Information about the Data

The dataset under consideration is a comprehensive collection of user-generated reviews from the Steam platform. This dataset includes a wealth of information, both textual and numerical, regarding user interactions with games. The fields encompass a range of data points: unique identifiers for both the game (AppID) and the review (recommendationid), user-specific metrics (such as the number of games owned and playtime), review content, sentiment indicators (voted_up), and engagement metrics (votes_up, votes_funny, comment_count).

One of the most intriguing aspects of this dataset is the inclusion of sentiment-related information, which not only covers binary sentiment classification (whether a review is positive or negative) but also quantifies user engagement through metrics like helpful votes and comment counts. The presence of timestamps allows for temporal analysis of review patterns over time, which is crucial for capturing trends and shifts in user preferences.

The dataset's characteristics suggest its potential for building a sentiment-based recommendation system. Such a system can leverage the nuanced understanding of user sentiment and engagement to recommend games that are not only popular but resonate well with users' expressed preferences and experiences.

Method of Data Collection

The primary method of data collection for this dataset is via the Steam Web API, which provides access to user reviews and associated metadata. The API allows for querying specific games using their AppID and retrieving batches of reviews, which include both textual content and various metadata. The dataset's construction involves iteratively fetching and compiling these details across the entire catalog of games available on Steam.

Data Volume Estimate

The raw dataset initially consists of approximately 500,000 instances, representing individual game reviews. After applying a filter to retain only reviews in English, the dataset is refined to around 200,000 instances. Each instance includes both structured data, such as numerical identifiers and categorical sentiment indicators, and unstructured text in the form of user-generated reviews.