Optimizing Storytelling, Improving Audience Retention, and Reducing Waste in the Entertainment Industry

Andrew Cornfeld, Ashley Miller, Mercedes Mora-Figueroa, Kurt Samuels
University of Virginia, cpm6gh, asm2fe, eqa7yg, nps3cs@virginia.edu

Abstract — **This project tries to tackle the task of accurately predicting a shows viability. Natural language processing (NLP) was performed on television scripts to calculate sentiment scores and were used along with temporal information in the dataset for our research. This data was used to make a model that determined which features of the selected show were most important to the show's success. A second model was then trained to accurately predict the viewership for that show. The insights gained from these will help networks better understand what aspects of a script lead to success within each show genre.**

Keywords — Machine Learning, Narrative Analytics, SARIMAX, Audience forecasting, Natural language processing, Television viewership prediction, XGBoost.

1. Introduction

Television networks face significant uncertainty in predicting show success, leading to costly cancellations and unaired pilots. Reducing such inefficiencies requires predictive tools that link content features to audience behavior. This project investigates whether emotional and linguistic attributes of TV scripts, quantified through NLP, can inform viewership forecasting. The analysis evaluates whether these narrative indicators, alongside traditional metadata, can enhance predictive accuracy or provide actionable insights into factors contributing to a show's success.

2. Related Work

Early approaches to television viewership prediction primarily relied on historical ratings and schedule-related variables, using statistical models such as ARIMA and SARIMA to capture trends and seasonality in audience data [1]. Regression analyses incorporating factors like time slots, lead-in programs, and competing broadcasts were also common [2], enabling modestly accurate short-term forecasts based on prior viewing patterns. While effective for established shows with consistent audiences, these traditional models often struggled with shifts in viewer behavior or for new content lacking historical data. In response to these limitations, researchers have increasingly adopted machine learning methods to enhance prediction performance [3]. Studies have applied models ranging from linear regression and K-nearest neighbors to random forests, XGBoost, and neural networks, with ensemble approaches frequently outperforming individual models. Combining historical ratings, demographic variables, and scheduling data in algorithms like gradient boosting or ridge regression has been shown to improve predictive accuracy, especially in complex, nonlinear viewing environments. Industry implementations have further demonstrated that blending diverse models (e.g., SARIMAX, Prophet, XGBoost) into ensembles can reduce forecasting error across different television genres and audience segments.

More recently, natural language processing (NLP) has emerged as a novel approach to viewership prediction by incorporating content-based features derived from scripts or subtitles. Studies have demonstrated that linguistic characteristics—such as sentiment, emotional arcs, thematic complexity, and narrative structure—correlate with audience engagement and ratings [4]. For example, researchers have used LIWC-style emotion dictionaries and network text analysis to quantify narrative originality and emotional tone, finding significant associations with subsequent viewership. By integrating these NLP-derived features with historical and contextual data, predictive models gain access to intrinsic narrative signals, offering a promising avenue for understanding and forecasting audience responses beyond traditional metrics.

3. DATA

A. Data Description

Our dataset consists of approximately **25,257 episodes spanning 219 unique television shows**. Each episode record contains **144 variables** spanning show metadata, narrative emotion scores, linguistic style, cognitive processes, motivational drives, and engineered features.

Key metadata variables include show title, season, episode number, air date, episode length, IMDb rating, genre, network, viewership in millions, and cancellation status.

Narrative and linguistic features were derived from episode sub caption data, scraped from *OpenSubtitles* using an API, with additional show-level metadata (e.g., parent network, number of episodes, season count) sourced from IMDb.

A central focus of the dataset is the set of 43 linguistic and emotion-related NLP-derived scores, computed per three narrative acts per episode, resulting in 129 variables per episode. These variables include measures of emotional tone (e.g., Anger, Joy, Fear), sentiment (Positive, Negative), cognitive processes (Insight, Cause), and narrative style (Analytic, Clout, Authenticity, Tone).

B. Data Preprocessing

Prior to modeling, we performed a series of preprocessing steps to ensure data quality, consistency, and suitability for analysis. First, genre values were reduced from 54 categories to 14 broader genres, and categorical variables such as Genre and Network were one-hot encoded.

