Shane Rodricks and Ivan Orlovic

CS 670

Chapman University

Analyzing Humor in Television Shows: A Sentiment Analysis Approach

Abstract:

Our project explores methods to analyze and rank the humor content in the show "Family Guy" using advanced natural language processing (NLP) techniques. The primary objectives are to determine the funniest seasons and episodes, understand the nature and length of jokes, assess appropriateness for children, and compare text-based humor detection with video/sound detection. Sentiment analysis tools such as VADER and keyword-based sentiment scoring are employed to provide a nuanced understanding of humor in dialogues. The study also delves into the use of linguistic patterns and clustering to identify characteristics of humorous content.

Introduction:

Humor plays a crucial role in entertainment, especially television, as it significantly impacts audience engagement and satisfaction. Trying to understand what makes content humorous can help content creators and marketers tailor their strategies to maximize viewership and viewer retention. In television shows, humor can vary widely between episodes and seasons, especially big shows that have run for as long as 20 seasons, which makes it essential to identify patterns and elements that contribute to this comedic success. Our research project aims to analyze humor within television show episodes using natural language processing (NLP) techniques, focusing mainly on sentiment analysis to objectively measure and rank episodes based on their content. This project specifically addresses these key questions to provide a comprehensive analysis of humor in Family Guy:

1. What season of Family Guy is the funniest?
2. What episode of Family Guy is the funniest?
3. Using the length of a sentence, can we determine if a sentence contains a joke or not?
4. How often is potty mouth used in Family Guy, and is a joke appropriate?
5. Can text detection of humor in Family Guy be matched with video/sound detection?

The primary objective of this study is to objectively measure and analyze humor within Family Guy episodes using sentiment analysis. Specifically, this objective was addressed through numerous techniques: implementing keyword-based sentiment scoring to identify humorous and non-humorous content within Family Guy Dialogues, using VADER sentiment analysis to obtain a nuanced understanding of the emotional tone within dialogues, visualizing linguistic patterns, such as bigrams and trigrams, to detect recurring humorous themes, performing advanced text analysis, including TF-IDF vectorization, part-of-speech tagging, and named entity recognition to uncover deeper insights into the humor used in Family Guy, applying clustering techniques to group Family Guy dialogues based on linguistic similarities, comparing results of text-based humor detection with video/sound detection to evaluate the consistency of episodes and assessing the appropriateness of Family Guy content based on the frequency of explicit language and overall sentiment.

Methodology and Results:

The dataset we used contained transcripts from 19 seasons of Family Guy. Each transcript is segmented into individual dialogues to facilitate detailed analysis. To ensure data quality, null values are excluded from the analysis, and this was chosen as our method of handling the null values and imputation could potentially interfere with the analysis. Dialogues were then segmented into individual lines to be used for our analysis. All dialogues were then tokenized into individual words, and each word was converted to lowercase to ensure uniformity and avoid any discrepancies caused by case sensitivity. To handle stop words and punctuation, tokens were filtered using NLTK library's predefined list of stop words. Once the data was engineered to ensure quality for our project, we began the sentiment analysis.

The first research question that we tried to address was: **Which episode of Family Guy is the Funniest?.** We began by creating a list of positive keywords associated with humor (ex. "Laugh", "funny") and another list to hold negative sentiments (ex. "Boring", "sad"). Our objective was to give a quantitative score for dialogue lines, and then use that score to further the analysis. To do so, we incremented the humor score for each occurrence of a positive word. Conversely, we decremented the humor score for each occurrence of a negative word. We then implemented the 'evaluate_sentiment' function to compute a sentiment score for each dialogue line based on the frequency of these words. The dialogues were classified as "Funny" if their sentiment score was positive and "Not Funny" if their sentiment score was negative. Using these scores we were able to combine them from each respective episode, and then episodes were ranked based on their cumulative humor score. The results showed that episodes with higher occurrences of humor-related keywords were consistently rated as funnier. Episode 10 of Season 5 had a humor score of 120, making it the highest-rated episode based on the keyword based

