Data Science - Fall 2022 - Mini Project III

Alex Montoya Franco
alex.montoyafranco@abo.fi
EDISS Master's Programme - Åbo Akademi University, Turku, Finland

1. Introduction

For a long time, going to a theatre to experience the latest movies was not only a very common activity, but it was consider a worthy one, that was the only way to watch the latest and biggest movies. Watching a movie at home was a lesser experience to say the least. Nowadays, the explosion of content produced by different streaming services, such as Netflix, Disney+, Amazon Prime, in addition to the recent pandemic, have shown audiences that there is plenty of quality content to be entertained at home. In some cases, even, the latest movies are now available at home for no extra cost just weeks after their release on cinemas.

However, one of the biggest differences between the theatrical release vs streaming releases, is that in the former, talent across the board (directors, artists, actors) can reach very good deals, which usually includes a percentage of the movie's income, while on the other hand, streaming services do not offer these types of deals because revenue of a specific piece of content is not shared by the streaming services. For the actors, this is not a problem, they get paid if not thousands, millions of dollars. But for all other talent, salaries are now based on estimates rather than the actual revenue of content. Sometimes this content might be generating hundreds of millions of dollars for years while the talent involved is just getting a fraction of it. The majority going to studio executives and actors.

An interesting fact to point out, is that the industry is quickly changing. Netflix, the company that started it all when we talk about streaming, is releasing more movies on cinemas than ever before (which allows them to participate in awards season and generate additional revenue) while companies like Disney, long known for its big and famous theatrical movies is now also focusing more on its streaming strategy. Therefore, we wonder, Should entertainment companies focus on movie going experiences or streaming experiences? What do costumers want? Have customer's definitely shifted to prefer at-home entertainment or is there still room for going to the cinema to enjoy the latest movies?

With this report, we seek to take advantage of Twitter's API to analyze where customer's attention is. Is there more attention around streaming shows or cinematic releases? How has this changed over the past years, and can we predict how is it gonna be in the next few years?

With this project we intend to answer the following questions:

1. Is the movie going experience still relevant in today's pop culture compared to the attention that streaming services are facing?
2. What is the evolution of consumer's behaviour, denoted by number of tweets, trending topics, hashtags, social interactions, in terms of theatrical experiences vs streaming services?
3. Is it possible to predict this behaviour for the next few years?

This report is organized as follows: (1) Data Gathering, which explains what data is captured from Twitter's API and how it is captured, (2) Data Analysis, which includes a descriptive analysis of the current social media attention around movie going experiences vs streaming experiences, a trending analysis of cinema vs streaming, and a forecast to predict whether cinema experiences will be relevant in the future or not, and finally (3) Conclusions are presented describing the scientific bottlenecks of this project and how we overcome them.

2. Data Gathering

For this project, the required data was the amount of tweets discussing our topics over the past few years. Twitter's API basic tier a.k.a Essential access just allow the search and count of tweets from the past 7 days. Given this limitation, Academic Research Access was requested in order to get historical data from Twitter and be able to build our time series and forecasting. The access was granted so this enable us to continue with the desired path.

For requesting the data to the API, we used Tweepy [1], an easy-to-use python library for accessing Twitter's API.

The following two endpoints were used to answer our research questions:

- Tweet Counts [2], which helped us understand the size of the conversations around our topics over the last few years, without actually having to pull the tweets themselves. The Academic Research tier gave us access to the full-archive Tweet counts endpoint (March 2006 to today).
- Search Tweets [3], which allowed us to get the tweets text and context annotations [4], in order to improve our search query and find the relevant tweets for our research.

[1] https://www.tweepy.org/
[2] https://developer.twitter.com/en/docs/twitter-api/tweets/counts/introduction
[3] https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction
[4] https://developer.twitter.com/en/docs/twitter-api/annotations/overview

As mentioned above, the data that we required for this project is mainly the count of tweets addressing both "movie going experiences.ªnd "streaming experiences". In this sense, we don't need to go into the detail of each tweet, but just want to analyze what people are discussing about on the social network. On the other hand, these two topics are not very specific. For instance, how do we identify that a tweet refers to a "streaming experience¿ We consider that the key for our project's success is in how we construct the queries that are gonna provide that the right"data.

For instance, the "movie going experienceçould be characterized by the mention of main cinema chains around the world, mention of big movies released on cinemas, or context annotations. For example, movie festivals.

