NLP Final Project Report: Sports Fans Sentiment Analysis

Jianfeng Ke   **Sameed Khan**   **Daheng Kuang**   **Sean Lynch**   **Anveshak Rathore**
02031747   01987789   01891533   01858462   02036112

Abstract

The project looks at the sentiments of sports fans in different contexts. Going with the assumption that different fans respond differently to their sports team's results, the project aims to quantify if and when these fans respond differently. Using a state-of-the-art model, GoEmotions (Demszky et al., 2020), the project quantifies the difference in sentiments of fans from losing, winning, and other teams in 4 different contexts: the favorite or the underdog winning either a one-sided or close game. The research explores a previously untapped context of sports fan responses to specific external events. This allows us to see how different psychological responses are manifested in online language over time. Researchers would benefit by being able to delineate the emotions that different groups have at different times, and practitioners would know better when to target which specific groups of people.

1   Introduction

Fans of sports teams are often known to exhibit strong emotions while supporting teams. These emotions tend to be expressed and amplified even more on social media, which is the context of our project as well. This project's primary motive is to classify sports team fans' sentiments - before, during, and after the matches. The project looks at matches with two different types of results: the underdogs winning or the favorites winning. Our primary result shows the prevalence of two very contrasting emotions: annoyance and admiration. Beyond the initial classification measure, we also try another prediction task - predicting the result from the comments, and a second classification task: classifying sentiments of users over time, beyond just their responses to match events. The work is important for multiple reasons. Firstly, it offers interesting data points for the marketers to use for their targeting campaigns - targeting to annoyed potential consumers is very different from targeting potential consumers in the state of admiration. Secondly, it provides researchers with more insights into the mental states of online con-

tributors split across an external event with known outcomes. Thirdly, it provides an understanding of how these mental states are manifested in online language over time.

2   Related Works

In the field of grief responses, we see that people react to losses in 5 stages: Denial and Isolation, Anger, Bargaining, Depression, and Acceptance. We can use these studies (Mughal et al., 2023) to see if sports fans fit in the same model and see how the emotions that we gather using Go-Emotions compare to the different stages of grief reaction.

There have been other models - for example, the Go-Emotion database, from software engineers and researchers at Amazon, Google Research, and Stanford. In the paper, they found that certain emotions were easier to model than others using the Reddit dataset(Demszky et al., 2020). This also provides us with a recently validated model for emotions using the same context (Reddit) and published at a top-tier conference after peer review, ACL 2020.

Reddit data itself has found use in similar studies. One such study uses text from nine-thousand users in order to classify them using Myers-Briggs personality type indicators(Gjurković and Šnajder, 2018). Another uses Reddit-sourced text to train a model to detect offensive language within text(Hada et al., 2021). A comparison of emotional distributions can be seen in a study that used Reddit to find suitable users who resigned from their jobs and compared their comments before and after that date(Ireland et al., 2023). By utilizing Reddit's API, we should no doubt be able to find sufficient data within our specified forums for any and all classifications of outcomes in sports games.

There has also been some limited work on sports fans' sentiments. A recent study (Patel and Passi, 2020) looked at the sentiment of fan posts at the 2014 Soccer World Cup. However, they only used Naive Bayers, SVM, and other similar approaches. The authors also did not categorize the different types of events: wins and losses, that could specifically alter the emotions of the fans.

3 Method, Data and Results

The method for the project comprised three different steps:

1. Data Collection and Preprocessing

2. Classification

3. Analysis

3.1 Data Collection and Pre-processing

The project required two different forms of data: match details and user comments. The match details were required to categorize the match into one of four categories:

1. Favorite wins a close match

2. Underdog wins a close match

3. Favorite wins a one-sided match

4. Underdog wins a one-sided match

The favorite and underdog were decided based on betting odds available across the websites. There was rarely any difference between different platforms in terms of who was considered the favorite, so the average odds from Wikipedia were taken. The second decision was on the definition of a "close" match compared to a "one-sided" match. Conversations with fans of the two sports we shortlisted (American football and soccer) provided a general idea that we felt was good enough: a game with less than a **10-point difference (American football)** or a **2-goal difference (soccer)** was considered a close game. Everything else was not. Given the scope of the project, Reddit was the primary source used for the user comments' data. There were multiple reasons for this. Firstly, Reddit's structure of subreddits allows concentrated conversations to happen on a specific topic. Targeting r/nfl and r/soccer allowed us to look at the relevant discussions around these topics. Specifically within these categories of discussions, the focus was on match data. Reddit's sports subreddits often follow a method relying on Match Threads (and often pre-match and post-match threads) that provide a singular platform for all discussion relevant to the match. To ensure adequate representation for all 4 categories mentioned above, we ended up gathering Reddit threads for 24 matches across NFL and soccer, resulting in data for 10 SuperBowls and 1 regular season game for NFL, and 8 finals, and 3 other knockout matches from the UEFA Champions League in soccer. From these match threads, the three important things gathered through the Reddit API were comment content, poster flair, and upvotes. Poster flair was significantly important because it showed the team supported by the commenter. This is also why a lot of the data had to be scraped by the team, and we couldn't rely on existing databases that did not have the comment flair. To provide some context, a sample comment with a commenter flair is shown below.
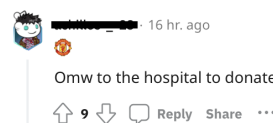

