![University of Hohenheim logo]

UNIVERSITY OF HOHENHEIM

5107-411
INTRODUCTION TO APPLIED DATA SCIENCE

December 28, 2023

Mandatory Assignment 2 - Sentiment Analysis of User Reviews

Name: Alina Weiss
Registration Number: 633613

Name: Linus von Klitzing
Registration Number: 964205

Study Program: Earth and Climate System Science

Contents

1 Introduction

The video game "New World" by Amazon Games is accessible through Steam, a digital distribution platform for video games. [1] Steam generates substantial amounts of data regarding games featured on their platform, and this data is accessible through an API. This study is founded on the raw data obtained from the prior research titled "User Reviews of Steam Video Games" specifically for the game "New World," which is identified by the Steam ID 1063730 on the Steamworks API. [2] To elaborate, the user review data is utilized for conducting sentiment analysis. Consequently, the data undergoes cleaning and analysis in two distinct ways, each with varying levels of depth.

Sentiment analysis is defined as "the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral" [3].

2 Data Cleaning

To conduct a sentiment analysis, the initial step involves cleaning and filtering the raw data, which is comprised of text strings from user reviews. The key procedures include eliminating all non-ASCII characters, converting all letters to lowercase to standardize the abbreviation "n't" to "not," and removing all stopwords, except for "not" as it can serve as a crucial indicator of sentiment in a review. Additionally, empty reviews are excluded from the analysis.

Finally, a review in its raw form "It had potential... but this game has an ungodly amount of bugs. Most of which have been around since launch. Amazon, I think stick to e-commerce, you're games are a joke." is modified to the following text: "potential game ungodly amount bugs around since launch amazon think stick ecommerce youre games joke". This illustration is based on the first review in the dataset, identified by the Recommendation ID 150109599.

3 Basic Sentiment Analysis

For a basic sentiment analysis, the words in a cleaned review are compared with lists of words expressing positive and negative sentiments. In this approach, a score is computed for each review by assigning a factor of +1 for every positive word and -1 for every negative word. If the word "not" precedes a positive or negative word, the factor is inverted.

The utilized word lists were compiled by Minqing Hu and Bing Liu, who conducted research on mining and summarizing customer reviews. [4]

The outcome of this simplistic analysis is illustrated in Figure 1. Clearly visible is the prevalence of neutral reviews, with a higher proportion of positive than negative reviews. Among the 9,684 total reviews, 24% are categorized as neutral, 57% as positive, and 19% as negative based on this analysis. The histogram further differentiates these values, revealing that the majority of reviews have scores close to 0, indicating they can be described as "slightly positive" or "slightly negative." Large values, representing "strongly positive" or "strongly negative" reviews, are infrequent.

It's important to note that values close to zero may also result from very short reviews.

Consequently, a more in-depth analysis would be necessary for more precise conclusions. Nevertheless, these findings can be compared with the distribution of positive and non-positive reviews in the data collected by Steam, which are 72% and 28%, respectively. Considering the differences in evaluation categories and the simplicity of the sentiment analysis, the deviation in values appears acceptable.

4 Deeper Sentiment Analysis using *syuzhet*

A more in-depth analysis is conducted using the *syuzhet* package for R, designed "for the extraction of sentiment and sentiment-based plot arcs from text" [5]. This analysis employs three dictionaries: "bing" "syuzhet" and "afinn" each generating a score for every review. Additionally, the emotional content of the reviews is explored across categories such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. This exploration is facilitated by the sentiment lexicon "nrc."

When comparing the scores obtained from the basic analysis and the *syuzhet* analysis using the "bing" dictionary, it is evident that their values are often quite close, with a mean deviation of 0.06 and a standard deviation of 2.40. However, their sign differs in 6% of the cases. Performing a similar analysis with the "syuzhet" and "afinn" dictionaries results in a mean deviation of -0.15 with a standard deviation of 2.79 for "syuzhet" and -2.02 with a standard deviation of 4.19 for "afinn." The sign differs in 20% of the cases for both libraries. One possible explanation for this discrepancy may be that all methods in *syuzhet* do not account for negation, as highlighted by Jockers and Jockers [6]. Negation is handled at a basic level in the manual analysis.

Figure 2 depicts the time evolution of sentiment scores calculated using different methods for 500 reviews in January 2023. The amplitude of the scores varies depending on the method, with the highest values achieved using "Afinn." Despite these variations, the sign of the scores generally aligns, indicating consistency in the fundamental tendency of reviews to be positive or negative. This observation is consistent with the results presented in the preceding passage.