On the other hand, the "streaming experienceçould be defined by the mention of a streaming service (Netflix, Disney+, HBO Max, AppleTV+, Amazon Prime Video), famous streaming shows (e.g., Stranger Things, The Crown), and also relevant context annotations (TV Shows, TV Episodes, Global TV Show).

These are some of the elements considered when building our queries.

- Query can be 1024 characters long (for academic tier)
- Standalone operators are: #hashtag, keyword or "phrase"
- Conjunction-required operators can be: is:retweet, place_country:US, has:geo, lang:en, -is:nullcast, context:domain_id.entity_id
- Boolean operators are available: AND (space between keywords), OR, NOT (-)

Initially, we searched some tweets in order to define the keywords for our queries, but we noticed that we were getting Tweets in a variety of different languages. In this situation, we decided to only include Tweets that are written in English with the operator 'lang:en', which facilitated the process of capturing relevant information from the tweets and might also be helpful for possible future NLP techniques.

We also include in our query -is:nullcast, which removes Tweets created for promotion only on ads.twitter.com. With this we hope to remove the "noise"that would represent self promotion of tv shows or movies on Twitter. We have not interest in advertisement tweets since this will lead to who is advertising more, instead of who is generating more interest from consumers.

We also include the query clause -is:retweet, which will not match on Retweets, thus matching only on original Tweets, Quote Tweets, and replies.

We finally tuned our queries by adding search with context annotations using 'context:domain_id.entity_id'. The purpose of this, is to find tweets that do not necessarily contain our specified keywords but that are still related to our topics.

Domain 131 (Unified Twitter Taxonomy) refers to Twitter's User Facing Interest Taxonomy. This taxonomy helps to power features on the platform such as, Topics.

The relevant contexts identified are:

- (Include) TV streaming services (domain: 131, entity: 1397573214239289349).
- (Include) Superhero films (domain: 131, entity: 1478371796449517568)
- (Exclude) Animated films (domain: 87, entity: 856976559529263104)
- (Exclude) Animation (domain: 131, entity: 1192082698568814592)
- (Exclude) Animation & Comics (domain: 131, entity: 1297951101136732160)

Twitter offers a great set of context annotations with a wide range of domains and entities to use. However, we don't use many of these since they could skew our results instead of tuning them. We won't use the domain "TV Shows"since it could also refer to tv shows from network television which we consider as another different discussion. Domain "Movies & TVïs also very general. Entities about specific tv shows (e.g. stranger things) were initially considered, but were ultimately removed since this analysis plans to take into account multiple years and generally specific shows are popular for a couple of them and audiences move from content to content very often.

We tuned our streaming query with the entity of TV streaming services.

Some of these domains collide with the movie going experience since they are covering movie or entertainment in general so for that reason they are not consider for the "streaming services query".

Domain 47 - Brands and Companies (e.g. Netflix - see tweets annotations page) is also not consider since this might bring tweets about the company itself and not their shows, tweets could be discussing market share, price, advertisement.

We did not find many relevant contexts (domains+entities) for our cinema query (also considering that the more complex we make our query, the more volatile our results might be).

Domain 117 (Movie Festival) was consider, although we did not need a specific entity (e.g., a specific festival) so we tried to use a wildcard, i.e., bring all entities from this domain using context:117.*. However, we got this error: "BadRequest: 400 Bad Request There were errors processing your request: Wildcards are not supported in Search when using the 'context' operator. Please resubmit without using a wildcard. (at position 1)". This was a setback for the project, since we discovered that a general search using context domains was not possible, we were only allowed to search with specific pairs of domains and entities.

Domain 86 (Movie: A film like Rogue One: A Star Wars Story) was also consider. For instance, both black adam and black panther were tagged with this context domain. The description made us believe it was related to star wars, but it ended up referring to movies of that size and scope (theatrical releases). Once again, we couldn't use this context, since we would have to input a specific entity (e.g., a specific movie), and that was not the intended purpose for context annotations.

To use contexts to find relevant tweets, we therefore need the domain and the entity. In most cases, the domain is interesting for our analysis but the entities are not. We are not looking for tweets talking about a specific movie, but movies in general. For example, 86.1297874929275113472 would be domain Movie, entity: Black Adam. A considerable amount of time was spent on finding relevant domains to tune our queries. However, as mentioned above, a general search of some contexts was not allowed by the API.