We engineered new temporal features by extracting the month, year, year-month, day of the week, and season from each episode's air date, along with a year-season combination variable to enable seasonal trend analysis. Additionally, we created a variable indicating whether an episode was a season premiere, season finale, or mid-season.

To capture viewership dynamics, we computed both the percentage change in viewership from the previous episode and a 3-episode moving average of viewership. The percentage change was explored as a potential alternative target variable over absolute viewership in millions, and the 3-episode moving average window was added as for its potential to enhance predictive performance allowing for longer term trends.

A key preprocessing step involved the temporal alignment of lagged features: we shifted each episode's explanatory variables so they corresponded to predicting the next episode's viewership (or its derived metrics). This alignment ensured that all information available at episode t was used to forecast outcomes at episode t+1, reflecting a real-world predictive scenario.

4. MODELING APPROACHES

3.1. SARIMAx Model

Since the data were sequential and we postulated that there was a cyclical nature to the TV shows over seasons, we first tried to use a seasonal autoregressive integrated moving average (SARIMA) model. Unfortunately, the data does not have consistent time steps between the episodes in any given show. Due to the nature of time series modeling, this posed a major limitation in how we were able to use this modeling approach. Instead of interpolating data points in between the existing ones, which would have introduced significant imputation error, we decided to try to use an exogenous variable to capture the information lost [5]. We then created a new feature to incorporate the time since the last episode to capture the influence of breaks in the show airing schedule. While this model type was able to give us reasonable estimates for viewership after a show had been airing for some time, it didn't work for shows that were just starting or had few episodes already. This was due to the modeling approach's minimum requirements for observations to converge. This meant that this approach was not entirely appropriate for the goals of our project. Since we are trying to argue that these data would help us make informed decisions about show direction and writing, we simply couldn't argue the utility of an approach that relies solely on the fluctuations in viewership to predict future measurements of the same. Likewise, incorporating the natural language scores into the model as exogenous variables was likely to introduce significant forecasting error beyond what the model was already exhibiting. We also could not fully justify their exogeneity since they were a part of the system which was driving the viewership.

3.2 Rolling XGBoost Model

Our second approach uses XGBoost which takes data solely from a given show and creates a sequentially retrained XGBoost model that predicts episode viewership using cumulative data from all preceding episodes. We chose to use XGBoost due to its efficiency as well as its ability to capture interactions between variables. This model removes temporal data such as air date, season and episode numbers. This model begins with Season 1 Episode 3 (to have enough data for training) and continues through all episodes of the show. One limitation of this model is that it requires some prior viewership data and thus cannot predict shows that have not aired yet. After completing some testing, we noticed that shows with large amounts of episodes can take significantly longer to run (several minutes). We considered other methods of reducing the time including setting variables for how often the model retrains, but we determined that the accuracy tradeoff wasn't worth the time that would have been saved.

3.3 Feature Selection Model

Since our original boosted model's performance was promising, we began to explore ways to improve upon this. Realizing that one of the biggest drawbacks of the dataset was that the natural language processing scores inflated the dimensionality of our data, we attempted to narrow the features being examined for each TV show.

To this end, we developed a nested boosted model that used a preliminary XGBoost regressor to determine the top 20 features using the built-in importance measure, gain. We then ran a second regressor on these features only. In addition to this discriminating step, we engineered other features to incorporate information from the previous episode's viewership and two moving averages with window sizes of 3 and 5 episodes. In this method we also explored cross-referencing the gain feature importance with Shapley Additive explanations (SHAP) for a more informed view of how each feature affected the model predictions [6]. Seeing that the rolling model also had trouble handling outliers when making predictions for the show The Office, this model also included a winsorizing step to reduce the influence of extreme values. The model was then evaluated using an 80:20 percent train test split at the show level. In addition to the regression modeling approach, this model also incorporated a similarity scoring system meant to compare the aggregate natural language scores from all episodes of a single show to all other shows, returning ones with the shortest Euclidean distance between the score vectors. We postulated that this represented shows with the most similar sentiment analysis present in their scripts.

