Algorithmic Analysis of YouTube Music Comments: measurement and applications

"This makes me cry but not out of sadness, just out of beauty"
(@chimbertocaviglia6919, on Thom Yorke's Bloom, 2021-11-06)

Stephane Gauvin (Laval University)[1]
stephane.gauvin@fsa.ulaval.ca
May 2024

Abstract

We develop and apply Transformer models to predict comments' sentiment on a large corpus of comments left on YouTube music videos. The model achieves an accuracy of 0.98, closely mirroring human consensus.

We then generate inferences on nearly 700 million English-language comments left on Official Artist Channels. Results show that, counterintuitively, the most popular artists receive below-average sentiment. This finding holds for superstars with more than 10 billion views and is robust across different operationalizations of sentiment.

We also find that most commenters have posted a single comment, and that prolific commenters tend to be less positive on average. The lower average sentiment towards superstars is consistent across commenters of all levels of activity.

Keywords: Transformers, sentiment analysis, music, comments

Introduction

The emergence of advanced information technologies is transforming the way we understand consumers by making available a vast corpus expressing opinions, beliefs and attitudes (Gunter et al. 2014). In parallel, new and powerful techniques can be used to automate processing, making it possible to analyze vast amounts of data (Rothman & Gulli, 2022), and recent advances have considerably increased the accuracy of algorithmic analyses.

In the early days of sentiment analysis, research focused on investor sentiment and reported a modest accuracy[2] of 0.6 (Das et al., 2001). Accuracy improved over time as Natural Language Processing (NLP) techniques evolved (Hutto and Gilbert's, 2014). Shukla et al. (2017) report accuracy in the neighborhood

[1] Full professor at Faculté des Sciences de l'Administration, Université Laval, Québec CANADA. Current research interests focus on marketing applications of language models.

[2] Accuracy if the proportion of correct predictions, i.e. true positives and true negatives. Precision is the proportion of true positives. The F1-score is the harmonic mean of precision and recall : F1 = 2 * (precision * recall) / (precision + recall), where recall is the proportion of actual values that have been correctly predicted. F1 varies between 0 (random) and 1 (perfectly accurate).

of 0.65 in their review of sentiment analysis in the field of music; Gómez & Cáceres (2017) report 0.80; Asif et al. (2019) report 0.75. Other recent papers have applied a variety of techniques on YouTube comments (Alhujaili and Yafooz, 2021; Ahuja et al., 2023; Deori et al., 2023) but they do not report accuracy or, when they do, report disappointingly low figures arguably because they rely on obsolete procedures.

A new generation of NLP models is now based on Transformers (Vaswani et al., 2017). A key feature of Transformer models is that they can be trained in parallel, making it possible to massively increase the size of the training corpus and model sophistication. In 2018, ELMo, a state-of-the-art recurrent neural network model, ran on 90 million parameters (Peters et al., 2018). The same year, Devlin et al. (2018) introduced Bidirectional Encoder Representations from Transformers (hereafter BERT) models having close to 340M parameters and in 2023 openAI launched GPT, built on 1.8T parameters (Raiaan et al., 2024). Albeit controversial (see Tedeschi et al., 2023), performance benchmarks such as the GLUE leaderboard report that Transformer models now routinely perform better than humans on a wide variety of tasks.

In one application of BERT models in the domain of sentiment analysis, Biswas et al. (2020) reported F1 scores of 0.91 and 0.78 for positive and negative entries in Stack Overflow.

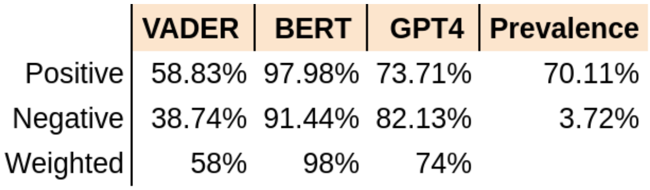
To put things in perspective, it could be argued that 5 years ago Hutto and Gilbert's VADER (2014) was still the most accurate tool for sentiment analysis. VADER's architecture was exquisitely simple: it worked on the basis of fewer than 10 000 lexical entities – including emoticons – whose valence had been empirically validated. This allowed for very fast processing (i.e. simply sum the valence of the lexical entities that are present in a text) with remarkable accuracy. In the "social media" category, VADER outperformed human agents (F1 of 0.96 vs 0.84 for humans) as well as all alternative procedures available at that time (F1 of 0.96 vs 0.66 baseline). While VADER systematically outperformed other algorithmic approaches, it was still no match for human judgment in many cases such as Amazon reviews (F1 of 0.63 vs 0.85), movie reviews (F1 of 0.61 vs 0.92) and New York Times editorials (F1 0.55 vs 0.65). (Hutto and Gilbert, 2014: 223)

These pre-Transformers benchmarks showed that while simple algorithms did outperform more convoluted machine-based computations, there was still a place for humans in most types of textual analysis when accurate classification mattered. This may no longer be the case.

At the risk of getting ahead of ourselves, consider Table 1 where we report the accuracy of various models predicting the sentiment of the validation sample used in this paper. That sample contained 744 comments, unanimously rated as positive, negative or neutral by 5 human raters. We can see that VADER scores a weighted accuracy of 58%, a value in line with those reported above.

Second, Transformer models, both in their BERT and GPT4 incarnations, clearly outperform VADER for all categories: if VADER was the best tool, this is no longer the case – Transformer models are vastly superior.

Table 1 : Predictive accuracy (F1-score)



	VADER	BERT	GPT4	Prevalence
Positive	58.83%	97.98%	73.71%	70.11%
Negative	38.74%	91.44%	82.13%	3.72%
Weighted	58%	98%	74%	

It is important to note that GPT4, a model with 1.8T parameters, is less accurate than BERT-Large, sporting fewer than 350M parameters. The reason is twofold. First, GPTs are trained over a vast corpus, across an almost infinite variety of contexts. These models have learned language patterns and developed an uncanny ability to predict what comes next. Something like "This cake is really very good […] is a positive comment". But if language patterns are atypical, GPTs may struggle. For instance, our 5 human raters unanimously deemed the following to be positive: "*Can someone tell me why did I just watched this 10 times and cry ???*". BERT-large flagged it as positive as well, just like human raters, but GPT4 opined that the comment could not be classified, despite being prompted to the fact that the comment was directed at a YouTube video. When the prompt made it clear that the comment was made in reference to a *music* video, GPT4 revised its opinion somewhat: "*The comment seems positive because the person has watched the video multiple times and had a strong emotional response to it. However, it's difficult to draw a concrete conclusion as crying can be associated with both positive and negative emotions.*" This justification is reasonable as one would expect "*I called customer service more than 10 times and cry"* to be negative. But as it happens, comments left on artist channels that included the word "cry" were vastly (27 times) more likely to be deemed positive by our human raters, something BERT was able to learn during its training, something to which GPT is oblivious.

