Perspective from Redditors - The Big Three in Tennis

L. Nguyen
Radboud University, Nijmegen, the Netherlands
lam.tungnguyen@ru.nl

ABSTRACT

Tennis fan all over the world has been witnessing the most dominant era of men's tennis - since the Big Three started taking over the ATP World Tour. The facts that 63 out of the last 77 Grand Slams tournaments belong to them and the top spot of the ATP ranking had been held by either of them for 17 years signify their absolute dominant on the Tour (Source). Thus, there sprouts a question of who is more supported by tennis fans more than the others.

1 INTRODUCTION

Being an avid tennis fan and since Roger Federer just announced his retirement from the game, I want to examine how the "Big Three"—Nadal, Djokovic, and Federer—compare to one another and determine who is more admired than the others. Facing the pressure of being famous with the media and millions of tennis fans, it's hard to conceal one's scandals. Djokovic is infamous for his denial to participate in the Australian Open and the US Open because of his vaccination status. Federer - who is known for his flawless display both on and off courts, incidents with the umpire are the only "scandals" that is well-known of. Similar to Federer, Nadal has captured the heart of many tennis fan. He is best known for his masculinity, strength and never give up attitude in both his physical and spiritual presence. Also like Roger, he hardly ever stirs a big controversy in tennis community. However, his 14th Roland Garros title is rumored by the pain-killing injections received by him to play the tournament.

Tennis enthusiasts generally agree that Federer, who has won six Laureus World Sportsman of the Year awards [7], has the least incidents on and off courts as well as having the most supportive fan base. In this project, I'm specifically interested in compiling posts and comments on each player's most well-known scandal and do sentiments analysis with Valence Aware Dictionary and sEntiment Reasoner - VADER and Roberta - a Pretrained model from HuggingFace and conclude who, by facing these scandals, receives more support from tennis fans

2 RELATED WORK

Due to the rapid expansion of Internet-based applications like social media platforms and blogs, daily activity-related comments and reviews have been produced as a result. Sentiment analysis is the process of compiling and examining the opinions, ideas, and impressions of people with reference to a variety of issues, goods, and services. In this particular topic, there hasn't been many related works about sentiments of tennis fans on the players, their tournaments, their daily life, their scandals or their interactions with the fans and the press.

TxMM'22-'23, Radboud University, Nijmegen, Netherlands

The most related paper that discusses similar topic is [8]. The paper explains through a regression model how sentiment bias can affect betting outcome in tournament finals and matches of the Big Four (Federer, Nadal, Djokovic and Murray) over data from Google Trends and betting sites.

3 APPROACH

My intended method would be in concordance with the research question: how the sentiments of Redditors on the scandals can reveal their relationships with each player then use the outcomes to conclude who receives the most support. The detailed procedure is mentioned in the subsections below.

3.1 Installing necessary Python packages

In this project, I am working with data from the subreddit r/tennis, hence the library 'PRAW' [1] - which stands for the Python Reddit API Wrapper - to crawl submissions and comments data on a particular subreddit is installed. To parse the scraped data into dataframes, 'pandas' is utilized. For string tokenization and sentiment ratings, I imported a class AutoTokenizer from the package 'transformers' and installed two subpackages 'nltk.sentiment' and 'scipy.special' for VADER and Roberta models. And finally, 'matplotlib' is implemented to plot the results,

3.2 Crawling data from Reddit

The most popular subreddit to discuss about tennis is r/tennis with over 900 thousands redditors, and I chose it to be the data source. First, to get authentication for accessing the API, I go to Reddit and create an app to get the security tokens. Next, I save the list of sentiment words (positive and negative words) from Github [5] to help filtering relevant posts and comments.

Having the required tools, the next step is to crawl the data. Here, the data is saved as an array containing posts and comments in the following format

```
dataset = [
    {
        "headline": {
            "title": "title",
            "id": "post_id",
            "score": "score"
        },
        "comment": [
            {
                "author": "author",
                "body": "comment_body"
            }
        ]
    }
```

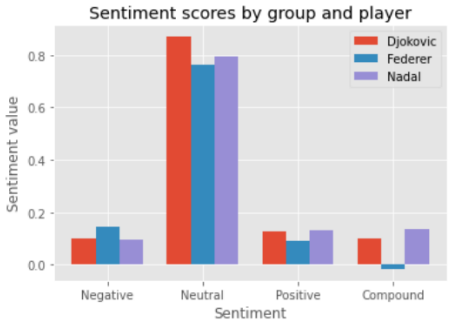

Figure 1: VADER results of sentiments of Redditors towards Djokovic, Federer and Nadal scandals

]

Next, I use the search term containing the name and the most famous scandals associated with each player'. For Djokovic, it's 'Djokovic vaccination', for Nadal, it's 'Nadal foot injections' and for Federer, it's 'Federer umpire'. The search is set to return relevant posts first with a limit of 10 posts.