3.4 Combined Model

After creating both the Rolling XGBoost model and the Feature Selection model, we attempted to combine the two concepts by running a model which first runs a full model to determine the top 10 most important features and then runs the rolling approach using only the 10 most important features. We noticed that this model was not as accurate as either model individually, so we ended up deciding to remain with the two models separately.

5. RESULTS

Due to the nature of many shows evoking different emotions and different genres speaking to different audiences, it is difficult to generalize results for a general model. We were able to more readily model single shows at a time to determine predictive capacity of our model. In doing so, we found that previous viewership was often a top predictor of future viewership. The presence of previous viewership and the moving average features in the top feature importance rankings implied that there was a strong relationship between the previous success and the future success of any given show.

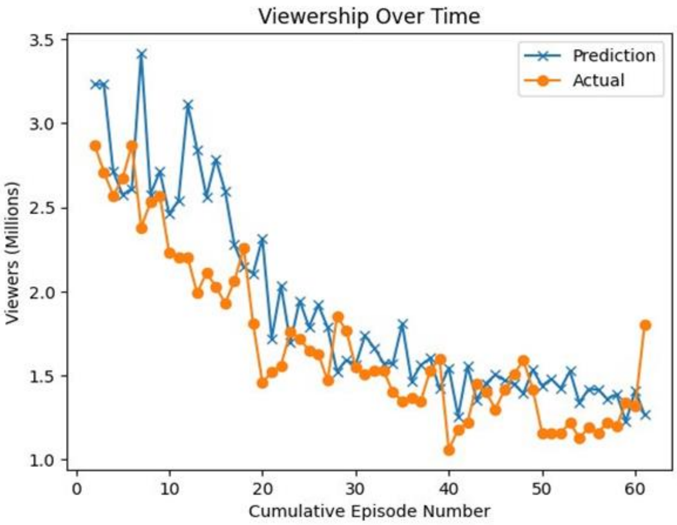

Figure 1: Graph of predicted vs actual viewership over time of Better Call Saul using the Rolling model.

The rolling XGBoost model was sometimes very predictive and sometimes not predictive at all. For example, it was very predictive for the show Better Call Saul, with an R^2 of 0.742 and an RMSE of 0.361. The figure above shows that the predictions were generally above the actual viewership but very much in line with it throughout the show's lifetime.

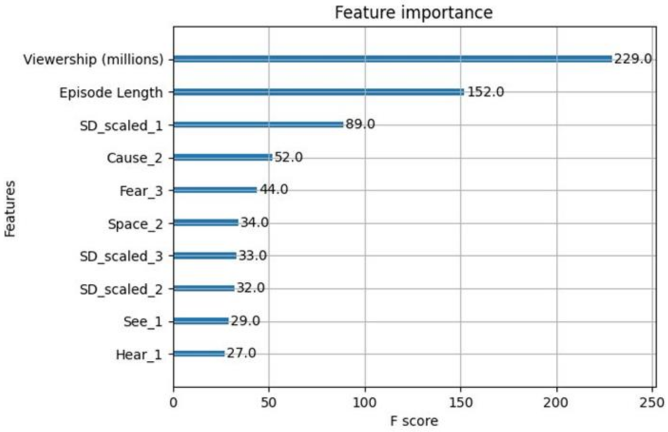

Figure 2: Feature importance plot for Better Call Saul using the Rolling model.

It is also clear based on the feature importance graph that previous viewership was by far the most important predictor of future viewership. This is unsurprising because often people will come back and watch the same shows repeatedly if they've been watching them for a while, often regardless of what may happen in one episode.

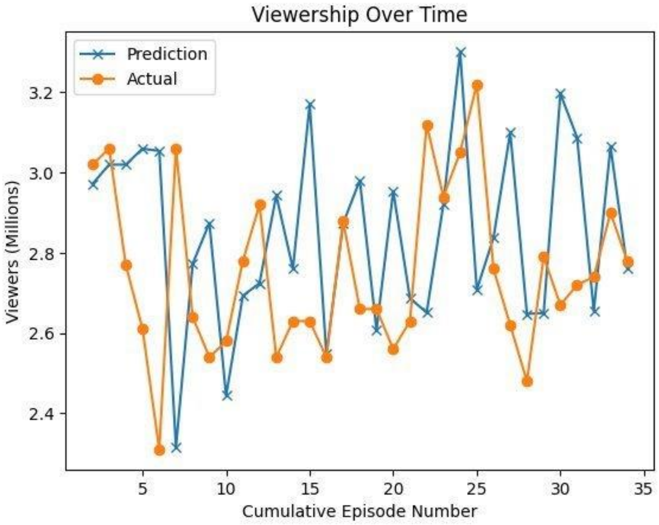

Figure 3: Graph of predicted vs actual viewership over time of Abbott Elementary using the Rolling model.

