Transforming the Tape

A Football Commentator's Report Card

The University of California, Berkeley: Department of Graduate Studies, Computer and Data Science.

This paper analyzes the game commentary transcripts of NFL and NCAA games over the last 60 years. It aims to develop an objective grading system via an overall "commentator report card" for the affiliated broadcasting networks.

Keywords: NFL commentary, BART, RoBERTa, Two-Tower modeling, NLP, sports media analysis

Abstract

The production companies televising National Football League and NCAA games incur massive revenue from viewership. Broadcast networks pay over $10 billion annually for the rights to produce the games in their respective leagues. The largest stakeholders currently broadcasting these games are ESPN, FOX, CBS, NBC, and ABC. These networks are not merely buying football games - they are buying the entire fan experience. Engaged fans stay longer, return to watch more games, and increase revenue for these respective production companies.

A significant portion of this investment by broadcast networks extends beyond the visual production to the voices of the game. Commentators play a critical role in delivering top-notch game analysis, play-by-play, and engaging conversations. These commentators vary in the style in which they present informational, dramatic, and entertaining components to viewers. Each network employs a dedicated set of commentators to cover their production of games. Like football teams compete against each other, networks compete for the best broadcasters. But how does one judge the makeup of a high-quality broadcast by the commentators? If commentating is a business asset, let's analyze it as one.

1 Introduction

This paper presents a framework for objectively evaluating NFL game commentators across three core dimensions: excitement, team bias, and information density. It utilizes natural language processing (NLP) techniques with pretrained transformer models analyze game commentary transcripts from the past 60 years. An objective scoring metric via the final *commentator report card* will serve as a benchmark for broadcasting companies. This paper hopes to benefit the production companies and fans by identifying the high-performing commentators and makeup to enhance viewer engagement and assist in strategic talent decisions.

Excitement plays a critical role in enhancing fan engagement and emotional investment. This paper will measure this dimension by analyzing the emotional intensity of the commentary and assigning higher scores to more engaging and enthusiastic delivery.

To evaluate bias, this paper first defines an unbiassedly, or neutral, commentator as one without preference or favoritism toward either team. A neutral commentator delivers a balanced narrative that leaves feelings of equity and satisfaction across different fan bases. Biased commentary creates resentment and subtracts from the general viewing experience. My scoring mechanism rewards neutrality and penalizes perceived partiality.

The third dimension is information density which captures a commentator's ability to provide insightful and game-specific analysis. This includes references to players, rules, statistics, and strategic context. Commentators' who enrich their broadcasts with relevant and well-prepared commentary will receive higher scores in this category.

Together, these three dimensions form a composite score that reflects the broader impact of their commentary on the viewer experience. This paper's goal is to translate qualitative aspects of live sports broadcasting into measurable and comparative insights across major television networks.

2 Literature Review

2.1 Excitement in Color Commentary

Color commentary plays a central role in shaping how fans emotionally experience a broadcast. Unlike objective commentary, which centers on factual, statistics-driven narration, color commentary adds personality, emotional storytelling, historical context, and insights between teams and players. In a study by Lee, Kim, Williams, and Pedersen (2016) titled *Investigating the Role of Sports Commentary: An Analysis of Media-Consumption Behavior and Programmatic Quality and Satisfaction*, the authors examined the effects of commentary style on audience perception and future viewing behavior. Participants were exposed to both objective and color commentary conditions. The results showed that color commentary significantly increased audience enjoyment and positively influenced re-watching intentions. Satisfaction with the broadcast was found to be a strong predictor of continued media consumption.

These findings are supported by Zillmann and Cantor's (1996) *Disposition Theory of Entertainment*, which ex-

plains how individuals' emotional responses to media are shaped by the expectations of their affiliated group (such as their favorite sports team). According to the theory, individuals experience positive emotions when their allies succeed or their enemies fail, and contrarily, feel negative emotions when their allies lose or enemies succeed. In this paper's context, a highly affiliated fan will feel happiness after a win and disappointment after a loss - vice versa feelings for their enemies' (opposing team fans) wins and losses. Emotional and expressive commentary amplifies these feelings reinforcing the fans' emotional alignment with the game.