3.3 Filtering the Data

For each player, the filtering words are set to "he", "him" and any names that the player are often called. For example, "Djokovic", "Novak", "Nole" for Djokovic. Next, the two filtering masks are created and combined, containing the sentiment words and the search term for each player. The 'dataset' array is filtered by applying the combined mask to the comment body, returning relevant comments that contain those specified words in the mask.

3.4 Sentiment Calculations

In this section, two models are used to calculate the sentiment, namely, VADER and Roberta models.

For VADER model, the desired sentiments are calculated for the post title, post body and the comments. A python dictionary 'player_sentiment' is initialized, with attributes 'neg', 'neu', 'pos' and 'compound', which stands for negative, neutral, postive and compound respectively and matches the output of the model. Then a series of for loops are ran over the list of the aforementioned data to gather the respective sentiments. Finally, an average value for each sentiment is calculated by dividing the total by the combined length of posts' titles, posts' bodies and their comments.

For Roberta model, the process is almost identical to VADER's, with an addition of polarity calculation for individual text by first tokenizing the text using the pretrained model generated by AutoTokenizer and fit the encoded text to Roberta. Finally, a softmax layer is fitted to the output to transform each output vector to a probabilistic one.

3.5 Plotting Word Cloud

An efficient way to display how the sentiments are distributed is by plotting a word cloud - a word visualization on how common a word appear in texts. This is done by implementing the existing 'WordCloud' library in Python and plot the comment body data for each player.

4 RESULTS AND ANALYSIS

In this section, I am presenting and comparing the outputs of the tennis fans' sentiments on each tennis player's most famous scandal for the two model, VADER and Roberta.

4.1 Result

Using the VADER model, the average sentiments of tennis fans opinion about Djokovic's vaccination status is 0.099 for negative, 0.87 for neutral and 0.126 for positive. The compound score, which is the normalization between -1 and 1 of the sum of positive, negative & neutral scores, is 0.101.

For Federer and his incidents with the umpires, the average sentiments of Redditors are slightly different. They have more negative views toward those scandals with an average negative of 0.143, a neutral of 0.765, a positive of 0.09 and a compound score of -0.0157.

Finally, for Nadal and his foot injections scandal during the 2022 French Open, the sentiments are similar to Djokovic's, with 0.095 for negative, 0.796 for neutral and 0.132 for positive and a compound score of 0.136. The VADER results for all three player can be seen in Figure 1.

The Roberta model, as in figure 2, however provides a contradiction to what the VADER gives, in which the results sentiments are more polarized, i.e less neutral. For Djokovic's, the model says that Redditors posts and comments are 0.454 negative, 0.463 neutral and 0.130 positive. For Federer's, the sentiments are similar, with 0.43 for negative, 0.465 for neutral and 0.134 for positive. For Nadal's, the sentiments are shifted towards positivity, with 0.375, 0.463 and 0.172 for negative, neutral and positive sentiment respectively.

The WordCloud output for Djokovic reveals that besides his name, the words that mention about the Vaccine and Djokovic his perspective toward those: "vaccinated", "vaccination", "someone", "even", "will", "getting", "want" are the most frequently mentioned. This is consistent with the output of Roberta model, that the sentiments of Redditors are generally not quite negative towards Djokovic's scandal.

The WordCloud for Federer reveals a different results, besides his names "Federer", "Roger", "Fed" are amongst the most mentioned, there are not many words that expresses Redditors opinions on

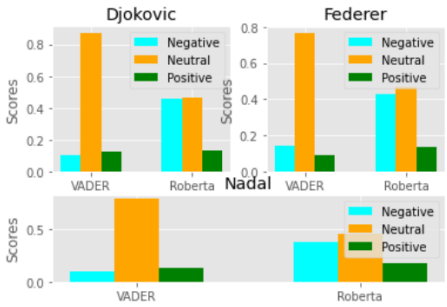

Figure 2: Comparison of sentiments between VADER and Roberta model of Redditors towards Djokovic, Federer and Nadal scandals

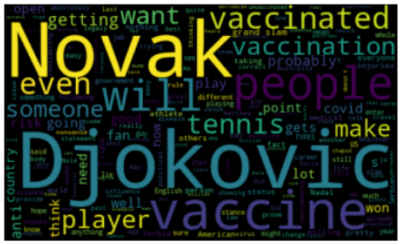

Figure 3: Word Cloud of Redditors towards Djokovic's scandal

his incidents with the umpires, which contradicts the result from Roberta Model, showing that statistically the sentiments are similar to Djokovic's.

