Tipo de documento: Tesina de Grado de Ciencias de la Comunicación

Título del documento: ¿Lo tóxico garpa?: un análisis sobre la asociación entre toxicidad y visitas e interacciones en videos humorísticos en YouTube

Autores (en el caso de tesistas y directores):

Joaquín López Calvo

Damián Fraticelli, tutor

Juan Manuel Pérez, co-tutor

Datos de edición (fecha, editorial, lugar,

fecha de defensa para el caso de tesis): 2022

Universidad de Buenos Aires

Facultad de Ciencias Sociales

Ciencias de la Comunicación

¿Lo tóxico garpa?

Un análisis sobre la asociación entre toxicidad y visitas e interacciones en videos humorísticos en YouTube

Autor:

Joaquín López Calvo | **DNI:** 39.463.124 | **Correo:** joaquinlcalvo@gmail.com

Tutores:
Dr. Damián Fraticelli | **DNI:** 24.405.148 | **Correo**: damianfraticelli@yahoo.com

Lic. Juan Manuel Pérez | **DNI:** 31.860.008 | **Correo**: jmperez@dc.uba.ar

Índice

Introducción

El presente trabajo explora los vínculos entre discursos tóxicos y visualizaciones e interacciones al interior de la sección “Comedia” de YouTube, bajo la hipótesis de que un incremento en la toxicidad de los comentarios se correlaciona con mayor cantidad de visionados y/o reacciones. Definimos toxicidad como: “lenguaje rudo, irrespetuoso o poco razonable que tiende a expulsar sujetos de la conversación” (Jigsaw, 2020).

Por otra parte, creemos que las enunciadoras femeninas podrían recibir más violencia que los masculinos. Por último, la forma en que se relacionan los comentarios tóxicos con el contenido de los videos también formará parte del análisis y las conclusiones. Al final, todas nuestras preguntas aúnan en un objetivo en común: determinar si la toxicidad aumenta la visibilidad de los enunciadores. En criollo: si lo tóxico garpa (en YouTube, en Comedia).

Creemos en la importancia de estudiar los fenómenos de violencia en redes por su probado efecto en influenciar actos concretos contra sujetos y colectivos minoritarios (Koehler, 2019). Por solo citar algunos ejemplos estudiados, en los últimos años fenómenos organizados de violencia en redes causaron: linchamientos de minorías étnicas en India (Poruthiyil, 2021), reuniones de neonazis en Charlottesville (Blout & Burkart, 2020) y el suicidio de una adolescente canadiense de 16 años (Lester, et al., 2013).

Si bien este estudio no se centra únicamente en discursos que generen violencia, sino en toxicidad como categoría más amplia, creemos que ambas caracterizaciones tienen puntos en contacto notables. Además, la evidencia empírica apunta a que las reacciones a comentarios con los que un individuo está en desacuerdo son más numerosas y enérgicas que las que provocarían opiniones consonantes (Röchert, et al., 2020).

A pesar de que en 2022 casi ninguna red social está libre de moderación, y de que todas las grandes plataformas optan por una estrategia mixta entre sistemas de Procesamiento de Lenguaje Natural (NLP de aquí en más, por sus siglas en inglés) y moderadores humanos, YouTube ha sido señalada en numerosas ocasiones como la más pasiva frente a la violencia en redes.

Según la Comisión Europea de Justicia (2021), la red de los videos quedó anteúltima en el ranking de remoción de contenidos denunciados, solo por encima de Twitter, que, también muy laxa con sus contenidos, ha seguido líneas propias ante los discursos tóxicos como el cierre de cuentas con contenido malicioso, con el caso del expresidente norteamericano Donald J. Trump como exponente más visible. Según el mismo informe, YouTube es la que más falla en comunicar a sus infractores la causa de la remoción, sólo 1 de cada 4 usuarios que vuelcan toxicidad en la red recibe

informaciones sobre su falta. En síntesis: la mayoría de las veces, los comentarios tóxicos eluden los filtros y controles de la plataforma.

No solo YouTube es una de las redes con menos moderación, donde la "limpieza" de comentarios es menos frecuente, además el tipo de contenido que se genera al interior de la plataforma habilita a interacciones muy divergentes: ¿cuántas respuestas diferentes puede motivar un video de 18 minutos con la narración constante de un locutor? Durante el análisis cualitativo exploraremos clasificaciones de esta divergencia, por lo pronto, partimos de la creencia de que encontrar discursos tóxicos al interior de una plataforma tan prolífica, heterogénea y a veces altamente permisiva como YouTube es más probable que en otras redes sociales.