In the context of this study, these three factors indicate a key finding relevant to this paper: *excitement* and even *informational richness*, which are staples of a color commentary style, are key components of a commentator report card. Commentators who use this *colorful* and emotionally engaging language are more likely to enhance viewer enjoyment and encourage long-term fan engagement benefiting their production of the games.

2.2 Bias in Commentary

Previous studies points to bias in sports commentary directly influencing both audience perception and fan engagement. In a paper by Lee et al. (2016), the research clearly demonstrated the style and tone of a commentator can significantly affect a viewer's enjoyment and likelihood of re-watching a broadcast. This notion makes intuitive sense to other sports too - how frequently does a local baseball fan find themselves watching their favorite sports team on the other team's local broadcast network? A perceived biased broadcast creates a reduction of perceived fairness or professionalism.

This aligns with the *Disposition Theory of Entertainment* (Zillmann & Cantor, 1996) as well. When a commentator's tone appears to favor an opponent, the emotional alignment of the fan, who experiences joy when an opposing team loses, is broken.

An NLP-focused study by Merullo et al. (2019) revealed that commentary bias is often systemic, shaped by language patterns that vary across demographic groups and team affiliations. The authors identified and quantified racial bias in NFL commentary using transformer-based models. Aligned with the objectives of this paper, their work offers a foundation for evaluating announcer neutrality through computational methods.

2.3 Information Density in Commentary

Recent trends in sports analytics show a growing emphasis on technical depth and real-time insight, especially in NFL coverage. This shift allows broadcasters to supply fans with richer game information and offers a proxy for assessing commentary detail.

The integration of player statistics, historical context, and predictive analytics has reshaped how fans consume football broadcasts. Oftentimes, fans loath to see the maximum run speed or catch probability of their favorite players. Detailed statistics in the form of new information enhances viewer engagement by providing understanding of rules, team strategy, and individual performance. From a linguistic perspective relevant to this paper, information density is approximated by measuring the frequency and diversity of named entities within a transcript, such as player names, teams, locations, and key events.

3 FOOTBALL Dataset

This paper analyzes transcripts of NFL Game commentary from FOOTBALL containing 1,455 games ranging from 1960 to 2019 with. The dataset columns are the unique game id, teams in the game (away team, home team), transcript of the game in text form, and the date of the game. Meanwhile, each row is a different game.

3.1 Exploratory Data Analysis

To understand the distributions, trends, and possible hidden features of the broadcasting networks, I explored different aspects of the dataset with visualizations.

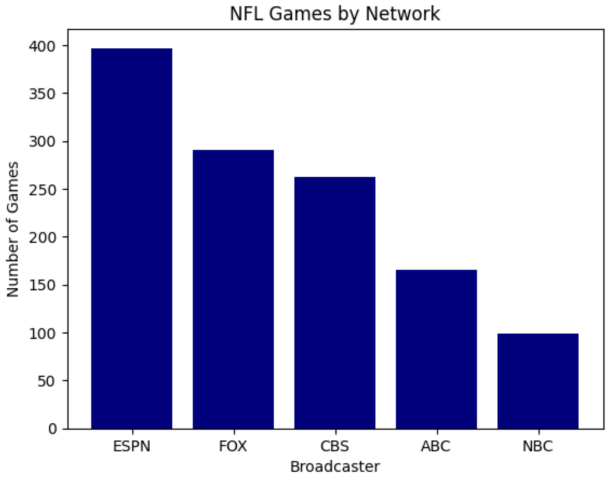

Figure 1: Distribution of dataset games across broadcasting networks.

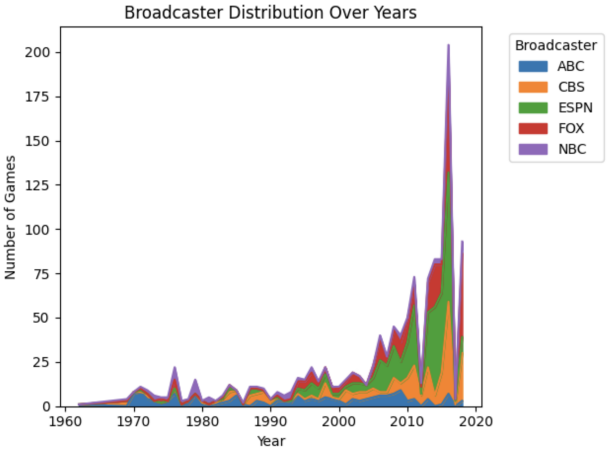

Figure 2: Distribution of broadcasters of games over time.