Figure 1: Sample Reddit comment with team flair

3.2 Classification

The classification of sentiments relied on two things: comment text from the matches specified above, and the Go-Emotions model. Given that the model was trained for sentiment classification on Reddit data, we did not need to further train or fine-tune the data. Go Emotion's model provides the probability for the following 28 emotions, which we predicted for 31,000+ comments.

Figure 2: Corpus of emotions in the GoEmotions model

3.3 Analysis

We show the final results by visualization and comparison between the emotion frequencies of different categories or groups. For example, we investigate how football fans react differently from soccer fans in general, how emotions in the population differ for matches in 4 different categories, and how people react differently when the match is in different phases.

Figure 3 shows the Stacked frequency of emotions of 3 types of viewers in 4 categories of matches. "F" stands for favored team supporters, "U" stands for underdog team supporters, and "N" stands for other neutral viewers. The definitions of 4 categories of matches are mentioned in section 3.1. In general, admiration, annoyance, and approval are the top 3 emotions with the highest frequencies. We noticed something interesting from the distribution plot, such as, admiration appears the most in the underdog team supporters when the underdog team wins a close match, and excitement appears the most when the underdog team wins a one-sided match.

200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249

250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
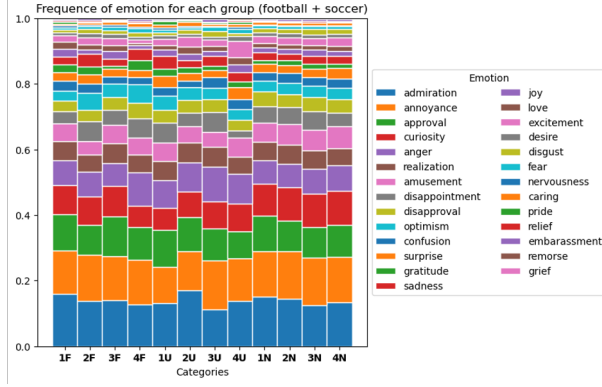288
289
290
291
292
293
294
295
296
297
298
299

Figure 3: Stacked frequency plot for emotions for different types of matches and different group of people
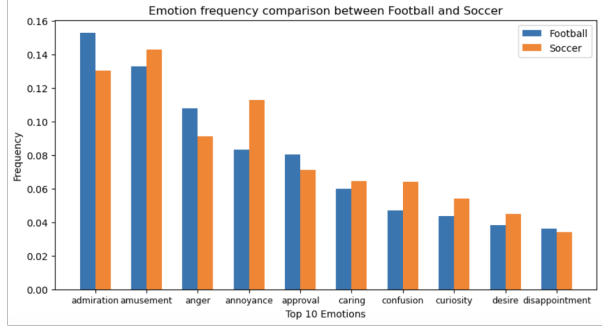

Figure 4: Top 10 emotion frequency comparison between football and soccer

Figure 4 shows the emotion frequency comparison between football and soccer for the 10 emotions with the highest frequency. Admiration and anger in comments appear more frequently in football fans, while amusement and annoyance appear more often in soccer fans.

In Figure 5, we show the emotion frequency comparison between matches in different categories for the top 10 emotions with the highest frequency. When the match is a close one and the favored team wins, emotions of admiration and anger appear most frequently. When the match is one-sided and the underdog team wins, the comments in the population show more annoyance, approval, and caring.

We also show the emotions change in the population when the match is in different phases in Figure 6. We investigate the comments in three phases: before matches, during matches, and after matches. In general, people have more anger and caring before matches than in the other two phases. More admiration and approval are observed during the matches than in the other two phases. And comparing to the other two phases, we see more
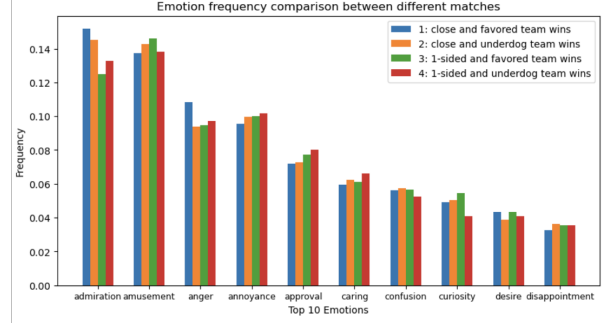