In Figure 3, up votes of reviews are plotted against sentiment scores derived from a basic sentiment analysis. All reviews with votes are included in this analysis. The majority of data points cluster within the range of a sentiment score of -5 to 10, and 0 to 3 up votes. Notably, there are instances of extreme sentiment scores, both positive and negative, yet they receive zero up votes. Conversely, there is a review with 46 up votes and a sentiment score of 4. Despite these variations, no discernible trend is visible. This suggests that reviews with extremely high or low sentiment scores are not inherently deemed more helpful by other users.

5 Summary

The report explores sentiment analysis of user reviews for "New World" by Amazon Games on Steam. Employing both basic and advanced methods, including the *syuzhet* package in R, the analysis uncovers consistent sentiment trends despite variations in amplitude and sign. Furthermore, the correlations between sentiment scores and up votes suggest that extreme sentiment scores may not necessarily indicate a high level of

helpfulness in reviews. This study offers a nuanced understanding of user sentiments, integrating both fundamental and sophisticated analytical approaches.
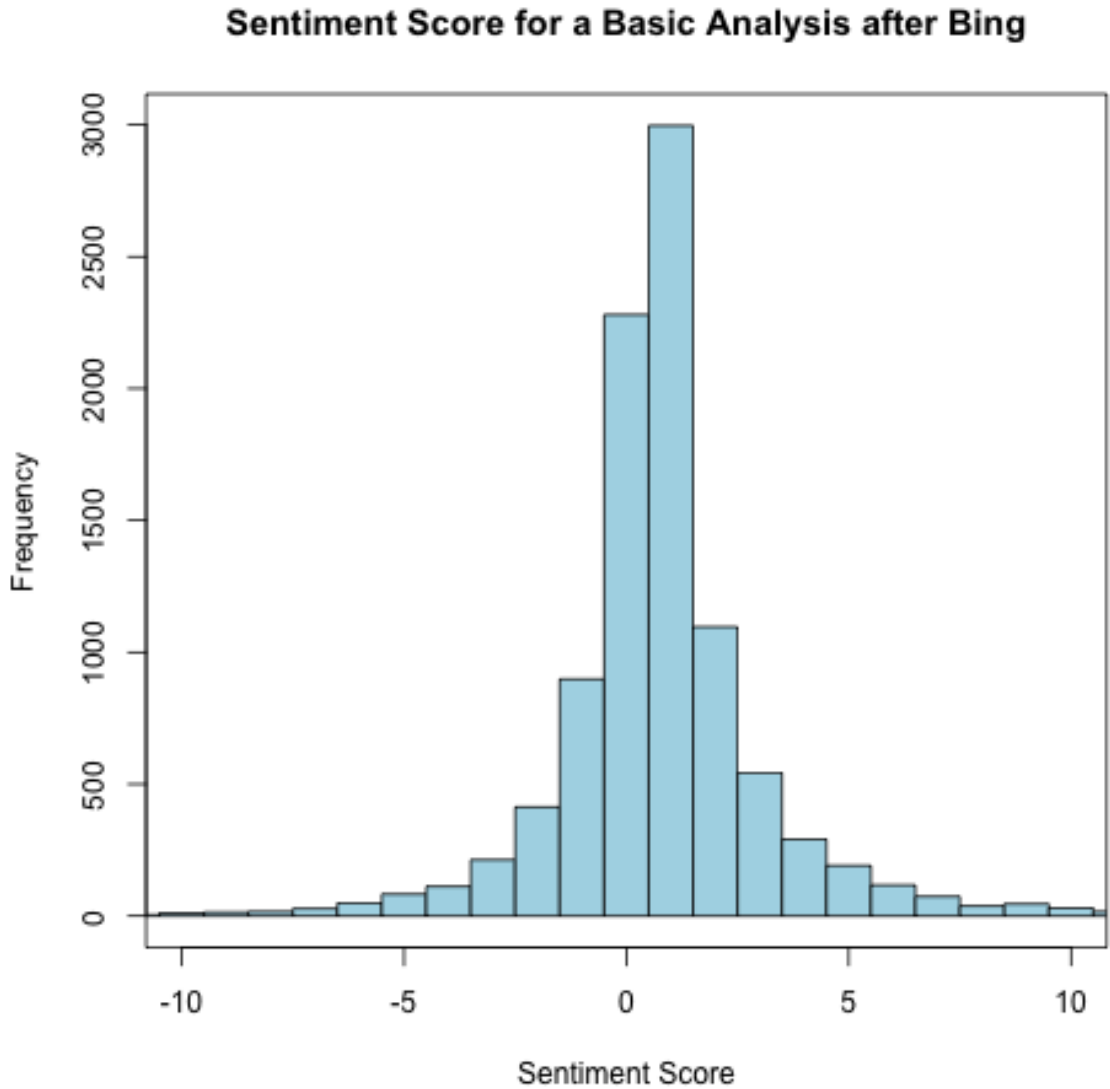
6 Figures

Figure 1: Frequencies of sentiment scores for a basic sentiment analysis. Negative values mean a negative sentiment and positive values a positive sentiment. The tails are cut of at -10 and 10 for better visualization.