Which brings us to the second, related, consideration: BERT models *require* training whereas GPTs *may* receive additional training on a specific corpus but are usually not, for a variety of reasons beyond the scope of this paper. Which brings us to the next section where we report on the training of our BERT models.

Data and model

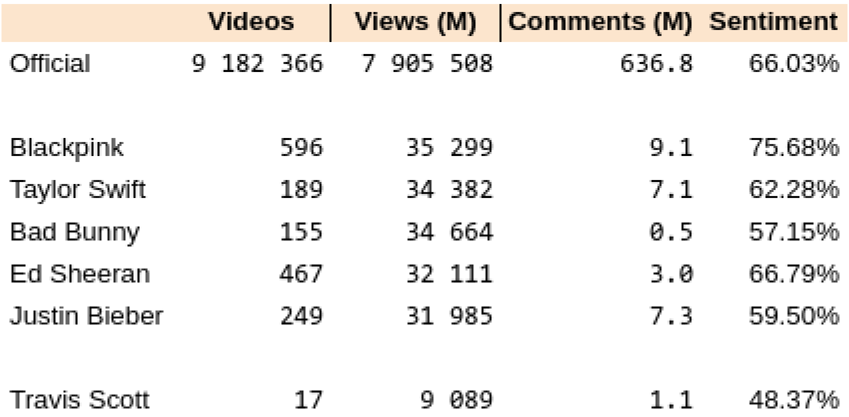
We focus on comments left on YouTube for two major reasons. First, the very nature of comments related to musical performances differs from what consumers write about most products. Music is a more engrossing experience. The comment quoted on the title page is one such example. The commenters are also using a different type of vocabulary (ex: Adorrrr😭❤️) that poses an interesting challenge to models trained on a more formal corpus. Second, YouTube's massive user base and first class infrastructure makes it relatively easier to gather vast amounts of data.

Data

At the time of this writing (April 2024), we are aware of the existence of more than 850M accessible music videos. From this corpus, we have identified 9.2M videos published in Official Artist Channels

(OAC[3]). These videos have been the target of 1.4B comments, of which close to 637M are expressed in English.

Table 2 : Music videos population



	Videos	Views (M)	Comments (M)	Sentiment
Official	9 182 366	7 905 508	636.8	66.03%
Blackpink	596	35 299	9.1	75.68%
Taylor Swift	189	34 382	7.1	62.28%
Bad Bunny	155	34 664	0.5	57.15%
Ed Sheeran	467	32 111	3.0	66.79%
Justin Bieber	249	31 985	7.3	59.50%
Travis Scott	17	9 089	1.1	48.37%

Official videos are from the artist's channel
Blackpink, Swift, Bad Bunny, Sheeran and Bieber are the artists with the most in-channel views
Scott is used to assess validity

Sentiment is the difference between the probability that a comment is positive vs negative. It takes values between -1 and 1. The population average of 66.03% is calculated at the comment level (637M). The average sentiment aggregated at the artist level (84 864 distinct artists) is 75.09%. This difference is due to the fact that artists with a large number of comments tend to have lower average sentiment. Averaging across artists therefore results in a higher value.

Table 2 briefly describes this population, including details for the top five artists as per the cumulative number of views generated by their official channels at the time of this writing: Blackpink, Taylor Swift, Bad Bunny, Ed Sheeran and Justin Bieber. We also report data for Travis Scott, used to assess the validity of the sentiment model

Model

The Sentiment model has been trained on 51 000 comments labeled by 5 research assistants (graduate students). Each assistant was tasked with the labeling of 11 000 comments, 1 000 among them being common to all raters. They were instructed to determine if a comment was more likely to be positive, negative or indeterminate (i.e. merely factual, ambiguous, unintelligible, lyrics from the song, etc.)

As we have seen in Table 1, a small proportion of the comments in the population is negative, so small in fact that in using a random sample, BERT might not have been exposed to enough negative comments in order to learn how to effectively recognize them. As we became aware of the massive imbalance in

[3] The OAC status is conferred by YouTube, upon request, if the channel meets criteria listed in https://support.google.com/youtube/answer/7336634?hl=en#zippy=%2Cprogram-criteria-and-eligibility

sentiment distribution at the model development stage, we ran VADER on a sample of 1M comments to generate a more evenly distributed mix of positive/negative comments. As mentioned earlier, VADER was quite adept at identifying positive comments, but barely better than a coin toss when it came to a negative comment, such that the final training corpus of comments ended up being labeled as 53%, 27% and 20% for positive, neutral and negative respectively by our human raters. The inter-rater agreement rate was 0.85 with a Cronbach Alpha of 0.96.

Armed with these data, we have trained two BERT-Large models following the procedure outlined in the Tensorflow classification tutorial[4]. We opted in favor of training two binary classifiers instead of a multinomial model due to the imbalance in the class distribution. Therefore, sub-model 1 determines whether a comment is positive or not (i.e. negative or neutral), and sub-model 2 determines whether a comment is negative or not.

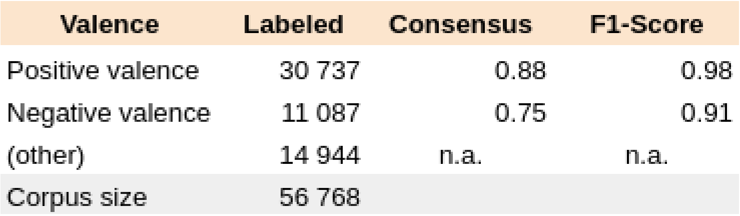
Models were trained on 50 000 labeled comments, using 80% of the sample for training and 20% for validation. Models were never exposed to the 1 000 comments labeled by all raters – this holdout sample was solely used to compute the models' F1 scores.

The weighted F1-score of BERT inferences on the holdout sample (744 comments where there was consensus across 5 human raters) is 0.98 (0.98 for positive sentiment and 0.91 for negative sentiment).

Reliability

In the social sciences, instruments are understood to be fallible and reliability is therefore assessed at the scale level, where several instruments are correlated. A scale will be deemed reliable when instruments yield closely related predictions and we usually summarize these with Cronbach's Alpha (Cronbach and Shavelson, 2004). In algorithmic analysis, the existence of an error-free corpus is implicit, such that the F1-score is the generally agreed-upon metric used to evaluate a model's reliability. This is controversial because this assumption the F1-score depends on the signal used to assess the value that is predicted (see Kofler et al., 2023). In our case the high degree of consensus between human raters suggests that the F1-score is actually a reliable indicator. (see Table 3)

Table 3 - Models accuracy



Valence	Labeled	Consensus	F1-Score
Positive valence	30 737	0.88	0.98
Negative valence	11 087	0.75	0.91
(other)	14 944	n.a.	n.a.
Corpus size	56 768		

Labeled is the number of comments used to estimate the model, split 80/20 between train/test
Consensus is the proportion of comments where all human raters agree on the label
F1-score = 2 x (precision x recall) / (precision + recall). F1 is reported for consensual comments

[4] The tutorial can be found at https://www.tensorflow.org/text/tutorials/classify_text_with_bert. The BERT-Large library is available here: https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4. Our code implementation is available upon request.