[0]Figure 8: Visualization generated using Matplotlib from the cleaned commentary dataset.

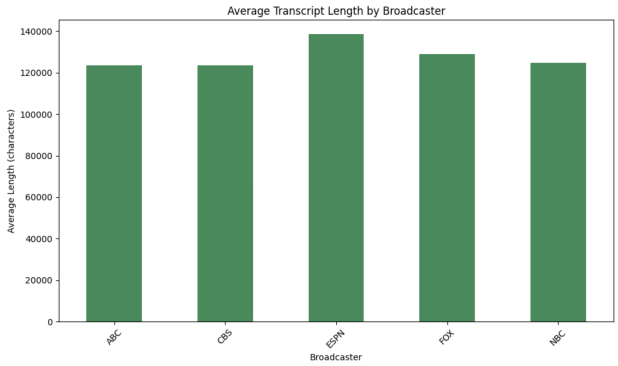

Figure 3: Distribution of transcript lengths (words spoken) by broadcast.

These figures provide contextual insight into some of the patterns across broadcasting networks. The dataset skews toward recent years with ESPN leading in the total number of broadcast games. Additionally, ESPN games show the highest average transcript length which is measured by total words spoken during commentary of the game. This observation could indicate a positive relationship between longer commentary and more informative and emotionally rich broadcasts.

4 Design & Development

The model was designed and aggregated by network on three different components: emotion, team bias, and information density with each equally contributing to the group's report card. Each aspect utilized a pretrained transformer. Several advantages of these models include a wealth of knowledge 'out-of-the-box' without requiring much training, capabilities for fine-tuning for specific tasks, and the ability to handle complex language patterns well.

4.1 Transformer Architecture

Transformer-based models have revolutionized natural language processing (NLP) over the past decade. These models allow computers to understand and generate human language with exceptional accuracy across tons of topics. Introduced in the famous paper *Attention is All You Need* Vaswani et al. (2017), transformers replaced traditional sequential architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks by introducing the concept of *self-attention*.

A self-attention mechanism allows the model to weigh the importance of each word in a sequence relative to all others. Similiar to our own brains, the model can undterstand which word(s) to 'attend' to the most for subsequent tasks. This allows transformers to capture long-range dependencies and complex contextual relationships. Pretrained transformer models such as BERT, RoBERTa, and BART are trained on massive text corpora (such as Wikipedia containing over 2.5 billion words) enabling them to perform tasks like classification, summarization, and zero-shot inference with minimal, if any, additional training.

In this study, I use pretrained transformer models to extract insights from the three commentary dimensions. The models' deep semantic understanding allows for a robust transcript evaluation performance without requiring labeled training data.

4.2 Preprocessing

First, commentary transcripts are processed by removing punctuation, converting text to lowercase, and tokenizing, or separating each word, on whitespace.

To evaluate commentator performance by network, a broadcaster label for each game is required. While certain networks tend to align with specific conferences, such as CBS usually with AFC, these trends are inconsistent. Therefore, I determine network affiliation by identifying the most frequently mentioned network name in each transcript. The network name is frequently mentioned during games for branding and legal requirements (Bruning, 2022). In rare cases of a tie, the game is excluded. This yields a clean dataset with reliable network labels for each game while maintaining a majority of the games in the original dataset.

5 Models and Results

5.1 Excitement and Emotion

To detect excitement levels, I utilized a distilled, performance-optimized version of RoBERTa, specifically the emotion classification model *j-hartmann/emotion-english-distillroberta-base*. This model is pretrained on a large corpus of English text and fine-tuned to classify emotions across categories such as joy, sadness, surprise, anger, and more.

To compute excitement scores, I used the model's output probabilities and applied custom weights to reflect each emotion's contribution to fan engagement on a scale from 0 to 1. Emotions such as joy and surprise were assigned the highest weights. Neutral or negative sentiments received lower values. These weighted scores were aggregated at the game level to generate a single excitement score per broadcast.