We still consider a domain to be relevant four our cinema search query given the number of tweets we believe do not include any mention to the keywords movie, theater, cinema but instead just talk about a movie watched in the theater. Even if these films are later promoted in streaming tweets, these are originally theatrical experiences.

The 130 domain (Multimedia Franchise - Franchises which span multiple forms of media like 'Harry Potter) was also consider but ultimately not included because these franchises now also include tv shows, animated shows, parks, so it would bump up the numbers without a clear correlation to cinema experiences.

Context 131.1002576732971384832 refers to the Marvel Universe, which was long known for its theatrical releases and big popularity, but in recent years it has expanded beyond movies so it might collide with streaming tweets / park experiences.

The context of "Superhero films"was finally included since there hasn't been many superhero films released first and only for streaming. None that come to mind. And these are the blockbusters that drive a lot of the conversation when discussing movie going experiences. Leaving them out would be an unfair comparison. Also, this context, provided a domain and entity which was neither too specific or too broad for our research.

Adding the Superhero films context bump up the numbers in a big way, concerned by the content of these tweets, we looked into the content of some of them, and determine that this context was also bringing tweets just talking about the comics that inspire these movies, or even animated series and films.

Since this are clearly not related to our topic we also filter out tweets that have those other contexts. Just keeping tweets about films, theatrical superhero films and removing the tweets only mentioning comics or animation entities. There has been very few animated superhero movies released at cinemas which might give more accurate results. Also removing tweets about superhero comics which might also drive some conversations.

This tweet shows that adding superhero film context to the query really help bringing relevant tweets. This tweet would have not been captured with our original query, since the only keyword that fits is movie.

«It's a fun movie man go support it #BlackAdamMovie»

Some of the captured tweets really show that the addition of contexts bring some very interesting tweets that we were not able to capture before, while also bringing some noise. Further data processing might help with this.

We decide to keep these results since it serves the purpose of our search: attention around the movie going experience. Again the logic behind including superhero genre is that even though some of these tweets don't specifically mention the experience of going to the cinema. A high percentage of these tweets started from that experience since for many years and just until recently that was the only way to watch the latest biggest superhero movies.

Listing 1 and **Listing 2** show the queries used to extract the required data. These were the result of an iterative process where tweets' text was analyzed in order to tune our queries and get as reliable information as possible.

For our streaming query, we consider the most popular streaming services. And we also consider tweets discussing original content, which is the most important content for each service.

Nowadays, streaming is also producing some high quality movies that caught people's attention, but this receives much fewer attention from the public than streaming shows. It would also produce a more complex query trying to differentiate streaming movies from theatrical movies that are now available to stream at home.

Some of the keywords chosen for the query are also based on looking at several tweets and seeing how people refer to these streaming services and shows.

Companies like Disney, long known for its theatrical releases and tv network, it's now a big player in the streaming wars. Careful consideration was taken when defining the query, in order to not bring disney tv shows from network, which have been part of pop culture for years, but never competing directly with the theatrical experience. Same with HBO. HBO shows have been very popular on tv network too, while hbomax is the streaming service for these shows.

Documentaries in streaming services have also been very popular (the keyword was added after searching tweets about apple tv plus and discovering that there were discussions going around their latest available documentary). This goes without mentioning how Netflix has had some very big documentaries over the past years.

An initial test was made to understand the magnitude of the conversations around our topics and the accuracy of our queries. In this test, we took what we considered was a -elevant"day both for streaming and cinemas.

For streaming, this was May 27, 2022, which was the day when Netflix released Stranger Things 4, while Disney+ released the first episodes of Obi-wan Kenobi. Two big streaming shows in this day which should be a pick for streaming overall, giving us correct domains/entities, and relevant tweets.

For cinema, this was October 27, 2022, which was the Los Angeles premier of Black Panther: Wakanda Forever, a highly anticipated movie. This day should represent a pick for the movie going experience since critics and some fans were able to watch the movie for the first time and start giving their opinions online without spoilers. After applying this query and discovering few instances of this movie, we applied a manual search on twitter and found out that most tweets in these type of scenarios don't mention theaters, premier, or

even the keyword "movie". We try to solve this issue using context annotations.