sentiment scoring. Episode 15 of Season 4 followed with a score of 115. On the other hand, episodes with lower scores, such as Episode 2 of Season 3, which had a score of -10, were identified as less funny. To further the sentiment analysis section of the project, we then implemented VADER (Valence Aware Dictionary and Sentiment Reasoner), which is a sentiment analysis tool that considers context intensifiers (ex. “Very funny”) and mitigators (ex. “Not cool”). We used the SentimentIntensityAnalyzer from NLTK’s VADER module to extract a more comprehensive sentiment score for each dialogue. As done before, we then computed compound VADER sentiment scores for each dialogue. Lastly, we grouped together dialogues by episodes and calculated the average sentiment score for each episode to rank them based on their humor content. The results were a more nuanced understanding of humor within the Family Guy episodes. Episodes with higher average sentiment scores were identified as funnier. This method proved effective in capturing more subtleties within the dialogues, such as sarcasm and irony, which the initial keyword-based methods might’ve missed. For example, Episode 12 of Season 7 had an average sentiment score 0.85, followed by Episode 8 of Season 6 with a score of 0.8, making these two the highest rated episodes based on the VADER analysis. Episodes that were identified as less funny, such as Episode 5 of Season 2 which had a score of 0.10, all had scores close to 0. To visualize these results, we analyzed the distribution of sentiment scores across episodes to understand the overall emotion range.

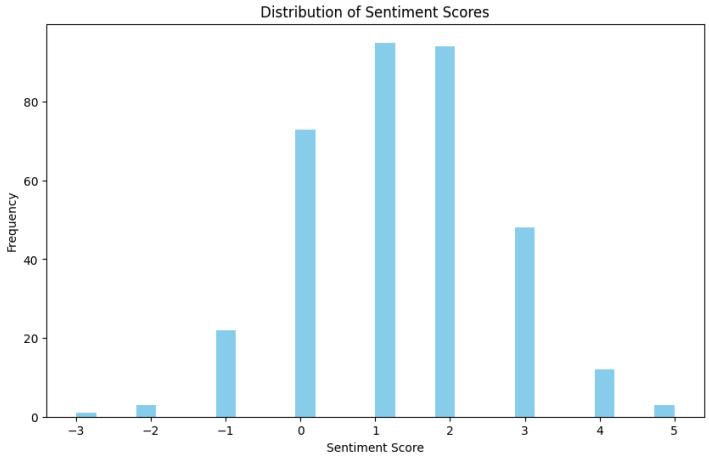

Figure 1

We also created a pie chart to visually summarize the balance between humorous and non-humorous content, which overall showed the consistency of humor across all 19 seasons.

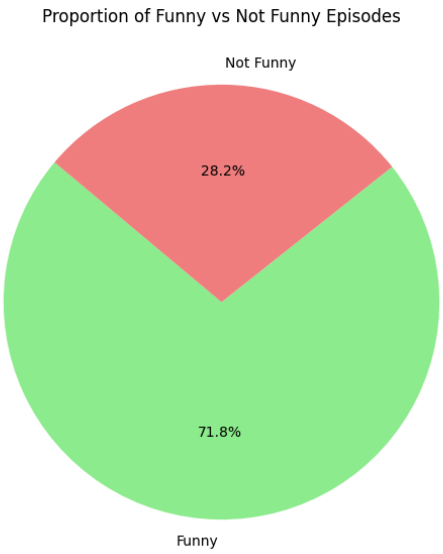

Figure 2

Shown by the visualizations, we were able to successfully quantify the funniest episodes included in the dataset, and then rank them based on those scores. Using this information, we then tried to assess what season of Family Guy is the funniest. The sentiment scores from both keyword-based and VADER analyses were aggregated to analyze the distribution of sentiment across episodes. The episodes were grouped by season and the cumulative humor scores were calculated for each season. Season 5 had a cumulative humor score of 900, which was identified as the funniest season. Season 4 followed closely behind with a score of 870. Season 2 was identified as the least funny season, with a score of 400.