Finally, excitement scores were grouped and averaged by network for easy comparison. The average excitement scores were close with ABC leading the group and CBS ranking at the end.

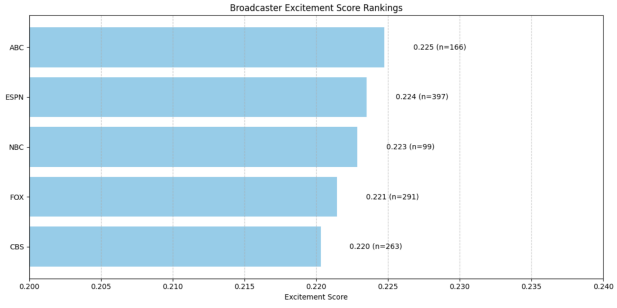

Figure 4: Average excitement levels by broadcaster

5.2 Team Bias

To detect directional bias in commentary between the two teams in each game, I used a RoBERTa-based model trained on the *Multi-Genre Natural Language Inference* (MNLI) dataset. This model specializes in evaluating logical relationships between sentence pairs determining whether one statement supports, contradicts, or is neutral to another. The model uses uses the teams labeled with 'home' or 'away'. For example, if the Ravens played at home and the commentator said "The Baltimore Ravens are my favorite team and I just love watching them play", the model supports the hypothesis of "The commentator favors home team" for this sentence and scores the game accordingly.

I used a zero-shot classification approach with Hugging Face's pretrained transformers. The model was provided with three candidate labels: "favoring the home team," "favoring the away team," and "neutral coverage." These labels were applied to segmented chunks of each game's commentary to assess directional bias.

Bias scores were computed based on the model's confidence in each label and then aggregated at the game level. Lower scores indicate more neutral coverage, preferred by fans, while higher scores reflect stronger favoring of a certain team in the game. Final scores were grouped by broadcasting network to support cross-network comparison.

A clear preference towards the home team in each broadcasts group emerged as a major trend in this component.

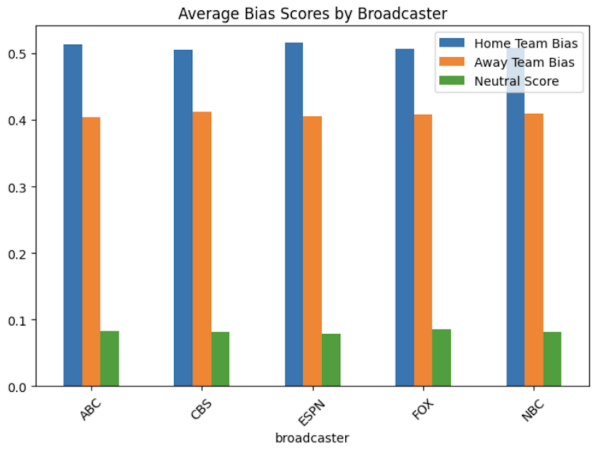

Figure 5: Average bias levels by broadcaster

5.3 Information Density

I used spaCy to measure the richness of information in each game's commentary - a pretrained NLP model featuring a powerful Named Entity Recognition (NER) framework. The model is purposefully trained to identify entities such as people, locations, organizations, and other proper nouns. Its contextual understanding allows it to disambiguate terms based on surrounding language. For instance, the word "Raven" could refer to the Baltimore Ravens NFL team, the poem by Edgar Allan Poe, or the bird. However, spaCy can correctly infer meaning based on context in surrounding sentences. Unlike older models, such as the *Bag of Words* approach, spaCy uses meaning to determine references synonymous words. In this manner, different words like "Ravens" and "Baltimore" are labeled as references to the same team.

The transcripts were tokenized and spaCy was used to count the number of named entities per game. To quantify this, an *entity density* metric was computed as the number of named entities divided by the total number of tokens in the transcript. A higher density indicates a more information-rich broadcast. A lower density suggests repetition or lack of detailed commentary resulting in a lower score.

Differences in information density may reflect excellent preparation for game commentary. While this research does not analyze the specific content or types of entities referenced, the summary results show that ABC leads all networks in median and most other summary statistics related to information density.