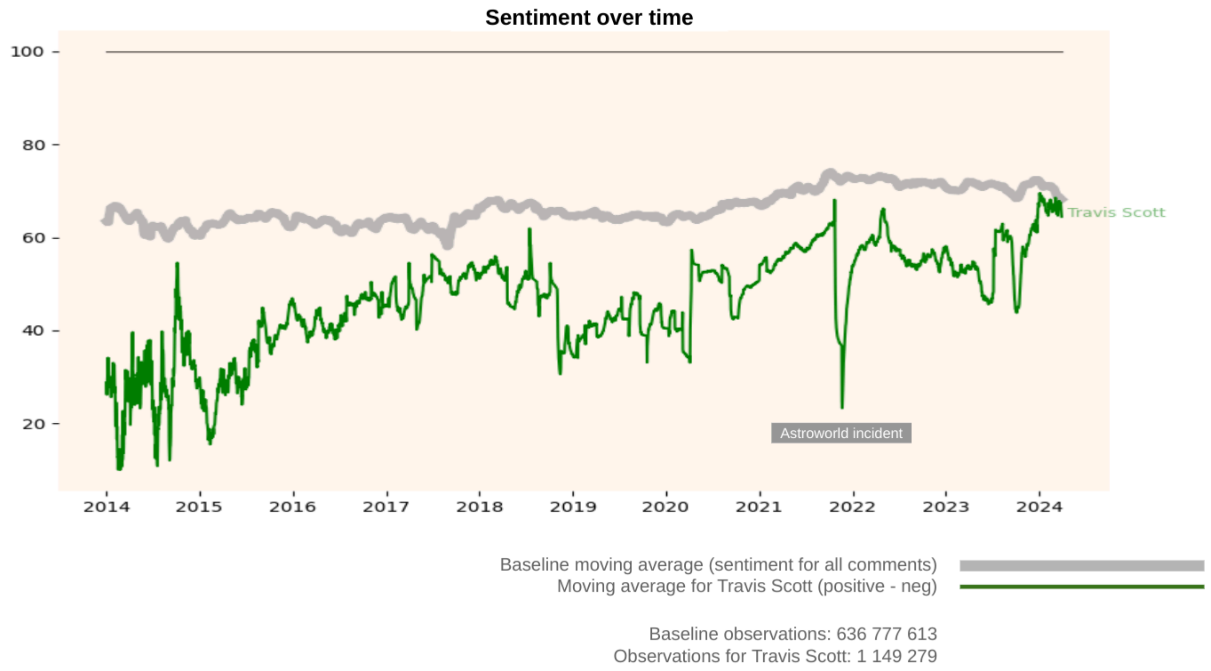
Validity

In order to assess validity, we have computed a 31-day centered moving average of the sentiment of comments left on Travis Scott's Official Artist Channel (OAC). Scott is an artist that has been at the center of several controversies, in particular one where he was said to have been slow to react to a crowd rush during his show at the Astroworld music festival, on November 5th, 2021, in which several festival goers lost their lives.

Figure 1 displays this time series dating back to the beginning of 2014. We see that it reacts on cue and with considerable magnitude to the Astroworld crowd rush tragedy. We can therefore reasonably conclude that the predicted sentiment is a valid measure of the actual sentiment towards a target.

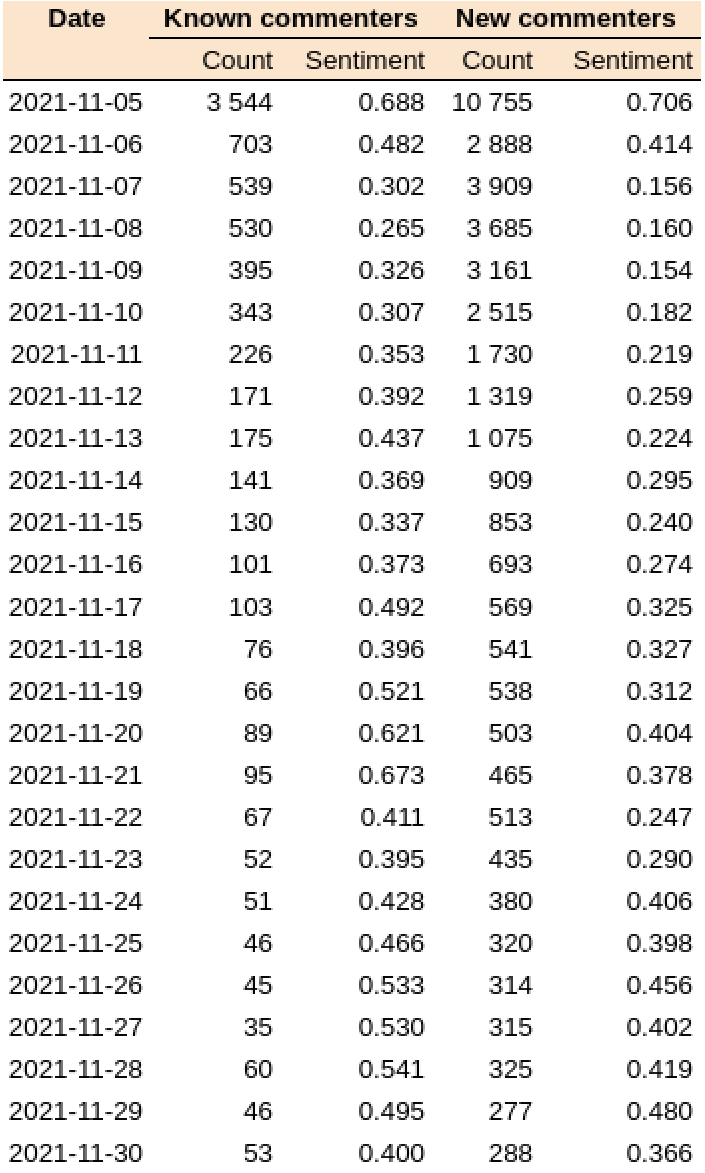
We can shed more light on the Astroworld event by focusing on the commenters' identity, in order to dispose of the argument that the drop was an artifact due to the prevalence of bots targeting Scott for a variety of reasons, and therefore may not be a reflection of the actual sentiment of human commenters. Table 4 shows how past and new commenters' sentiments have evolved at the time of the incident. Prior to the Astroworld incident, 184 000 distinct commenters had left at least one comment on Scott's OAC. We refer to these commenters as 'known commenters'. Close to 3 500 of them have left a comment on the day of the incident with a sentiment averaging 0.69 vs the 0.47 all-time average sentiment towards Scott, prior to the incident. The uptick in sentiment, clearly visible in Figure 1, was likely due to the anticipation of seeing Scott perform that night, at the festival. The sentiment expressed by 'known commenters' is almost identical to the 0.71 measured amongst the 'new commenters' who were leaving their first comment(s) on Scott's OAC, on that day.