Listing 1: Streaming query

```
# Without context
(netflix OR disney+ OR disneyplus OR hbomax OR hulu OR
  appletvplus OR peacock OR primevideo) (show OR series
  OR episode OR streaming OR original OR season OR
  documentary) -is:retweet lang:en -is:nullcast
# Number of tweets for May 27th, 2022: 15.451

# With context
(netflix OR disney+ OR disneyplus OR hbomax OR hulu OR
  appletvplus OR peacock OR primevideo) (show OR series
  OR episode OR streaming OR original OR season OR
  documentary) -is:retweet lang:en -is:nullcast context
  :131.1397573214239289349
# Number of tweets for May 27th, 2022: 10.859
```

Listing 2: Cinema query

```
# Without context
(watch OR IMAX OR Dolby OR 4DX OR theatrical OR drink
   OR snacks) (theatre OR theater OR cinema OR amc OR
  cineworld OR cinemark OR cinepolis OR cineplex OR vue
  OR cgv OR cinemex OR wanda OR pvr OR kinepolis OR
  finnkino OR odeon) (film OR movie OR screening OR
  showtime OR premier OR franchise) -is:retweet lang:en -
  is:nullcast -bluray
# Number of tweets for October 27th, 2022: 489

# With context
(((watch OR IMAX OR Dolby OR 4DX OR theatrical OR
  drink OR snacks) (theatre OR theater OR cinema OR amc
  OR cineworld OR cinemark OR cinepolis OR cineplex OR
  vue OR cgv OR cinemex OR wanda OR pvr OR kinepolis
  OR finnkino OR odeon) (film OR movie OR screening OR
  showtime OR premier OR franchise)) OR (context
  :131.1478371796449517568 -context
  :87.856976559529263104 -context
  :131.1192082698568814592 -context
  :131.1297951101136732160 -context
  :131.847521144173350912)) -is:retweet lang:en -is:
  nullcast -bluray -trailer
# Number of tweets for October 27th, 2022: 17.315
```

The count of tweets for May 27th in relation to streaming gave us a result of 15.451 tweets, which set the magnitude for the conversation about this topic. On the other hand, the count for October 27th in relation to the cinema experience gave a result of just 489 tweets. Thus number seemed quite wrong for us, so we decided to analyze it further, and therefore discovering how the conversation is quite different from the streaming one. In streaming there is always mention of the streaming service or the keywords .ºriginal or show or series"while in the movie going experience, the tweet might just even have the word "movie", but filtering tweets just by having that keyword would be highly misleading. Here context annotations are used to improve the results.

After adding the context annotations, the number of tweets for streaming went down to 10.859 which might have helped removing some spam tweets or mention of popular tv shows that are not available on streaming. On the other hand, the number of tweets discussing the cinema experience bump up to 17.315, we consider this to be above the expected number but closer to the expectations that the initial value. Nowadays, superhero movies and specially, Marvel movies are usually the ones capturing the attention of many viewers around the world, so having a low number of tweets in the day of one of their premiers did not seem quite right. This number was higher, but the removal of comic mentions, animation shows, and other non related film tweets gave this final number.

We believe this context (superhero films) is reliable since once again very few super hero movies are release for streaming and when they do get to streaming services the attention is not as high as when they released on cinema. This result also gives us an interesting thing to discuss. Maybe the was for consumer's attention is not even between streaming shows and theatrical releases, but between general entertainment vs superhero content? - This is not a new discussion. Many directors have given their opinions about marvel movies, star wars and so on and how franchises are just flooding the market with content while other movies cannot grab audiences attentions. Are they right? Are superhero movies saving the cinema industry or are they just reducing the competition?

Once again, this result for cinemas might be a bit skewed but even if they are not talking specifically about their experience in the cinema, they are talking about a pop culture topic that starts at the cinema.

We will keep the query this way since shows a more expected behavior than the previous 492 value. These two days chosen for our queries (streaming and cinema) were carefully consider as two big days for the two industries so such a huge difference in attention was clearly a concern. Let's visualize the behavior where we should expect more ups and downs (if movies are always above streaming we need to consider further improvements to the query).

The initial results for the cinema query was a bit higher (around 800 tweets). However, further analysis of some tweets showed a considerable amount of "spam"tweets promoting bluray to watch the movies. For this reason we added -bluray. The new set of tweets after removing tweets mentioning bluray seemed a success.

We are also removing tweets mentioning trailers since they are not always a reliable source to whether or not somebody indeed goes and watches and movie or a tv show. Also some big movies have a lot of attention for the trailers so this might skew the analysis.

