trust Pilot:**

Scrapped around 150 comments on Henry Harvin education

Then we merged all the above comments into one single excel file.

**Data cleaning:**

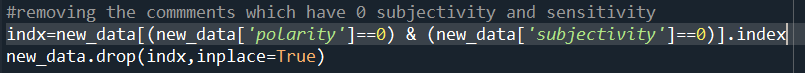
The steps used to complete preprocessing our data were:

1. Make text lowercase
2. Remove punctuation
3. Remove emojis
4. Remove stopwords
5. Lemmatization - converting to root words

**Predicting the sentiments:**

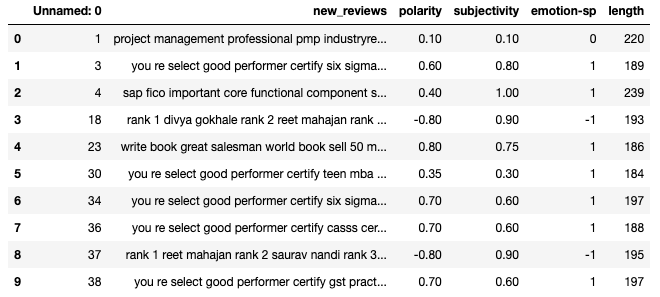
\*Used text blob library to predict the sentiments i.e polarity and subjectivity.

\*Then we set the polarity and subjectivity threshold as 0.3 and 0.2 .

\*Then removed all the comments which had 0 subjectivity and 0 polarities, since these didn’t give any useful insights and most of these comments were just tweets by henry Harvin itself.

**Data visualization and interpretations:**

The data that has been collected and run through an ML model(unsupervised learning approach- Using TextBlob library can be represented as below:

Fig: Screenshot of the dataset 

The polarity, subjectivity, and emotion are represented along with the length of each of the reviews. The descriptive analysis on the above parameters is as follows:

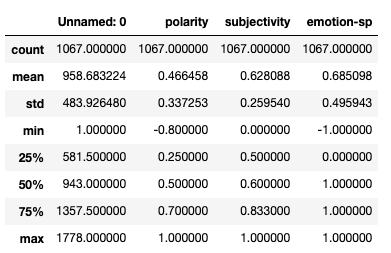


Fig: Screenshot of the descriptive statistics

Polarity: The mean is nearly 0.5 showing more positive reviews hence indicating a positive presence among the reviewers and indicating users are satisfied with the course curriculum.

Subjectivity: The mean is about 0.63 showing the reviews are quite subjective that are opinionated in nature rather than factual. This makes drawing conclusions solely on the basis of reviews difficult.

Using seaborn and matplotlib the data visualizations are carried out.

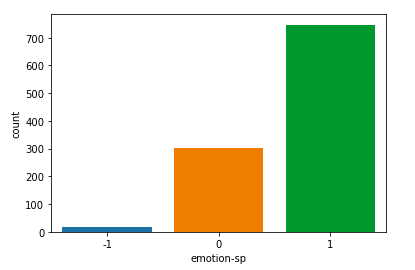
On getting the countplot we find out there are huge no of positive reviews whereas the negative reviews are very less in number.

Fig: The countplot showing the no of negative (-1) , neutral(0), positive(1) reviews.

Sentiment Length:

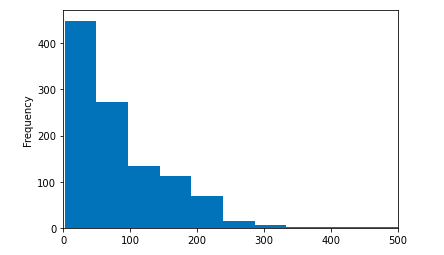


Fig: Length of the sentiments

The above histogram shows that the majority of the reviews have a length of 0 to 50 characters and the number of reviews has reduced as the length increased.

Digging deeper, analyzing the length with respect to the category(positive, negative and neutral), we find that majorly the negative reviews are longer and positive ones are shorter.

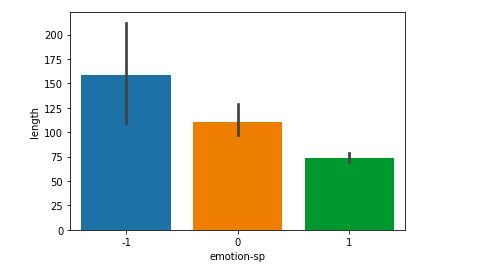


Fig: Category vs Length of the reviews

Density curve of Polarity

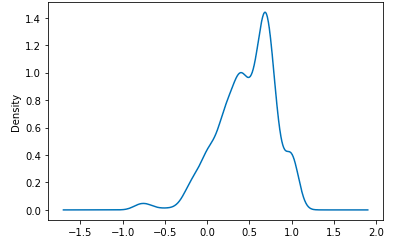


Fig: Density curve of polarity

The above graph depicts that most of the reviews are having the polarity of 0.6 to 0.75 hence indicating users are satisfied with the courses. But the satisfaction level is not very high.

Density curve of Subjectivity

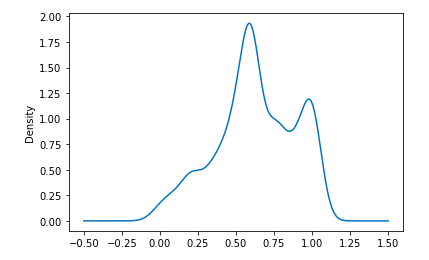
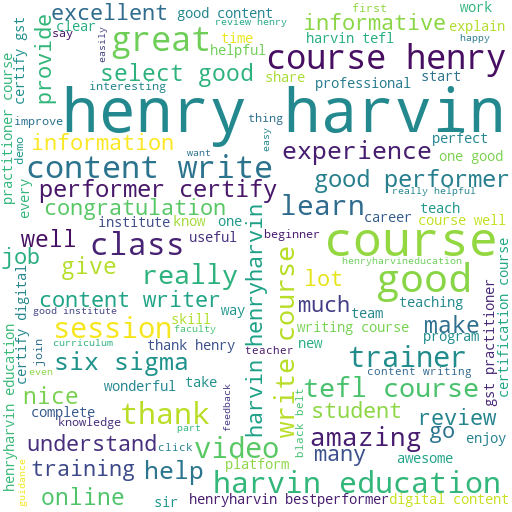


Fig: Density curve of subjectivity

The above graph depicts the majority of the reviews are opinionated and based on emotions (having high subjectivity between 0.5 to 0.6 and 0.9 to 1). This makes making conclusions based on these opinions solely becomes difficult.

**Word Cloud:**

The word cloud below shows the frequently used words in the reviews given by the users. The words in larger font size are more frequent in the reviews such as informative, content writer, six sigma,tefl course, training, and best performer.



**Suggestions:**

* Need to increase the reach - covering more institutions
* Diversify on the courses.
* There are many government colleges and universities that do not have such exposure. Approaching them might help.
* Content writing, six sigma and tefl courses are much popular among Henry Harvin users. So concentrating on these courses will make a good impact.
* Congratulating the best performers is also appreciated by the users. So it can be done extensively for all courses and scholarships can be provided to them.

**TEAM MEMBERS AND THEIR CONTRIBUTION:**

PRANESH PM - youtube reviews and predicting sentiments using text blob.

Sangeetha Ray - trust pilot reviews and data visualization

HIMANCHAL SINGH - Twitter comments, data cleaning, and word cloud.

GOUDA SRIKANTH - google reviews.