SportsOri: A Novel Dataset for Analyzing Public Sentiment on Controversial Sports Events in YouTube Comments

Yuvraj Singh
IIIT Bhubaneswar
Bhubaneswar, Orissa, India
yuvraj.mist@gmail.com

Devadripta Jadhav
Savitribai Phule Pune University
Pune, Maharashtra, India
devadripta@gmail.com

Samiksha Boduwar
Savitribai Phule Pune University
Pune, Maharashtra, India
samikshaboduwar@gmail.com

Kripabandhu Ghosh
IISER Kolkata
Mohanpur, West Bengal, India
kripaghosh@iiserkol.ac.in

Abstract

This paper presents an analysis of YouTube comments on famous and controversial Public Sports Events. We explore public stance (stance detection) on a total of 6 famous controversial sports incidents by extracting and processing YouTube comments. Stance detection is performed on multiple events, including *The Underarm Incident*, *Jonny Bairstow's Run-Out* Incident, *Ashwin's Mankadding* Event, *Luis Suarez Handball* Event etc. The complete event details, results and evaluation metrics will be discussed in detail in subsequent sections. Our models can be found here [1] [2] [3]. Our entire pipeline can be found here [4]

CCS Concepts

• **Information systems → Web mining**.

Keywords

Stance Detection, YouTube Comments, Social Media Analysis, Sports Controversies

ACM Reference Format:
Yuvraj Singh, Devadripta Jadhav, Samiksha Boduwar, and Kripabandhu Ghosh. 2018. SportsOri: A Novel Dataset for Analyzing Public Sentiment on Controversial Sports Events in YouTube Comments. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

[1]https://huggingface.co/YuvrajSingh9886/Llama3.1-8b-Maradona/
[2]https://huggingface.co/YuvrajSingh9886/Llama3.1-8b-Frank-Lampard/
[3]https://huggingface.co/YuvrajSingh9886/Llama-3.1-8b-Luis-Suarez
[4]https://github.com/YuvrajSingh-mist/Public-Sports-Controversy/tree/master/data/PDFs

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Sports engages billions of followers worldwide[5] and impacts the economy [?]. Sports controversies often ignite passionate discussions among fans, analysts, and players. With the rise of social media, platforms like YouTube have become central to these discussions. This study aims to analyze the stances or perform opinion mining namely for, against, and neutral on comments from famous social media platforms – YouTube, focusing on events such as Jonny Bairstow's Run-Out Incident, Luis Suarez Handball Event etc. To our knowledge, the first-ever study of civic engagement in controversial sports events (cricket and football) spans around 40 years. LLMs (Llama3 family) were used for initial annotations (stance) of comments and later fine-tuned for comparative performance analysis (30% boost in accuracy).

Our study stands apart from its counterparts by focusing on public sentiment analysis surrounding controversial sports events, specifically through the lens of YouTube comments. Papers like SportQA [?] aims to evaluate how well large language models (LLMs) understand sports knowledge through a benchmark dataset while Run Like a Girl! [?] delves into gender bias in sports-related datasets, highlighting underrepresentation and naming disparities, we shift the focus to how people react to contentious moments in sports. It uses stance detection techniques to analyze public opinion, offering insights into the emotional and polarized responses to events like the Underarm Incident or Jonny Bairstow's Run-Out. Moreover, studies like Generating Sports News from Live Commentary [?] are centred on automating sports news generation from live commentary, emphasizing summarization and natural language generation. In essence, while the other studies explore sports understanding, bias, and news automation, our study uniquely examines the social media-driven public discourse around sports controversies, making it distinct in its focus on human reactions and sentiment rather than dataset creation, model evaluation, or bias analysis. Thus, after a thorough human verification, we are releasing a dataset of 40K+ opinion labelled comments (Section 2) and discuss the results in Section 3.

[5]https://www.statista.com/chart/14329/global-interest-in-football/

2 Dataset Creation: SportsOpi

2.1 Data Collection

We first identified famous public sports controversies by randomly picking 100 such events (football and cricket) from Wikipedia [6] and fetching 100 YouTube videos, sorted by "Most relevant" filter, related to each such controversy. Subsequently, the comments were fetched through YouTube Data API [7], and sorted in decreasing order of number of comments. Thus, the final controversial events were chosen by considering the total number of samples each event had followed by quality, engagement and balance of polarity. Our methodology focussed on the creation of a curated playlist for each event, ensuring diverse opinions. Since we are looking for public engagement on historic sports controversies, we look for events that have opinions of diverse classes viz. **Favour**, **Against**, **Neutral** and **Irrelevant**. The respective definitions for these labels can be found here.