Contrary to the *Better Call Saul* example, another show analyzed was *Abbott Elementary*, where the model was almost no more predictive than simply guessing viewership numbers. Predictions were very accurate with the model, with a similar RMSE of 0.330, but with an R^2 of 0.039, the model benefitted from all the actual viewership numbers being roughly between 2.4 and 3.2 million viewers. Notably, viewership was not one of the top 10 features of the model when trained on *Abbott Elementary* data.

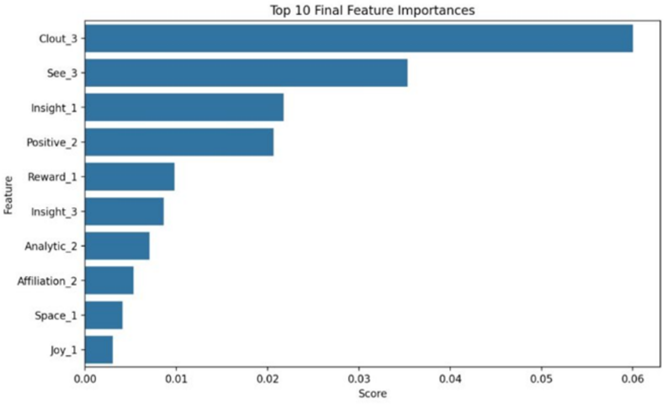

Figure 4: Feature importance plot for Abbott Elementary using the Rolling model.

The feature selection model allowed us to reduce the dimensionality of our feature space and focus on the most important features for a given show. It could also inform the writers of which emotional scores were highly influential in predicting the success of the TV show, which would allow them to focus on these pain points when evaluating the content of their product. When comparing the results of this model on the show Better Call Saul, we had an RMSE of 0.208 and an R^2 of 0.763. This implied that both models were performing similarly on this show.

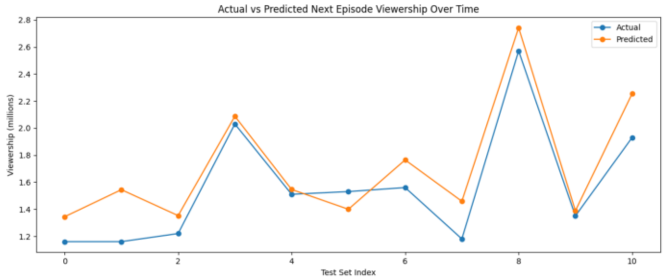

Figure 5: Graph of predicted vs actual viewership over time of Better Call Saul using the feature selection model.

The top three features by gain importance were previous viewership and the two moving averages of viewership with window sizes of 3 and 5 episodes. This corroborated the results of the other model and helped support the claim that previous viewership is a significant driver of next episode viewership.

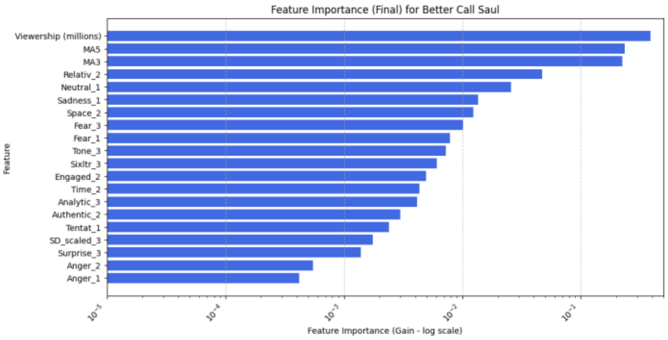

Figure 6: Final feature importance (log-gain scale) for Better Call Saul using the feature selection model

Abbott Elementary did not perform as well using this model with an RMSE of 0.420 and an R^2 of –0.4780, implying that this model is performing more poorly than a model which simply uses the mean of the target values as predictions for all observations. This could be because the predictions are only validated on 20% of the values – 7 observations in this case, compared to the rolling model’s iterative testing. Regardless, this show doesn’t seem to fit the model approach well either way.

