Lets have a basic data exploration focuses on the properties of a single component of the dataset.

50k movie reviews with both positive and negative sentiments were extracted from IMDb, allowing no more than 30 reviews per movie.

50% of the dataset is fed into the training model while the rest are considered as testing datasets.

IMDB:

IMDB is one of the largest movie databases, which recommends shows and movie and provide insights into popular taste. It also frequently used as predict sentiment.

Data exploration:

null objects, duplicated data, sentiment annotations, and review samples with positive and negative sentiment are scanned in this section. This provides the direction to support further data exploration and pre-processing.

No null values are detected in the dataset.

It is found that 418 reviews are the same. These duplicated data will not cause data imbalance in this project as it is an insignificant portion of the dataset (418\*2/25000 = 1.6% < 2%).

Both positive and negative reviews are equally distributed across the dataset.

There are some examples of bad data in the dataset.

Moreover, there is an example observed that a single review contains a mixture of positive and negative wordings. In addition, sentiment-unrelated paragraphs are noticed within these comments. This may impact the sensitivity of the model.