Figure 1 - Sentiment timeline for Travis Scott



Once fans, known and new, became aware of the crowd rush, sentiment fell markedly, especially among new fans. Known fans' sentiment rebounded to pre-Astroworld levels within 2 weeks before fading to values slightly below the historical average. Sentiment amongst new fans, some of them presumably no fan at all, cratered at 0.15 a few days after the event, and remained consistently below the average sentiment observed among 'known commenters'.

Table 4. Sentiment Towards Travis Scott 2021-11-05 / 2021-11-30



Date	Known commenters		New commenters	
	Count	Sentiment	Count	Sentiment
2021-11-05	3 544	0.688	10 755	0.706
2021-11-06	703	0.482	2 888	0.414
2021-11-07	539	0.302	3 909	0.156
2021-11-08	530	0.265	3 685	0.160
2021-11-09	395	0.326	3 161	0.154
2021-11-10	343	0.307	2 515	0.182
2021-11-11	226	0.353	1 730	0.219
2021-11-12	171	0.392	1 319	0.259
2021-11-13	175	0.437	1 075	0.224
2021-11-14	141	0.369	909	0.295
2021-11-15	130	0.337	853	0.240
2021-11-16	101	0.373	693	0.274
2021-11-17	103	0.492	569	0.325
2021-11-18	76	0.396	541	0.327
2021-11-19	66	0.521	538	0.312
2021-11-20	89	0.621	503	0.404
2021-11-21	95	0.673	465	0.378
2021-11-22	67	0.411	513	0.247
2021-11-23	52	0.395	435	0.290
2021-11-24	51	0.428	380	0.406
2021-11-25	46	0.466	320	0.398
2021-11-26	45	0.533	314	0.456
2021-11-27	35	0.530	315	0.402
2021-11-28	60	0.541	325	0.419
2021-11-29	46	0.495	277	0.480
2021-11-30	53	0.400	288	0.366

There are no obvious signs of bot activity, such as an inordinate number of comments made by a single commenter (more about this later) or a sharp increase in sentiment polarization (we wouldn't expect bots to make nuanced comments).

Overall, we feel that we can make a strong case about the prima facie validity of the algorithmic sentiment measure. It correlates strongly with humans' interpretation and the analysis of the Travis Scott sentiment time series behaves as one would expect, on time, in the right direction and with force.

In the next sections we illustrate how this tool can be used to gain insights in the rather complex domain of affect towards artists.

Applications

Corpus

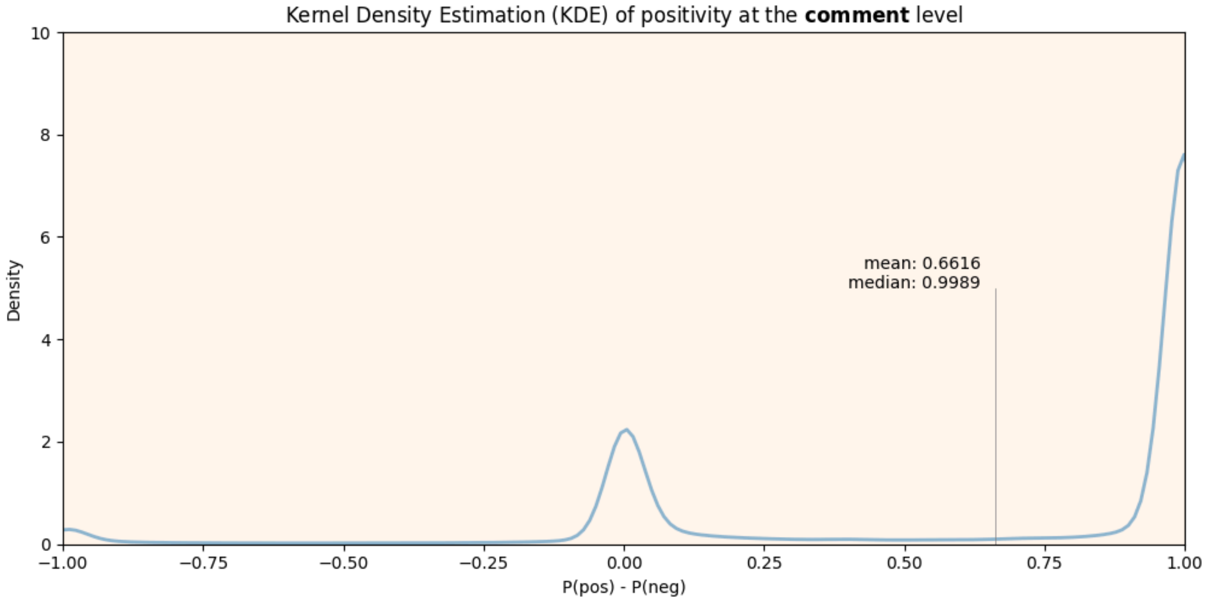
As indicated in the previous section, we have harvested more than 600M comments written in English, left on Official Artist Channels (OAC). Comments can be positive, negative or neutral (factual, lyrics, ambiguous, nonsensical, etc) such that we have trained two BERT-Large models, one to assess the probability that a comment is positive, and another to assess the probability that the comment is negative. Our analyses have been updated with comments published prior to April 30th, 2024.

Two things must be kept in mind. First, inferences indicate the *probability* that a comment is either positive or negative and have little if anything to do with the comment's *intensity*. The comment with the highest positive inference (0.99999493) reads “Nice beautiful and lovely.cute”, arguably weaker than “This makes me cry but not out of sadness, just out of beauty”, left on Thom Yorke's Bloom. While the latter is also inferred with a high probability of being positive (0.9999675) more that 140M comments show higher probability of being positive, including mild praise such as “Refreshing. I like this a lot!”. Recall that chatGPT 4 didn't call a comment as being positive because it included the word ‘cry’, often associated with negative emotions. While BERT was trained to mitigate this negative association in the context of music videos, one may argue that the presence of the word “cry” in the aforementioned comment still has a residual impact.

Second, a low probability of being positive provides an ambiguous signal. It may reflect a high probability of being negative, but it may also be that the comment was neutral such that the model predicted low probabilities for both positive and negative valence. We therefore compute a positivity score by taking the difference between the probability of the comment being positive and the probability of it being negative, which removes the effect of neutral comments. The positivity score can take any value in the -1 .. +1 interval.

Figure 2 shows the distribution (kernel density estimate (KDE)) of these scores. The average stands at 0.66 and the median is surprisingly high (greater than 0.99), signaling that (1) most comments are positive, and (2) BERT is highly confident in its inferences. The KDE shows another peak close to 0.0, due to neutral comments, and a smaller one at -1.0, i.e. where the comment is inferred to be negative.