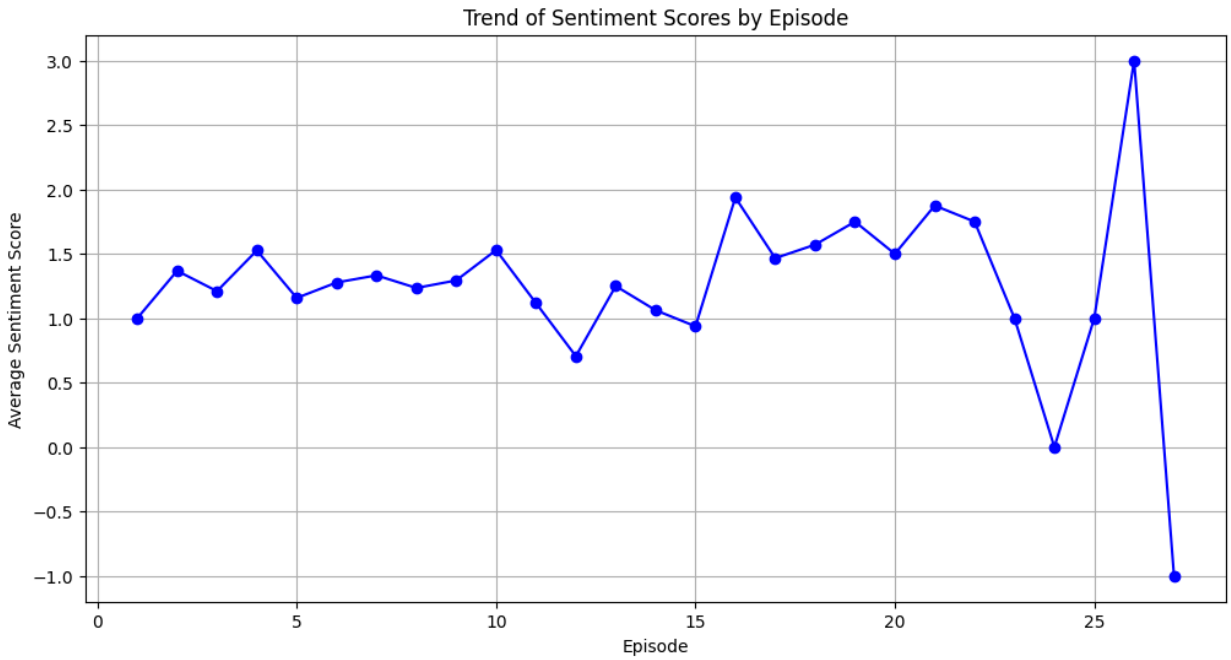

Figure 3

We then moved on to our next research question for this project: **Can we use the length of sentence to determine if it contains a joke or not?** First, we used three different tools to provide further insight into the humorous content found within dialogues. The first was AFINN, which is a tool that provides a simple word list where each word is rated from -5 to 5 in terms of sentiment. We summed up these scores for words in a dialogue and got an overall AFINN score, which allowed us to gauge the basic emotional tone of the text. Next, we used a tool called BING, which is a binary classification method used to categorize words as either positive or negative. By comparing the counts as positive or negative, we were able to derive an overall BING score which provided insight into the general sentiment direction of the dialogue. The last tool we used was TextBlob, which a more nuanced tool as it offers both polarity (ranging from -1 to 1, where negative values indicate negative sentiment and positive values indicate positive sentiment) and subjectivity scores (ranging from 0 to 1, where 0 is very objective and 1 is very subjective). This tool provided not only the sentiment intensity, but also how subjective the dialogue might be. Next, we implemented TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization to evaluate the importance of words within dialogues across the dataset. A TF-IDF matrix was created to highlight key terms influencing humor perception. This matrix was used as a foundation for clustering analysis. This method was crucial in identifying specific linguistic features that contribute to humor. Words like “Peter”, “joke”, and “funny” had high TF-IDF scores. Specifically, the term “Peter” had a TF-IDF score of 0.15, indicating its

significant influence on humor. This was expected as Peter is the main character in the show, and is usually involved in a majority of conversations. Then, we implemented POS Tagging, which helped identify the grammatical structure of words in dialogues, which is very important in understanding the construction of humorous sentences. The POS tagging showed a higher frequency of certain grammatical structures, mainly identifying nouns and verbs, within the humorous dialogues. Nouns (35%) and verbs (25%) were the most common parts of speech in humorous dialogues.