Movie going experiences are also denoted by the additional services the movie chains provide such as drinks, snacks or premium screenings, for that reason they are included in the query. For example, this tweet is a clear indication of an experience that we are able to capture. «I rarely drink as it's like once in a blue moon, but earlier- I got myself this neat drink from the AMC bar before the One Piece movie started. I adore the juice cocktail as I completely drank all of it during the movie. C: Thanks to the barrista for the recommendation."Negi - @Negilarious»

These tweets show that our queries brought interesting and accurate tweets. These tweets represent the conversation and attention of consumers in respect to our two topics at hand.

Streaming: «At $30M per episode, I'm pretty sure S4 of #StrangerThings is the season finale of Netflix»

Streaming: «So I finally got the Netflix password and saw Love Death+ Robots....Yeah, this the final season isn't it?»

Cinema: «@WhyYaseen can't wait for the film! pretty sure i can remember going to the cinema to watch the first one that came out :)»

Cinema: «I'm at the movie theater about to watch Black Adam, this is my first time in a movie theater in literal months so I'm super hyped! #BlackAdam #BlackAdamMovie»

This last cinema tweet, brought up an interesting entity "Superhero films"which we did not consider at first. The content of the tweet and pop culture in general has been drawn to the spectacle of superhero movies, they are indeed bringing a lot of fans to cinemas. Therefore, we included this to the query to balance the magnitude of the conversation.

We checked the context of these tweets and improved our search so that we find tweets that don't explicitly mention a movie theater but just talk about a movie that was released in theaters.

The gathered data is the count of tweets for the past 15 years to help us answer our project questions. The counts endpoint paginates at 31 days per response. So we calculated to number of pages needed to access data from multiple years.

Is important to remark that the API does not always behave in the same way, sometimes some days are missing, but given the nature of the project and the analysis over multiple months and years we believe this is not an issue for us. Although some attention needs to be paid to this.

3. Data Analysis

We proceed to answer our project questions based on the data captured and visualizations produced. For time series analysis and forecasting we use Darts [5], a python library for easy manipulation and forecasting of time series. This library was really helpful to simplify the process of loading and visualizing time series data, which otherwise would require more work to train with sci-kit learn functions and more effort to visualize with matplotlib. This library was discovered during the development of the data science lab 3 and we decided to use it here given it versatility and easy-to-use interfaces.

3.1. Project Question 1

Is the movie going experience still relevant in today's pop culture compared to the attention that streaming services are facing?

Yes, the movie going experience is still relevant in today's pop culture (using a count of this year's tweets). We can even see how in some periods the attention around the movies is huge compare to streaming. Although some days streaming captures more attention. This might be skewed with the inclusion of superhero movies, but that is not a problem, this is an answer. We believe some of the spikes in the graph related to movie going experience might be the premieres of some of the biggest movies of the year (superhero movies?).

One assumption about these results might be that for a vast amount of shows the audience is divided, and therefore they do not generate as much social media attention as we would expect. Additionally, for many tv shows, audiences simply start them and watch them at different times while movies are usually very big events were most of the discussion is concentrated in their opening weekends on cinemas.

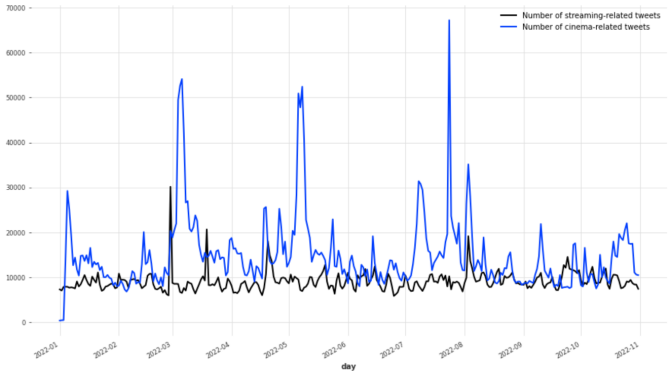

Figura 1: **Social media attention around streaming and cinemas - 2022**

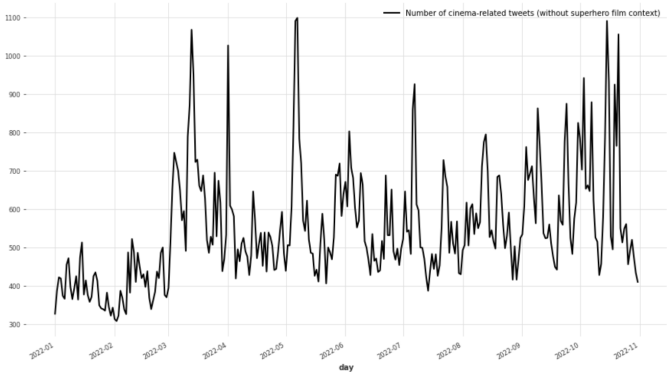

Figura 2: **Attention around cinema without using superhero film context - 2022**

