SONAL, Spring'23

SENTIMENT ANALYSIS ON GAME OF THRONES

INTRODUCTION

Game of Thrones (GoT) is an American fantasy drama television series based on George R.R. Martin's A Song of Ice and Fire novels. The series premiered on HBO in 2011 and ended in 2019, with a total of eight seasons and 73 episodes. The show has been widely acclaimed for its storytelling, characters, visual effects, and production values. It has also been a commercial success, with a massive fan following around the world.

GoT has generated significant interest in academic circles, with researchers exploring various aspects of the show. These include its representation of gender, race, and sexuality, its use of medievalism and fantasy tropes, and its impact on popular culture. Moreover, GoT has been analyzed for its political themes, such as power, conflict, and governance, and its relevance to contemporary political issues.

However, despite the critical acclaim and commercial success of the show, it has been subject to criticism, particularly regarding its final season. The ending of the series was met with mixed reactions from fans, with some expressing disappointment and frustration. This has led to debates about the quality of the show's writing and its ability to deliver a satisfying conclusion. Therefore, this proposed study seeks to explore another aspect of the show's popularity, which is its re-watchability. This study aims to use sentiment analysis on comments under a Twitter post to determine which season of GoT is the most re-watchable. This research question is relevant as it would provide insight into which season resonates the most with the viewers and has a lasting impact.

RESEARCH QUESTION

Which season of Game of Thrones is the most re-watchable based on sentiment analysis of comments under a Twitter post?

DATA

Data collection for sentiment analysis on "Which season of Game of Thrones is the most re-watchable?" from Twitter was done using a python module called ' snscrape' to search for tweets that mention the TV show Game of Thrones and specifically ask which season is the most re-watchable. The data is extracted from a Twitter post saying "Which season of Game of Thrones is the most re-watchable?". Since, the tweet's url is already known, all the comments under the tweet were extracted. These comments were anywhere from the data of the tweet tweeted (i.e., 2023-03-05) till the present day. In this manner, number of tweets/comments under the tweet saying "Which season of Game of Thrones is the most re-watchable?" were collected and saved

into a file, which were then analyzed to determine the overall sentiment of Twitter users towards the various seasons of the show. This sentiment analysis helps to determine which season is the most popular and provide insight into why viewers find certain seasons more re-watchable than others.

We first define the tweet URL, and then use a python module called ‘snscrape’ to scrape all the comments for the tweet. It does this by searching for tweets that mention the tweet URL. We check if a tweet has no replies for it (i.e. it's a top-level comment), (or) has any further replies for it (i.e. it's a reply to a comment). These comments and replies are then converted to Pandas Data Frames and saved to separate CSV files. In this case, there were no replies found, therefore the CSV file for the replies will return empty. Therefore, the dataset we will be considering for the sentiment analysis is the “comments.csv”, which consists of 688 rows × 29 columns.

As, we don’t require all the columns from this dataset, We further preprocess the dataset to clean and transform it into a format that is suitable for this analysis. The raw data that we have collected may contain irrelevant or noisy information, such as special characters, stop words, or inconsistencies in formatting, which can interfere with the accuracy of our analysis. By preprocessing the data, we can remove or modify these elements and standardize the format, making it easier to analyze and extract meaningful insights. Additionally, preprocessing can help improve the performance of machine learning models that are trained on the data, by reducing noise and improving the quality of the features used for analysis.
Since, data preprocessing is an important step in performing sentiment analysis on textual data, the following steps have been performed to preprocess the data ("comments.csv" dataset)for sentiment analysis:

1. Dropped irrelevant columns: Some columns in the dataset are not relevant for our sentiment analysis. The columns: "url", "id", "user", "conversationId", "lang", "source", "sourceUrl", "sourceLabel", "links", "media", "retweetedTweet", "quotedTweet", "inReplyToTweetId", "inReplyToUser", "mentionedUsers", "coordinates", "place", "hashtags", "cashtags", "card", "viewCount", "replyCount", "retweetCount", "likeCount", "quoteCount" and "vibe" will not be needed for sentiment analysis.
2. Removed duplicate comments in the dataset. These comments had to be removed to avoid bias in the sentiment analysis.
3. Removed special characters and punctuations, as they won’t be required for sentiment analysis.
4. Converted the text to lowercase for more consistency.
5. Performed Tokenization: The text data has been tokenized into individual words, which can be used for further analysis.
6. Removed stop words: Stop words like "the", "a", "an", "in", "at", etc. As they won’t be needed for sentiment analysis. Therefore, these can be removed from the text data.
7. Performed Lemmatization: Words that may be in different forms like "like", "liked", "likes". Such words were removed, as they can be normalized using lemmatization (also known as stemming)
8. After performing the above steps, the processed text data can further be used for sentiment analysis.