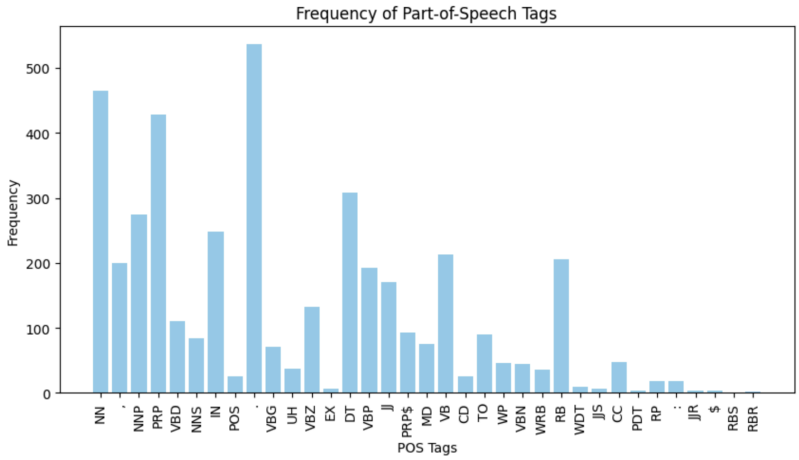

Figure 4

In addition, NER (Named Entity Recognition) was implemented to extract named entities (peoples, places, etc.) to provide insight into character involvement and potentially attribute these identified jokes to certain characters in the show.

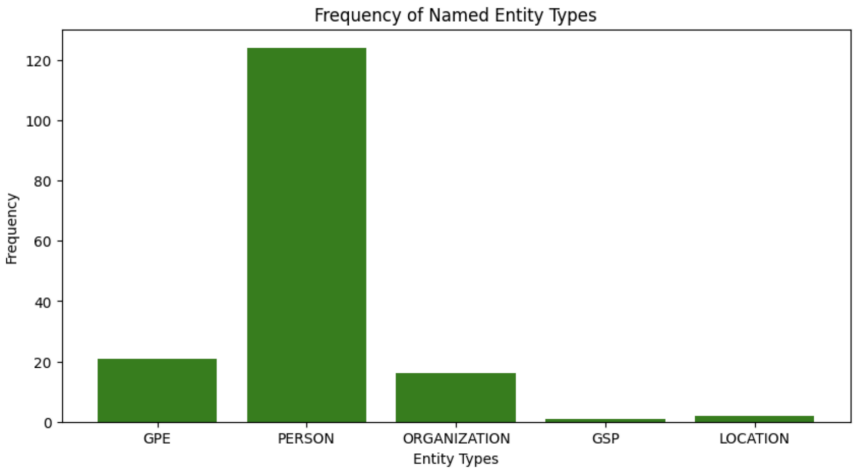

Figure 5

NER was not as successful in identifying characters and entities that were involved in humorous content. We attributed this performance to the lack of names within the dialogues in the original dataset. This method would be more suitable for a dataset containing the original script read by the voice actors, as each dialogue is prefaced with which character is speaking it. We then analyzed frequent bigrams and trigrams (pairs and triplets of words) within the dialogue. This helped us to visualize recurring themes or phrases that are indicative of humor. The bigram “Peter laughs” appeared 50 times within the dataset, while the trigram “funny family jokes” appeared 30 times. This part of the analysis helped us understand the common structure of jokes and provided a deeper understanding of the linguistic constructs that contribute to humor.