However, coming back to our discussion about superhero films, we cannot deny that our inclusion of the superhero film context gave a huge boost to the cinema numbers, at some extent we believe this might be skewed by non-relevant tweets, but after careful consideration, we determine this might partially explain today's pop culture status.

Is worth nothing that when preparing the initial query to capture this data, without the explicit inclusion of the superhero film context, the results were very low compared to streaming. Is this a sign that people now just consider as worthy going to the cinema, the big superhero movies? Are now other movies being left out to watch at home?

To give a more accurate answer to our first question, we adventure to say, that big blockbuster movies such as superhero films are not only relevant in today's pop culture, but they drive the conversation. On the other hand, today's cinema without the boost of these movies seem to struggle to compete with today's competitive landscape where streaming provides audiences with high quality 8-10 hours shows that can clearly compete with the latest movies available.

[5] https://unit8co.github.io/darts/index.html

This behavior is also quite interesting, while we might have expected a çlear"winner between streaming and cinema, the truth is that they both have highs and lows, completely understandable and probably related to the release of hot shows and movies.

3.2. Project Question 2

What is the evolution of consumers' behaviour, denoted by number of tweets, trending topics, hashtags, social interactions, in terms of theatrical experiences vs streaming services?

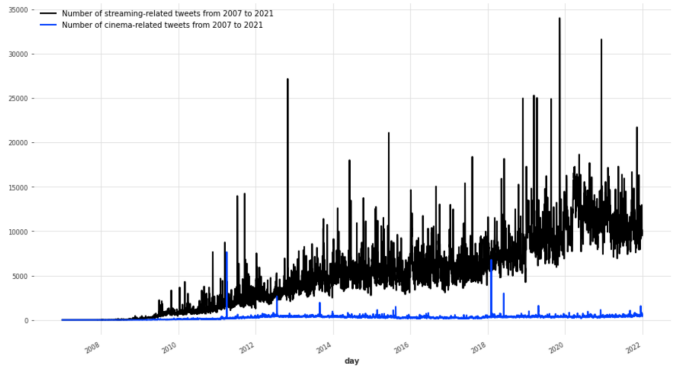

Figura 3: **Social media attention around streaming and cinemas - 2007 to 2021**

We get 15 years worth of tweets. From 2007 to 2021. The Full-Archive Search API provides complete and instant access to the full corpus of Twitter data dating all the way back to the first Tweet in March 2006.

The initial search to answer this question included the context annotations previously discussed. However, we got a lot of missing values. It seems that if a tweet was not annotated by Twitter, which was the case for all the tweets before the feature was introduced, then these tweets were not included in our results.

As a consequence of this, this search is done with our initial queries which did not included context annotations.

The streaming query without context annotation is still relevant. Given that we used the context to tune our results, but initial results seemed relevant to our analysis.

On the other hand, our cinema results without the context annotation was clearly impacted showing a similar scale that the analysis made for our first project question. We believe that we are either missing a lot of tweets that do not fit our keywords or this is truly the situation for the movie going experience.

We see how streaming has been growing over the past few years, which could be related to the amount of content that there is now available and also the creation and establishment of new streaming services. Finally, let's not forget how the recent pandemic made theaters closed and boosted the streaming services subscriptions.

So what is the evolution in consumers behavior? Streaming attention has grown, denoted by some clear factors, while the movie going experience has pretty much stay the same. We believe there is some reason to this, since the theatrical experiences have not changed that much apart from new formats available, while the streaming experience has seen major changes. For example, better entertainment systems at home, better loading times, better recommendations algorithms, and so on. If the tendency remains we could argue that cinemas are safe, there would always be people going to the cinema, but the problem is that the more streaming grows, the less time people might have to go to the theater, might this tendency change in the future? Let's see.

3.3. Project Question 3

Is it possible to predict this behaviour for the next few years?