Characteristics of our final dataset:

1. date: date when the comment was posted
2. rawContent: text of the comment extracted
3. processed_content: text of the comment after removing all punctuations, special characters, converting into lowercase
4. tokenized_content: list of individual words after being tokenized from that specific comment
5. season: Specifies the season number extracted from the rawContent of the dataset

METHOD

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool that can analyze the sentiment of a piece of text by assigning scores to each word based on its sentiment polarity. The sentiment scores for the words in a text are then aggregated to calculate an overall sentiment score for the entire text.
To perform VADER sentiment analysis on a piece of text, we first need to tokenize the text into individual words and remove any stopwords or other noise that is not relevant to sentiment analysis which has already been done during data preprocessing. Then, we have used the SentimentIntensityAnalyzer() function from the vaderSentiment package which contains sentiment scores for thousands of words, to assign sentiment scores to each word in the text. Finally, we can aggregate the sentiment scores to calculate an overall sentiment score for the entire text.

RESULT

The line plot from the Vader sentiment analysis for the Game of Thrones seasons shows the sentiment scores for each season on the y-axis and the corresponding season numbers on the x-axis.
The sentiment scores range from -1 to 1, with -1 indicating a highly negative sentiment, 0 indicating a neutral sentiment, and 1 indicating a highly positive sentiment. The line plot will have a point for each season, with the sentiment score for that season plotted on the y-axis.

The following line plot shows the visual representation of the sentiment trends for each season. If a season has a higher sentiment score, it indicates that the sentiment for that season was more positive overall. If a season has a lower sentiment score, it indicates that the sentiment for that season was more negative overall.

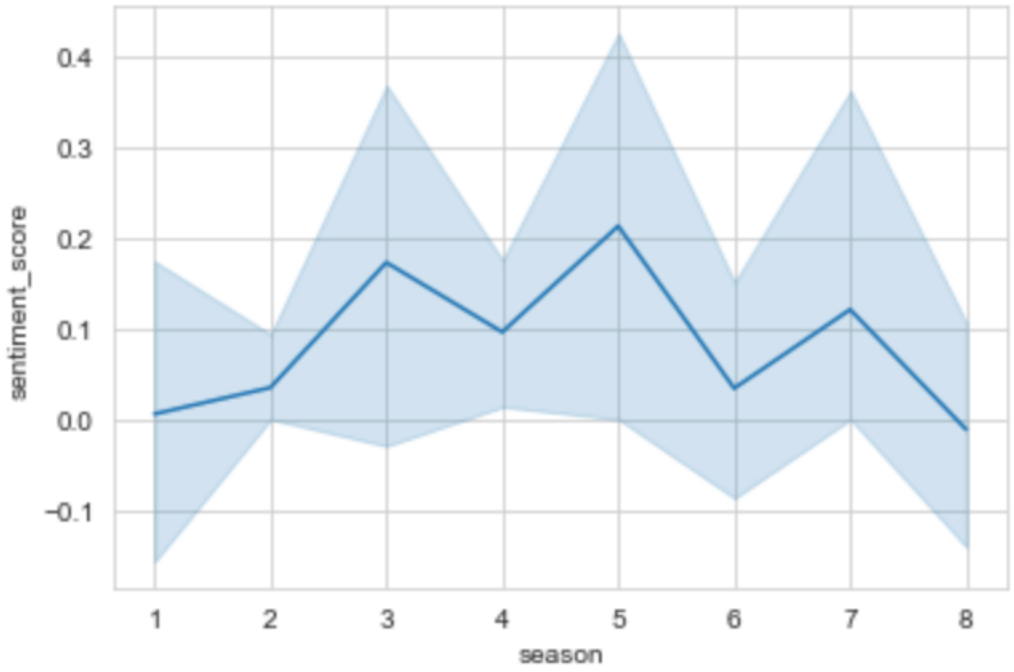

Figure: Sentiment Analysis on Game of Thrones

From the above figure, Season-5 has the highest sentiment-score followed by Season-3, 7, 4, 6, 2, 1, and 8.

CONCLUSION

Answering to the question "Which Game Of Thrones season is the most re-watchable?", People seem to like the Season-5 of Game of Thrones more and find it to be the most re-watchable season. While season-1 & 8 seemed to be the least liked/re-watchable ones. Whereas the third season seemed to be the second-most worthy to be rewatchable.

LIMITATIONS

Conclusions cannot be accurate due to the following:

1. Unable to interpret the exact context of a text. For instance, the word "kill" would typically be interpreted negatively, but it may not mean it in the context. For example, "You guys were killing this performance", the word "kill" is a positive expression here.
2. Since, the sentiment analysis was performed based on the intensity scores of the English words. Performing sentiment analysis with any foreign languages other than English would produce inaccurate results.

REFERENCES

1. Hutto, C. and Gilbert, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*. 8, 1 (May 2014), 216-225.
2. Frangidis P, Georgiou K, Papadopoulos S. Sentiment Analysis on Movie Scripts and Reviews: Utilizing Sentiment Scores in Rating Prediction. Artificial Intelligence Applications and Innovations. 2020 May 6;583:430–8.