

Small Text Telling you my Journey

What did you start by looking at?

1. I try to understand the data.
2. I look at the data type of each column.
3. I check for missing values.
4. I look at the columns and try to relate them to get other meaningful columns (feature engineering).

Which frameworks/technologies did you use?

I used Python because It has a vast collection of libraries for numerical computation and data manipulation and provides libraries for graphics and data visualization to build plots. Here are the libraries I used to make data tasks easier [Pandas, Matplotlib, Seaborn, and Plotly]

Why did you use a certain package?

Pandas: Pandas provide functions to handle missing data, perform mathematical operations, and it is the most advanced and fast-growing tool for data processing and manipulation. It enables data structures to be converted into DataFrame objects, missing information to be processed, DataFrame columns to be added/removed, and missing files to be added.

Matplotlib: Matplotlib is a standard two-dimensional data visualization library. It is flexible and easily configurable. Matplotlib library is commonly used for plotting data points and creating interactive visualizations of the data by writing a few lines of code.

Seaborn: Seaborn is better for more advanced statistical plots and provides more attractive default color palettes, It helps in compiling whole data into a single plot.

Plotly: Plotly supports dynamic charts and animations as a first principle and this is the main difference between other visualization libraries like Matplotlib or Seaborn. Plotly allows you to generate graphs offline.

Why did you do something by hand? Why did you script it?

By Scripts

- I wrote scripts to clean the data, add new columns (features), remove stopwords.

Manually

- I extended the stopwords list..