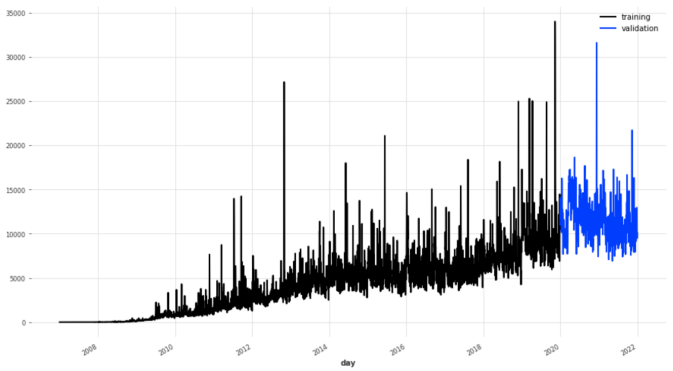

Figura 4: **Spliting timeseries into training and testing with Darts**

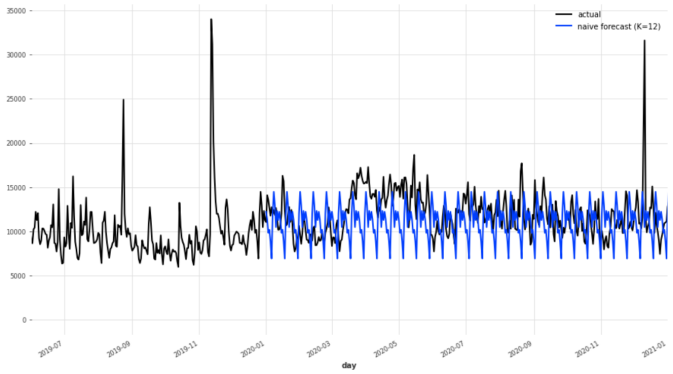

Figura 5: **Forecasting of streaming attention - View 1**

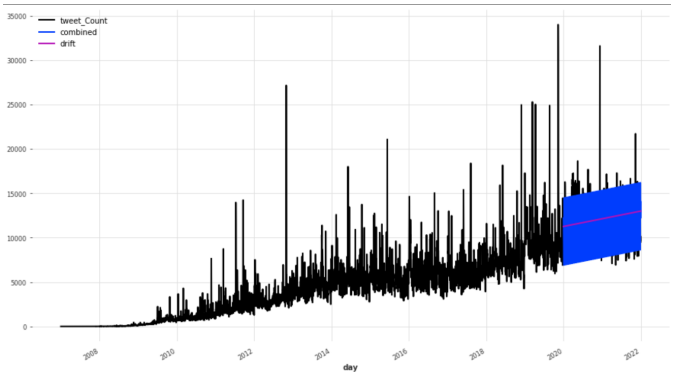

Figura 6: **Forecasting of streaming attention - View 2**

Yes, it is possible to predict this behavior, although more time was certainly needed to come up with a good prediction model for our time series and perhaps to improve our queries and get a better sense of the two topics. Also, forecasting of time series analysis required further knowledge of seasonality and other concepts that do not necessarily apply to common predictions based on regressions and classifications.

We divided our test set into training and testing using Darts and proceeded to predict the behavior around the streaming discussions. We feel that predicting cinema behavior given our current results would not be correct, we instead focus on streaming tendency to predict since it shows a clearer tendency. Perhaps the grow in streaming attention is driven by new actors appearing, and the explosion of content, while the movie going experience has been actually reduced to the success of a number of high level movies.

This analysis can clearly be improved. However, **Figure 6** shows a tendency of streaming growing over the next few years.

4. Conclusion

For this project, we spent a significant amount of time in the building process of our queries, in order to get the most accurate results and represent reliable time series analysis.

This effort was successful at certain extent but the limitations provided by the API limited some of actions that we could have taken to improve our data.

However, with all things consider we believe the analysis was still interesting and it gave us some highlights. For instance, streaming is certainly growing and while cinema is still part of today's pop culture, we believe it does not have that place reserved for the coming years.

Our initial explorations showed an interesting division between positive tweets or negative reactions about a show/-movie, this is not the focus of our project but is nonetheless interesting and might represent future work.

The cinema industry might need to pay attention to these results, social media attention might not be going down, but is also not going up. While streaming is growing, at some point audiences might not have enough time for both. Which one will the choose?