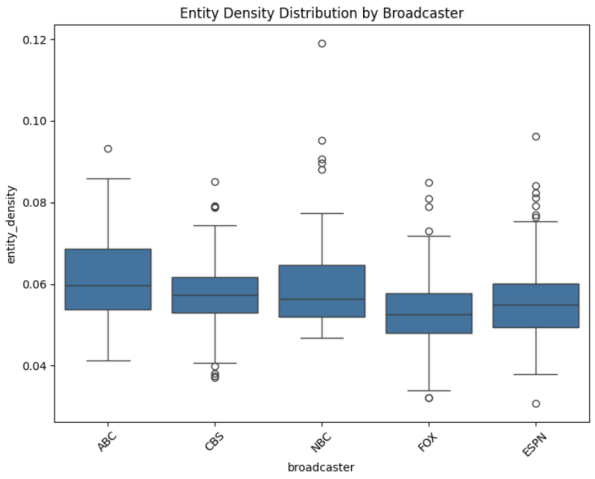

Figure 6: Distribution of broadcaster information density statistics

5.4 Report Card

The *Broadcaster Report Card* synthesizes the three key dimensions of my commentary analysis: emotional excitement, team bias, and information density. Due to varying scales of data in each component score, I normalized the outputs with MinMaxScaler to a 70–100 (the scale fo standard schooling letter grades). For bias scores, lower values represent better (more neutral) performance, unlike the direction of the other scores. Therefore, the scale was inverted to ensure alignment with the scoring direction of the other metrics.

I assigned final letter grades based on percentile rankings using a weighted distribution shown in the table below. The final visualization includes overall broadcaster grades and a heatmap illustrating performance across individual metrics.

Table 1: Commentator Grade Distribution

Grade	Percentile Range
A	Top 15%
B	Next 30%
C	Next 35%
D	Next 15%
F	Bottom 5%

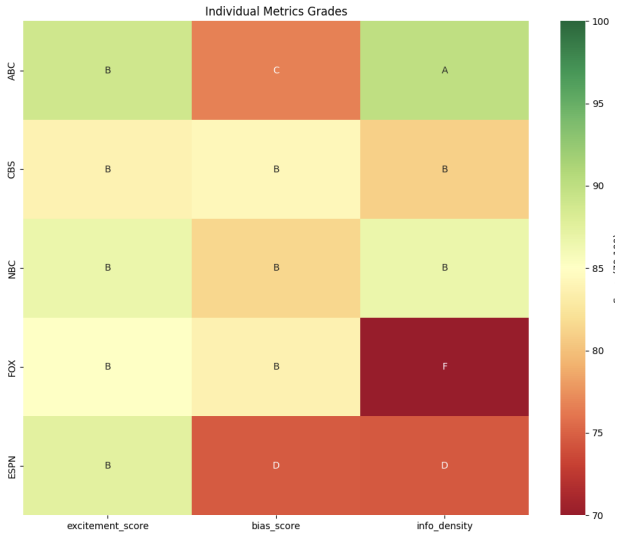

Figure 7: Distribution of broadcaster information density statistics

The following table shows the final scores for each broadcasting network for the respective performances:

Table 2: Overall Commentator Score Rankings by Network

Rank	Network	Average Score
1	ABC	85.1
2	NBC	84.9
3	CBS	82.9
4	ESPN	78.8
5	FOX	78.7

6 Conclusion

This paper presents a objective approach to evaluating NFL game commentary using natural language processing techniques and transformer-based models. With my *commentator report card* based on three fundamental concepts directly influencing fan retention in excitement, bias, and information density, I demonstrate the power of NLP tools to assess commentator quality for the benefit of both audiences and network companies.