Figure 4: Word Cloud of Redditors towards Federer's scandal

For Nadal and his incidents with foot injections during the 2022 French Open, similar to Federer's, not many sentiment words are expressed by Redditors, most are about Nadal's incidents: "pain", "feet", "injury". This is consistent with output of Roberta Model, which reveals that the sentiments are mostly less negative and more neutral than Djokovic's.

4.2 Analysis and Discussion

Figure 1 shows that the sentiments of Redditors towards Djokovic, Nadal and Federer's scandals are mostly neutral. This can be explained by the fact that VADER uses "bag of words" approach - taking all words in a sentence and evaluate each word as positive, neutral and negative, then combined the values to a total score for each sentiment. We can see the same pattern in the WordCloud of

Figure 5: Word Cloud of Redditors towards Nadal's scandal

each player, as there are not many words that reflect the sentiments of Redditors. A more accurate calculation would come from Roberta model, which is shown in figure 2.

The results from Roberta model shows an increase in negative sentiments compared to the results of VADER's. This is because VADER doesn't account any relationship between words, which is important in analyzing text. On the other hand, Roberta is pretrained on a large corpus of data from Twitter and can pick up the underlying relationships between words in a text. For example, not every word in the text is used with its intended meaning - a negative word can be used for sarcastic reasons. Therefore, Roberta can reveal the underlying sentiment of a written text, in this case, opinions and comments of Redditors toward a tennis player's scandal.

Also, from the output of both model, we can see that the sentiments of tennis fans on Reddit are quite similar between the three players, although this also depends on the scandal that each person is facing. For example, Federer situation only involves himself and the umpire, which is more forgivable and less personal than Djokovic - whose incidents have stirred up controversy in not only the tennis community but to science and the people fighting Covid, the gap in sentiment still reflects that tennis fans are not biased toward favoring Federer. They stand for each player's right from the media and respect their opinion.

4.2.1 Internal Validation. To evaluate the accuracy of one of the model, I am taking an example of a comment on Djokovic's vaccination status and his denial to enter the Australian Open in 2022. The comment can be summarized as Novak and his decisions can influence millions of his fans, and some could be in serious

```
[33] # Run for Roberta Model
     encoded_text = tokenizer(example, return_tensors='pt')
     output = model(**encoded_text)
     scores = output[0][0].detach().numpy()
     scores = softmax(scores)
     scores_dict = {
         'roberta_neg' : scores[0],
         'roberta_neu' : scores[1],
         'roberta_pos' : scores[2]
     }
     print(example)
     print(scores_dict)

     > “Because it’s terrible for tennis, not good for him – and, really, he does lead by example. The fact that he doesn’t trust it, it does lead a lot of peopl

     This is why Novak, and others, who say it's a personal decision are crazy. It affects tons of other people. I mean, this whole epidemic started with ONE PER

     So great, if Novak gets sick, he's worth 200 million or whatever, he can afford and get quick access to the state of the art covid treatment. But the people

     This is so dumb. My whole life I've received so many vaccines. When I was about to go to Thailand, I had to get like 3 or 4 different ones in the same day.
     {'roberta_neg': 0.7899501, 'roberta_neu': 0.1769149, 'roberta_pos': 0.033134975}
```

Figure 6: Snippet of the comment toward Djokovic's scandal and output from Roberta Model

health problems because they don't have access to medical treatment like he does. The output of Roberta reveals that the comment is 79 percent negative, which accurately depicts the disagreement of a Redditor toward Djokovic's decision. With the accuracy of Roberta model for Djokovic's scandals, we believe that the outputs will be similar toward Federer and Nadal's scandals.

4.2.2 External Validation. Browsing across several news platforms, it is not uncommon to see articles that express negativity towards Djokovic vaccination status and his denial of entry to the Australian Open and US Open Grand Slam tournaments. In this BBC article [4], it is claimed that as of February 2022, more than 60 percent of the world population had been vaccinated and the vaccine has saved over 500 thousands live in Europe and that it has been under rigorous testing and clinical monitoring. To us, these facts alone are ample to convince one to take the vaccine, hence it is understandable that the tennis world has been unsupportive of Djokovic's decision. Tim Henman - a renowned British tennis player, says that Djokovic "is certainly jeopardising his chances of being the greatest male player of all time" and Pam Shiver - a five time Wimbledon double champion was hopeful that some medical authorities would discuss with him and gain his trust on the vaccine.