Aunque la negatividad en redes no es exclusiva de ningún género, estudios en psicología conductual señalan que muchas de las respuestas risibles se basan en violaciones a normas o expectativas carentes de una amenaza real de violencia (McGraw & Warren, 2010). Y si buena parte de los mensajes humorísticos contienen una falsa agresión, es de suponer que los contenidos risibles en redes sociales –donde el contenido público vuelve imposible predefinir un enunciatario de similares valores- atraigan gran cantidad de respuestas tóxicas. Por eso decidimos investigar la toxicidad en el marco de videos de la sección "Comedia". Sobre los doce youtubers más populares del género en Argentina, recortamos la totalidad de lo producido entre el 20 de marzo de 2020 y el 20 de marzo de 2021.

Para responder nuestras preguntas recolectamos 1.562.810 comentarios y los analizamos con Perspective.API, un modelo de NLP que identifica el nivel de toxicidad en textos. Luego comprobamos las correlaciones con el coeficiente de Pearson para, por último, analizar los casos con más presencia de toxicidad bajo el modelo semiótico de análisis enunciativo del humor.

Si bien en los últimos años un número considerable de estudios abordaron la violencia en redes desde una perspectiva interdisciplinar, conjugando las herramientas de Inteligencia Artificial con marcos de la psicología o las ciencias sociales, al menos tres condiciones del presente trabajo escasean entre la literatura.

En primer lugar, los contenidos de medios de comunicación masiva y los conflictos de la política partidaria han recibido gran parte de la atención académica de los especialistas en discursos del odio y discursos tóxicos durante los últimos años. Mientras, campos como el humor -entre otros-, que representan una porción considerable del tiempo de recepción en redes sociales y donde la construcción de sentido merece al menos la misma atención, han quedado relegados.

Luego de meses de búsqueda concluimos en que no hay investigaciones sobre la relación entre toxicidad y humor en redes sociales, al menos no en la escala que nos planteamos en este caso. A pesar de que nuestros métodos no coincidan con los más usuales en Estudios Culturales, sí somos herederos de cierta mirada inquisitiva hacia fenómenos que las investigaciones "serias" pasarían por alto.

Segundo, la vasta mayoría de los estudios se centran en Twitter. Esto podría deberse a su diseño como arena para el debate, la confrontación y los discursos políticos, o a que cuenta con una API mucho más amigable al *crawling*[1], reduciendo por mucho las labores de captura de datos. Además, se han desarrollado marcos analíticos específicos para Twitter que han marcado directivas claras y concisas para su investigación (Pak & Paroubek, 2010). Sin embargo, la certidumbre de que la violencia en redes excede a Twitter nos lleva a estudiarla en otras plataformas, aun cuando esta decisión suponga (grandes) esfuerzos adicionales.

En el ranking de las webs globales con más visitas, YouTube se ubica segunda con 22.8 mil millones de visitantes mensuales solo detrás de Google, mientras que Twitter es la novena (Statista, 2021). De acuerdo SimilarWeb (2021), sitio de referencia en estudios de mercado, las visitas a YouTube equivalen aproximadamente a las de todo el ecosistema Meta sumado (Facebook, Instagram y WhatsApp), con 3.6 mil millones. Aunque se pueda suponer una relación dispar entre visita y comentario a través de las plataformas, la suma de aproximadamente mil millones de vistas diarias en YouTube (Chen, et al., 2022) da cuenta de un fenómeno de dimensiones que no pueden ser desatendidas.

Mientras que en volumen de visitas YouTube es un gigante al lado de Twitter, la relación parecería directamente inversa en cuanto al número de investigaciones publicadas para cada plataforma. La subrepresentación de estudios sobre los comentarios de YouTube sumado a la muy suave moderación de la red, abre un campo más que fértil para nuevas investigaciones en este marco.

Por último, sobran los dedos de una mano para contar los intentos por conjugar la Inteligencia Artificial (IA) con la semiótica contemporánea. A nuestro juicio, esta distancia se debe más a una pertenencia a culturas epistemológicas antaño muy alejadas que a una incompatibilidad en los métodos. Esta investigación también busca ser prueba de que los métodos de la semiótica pueden llenar de sentido la inconmensurable cantidad de datos y la eficiencia en el análisis que hoy nos permiten las ciencias de la computación.

[1] De aquí en más nos referiremos con este término inglés a la recolección automática de datos a través de métodos computacionales.