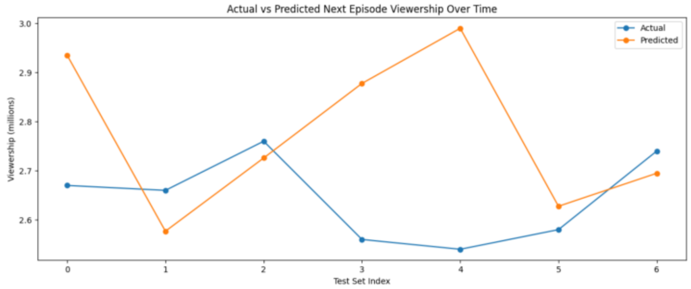

Figure 7: Graph of predicted vs actual viewership over time of Abbott Elementary using the feature selection model.

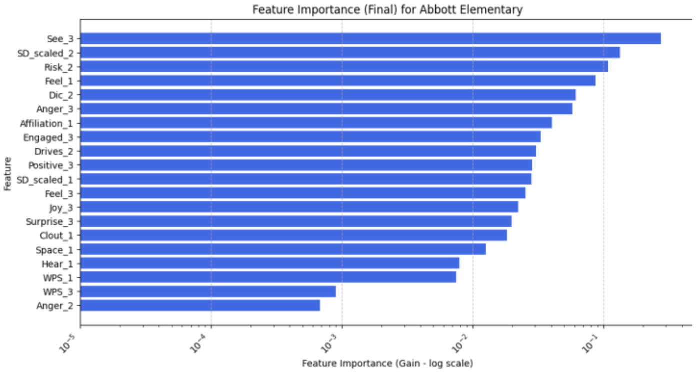

Figure 8: Final feature importance (log-gain) scale for Abbott Elementary using the feature selection model

When cross-referencing the SHAP values to the gain importance, we found that the top NLP features were often included in both rankings, though not always in the same order. Also, since the SHAP values are bivariate and they associate an upwards or downwards influence to the predictions in tandem with a color indicating a high or low value of the predictor itself, we can more precisely explain the effect that they have on viewership (Ergün, 2023). For example, Abbott Elementary saw a significant net negative effect on the next episode's viewership when the language in the script referring to things that could be seen or imagery in the third act had low values or appeared less often.

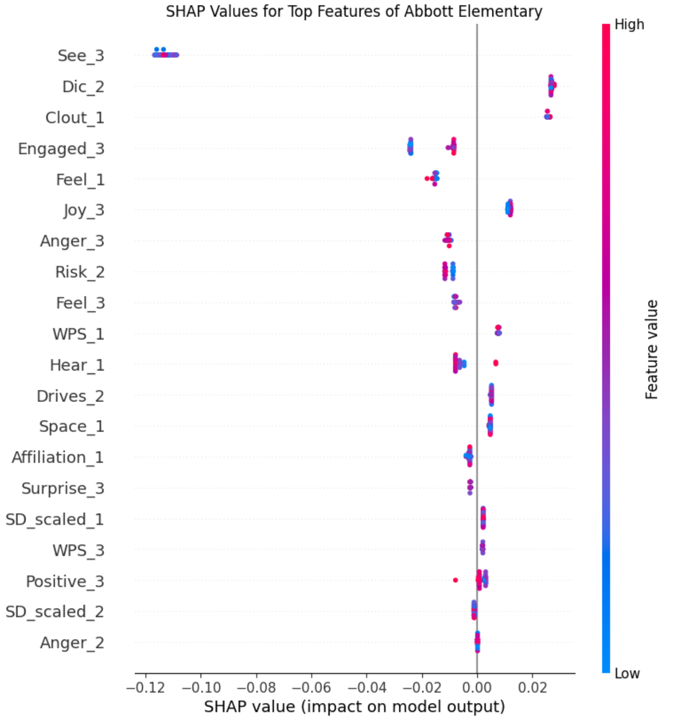

Figure 9: SHAP values for Abbott Elementary (excluding viewership and moving averages)