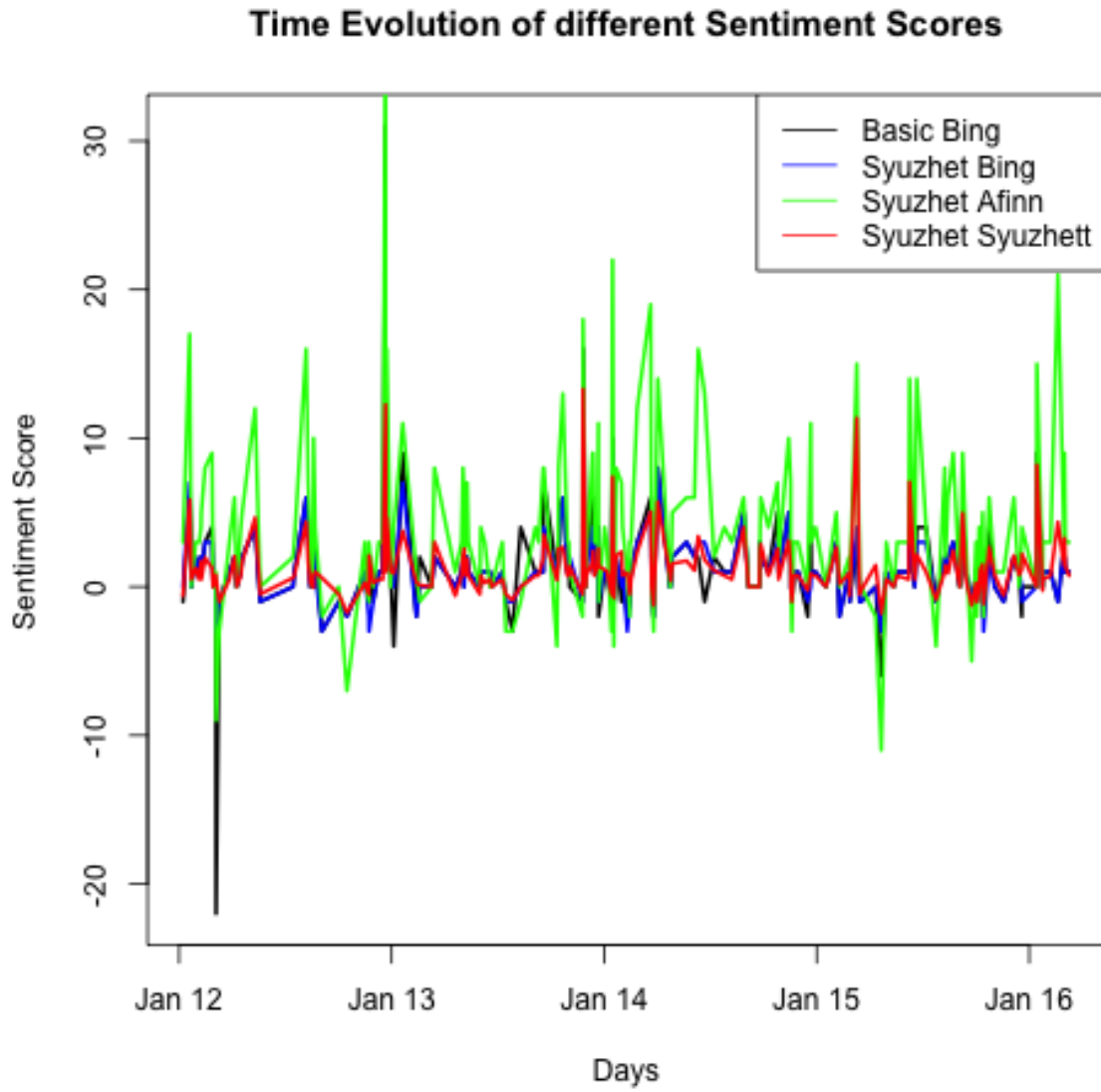# Time Evolution of different Sentiment Scores