En el primer capítulo estableceremos las bases teóricas de nuestra experimentación, centrándonos en dos ejes: por un lado, la semiótica de la hipermediatización y las adaptaciones contemporáneas de la teoría de la circulación del sentido de Eliseo Verón, por el otro, sentaremos los parámetros de nuestra aplicación de modelos de Inteligencia Artificial centrada en problemas de lenguaje.

En el segundo capítulo presentaremos el estado del arte de trabajos sobre comentarios en YouTube, las tácticas de los enunciadores al interior de un medio que no controlan, así como las clasificaciones de las diferentes formas de violencia en redes y lo que se conoce sobre la moderación algorítmica de los contenidos y comentarios de la plataforma propiedad de Alphabet. Nos centraremos en los youtubers dada la relevancia de estos actores al interior de la red, medible tanto en el número de suscripciones y visitas como en la intensa actividad de las comunidades que se forman en torno a sus canales.

El tercer capítulo está dedicado a establecer los límites de nuestra investigación. Por un lado, encontraremos el desarrollo de las hipótesis, sus razones; en la línea de qué otras investigaciones del pasado nos encontramos, punto que retomaremos en las conclusiones para sugerir futuras vías investigativas. Por el otro, detalleremos la metodología adoptada: la implementación de IA con Perspective, el análisis estadístico, nuestro modelo semiótico para el análisis del humor en redes sociales y, al final, una caracterización de nuestro conjunto de datos y sus formas de recolección.

En el cuarto capítulo encontramos los resultados de nuestro análisis que se desarrollan con dos lógicas diferentes. Primero, la metodológica, divida en etapas cuantitativas y cualitativas. Pero otra forma de mirarlo sería al nivel de los enunciadores, aquí nuestra mirada oscila de lo general, a lo particular, y de vuelta a lo general: definimos una escena, nos adentramos en el comportamiento de algunos enunciadores o videos, para volver a una mirada panorámica que contextualice esas observaciones particulares.

Por último, los dos apartados finales corresponden a las discusiones que motivan nuestros descubrimientos y las conclusiones que buscamos establecer. Aquí es donde buscaremos sugerir algunas posibles líneas de investigación a futuro, no solo en base a nuestras conclusiones sino también al conjunto de herramientas que desarrollamos y liberaremos para el uso de la comunidad científica.

2. Objeto de estudio y enfoque teórico

2.1. Marco Teórico

2.1.1. Semiótica de la hipermediatización

La teoría de la semiosis social que Eliseo Verón (1987) expuso tenía como objetivo explicar la circulación de los discursos en una sociedad dando cuenta del peso específico de la dimensión discursiva en la construcción social de lo real.

Entre dos conjuntos de condiciones, las de producción y las de reconocimiento, circulan los discursos sociales y sus efectos:

> "Las condiciones productivas de los discursos sociales tienen que ver, ya sea con las determinaciones que dan cuenta de las restricciones de generación de un discurso o de un tipo de discurso, ya sea con las determinaciones que definen las restricciones de su recepción. Llamamos a las primeras condiciones de producción y a las segundas, condiciones de reconocimiento."
> (Verón, 1987, p. 127)

En las postrimerías del siglo XX, Verón (1999, 2001) señalaba que las sociedades occidentales, modernas y altamente industrializadas, eran también sociedades mediatizadas. La frontera entre los órdenes de lo "real" y de la representación estaba entonces a cargo de los medios de comunicación.

Medios que no eran dueños de la hegemonía de la mediatización. Así, el funcionamiento de las instituciones, de las prácticas, de los conflictos, de la cultura, se estructuraba en relación directa con la existencia de los medios. Medios que necesariamente comportaban una dimensión colectiva y tenían una característica central: el sistema broadcast generó un nuevo espacio público que "permitió a pocos, principalmente profesionales, hacerle llegar sus discursos a muchos" (Carlón, 2015, p. 33).

Los actores en los procesos de comunicación eran cuatro: medios, instituciones, actores individuales y colectivos. Mientras que los primeros dos eran capaces de producir "colectivos" a través de mediatizaciones, los actores individuales y colectivos permanecían en el campo de la recepción.

Sin embargo, el rígido muro moderno que separaba lo público de lo privado acabó cediendo a las demandas de la posmodernidad. Y, aunque Verón llegó a ver algunos de estos cambios antes de fallecer, su teoría debió ser ajustada a razón de los cambios

contemporáneos. Mario Carlón fue uno de los académicos que aventuró un camino para esta adaptación.