Figure 2. Kernel Density of the Distribution of Positivity Across All Comments



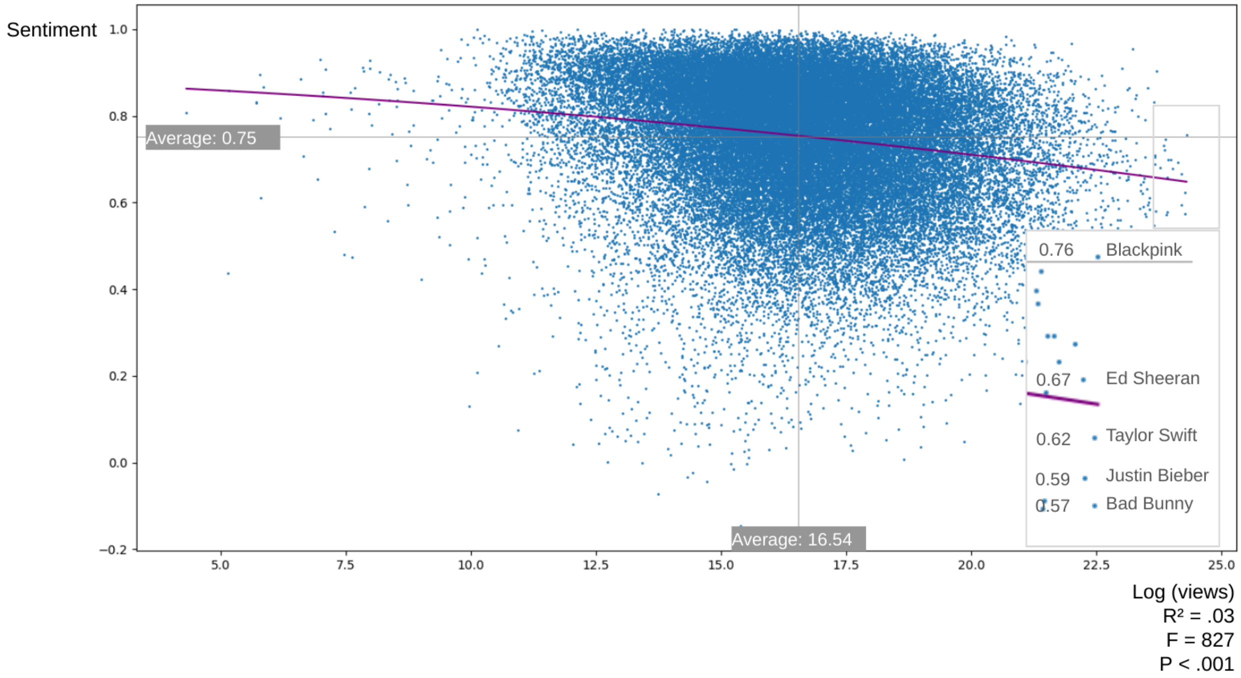
Positivity = P(pos) - P(neg), where P(pos|neg) is the probability that a comment is positive|negative

Artists

Having scored individual comments, we are now in a position to aggregate them at the OAC level. We did compute an average sentiment score for every artist having received at least 100 comments, which gives us a pooled sentiment metric for a total of 53 592 distinct artists. The average sentiment towards an artist is 0.75 (median of 0.78), with the distribution of values close to a slightly skewed normal. The positive shift in mean (the average mean value at the artist level is higher than the mean of individual comments) is indicative of what will become clearer in a moment – lesser artists, targeted by fewer comments, generally receive a higher sentiment score.

Figure 3 plots artists over the sentiment x log(views) space. Whereas one would think that there should be a positive relationship between sentiment and views – after all, we watch what we like, don't we – the data says otherwise. The more conservative interpretation would be that there is no meaningful relationship between these two metrics, however counterintuitive it may seem. A statistical test confirms a small but significant negative relationship ($R^2 = .03$; $F = 827$; P-value $< .001$) if we fit the relationship with a quadratic model (i.e. curvilinear).

Figure 3. Distribution of OAC Average Sentiment x Log(views)



Each dot marks an artist in the sentiment x logs(views) space. Average sentiment (0.75) and log(views) (16.54) are marked by gray lines. A rectangle, top right, focused on the region of artists with the most views, is expanded below. The regression line (purple) shows predicted sentiment as a function of log(views), based on a quadratic model.

Blackpink is the only band with above average sentiment. Ed Sheeran is below average, but above expectation. Taylor Swift, Justin Bieber and Bad Bunny are below expectations.

On closer examination, focused on the top-five artists identified above, we see that Blackpink is barely above average sentiment while the four other leading artists are clearly below average, as we already saw in Table 2. The scatterplot merely puts these results into perspective and suggests that this is systematic rather than anecdotal. If we were to consider only artists with more than 10 billion views – we count 69 such superstars – we would find an average sentiment score of 0.66, considerably lower than the 0.75 observed across all artists. The superstar with the highest average sentiment is Gaín Guzmán (an impressive 0.95 with more than 12 billion views; but derived from a surprisingly small number of English comments: 337). At the other end of the superstar spectrum we find Drake, with a sentiment of 0.41, more than 17 billion views and more than 450 000 English comments.

This naturally begs the question of ‘why’? Why is it that the most popular artists are not also those enjoying the most positive comments? This is what we explore in the next section.

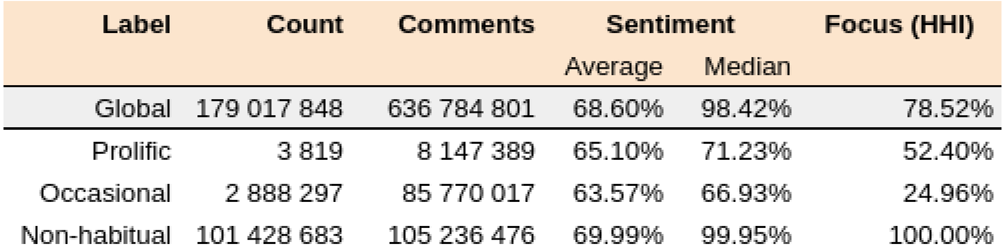
Commenters

We have arbitrarily divided the commenters population into 3 disjoint groups: (1) Non-habitual commenters are those who have left a single comment on any artist’s OAC. This is a large group representing more than half of all commenters. (2) Occasional commenters are defined as those who have left between 20 and 40 comments in the music space. (3) Prolific commenters have left more than 1 000 comments.

We report several indicators for each group on Table 5: (1) the count; (2) the total number of comments posted by members of the group, the (3) average and (4) median sentiment of these comments, and (5) the Hirshman-Herfindhal Index (HHI) of concentration, an indicator ranging from close to zero, when comments are spread evenly across targets, to a maximum of 1 indicating that all comments left by a commenter were focused on a single artist. The HHI is trivially 1 for non-habitual commenters but, interestingly, lower for occasionals than for prolific commenters who tend to be relatively more focused.

If we consider average sentiment, we find that all three groups have fairly similar average valence but that Non-habitual commenters make slightly more positive comments.