2.2 The Events

Following the data collection pipeline, a total of six events were chosen, namely, **Frank Lampard Ghost Goal** [8], **Maradona Hand of God** [9] , **Luis Suárez's deliberate handball** [10] , **The Jonny Bairstow Ashes Runout** [11], **Ravichandran Ashwin's Mankading** [12], **The Underarm** [13].

A summarized version of the extraction code is available, and the full code repository link will be available at our **Github repository**.

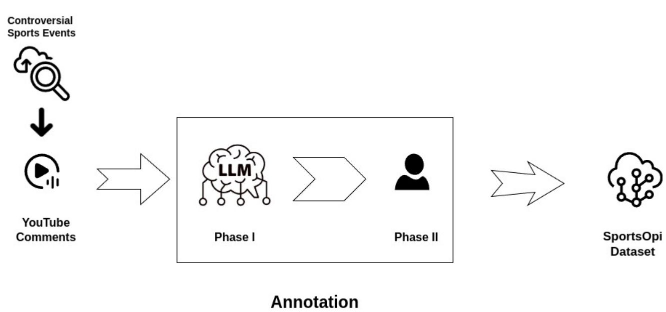

Figure 1: Opinion Annotation + Data Collection pipeline

2.3 Opinion Annotation Pipeline

Figure 1 shows the process of our annotation pipeline for opinion mining. The process for stance detection on our curated dataset consists of two stages as follows -

[6]https://en.wikipedia.org/wiki/Category:Sports_controversies
[7]https://developers.google.com/youtube/v3
[8]https://thesefootballtimes.co/2016/02/28/diego-maradona-and-the-reality-behind-the-hand-of-god/
[9]https://thesefootballtimes.co/2016/02/28/diego-maradona-and-the-reality-behind-the-hand-of-god/
[10]https://www.skysports.com/football/news/12040/12759389/uruguays-luis-suarez-says-he-will-not-apologise-to-ghana-for-his-handball-that-knocked-them-out-of-2010-world-cup
[11]https://www.espncricinfo.com/story/jonny-bairstow-reignites-ashes-stumping-row-1405100
[12]https://www.espncricinfo.com/story/jonny-bairstow-reignites-ashes-stumping-row-1405100
[13]https://www.espncricinfo.com/story/trevor-chappell-s-underarm-delivery-498574

Stage I: After the **Data Collection Pipeline**, a dataset of comments from the chosen 6 sports controversies was created from which we sampled 200 random comments on which initial annotation was done using zero-shot prompt using Llama 3.1-8b Instruct model. LLMs were used as initial annotators [**? ?**] due to the popularity and success of the same in the synthetic data generation domain. Subsequently, we, to do the initial annotation (*Favor*, *Against*, *Neutral*, *Irrelevant*). It is important to keep in mind that the opinion labels were made with the subject of the controversy in mind, like Maradona from *Maradona Hand of God* etc. We did the human verification with this idea in mind [14].

Stage II: Next, we humanely verified the labels as a result of **Stage-I**. Thereafter, a sample of 20 comments were chosen, which became the basis of the few-shot prompt [15]. Only a few samples were used, since LLMs are prone to "overfit" to a specific type of data/samples provided in the k-shot prompt (if in excess).

This few-shot prompt was then used to annotate the entire dataset of comments with opinions. This was followed by thorough human verification.

Table 1 shows the total number of comments, segregated into class-wise number of samples as well. **Table 2 shows the result of our annotation process. It consists of sample of each of the four labels from our dataset.**

The following details the above-mentioned pipeline for each of the controversies used to constitute our dataset KRIPA: there should NOT be separate strategies for different events. If it is done, it needs to be justified. -

Event	#C	F	N	I	A
Ashwin Mankading	3785	205	414	1734	1424
Frank Lampard Ghost Goal	13520	7000	1800	1100	3200
Johny Bairstow Runout	6073	331	1936	1786	1987
Luis Suarez Handball	11546	2400	2200	4200	2600
Maradona Hand of God	5159	2100	900	1500	500
The Underarm Incident	3676	330	126	1063	2113
Total	**43759**	**12336**	**7376**	**11356**	**11824**

Table 1: Name of comments, and class-wise distribution of comments.

3 Results and Discussion

Preliminary analysis indicates a significant division in public opinion across the six events within our dataset. Fine Tuning on our dataset improves the accuracy of the labels by a drastic margin along with other metrics such as F1 score, recall and precision as compared to the base instruct model.

Table 3 shows the result of fine-tuning models on each of the six events. Detailed results, including the distribution of stances (For, Against, Neutral, Irrelevant) and evaluation metrics (accuracy, precision, recall, F1-score).

