**3. Methodology:**

3.6 Write about Sentiment Score Calculation:

Vader, a popular sentiment analysis model was used to calculate the sentiment scores of each tweet extracted from Twitter. Vader’s Sentiment Intensity Analyzer library produced a dictionary output that displays sentiment(key) and polarity scores(value). To determine the possible sentiment of each tweet, the value of compound can used as indicator. Scores between -0.05 to 0.05 can be considered as neutral and anything above 0.05 is positive, otherwise negative.

Average (compound) score for each movie was calculated using Pandas groupby aggregate mean (agg) function.

**4. Experiments:**

* 1. Experiment Design: Here, you want to write down the steps you have taken to design the experiment. Please include the machine configuration, different APIs (such as APIs for Twitter, Vader, Pytorch, etc. )

Tweepy, a python library for accessing Twitter API was utilized to extract tweets related to given movie titles. Twitter developer credentials (Access token and Consumer API keys) were defined initially to obtain Twitter function access. Subsequently, Twitter authentication handler and API objects were created to be able to call Twitter cursor function.

* 1. Dataset Preparation: Here, you want to write the steps you have taken to prepare the datasets. You have to include the quantifiable Information (such as average tweets per movie, average ratings per movie, the highest number of ratings for a movie, the lowest number of ratings for a movie) of the datasets used in this experiment.

1. Sentiment Analysis - The dataset used in this section are (1) u.item of MovieLens 100k dataset to get the movie list and tag words and; (2) twitter data to get sentiment scores.

**To prepare the tag words**, a series of steps were performed, such as the following:

1. Removed the trailing (year) for each movie titles
2. Reformat the movie titles with incorrect construction (e.g “Birgcage, The” and “Time to kill, A”)
3. Removed space for each formatted movie titles.

**To clean the twitter data,** the following pre-processing steps were performed:

1. Dropped null and duplicate values
2. Removed alphanumeric characters including emojis.
3. Removed links
4. Removed stopwords by importing nltk.corpus stopwords library.
5. Lemmatized each words using nltk.stem.WordNetLemmatizer.

**Average tweets per movie:** Limit of 5 tweets per movie was enforced during extraction. Despite several attempts to extract all the tweets, we were only able to fetch valid and non-null tweets for 1182 movies out of 1682 original list. Hence, movies with missing tweets (total of 500 movies) were filled with 0 sentiment score value as advised.

**Top 10 movies:** Based on the extracted tweets,“Courage Under Fire” topped the movie with highest positive review of 97.5%. Followed by “The Big Squeeze” with 93.7% average sentiment score.

**Chart

Description automatically generated**

**Bottom 10 movies:** Movie with lowest sentiment score turned out to be “Killer: A Journal of Murder (1995) with -84% followed by “Faster Pussycat! Kill! Kill! with -82%.

Chart, funnel chart

Description automatically generated