Table 5 - Commenters descriptives



Label	Count	Comments	Sentiment		Focus (HHI)
			Average	Median	
Global	179 017 848	636 784 801	68.60%	98.42%	78.52%
Prolific	3 819	8 147 389	65.10%	71.23%	52.40%
Occasional	2 888 297	85 770 017	63.57%	66.93%	24.96%
Non-habitual	101 428 683	105 236 476	69.99%	99.95%	100.00%

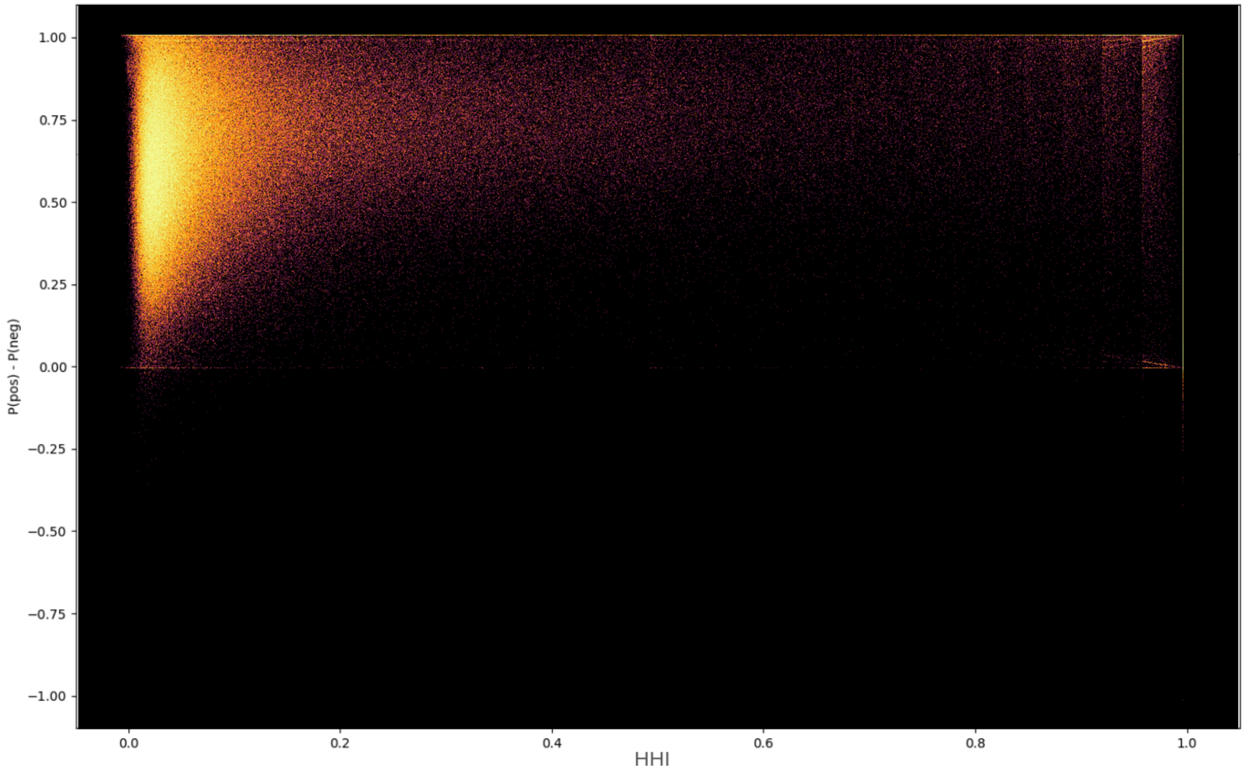
Prolific commenters have left more than 1000 comments
Occasional have left between 20 and 40 comments
Non-habitual have left one comment.
Focus is the HHI statistic (sum of squared shares). By definition, non-habitual commenters show HHI of 1.

Figures 4a and 4b plot the distribution of commenters' sentiment across focus, for commenters having left more than 50 comments. We have removed commenters with fewer comments for clarity – commenters with low comment counts tend to be trivially focused, which hides an important characteristic of commenters behavior as we will see in a moment.

Figure 4a is a shaded plot, better at rendering the overall distribution of data points but difficult to appreciate on a static medium because we cannot zoom into apparently empty regions; Figure 4b is a conventional scatter plot that suffers from overplotting – in dense regions dots coalesce into a uniform surface, failing to properly render differences between medium and high density regions as we can readily see on Figure 4a, but nonetheless useful on a static medium like this research paper because it allows us to see what happens in low-density regions.

Figure 4a shows an interesting pattern that looks like an hourglass lying on the side, where the distribution of sentiment is more or less normal for less focused commenters and more polarized in the case of highly focused commenters – highly focused commenters show a sentiment distribution that is far from normal. Instead, we see a concentration of highly positive, followed by a long tail of less positive commenters. And if we look at the bottom right of the scatter plot on Figure 4b, we find a few highly focused, very negative and sometimes highly prolific commenters, such as Ug.

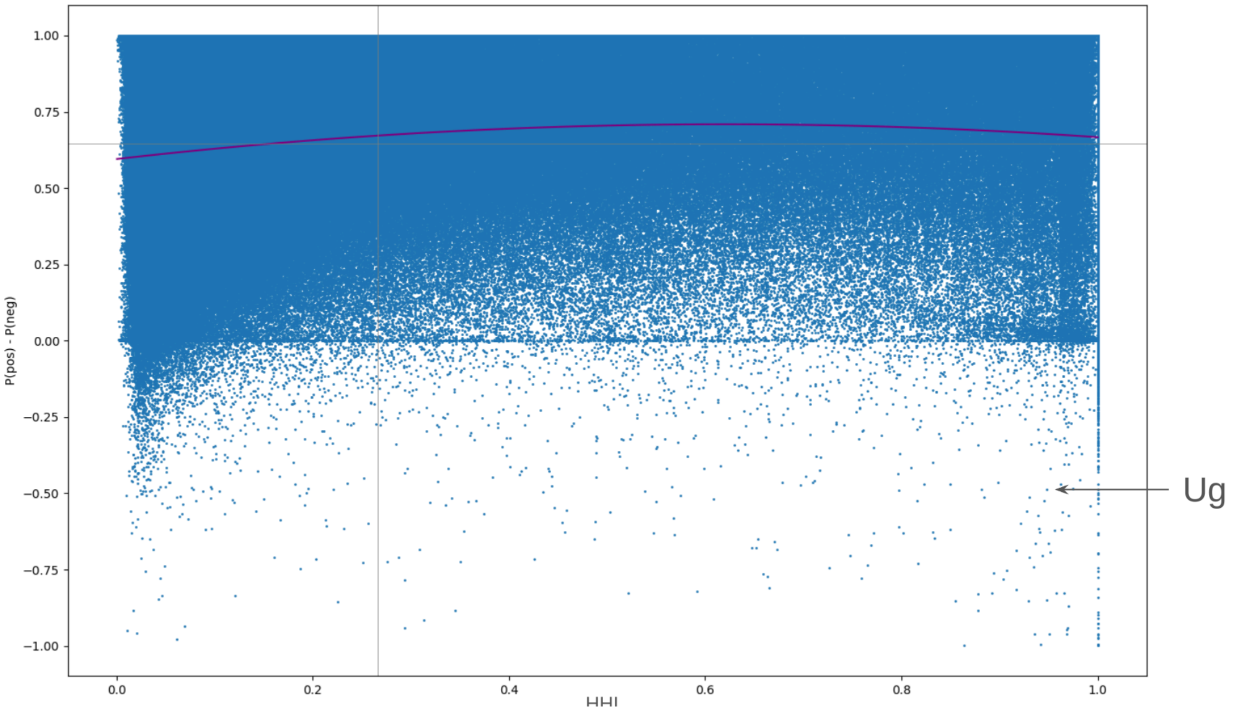
Figure 4a. Distribution of Commenters Average Sentiment x HHI (Datashader)