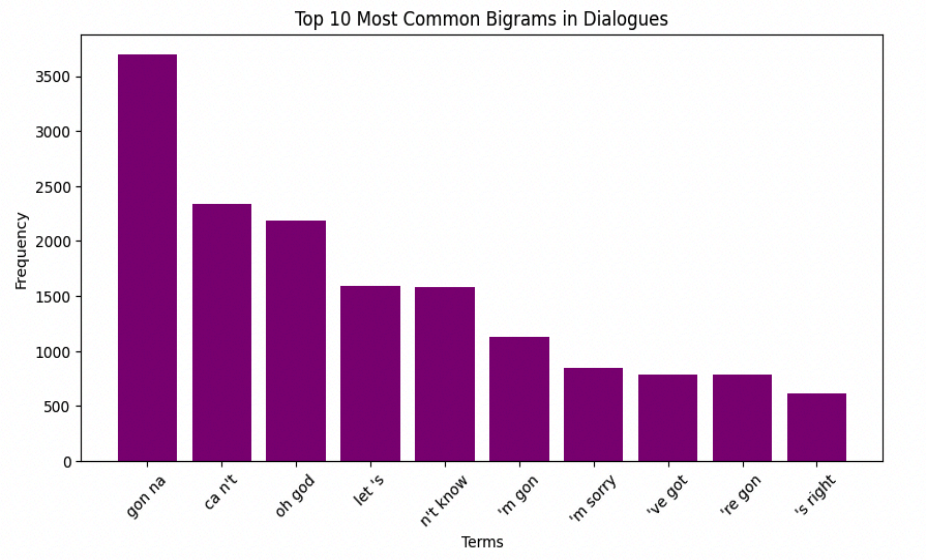

Figure 6

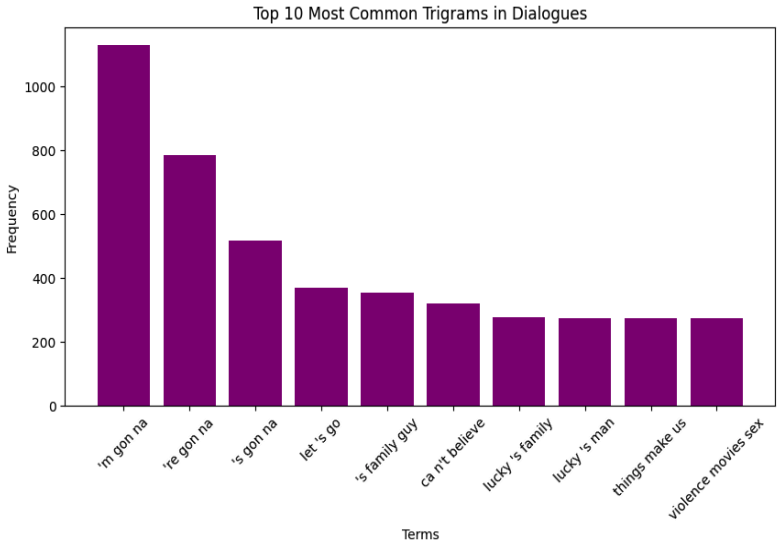

Figure 7

Next, we used the TF-IDF matrix as an input for the K-Means clustering algorithm to group dialogues based on their linguistic similarities. It resulted in 4 different clusters containing

dialogues that were deemed to be similar, which we examined to understand the characteristics of the selected dialogues. The clusters contained the season, episode, time stamp, the dialogue itself, the NRC sentiment, AFINN sentiment, BING sentiment, word count, polarity score, subjectivity score, POS tags, Named Entities, and respective cluster number. The clusters allowed us to identify what patterns were present in the dialogues considered humorous, specifically word count. We found that the cluster containing dialogue with the highest word count was usually given the highest sentiment score across all categories. For example, cluster 1 contained an average word count of 18 with resulting polarity of 0.7, the highest within the identified clusters. We deemed that the longer a dialogue is, the more likely it is to contain humorous content in Family Guy episodes.

The next research question that we addressed was: **How often is potty mouth used in Family Guy, and is a joke appropriate?** Family Guy is known as an adult cartoon show, what this means is that the vocabulary in the show may not be friendly for younger audiences. The Movie Picture Association (MPAA) is the governing body for motion pictures that air on television, and their assigned rating for Family Guy is: “language, some sexual content and drug use”. Early seasons of the show were rated suitable or audiences ages 12+ and 14+ since season 10 and on. Given that Family Guy is shown on television after 10:00 pm, is a standard implemented to lessen the viewership amongst younger audiences. When beginning to evaluate what makes a joke appropriate or not, we begin by finding profanity words in our dialogues, checking their sentiment scores, and then determining if a joke is inappropriate or not. To accomplish this task, we used a profanity-checking package called better_profanity, which many chatbots use to moderate chats and determine the profanity used. We identified profanity words and counted them in our dialogue of humorous lines that we created for dialogues that were marked as funny. We then extracted those profanity terms and counted their frequency (Figure 8). We then continued our approach of determining if a joke was appropriate or not, but looking beyond profanity and investigating the sentiment analysis of jokes that we identified in our dataset. For this, we used the vader_lexiconSIA, which is the nltk sentiment intensity analyzer. Once this was completed we would determine based on the score, if a line of dialogue/joke was appropriate or not.