[14]https://github.com/YuvrajSingh-mist/Public-Sports-Controversy/tree/master/data/PDFs
[15]https://github.com/YuvrajSingh-mist/Public-Sports-Controversy/tree/master/data/Prompts

Event	Favor	Against	Neutral
Frank Lampard Ghost Goal	Germans can't say anything about unsporting behavior.	The 1966 ghost goal had to be paid for.	This was way more clear-cut than 1966.
Luis Suárez Handball	Morality always loses, and nice guys finish last.	Hand of Satan.	He will never step foot in Ghana.
Maradona Hand of God	Number 15 goal is something else...my favourite. Bravo.	The most cheating player in football history.	You will never be able to pick one between Maradona and Messi.
Jonny Bairstow's Run-Out	That's not cheating, that's the way of winning.	Same old Aussies, always cheating.	The lesson for the players is "pay attention".
Ashwin Mankading Event	If a bowler can keep his foot inside the crease, a batsman can wait with the bat inside the crease until the ball is bowled. What's wrong with that?	If you Mankad, you should be ashamed of yourself. That means you don't have the skill to take the wicket by bowling.	Ashwin merely expressed his disappointment but never wanted a wicket that way. Team decision reflects it.
The Underarm Incident	That time it was legal to bowl underarm according to rules.	That was against the rules!! Couldn't they just ball a normal delivery? I mean there was no way a six would have been surely hit... well there could be...	What were the exact rules for underarm deliveries back then? Were you allowed to bowl as many as you wanted, and if so, why didn't they do it all the time?

Table 2: Examples of Favor, Against, and Neutral Comments for Controversial Events

Model	Event	Accuracy	Recall (micro)	Precision (micro)	F1 (micro)
DeepSeek-R1-Distill-8B (Not Fine Tuned)	*Maradona Hand of God*	22.76%	23%	31%	9%
	Luis Suarez Handball	33.7%	34%	49%	24%
	Frank Lampard Ghost Goal	223 %	23%	31 %	9%
DeepSeek-R1-Distill-8B (Fine Tuned)	*Maradona Hand of God*	76.20%	76%	76%	76%
	Luis Suarez Handball	78%	78%	78 %	78%
	Frank Lampard Ghost Goal	76.2 %	76%	76 %	76%
Llama 3.1-8b (Not Fine Tuned)	*Maradona Hand of God*	46.8%	46%	56%	40%
	Luis Suarez Handball	22.8 %	23%	31 %	9%
	Frank Lampard Ghost Goal	26.3 %	26%	39 %	25%
Llama 3.1-8b (Fine Tuned)	*Maradona Hand of God*	79.04%	79%	78%	77%
	Luis Suarez Handball	79.5%	80%	79%	79%
	Frank Lampard Ghost Goal	71.6 %	72%	72 %	72%

Table 3: Comparison of models with/without fine-tuning on our constructed dataset

3.1 Detailed Analysis

(1) *Number of samples* (**Favor**, **Against**, **Neutral** and **Irrelevant**)
   (a) The labels, *Favor* and *Against* is significantly higher for *Frank Lampard Ghost Goal* compared to other events with *Favor* being comparatively higher, followed by *Luiz Suarez Handball* event.
   (b) The *n*umber of samples for Neutral label is higher for *Luis Suarez Handball* event.
   (c) The label *Irrelevant* is significantly higher for *Luis Suarez Handball* event meaning the majority of the comments couldn't be classified into the other three labels.

(2) *Variations of Praise and Criticism*
   (a) Instances of *Direct Criticism* is highest for *Maradona Hand of God* event.
   (b) *Direct Praise* accounts highest for *Luis Suarez Handball* with equal instances for *Maradona Hand of God* and *Frank Lampard Ghost Goal*.
   (c) For *Indirect Criticism*, *Frank Lampard Ghost Goal* is highest followed by *Maradona Hand of God*.
   (d) In terms of *Favor* label (Direct + Indirect Praise), Frank Lampard event is highest, followed by Maradona and then by Luis Suarez events.
   (e) Similarly, for *Against* label (Direct + Indirect Criticism), Maradona event's count is highest followed by Frank Lampard and then Luis Suarez events.

Overall, *Frank Lampard Ghost Goal* event is highly favoured as well as resented by the public. A balance between the three opinions can be found in *Luis Suarez Handball* event.

(3) *Indepth analysis of stance labels*
   (a) We further investigated the primary stance labels, especially **Favor** and **Against**, by introducing more granular sub-categories: **Direct Praise**, **Indirect Praise**, **Direct Criticism**, and **Indirect Criticism**, along with tracking **Slang Use** and **Racial Abuse**. This allowed

for a finer understanding of how different types of expressions correlate with overall sentiment across the datasets.