According to our data, Ug has published close to 13 000 comments, including more than 12 500 targeting Taylor Swift, between December of 2009 and July of 2014. In addition to Swift, Ug has also published close to 300 comments targeting Selena Gomez and 80 targeting a dozen of other artists. Ug usually left negative comments, averaging -0.46 with the exception of a few artists where he left positive reviews (ex: Justin Bieber, Psy). Ug's activity has been enough to lower the average sentiment towards Swift by a full percentage point even though she had received more than 1.2M comments during that time period.

While Ug is an extreme case, Figure 4b shows a fairly large number of highly focused and negative commenters. In the aggregate their weight is enough to suggest that they drive down average sentiment towards superstars.

Figure 5b. Distribution of Commenters Average Sentiment x HHI (Scatter plot)



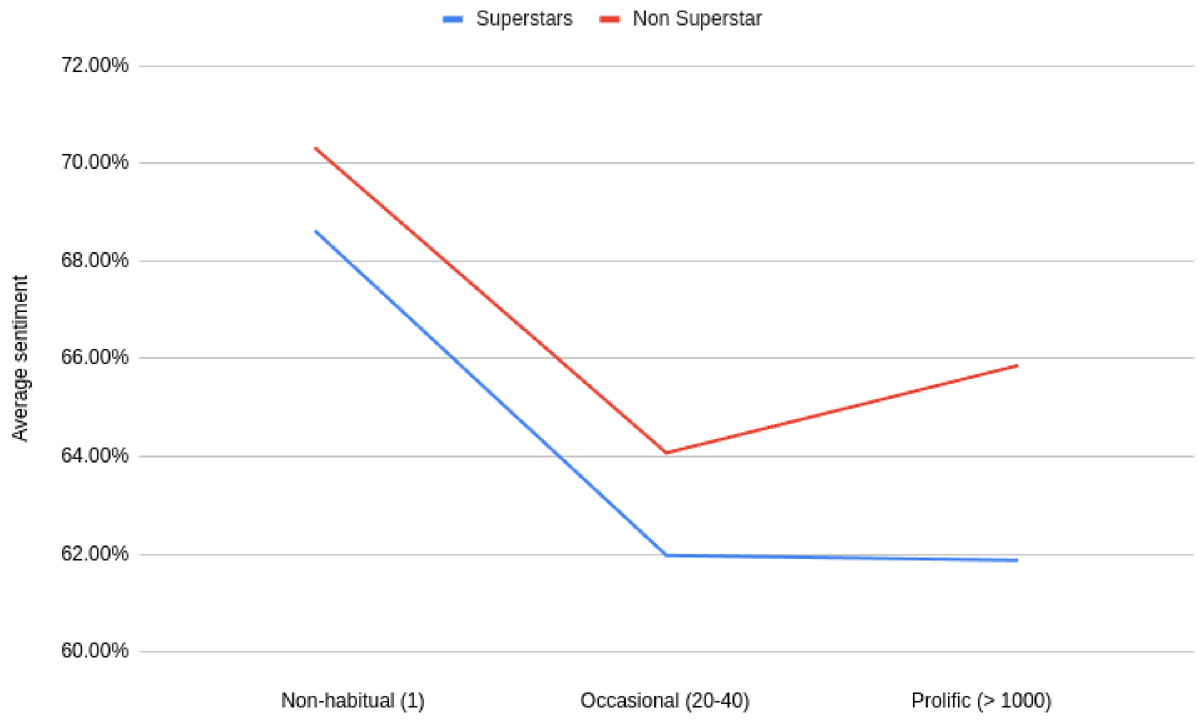
This scatterplot allows us to locate commenters in low density regions of the plot
Commenter 'Ug' stands out as a very prolific commenter with more than 12 500 comments, mostly negative and targeted at Taylor Swift

The overall pattern is presented in Figure 5. where we display the average comment score, by commenter (Non-habitual, Occasional or Prolific) and artist (Superstar or Non Superstar) profile. We see that the sentiment curve of Superstars is lower across all types of commenters, a pattern consistent with the Mona Lisa curse (i.e. high expectations leading to disappointment).

Evidence therefore suggests that the main driver for the lower sentiment towards the most popular artists is largely a consequence of a general, diffuse, increase in negative comments. In addition, a few focused prolific commenters such as UG do occasionally target superstars. It may be worth noting that this plague seems to have had a greater impact on Taylor Swift, where 4 prolific commenters posted mostly negative comments, vs none for Ed Sheeran, Justin Bieber and Bad Bunny. The case of Blackpink is in a category by itself. Whereas the four previously mentioned artists have been targeted by a handful of prolific commenters, we count no less than 540 Blackpink-prolifics where the vast majority are very positive. There are, however, four prolific negative commenters.

These patterns will be investigated in more detail in a forthcoming paper.

Figure 5. Sentiment by Type of Commenter and Artist



Superstars' OAC have more than 10 billions views
Commenters are split in three groups of different prolificity.
Average **sentiment** towards Superstars is lower across all types of commenters

Discussion

We briefly discuss takeaways and directions we plan to take in future research.

With respect to methodology; (1) We have confirmed the superiority of Transformer models in predicting sentiment, thus replicating Biswas et al. (2020). (2) We achieved superior accuracy, due to our training procedure where we relied on several human raters to label training items, thus reducing the risk of training bias, and we have strived to provide a balanced training corpus, by using alternative models to increase the prevalence of uncommon categories in the training sample. We believe that the accuracy levels we achieved are the highest in the extant literature on sentiment analysis at the time of this writing. (3) We have shown that a properly trained parsimonious model (BERT-Large) performs better despite being orders of magnitude smaller than Large Language Models.