When examining the similarity scores calculated for each of the shows by Euclidean distance between the aggregate NLP scores, we had to rely largely on human feedback to confirm the results as feasible. For some shows this was easier than others. Breaking Bad appeared as the top similar show for Better Call Saul, which is fitting as it is the parent show that was spun off to create Better Call Saul. The top result for CSI: Miami was CSI: Crime Scene Investigation, which we concluded was certainly true, given they were also created by the same production company and from the same concept. Other results from this method were not as easy to justify. In the results for The Shield, we were surprised to see that the top five similar shows included Regular Show, Adventure Time and Rick and Morty.

6. DISCUSSION

When comparing an RMSE plot of both the rolling XGBoost model and the feature selection model, both models frequently can achieve an RMSE of under 1 million viewers. The rolling model has slightly more tailing towards higher RMSE values than the feature selection model does, making the feature selection slightly more consistent. When trying to analyze what factors determined high or low R^2, we plotted distributions for both models by genre, and we determined that there was no correlation between genre and R^2 value.

Interpreting the importance of features using the built-in importance was helpful in trying to determine whether the model was relying significantly on autoregressive features, but SHAP values allowed us to explain more about the effect these features had on viewership. This more comprehensively achieves the goals of this project by allowing us to home in on what elements of scripts drive viewership up or down.

When examining the results of the function to rank TV shows by the distance between vectors, we were surprised at some of the shows that were ranked as highly similar. While these results may have been unexpected, we argue that this is able to draw insights not obvious to professionals in the industry and could prove valuable in this way. It would allow a writing team to contextualize their work with respect to existing TV shows to justify creative choices while using previous industry ventures as a litmus test.

7. CONCLUSION

It may not come as a surprise to those working in the entertainment industry that previous episode viewership is often the best predictor of next episode viewership, but this approach also gives writers the ability to see what features are dominant in shows like The Office and Abbott Elementary, helping them when trying to construct the next great mockumentary. It provides us with good metrics to compare shows, giving network executives historical context, so they can say with greater confidence which pilots or airing shows will be a success.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Anthony Palomba for bringing this project idea forward and for his invaluable insights, feedback, and industry expertise that shaped the direction of this research. We also thank Tom Hartvigsen for his dedicated mentorship and guidance throughout the project, providing critical support in both technical and strategic aspects of our work.

REFERENCES

[1] H. Vo, “Forecasting television viewership: A machine learning approach using narrative content and metadata,” *Proc. Int. Conf. Data Sci. Media Analytics*, pp. 45–51, 2022. [Online]. Available: https://hvo.github.io/papers/viewership.pdf

[2] X. Liu and J. Zhang, “Prediction of TV program ratings based on machine learning models,” *J. Data Anal. Appl.*, vol. 18, no. 3, pp. 201–210, 2021. [Online]. Available: https://www.booksci.cn/literature/114012607.htm

[3] Nielsen, “Using machine learning to predict future TV ratings in an evolving media landscape,” *Nielsen Insights*, Mar. 2016. [Online]. Available: https://www.nielsen.com/insights/2016/using-machine-learning-to-predict-future-tv-ratings-in-an-evolving-media-landscape/

[4] M. Griffin, “Predicting Nielsen ratings from pilot episodes’ scripts: A content analytical approach,” *Academia.edu*, unpublished. [Online]. Available: https://www.academia.edu/85926128/Predicting_Nielsen_Ratings_from_Pilot_Episodes_Scripts_A_Content_Analytical_Approach

[5] J. Korstanje, “The SARIMAX model,” in *Advanced Forecasting with Python*, Berkeley, CA: Apress, 2021, pp. 131–150.

[6] S. Ergün, “Explaining XGBoost predictions with SHAP value: A comprehensive guide to interpreting decision tree-based models,” *New Trends Comput. Sci.*, vol. 1, pp. 19–31, 2023.