Figure 2: Time evolution of the sentiment scores of the basic analysis after Bing and the three different analyses with *syuzhet* for 500 reviews in January 2023. Restriction to 500 reviews for better visualization.
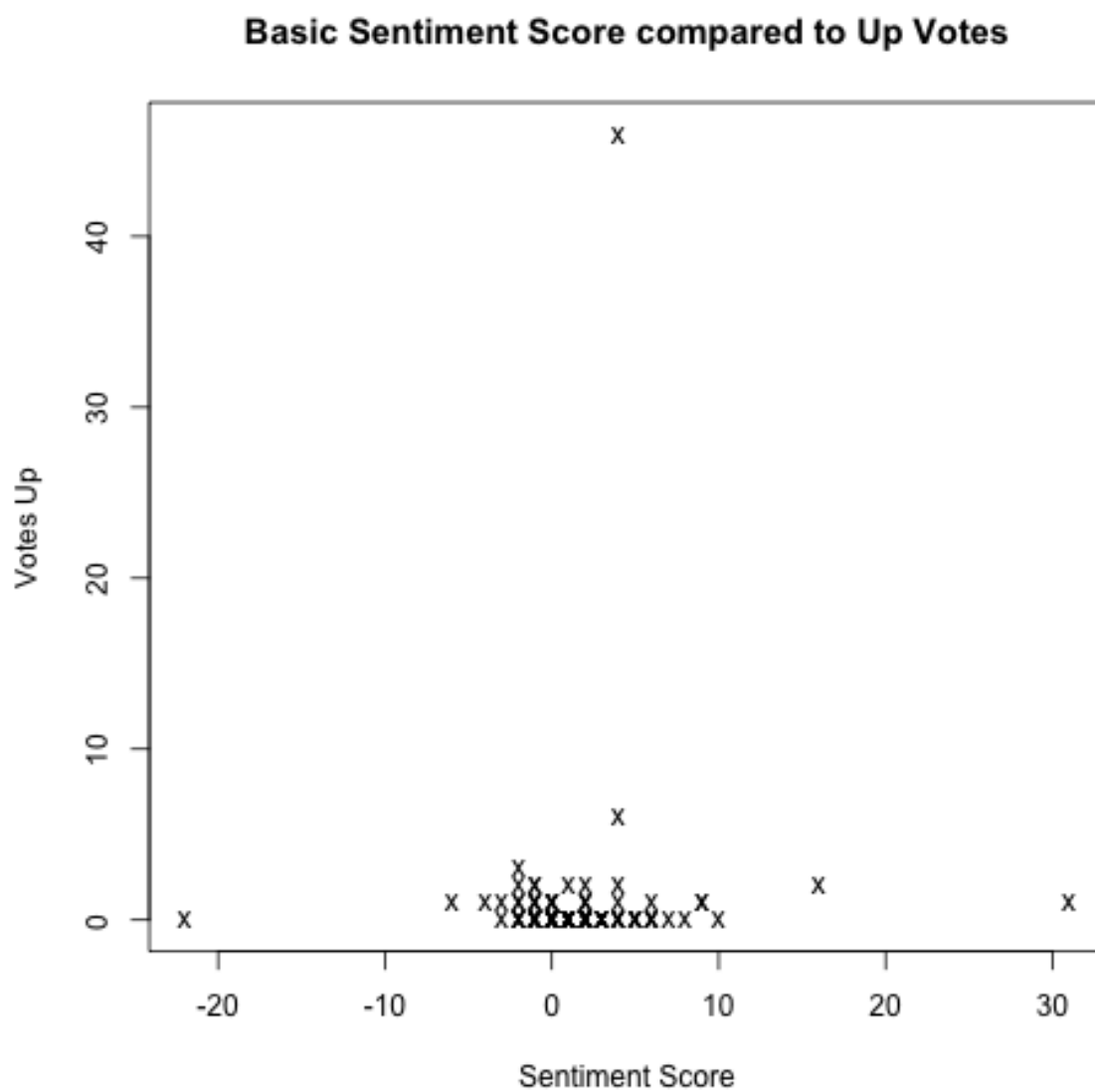
Figure 3: Votes up plotted against the sentiment scores of the basic analysis after Bing. All reviews with votes are considered.

References

1. *Steam, The Ultimate Online Game Platform* Nov. 2023. `https://store.steampowered.com/about/`.

2. Von Klitzing, L. & Weiss, A. Mandatory Assignment 1 - User Reviews of Steam Video Games (2023).

3. Dimpfl, T. *Text Mining and Sentiment Analysis, Lecture Introduction to Applied Data Science, University of Hohenheim* (2023), 11.

4. Hu, M. & Liu, B. Mining and Summarizing Customer Reviews.Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004) (Aug. 2004).

5. Jockers, M. L. *Syuzhet: Extract Sentiment and Plot Arcs from Text* 2015. `https://github.com/mjockers/syuzhet`.

6. Jockers, M. L. *Matthew L. Jockers* `https://web.archive.org/web/20190708100723/http://www.matthewjockers.net/page/2/`.