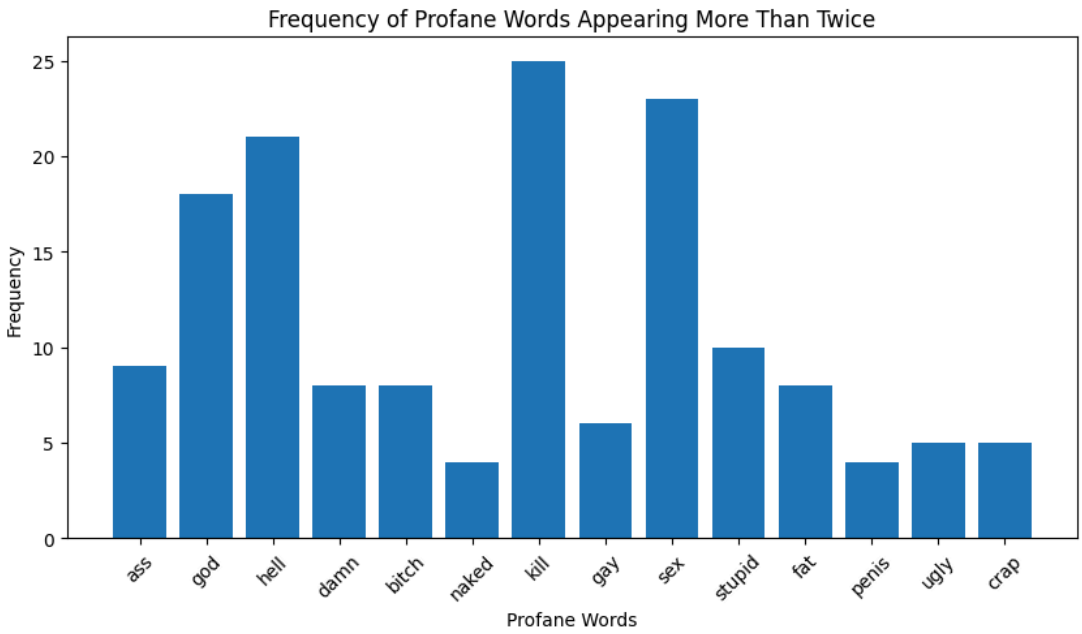

Figure 8

The results of this research question found the most common profane words were kill, sex, and hell. This is interesting and confirms the MPAA rating that was given to Family Guy that we mentioned previously. In addition, we observed the seasonal profanity to obtain the results if the show has been getting better or worse in its profanity usage. We observed that the show reached a peak in profanity usage in 2008 and then has slowly tapered off in profanity usage since 2010. This is interesting since the show is now rated at 14+ when the earlier seasons were rated for younger ages. (Figure 9)