This being said, rapid progress in the development of LLMs raises questions about the importance of context-specific training, a time-consuming and costly process. In this paper we noted that LLMs, being trained on a more universal corpus, may struggle in the context of specific applications, such as inferring the sentiment of a comment left on a music video, where the word "cry" is present. Out of curiosity, the day before filing this paper, we've submitted the following prompt to chatGPT 4o, Claude's Opus, Meta.ai and Google's Gemini 1.5. :

> *Below is a comment left on a music video. Please tell me if the comment is positive, negative or neutral/indeterminate.*
>
> *"Can someone tell me why did I just watched this 10 times and cry ???"*

This time, all four LLMs returned an unambiguous positive signal, similar to what our human raters felt, similar to what the trained BERT model infers, but unlike the chatGPT 4 indeterminate call of just a few weeks earlier. We plan to rerun our entire validation sample on these four leading LLMs to assess the current state of the art in generative NLP models and derive the value of training. If accuracy is no longer a consideration, latency and cost might still be critical decision factors. In our case, the local model took several months to generate close to 6 billion inferences used in our work. We had run a quick informal experiment on chatGPT 3.5 and estimated that an API-based inference process would have required more than a year and cost millions of dollars. Intensifying competition between solution providers might have changed the decision calculus.

With respect to the substantive issue of the nature of comments' sentiment in YouTube music space, we have made the surprising discovery that more popular artists do not receive more positive comments. This is not trivial as the literature has established a robust positive relationship between sentiment and sales, for all kinds of product categories, including movies, pre and post Internet. (De Maeyer, 2012; Eliashberg and Shugan, 1997; Wu and Dang 2018). Our findings suggest that while the positive relationship is generally true, it breaks down in the case of superstars, where high expectations lead to disappointment (Mona Lisa curse).

We have also shown that the reduction in positivity is robust across commenters' prolificity and that while there are intensely negative commenters in this space, their impact is generally minor for most artists. But this question would certainly benefit from more research.

It should be re-emphasized that sentiment analysis refers to the probability of a comment being positive or negative, rather than to the intensity of the comment's valence. As a consequence, interpreting this signal might be misleading. Although we did not report our investigation on the prevalence of emotions, suffice it to say that ongoing research has found the same pattern in the expression of *love*, that is less prevalent among comments left on superstars' channels.

Last, we believe that Transformer models will have a very significant impact on research interested in interpreting consumers' comments. At a superficial level it is obvious that a procedure that is at least as accurate as humans and orders of magnitude faster and economical does open new doors in the field.

Yet, a more fundamental shift is likely to occur. Figure 4a evokes a starry night where each dot, depending on what is our focus, stands for a single comment, an artist, or a commenter, somewhere in the space that we have defined. Figure 4b makes it clear that as precision increases, each dot can be made smaller and patterns that were impossible to discern with a blunt instrument progressively appear "in the sky". Accurate models, massive datasets and powerful visualization techniques make it possible to grasp general patterns and find subtle phenomena, or to zoom in exceptional behaviors, the social equivalent of black holes. Both interpretative analysis and classical quantitative models should greatly benefit from these new and powerful exploratory techniques.

This toolset (highly accurate inference, massive datasets and powerful visualization) will allow in depth exploration of the birth of superstars by looking at the evolution of fans' comments over time. Our findings suggest that superstars, who obviously start their career as artists with lesser profiles, attract more and more fans over time with the regrettable consequence of lowering average sentiment. Justin Bieber had a sentiment score of 0.86 in 2007, his debut year, 0.26 *above* average. In 2024, Bieber is 0.14 below average. A similar, but perhaps more muted pattern, is found in Taylor Swift who premiered in 2008 with a close to 0.10 above average sentiment over other artists. In both cases, Bieber and Swift were receiving below average sentiment at the start of their third year.

But anecdotes do not make a rule. Blackpink has always been above average, Bad Bunny always below, Ed Sheeran oscillates around the mean. At this point in time we have no theory, not even strong intuitions to explain these patterns. All that we have is questions, but perhaps the right tools to find answers.

References

Ahuja, H., Kaur, N., Kumar, P., & Hafiz, A. (2023, October). Machine Learning based Sentiment Analysis of YouTube Video Comments. In *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)* (pp. 1-6). IEEE.

Alhujaili, R.H. and W. M. S. Yafooz, (2021). Sentiment Analysis for Youtube Videos with User Comments: Review. International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, pp. 814-820

Biswas, E. et al. (2020). Achieving Reliable Sentiment Analysis in the Software Engineering Domain using BERT, 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 162-173 (link)

Cooper, Andy (2024), The World's Most Disappointing Masterpiece to Visit. Couponbirds (link)

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391-418.

Das, Sanjiv Ranjan and Mike Y. Chen (2001). Yahoo! For Amazon: Sentiment Parsing from Small Talk on the Web, EFA 2001 Barcelona Meetings, 45 pages (link)

De Maeyer, P. (2012). Impact of online consumer reviews on sales and price strategies: A review and directions for future research. *Journal of Product & Brand Management*, *21*(2), 132-139.

Deori, Maya, Vinit Kumar and Manoj Kumar Verna (2023). Analysis of YouTube video contents on Koha and DSpace, and sentiment analysis of viewers' comments. Library Hi Tech 41(3): 711-728 (link)

Demszky, Dorotta, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade and Sujith Ravi (2020).GoEmotions: A Dataset of Fine-Grained Emotions. Association for Computational Linguistics (ACL) 2020 (link)

Devlin, Jacob et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint (link)

Eliashberg, J., & Shugan, S. M. (1997). Film critics: Influencers or predictors?. *Journal of marketing*, *61*(2), 68-78.

Gómez, L.M., Cáceres, M.N. (2018). Applying Data Mining for Sentiment Analysis in Music. In: De la Prieta, F., et al. Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017. PAAMS 2017. Advances in Intelligent Systems and Computing, vol 619. (link)

Google. (2024-04-30). *Transparency Report [https://transparencyreport.google.com/youtube-policy]*

Gunter et al. (2014). Sentiment analysis: A Market-Relevant and Reliable Measure of Public Feeling?, International Journal of Market Research, 56(2): 231-247 (link)

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225. (link)

Kofler, Florian, Johannes Wahle, Ivan Ezhov, Sophia Wagner, Rami Al-Maskari, Emilia Gryska11, Mihail Todorov, Christina Bukas, Felix Meissen, Tingying Peng, Ali Ertürk, Daniel Rueckert, Rolf Heckemann, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, Bjoern Menze and Marie Piraud. (2023, April). Approaching peak ground truth. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-6). IEEE.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. arXiv preprint arXiv:1808.08949.

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access.

Shukla, Stuti, Pooja Khanna, and Krishna Kant Agrawal (2017). Review on sentiment analysis on music. *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS)*. IEEE, 2017. (link)

Tedeschi, Simone, Johan Bos, Thierry Declerck, Jan Hajic, Daniel Hershcovich, Eduard H. Hovy, Alexander Koller et al. (2023). What's the Meaning of Superhuman Performance in Today's NLU?. *arXiv preprint arXiv:2305.08414*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wu, J., L. Du and Y. Dang, "Research on the Impact of Consumer Review Sentiments from Different Websites on Product Sales," *2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, Lisbon, Portugal, 2018, pp. 332-338, doi: 10.1109/QRS-C.2018.00065.