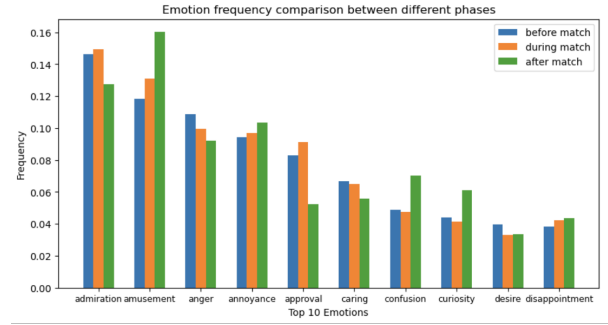Figure 5: Top 10 emotion frequency comparison between football and soccer

Figure 6: Top 10 emotion frequency comparison between different phases

amusement and annoyance after matches.

4 Experimentation

4.1 Match Outcome/Sport Prediction

We created a classification model using Decision-Tree, which we settled on after testing with other models and finding it to be the best for predicting category/sport using only phases and emotion-labels(no comment text data used as feature).

The result when predicting just the 4 victory outcomes(1-sided vs. close; favored team vs. underdog team) gives us an accuracy of around 0.40, but what is interesting about the results is that the F1 score is higher for games in which the favored team won vs. underdog team by around 0.10. When splitting the target into more classes of the emotion coming from neutral, favored, or underdog team fans. We can see that most of the accuracy comes from neutral members, as they are the largest group. However, the most interesting is that the highest precision comes from not the neutral comments but from favored team fans in which the match is 1-sided and the underdog team wins.

When using the model to predict between the 2 sports(Football vs Soccer), the model does better than random chance with a score of 0.70 accuracy. Some interesting findings from this are that recall

3

is very low for Football at 0.57 but a lot higher for Soccer at 0.82. The F-1 scores for both are 0.65 and 0.74.

4.2 Reddit Thread Generation

We sought to use our data to predict how a theoretical game, like an Eagles vs Rams Super Bowl would look like. This would mean training a Sequence to Sequence model that inputs and outputs English. We tried doing this by using two EncoderDecoder models. The encoder for our new model was taken from Google's BERT to BERT English to German transformer model (Rothe et al., 2019). We give it a context, which contains the game information like the team names, scores, match MVP etc. The encoder performs self-attention on this (tokenized) context to give us the attended tokens. We then use these tokens as the Key and Value in our decoder to perform cross-attention, which is the mirrored version of the encoder model, i.e., German to English.

During the fine-tuning process, the decoder predicts the next token given a tokenized Reddit comment using the aforementioned cross attention. The loss is the Cross-Entropy loss of the predicted token and the target token.

After the model has finished training, the model is used for generation. We give it a custom context and a starter token(s). We then let the model generate infinitely till it either hits an EOS token or till it hits a predefined generation length.

5 Conclusion

The project highlights that there are indeed significant differences between the types of emotions exhibited by the supporters of winning and losing teams based on their winning odds and the manner of victory/loss. These emotions change from the type of loss to the time at which they're captured as well. These findings are significant for researchers, marketing practitioners, and consumers themselves for multiple reasons.

6 Contributions

Sameed Khan: Project idea and conceptualization, data gathering, some pre-processing, report writing, analysis of individual commenters' data, and presentation.
Jianfeng Ke: Majority of the visualization of the main results.
Sean Lynch: Majority of data collection and pre-processing, some visualization of results, input in analysis of results.
Anveshak Rathore: Creating and finetuning the Sequence to Sequence transformer model on our analysed data.
Daheng Kuang: Wrote the base code for loading the pre-trained Bert model and also the code for scraping the Reddit API data. Created and tested the Decision-Tree model.

References

Dorottya Demszky, Dana Movshovitz-Attias, Jeong-woo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2700–2717, Online. Association for Computational Linguistics.

Molly Ireland, Micah Iserman, and Kiki Adams. 2023. Sadness and anxiety language in Reddit messages before and after quitting a job. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 467–478, Toronto, Canada. Association for Computational Linguistics.

Saba Mughal, Yusra Azhar, Margaret M Mahon, and Waqas J Siddiqui. 2023. Grief reaction and prolonged grief disorder. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

Ravikumar Patel and Kalpdrum Passi. 2020. Sentiment analysis on twitter data of world cup soccer tournament using machine learning. *IoT*, 1(2):218–239.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2019. Leveraging pre-trained checkpoints for sequence generation tasks. *CoRR*, abs/1907.12461.