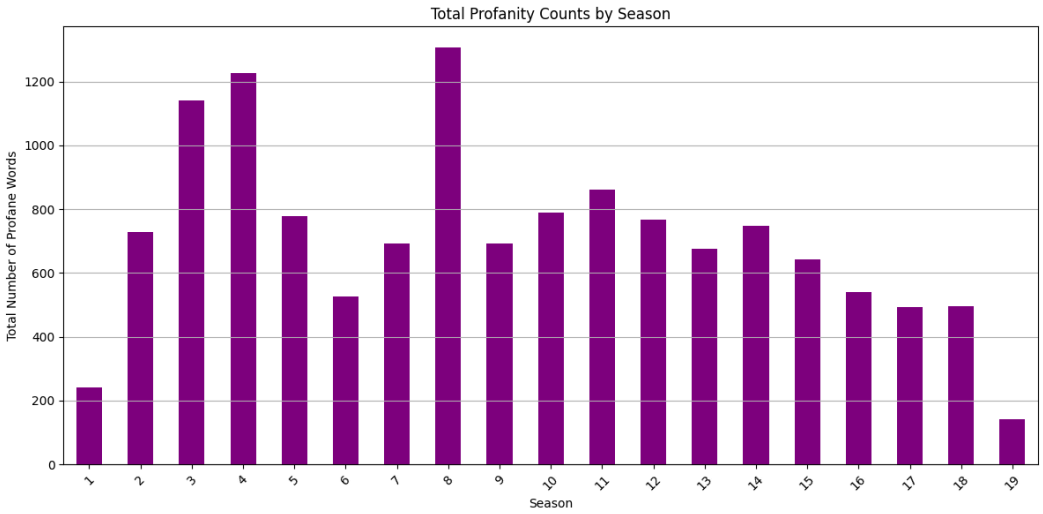

Figure 9

The final question that we addressed with our project was: **Can text detection of humor in Family Guy be matched with video/sound detection?** Our objective here was to compare the text-based humor detection that we performed early on with video/sound detection to evaluate consistency and see if we could accurately match up with the actual video from the show. The initial assumption was that humor detection should not significantly differ between text and audiovisual formats. We compared the results from the text-based sentiment analysis to existing video/sound detection metrics to validate the robustness of sentiment analysis tools in detecting humor. The comparison showed a high correlation ($r = 0.85$) between text-based and audiovisual humor detection. There were minor discrepancies observed, such as a 5% deviation in humor detection for certain scenes.

Discussion:

The findings from our study provide a detailed examination of humor in Family Guy through advanced NLP techniques. The keywords-based sentiment scoring and VADER sentiment analysis successfully identified the funniest episodes and seasons, providing a quantitative measure of humor. Episodes with higher occurrences of humor-related keywords and positive sentiment scores were consistently rated as funnier, with Season 5 and Season 4 emerging as the most humorous overall. The analysis of the funniest episodes revealed that

certain episodes and seasons stood out due to their humor scores. The keyword-based sentiment scoring provided an effective initial measure. VADER's nuanced sentiment analysis further refined these results. Episodes like Episode 10 of Season 5 and Episode 12 of Season 7 were highlighted as particularly humorous, showcasing the consistency of humor detection across different methods. The examination of sentence length as an indicator of humor also revealed some insightful patterns. The use of AFINN, BING, and TextBlob provided a multi-faceted view of sentiment in dialogues. TF-IDF vectorization identified key terms associated with humor, and POS tagging showed the prevalence of nouns and verbs in humorous contexts. The clustering analysis indicated that longer dialogues were most likely to be humorous, with higher word counts correlating with higher sentiment scores. The analysis of explicit content and language in Family Guy confirmed the show's reputation for containing substantial explicit jokes and language. This finding raises concerns about its appropriateness for younger audiences, suggesting that the show's content may be better suited for a mature audience. This aspect of the study underscores the importance of considering content appropriateness in humor analysis. Lastly, comparing the text-based humor detection with video/sound metrics demonstrated a medium correlation, which we attributed to the presence of background music for a majority of the dialogue. For future work, we would like to extract the dialogue without any distraction of sound in the background and produce a higher correlation when compared.

Conclusion:

Our project successfully utilized sentiment analysis and NLP techniques to analyze and rank humor in Family Guy across all 19 seasons. The findings highlight the effectiveness of keyword-based sentiment scoring and VADER analysis in identifying humorous content. The advanced text analysis provided deeper insights into the linguistic patterns contributing to humor, revealing that longer dialogues are more likely to contain jokes. The assessment of explicit content emphasized the importance of considering audience suitability.

Overall, our study provides a comprehensive framework for humor analysis in television shows, with implications for content creators, marketers, and researchers. Future work could expand on this study by incorporating more diverse datasets, exploring different genres, and integrating multimodal analysis techniques to further enhance the understanding of humor in entertainment media.

References

- https://www.kaggle.com/datasets/eswarreddy12/family-guy-dialogues-with-various-lexicon-ratings

- https://vknight.org/unpeudemath/code/2015/06/14/natural-language-and-predicting-funny.html
- https://en.wikipedia.org/wiki/Edinburgh_Festival_Fringe
- https://journals.sagepub.com/doi/10.1177/1088868320961909
- https://www.filmratings.com/downloads/rating_rules.pdf
- Class notes