This paper yields a comprehensive evaluation of broadcaster performance for both NFL and college football games. To add to the overall grade of metrics, I combined the three metrics with an even weight. ABC emerged as the top-performing network, consistently scoring highest in information density and maintaining strong excitement

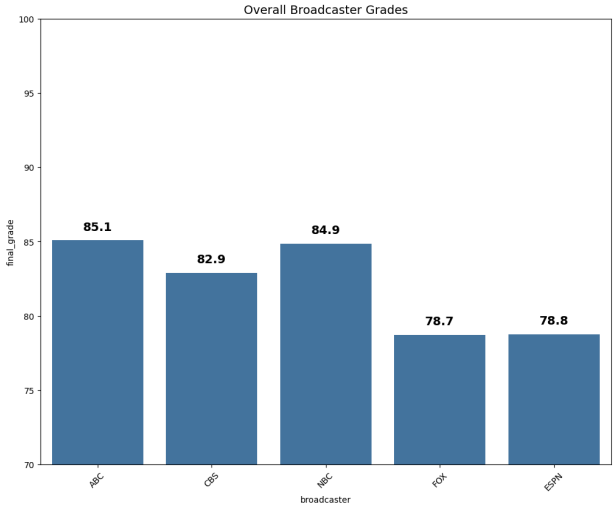

Figure 8: Broadcast overall grade

metrics. NBC followed closely, showing balanced performance across most dimensions with particularly strong emotional range. CBS resulted average scores across each of the three metrics. ESPN performed well in excitement but severely lagged in detail and neutrality. FOX received the lowest average scores across the multiple categories. My results suggest ABC and NBC provide the most engaging and informative commentary experiences while lower-performing networks may benefit from emphasizing neutrality and detailed color commentary. These results may benefit the NFL and NCAA to sign contracts with ABS and NBC maximizing viewership. I also show the sub-performing companies methods to reevaluate strategic decision-making and improve their commentators' perfomances.

The results suggest meaningful differences across networks and games raising important questions about the role of commentary in shaping viewer experience and engagement. With billions invested in broadcasting rights, the ability to evaluate and improve commentator performance through data-driven methods is both highly beneficial and timely.

7 Further Study

Future work could incorporate more granular analysis at the sentence or speaker level. I grouped the commentators' by company - perhaps finding the top performer would prove useful to maximize his/her air time in the biggest games.

The clear trend of a home-team bias is intriguing. I could extend research into this trend by implementing a *Two-Tower Model*. One tower processes commentary related to the home team and the other processes away team references. This approach could add modeling sophistication and allow for a deeper comparison of how each team is discussed in different game contexts. It may also help quantify whether coverage is balanced when the same team appears as both home and away across different broadcasts.

8 Appendix

For the curious reader, this appendix provides additional background on several of the transformer models employed in this paper.

Transformers are composed of *encoder blocks*, which process input sequences *decoder blocks* (optional) which generate output. These models also include *multi-head self-attention layers* that allow the network to focus on different parts of the input simultaneously, as well as *feed-forward layers with layer normalization*, which stabilize and accelerate learning processes.

8.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) was introduced by Google in 2018. It was the first model to process entire text sequences bidirectionally capturing relationships between *all* words in context. Earlier models, such as RNNs, read text sequentially and lacked this holistic view.

BERT is pretrained using a *Masked Language Model* where words are randomly masked, or hidden, in a sentence and the model is trained to predict them. It also uses *Next Sentence Prediction* to determine whether two given sentences are adjacent in the original text. This provides beneficial for the model to understand sentence-to-sentence relationships. With its powerful encoder and context-awareness, BERT is well-suited for classification tasks and serves as a strong baseline in many NLP applications (Devlin et al., 2019).

8.2 RoBERTa

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is an enhanced version of BERT developed by Facebook AI Research in 2019. It improves its recent predecessor BERT by training on a larger dataset, using dynamic masking rather than static, and eliminating the Next Sentence Prediction objective. RoBERTa also supports longer input sequences and proves to outperform BERT across many NLP benchmarks due to its better future generalization (Liu et al., 2019).

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Lee, M., Kim, D., Williams, A. S., & Pedersen, P. M. (2016). Investigating the role of sports commentary: An analysis of media-consumption behavior and programmatic quality and satisfaction. *International Journal of Sport Communication*, *9*(3), 337–353.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Merullo, J., Yeh, L., Handler, A., Grissom II, A., O'Connor, B., & Iyyer, M. (2019). Investigating sports commentator bias within a large corpus of american football broadcasts. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6355–6361. https://doi.org/10.18653/v1/D19-1666

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html