As can be seen from Figure 2, the negative sentiments of Redittors about Federer's incidents with the umpire is comparable to the ones of Djokovic, however, there are not many sources on the internet that confirm this sentiment. In an article by Eurosport [3], when asked about Federer being too slow between points because of having to grab his towel thus when on a heated exchange with the umpire, Mats Wilander - a 7 times Grand Slam champion, defends Federer by saying "You have to go and get your own towel and Federer plays so quick between points - always has". This is due to the fact that Federer is generally less known for his incidents on and off the tennis court thus receiving more benign reactions from the tennis world.

Figure 2 also tells us that Nadal doesn't receive as many support from Redditors with a near 0.5 negative sentiment. Unlike Federer, there have been polarized reactions of the sports community towards Nadal's foot injections. In this article by Eurosport [6], Chris Evert - who holds 18 Grand Slams, praised Rafa's fighting spirit going through adversity and said that the 2022 French Open title is "the most meaningful ever" to him. On the other hand, from [2], Pinot and Martin - two French cyclists have expressed dissatisfaction with the foot injections Nadal received, stating that such treatment is not approved in cycling and one should not have participated having such injury.

The project is done under a few assumptions. First, the data for all three players includes comments towards others not in the big 3 - others tennis players and possibly trainers, coaches and people who are involved in the tennis community. This is because the applied filter mask does not consider comments that mentions both the big 3 player and those others people. This could directly influence the sentiments calculation by comments targeting the wrong person. For example, there could be comments about Federer's coaches whose tantrums towards umpires are never heard of, in this case, this would still be counted as a positive comment towards Federer, which is related to our second assumption.

Our filtered data also assumes that the comments are reflected towards the players, and not the other party involved in the scandal. All too often, there are a lot of comments targeting an opposite target of the player, but contain the same sentiments as the direct one to the player himself. For instance, a comment could criticize the Australian Government for not letting Djokovic enter the country despite having an exemption for vaccine, while others blame him for not vaccinating and even hypothesize him winning the Australian Open had he been given entry. Once again, this assumption could influences the sentiment calculations, however, the polarization opinions of tennis community would also be based on common sense of what is right and who is at fault, hence there would not be a huge difference between number of comments who are supportive and unsupportive of a player.

5 CONCLUSION

The project is challenging but quite rewarding. Finishing the analysis and discussion, we cannot conclude whether Federer, Djokovic or Nadal has a more supportive fan base. The fact that the outputs from both VADER and Roberta models reveal similar sentiments cements that tennis fans are not predisposed toward any player. They support each player's right toward the scandals and value their viewpoint. In future work, we think it would interesting to do the same experiment with a more representative dataset, an improved filtering and alternative statistical methods besides averaging - which could balance out each sentiment's nuances, such as precision and recall. This would rule out our current assumptions and provide readers with an improved analysis.

REFERENCES

[1] [n. d.]. The python reddit api wrapper. https://praw.readthedocs.io/en/stable/

[2] Patrick Fletcher. 2022. Pinot and Martin speak out against Nadal's injections. https://www.cyclingnews.com/news/pinot-and-martin-speak-out-against-nadals-injections/

[3] Dan Quarrell. 2021. 'big mistake!' - Roger Federer defended over towel row with umpire and Marin Cilic at French Open by Wilander. https://www.eurosport.com/tennis/roland-garros/2021/french-open-roger-federer-defended-over-towel-row-with-umpire-by-mats-wilander-big-mistake_sto8353978/story.shtml

[4] Amol Rajan. 2022. Novak Djokovic willing to miss tournaments over vaccine. https://www.bbc.com/news/world-60354068

[5] Shekhargulati. [n. d.]. Sentiment-analysis-python/opinion-Lexicon-english at master · Shekhargulati/sentiment-analysis-python. https://github.com/shekhargulati/sentiment-analysis-python/tree/master/opinion-lexicon-English

[6] Ben Snowball. 2022. 'the foot was asleep' – Rafael Nadal played French open final with 'no feeling' in left foot after injection. https://www.eurosport.com/tennis/roland-garros/2022/the-foot-was-asleep-rafael-nadal-played-french-open-final-with-no-feeling-in-left-foot-after-injecti_sto8979017/story.shtml

[7] ATP Staff. [n. d.]. Roger Federer's twin wins crash Laureus web site: ATP tour: Tennis. https://www.atptour.com/en/news/federer-laureus-awards-2018-tuesday

[8] S. van Rheenen. 2017. *The Sentiment Bias in the Market for Tennis Betting*. Master's thesis. http://hdl.handle.net/2105/38118