(b) **Maradona Hand of God Event (Ref: Table 4 [16]):** *Clear Alignment:* Direct Praise (I=1) aligns 100% with **Favor** (E=0); Direct Criticism (H=1) aligns 100% with **Against** (E=1). *Ambiguous Alignment:* Indirect Praise (K=1), Indirect Criticism (J=1), and Slang Use (L=1) are spread across multiple labels. *Rare Instance:* Racial Abuse (M=1) appeared infrequently under both **Favor** and **Against**.

(c) **Luis Suarez Handball Event (Ref: Table 5):** *Strong Alignment:* Direct Criticism (I=1) shows 94% alignment with **Against** (D=1); Direct Praise (J=1) shows 78% alignment with **Favor** (D=0); Racial Abuse (N=1) shows 71% alignment with **Against** (D=1). *Weak Alignment:* Indirect Criticism (K=1) and Slang Use (M=1) lacked strong correlation with a single label.

(d) **Frank Lampard Ghost Goal Event (Ref: Table 6):** *Clear Alignment:* Direct Criticism (I) aligns 100% with **Against** (H=1); Direct Praise (J) aligns 100% with **Favor** (H=0). *Notable Trends:* Approx. 54% of Indirect Criticism (K) comments were labeled **Favor** (H=0); Approx. 57% of Slang Use (M) comments were labeled **Against** (H=1).

(4) *Probing Analysis of LLM outputs*

We ran probing analysis on the attention outputs of the fine-tuned (denoted as positive class) and non-fine-tuned (negative class) versions of the LLama 3.1-8 b-Instruct model, denoted by Fig 2 and 3, respectively.

The attention outputs is particularly high for the tokens - `exact_answer_first` and `exact_answer_last` for fine tuned model for majority of the layers, while for the non fine tuned one, it was high for all the tokens and layers, not consolidating to tokens particular to the answer to be generated, or the labels.

These heatmaps were generated by resp. F1 scores of a logistic classifier's accuracy were 82Due to the high metric (f1) for the fine-tuned model, we termed the labels generated to be a 'positive class' and for the other model to be a 'negative class' (low f1 score).

We can see that the heatmap for the positive class model has high attention outputs for the tokens corresponding to the labels to be generated or the answer in particular, while it attends to every token or haywire for the non-fine-tuned negative class model.

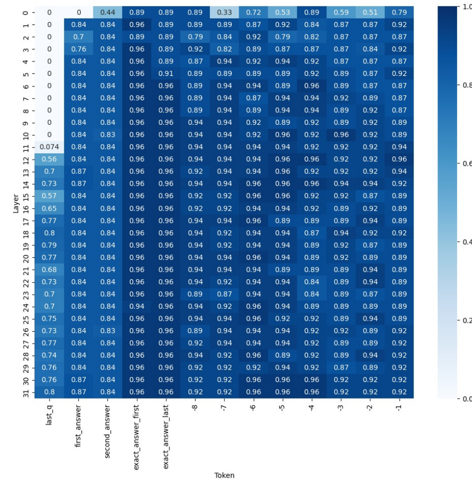

Figure 2: Frank Lampard

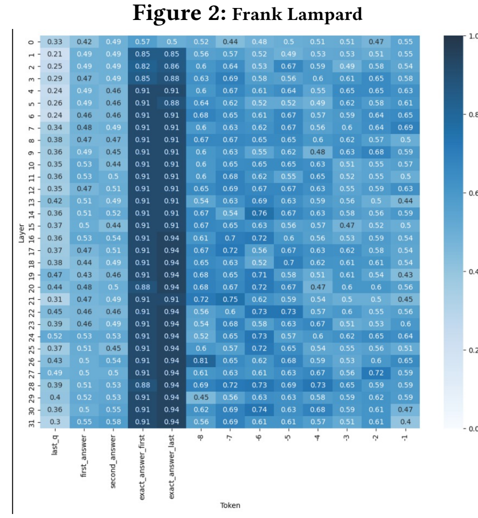

Figure 3: Suarez

Figure 4: Attention heatmaps for fine-tuned LLama model.

4 Conclusion

This study highlights the role of social media in shaping public perception of sports controversies. The integration of automated data extraction and stance detection provides a comprehensive view of audience sentiment. Future enhancements will aim to improve accuracy and broaden the scope of analysis.

[16] https://github.com/YuvrajSingh-mist/Public-Sports-Controversy/blob/master/data/PDFs/Tables.pdf