La llegada de la web 2.0 y sus plataformas significó para Carlón (2020) el pasaje a una sociedad hipermediatizada en la que “cada uno de nosotros administra un propio medio de comunicación” a través de redes sociales. En la sociedad hipermediática, los dos sistemas están en permanente relación y cada punto de contacto entre ellos establece una relación intermediática que desencadena un proceso de incremento de la complejidad. El ejemplo de Carlón son los hashtags de Twitter en la pantalla de televisión, que introducen otro medio -y otra discursividad- en los medios de comunicación masiva.

Antaño, medios poderosos capaces de “instalar una agenda” o “construir acontecimientos” se comunicaban de forma descendente: desde los medios de comunicación a los receptores masivos, una relación asimétrica.

Desde el punto de vista de la comunicación, la sociedad hipermediatizada trae consigo dos nuevas direcciones para los mensajes: la ascendente, en la que un enunciador amateur desconocido genera a partir de sus redes sociales un ‘colectivo de comunicación’ y despierta el interés de los medios de comunicación masiva, y la horizontal, la de los diálogos entre enunciadores que ya no se ven afectados por asimetrías de poder, sino que se encuentran en una situación de paridad. Esta última trae consigo la novedad de la comunicación intra-sistémica, central a los objetivos de nuestro trabajo.

Las direcciones de la comunicación son ahora variadas. El crecimiento de la complejidad que estalla cuando se produce una relación intermediática se manifiesta a nivel comunicacional. De ese modo, una discursividad descendente comienza a convivir casi siempre con tensiones con otra que es ascendente.

Carlón hablará de “saltos de escala” en la mediatización cuando “un proceso intra-sistémico que se venía desarrollando en los medios con base en Internet pasa a otro intersistemico, entre redes y medios masivos”. En el presente, medios e instituciones dejaron de ser los únicos dispositivos sociales capaces de producir mensajes con llegada masiva, y actores individuales y colectivos han pasado a compartir el ágora de la web 2.0. Todavía no hay paridad entre los nuevos actores y el sistema heredado de las sociedades occidentales, todavía las cadenas y medios masivos se ubican en un lugar de poder superior al de los influencers, y por eso se habla de “ascendente” y “descendente” en la circulación intersistémica.

Pero no todo es cambio para Carlón. La circulación veroniana sigue operando en forma de cadena:

> …a una operación en reconocimiento la siguen operaciones en producción que son seguidas por operaciones en reconocimiento, etcétera. Así lo que muestran los casos que hemos estudiado es que una vez que se han establecido las identidades, soportan toda la comunicación. Es un hecho que acontece en distintos niveles: en el intra-mediático, es decir, entre pares en la red, y en el nivel inter-sistémico, en la circulación de contenidos entre las redes sociales mediáticas y los medios masivos.
> (Carlón, 2020, p. 124).

Uno de los eslabones de reconocimiento de esa cadena infinita es el que buscaremos analizar en nuestra investigación: las respuestas y los comentarios. Mientras la circulación como concepto aparece intacta, en nuestro recorte de los 12 enunciadores hipermediáticos más populares de Comedia en Argentina ya no hay medios masivos, sino solo actores individuales emergentes en el medio digital. Un escenario novedoso y aún poco estudiado que la sociedad hipermediatizada presenta a las ciencias de la comunicación.

Retomaremos dos puntos más de la teoría veroniana. Primero, la noción de contrato de lectura, la articulación de las "expectativas, motivaciones, intereses y contenidos del imaginario de lo decible visual" (Verón, 1985, p. 172). Aunque los enunciadores busquen modelarlo en producción, para (Verón, 2004) el nexo entre la situación de enunciación de un soporte -las formas de comunicar- y sus enunciatarios se consuma en el momento de la lectura, en reconocimiento. Hasta donde sabemos, aún no se ha propuesto una adaptación del contrato veroniano de lectura para la enunciación en redes. En las redes sociales, la constante retroalimentación permanente entre la instancia de producción y reconocimiento altera las lógicas del contrato de lectura (Carlón, 2020).

Por último, Verón (2013) actualiza su sociosemiótica con una perspectiva no antropocéntrica, reconociendo en los dispositivos y nuevas formas técnicas a productores de sentido. Además de registrar realidades, las máquinas las crean y delimitan las fronteras de qué y cómo se puede decir. Las interfaces y las puertas que abren (o cierren) deberán entonces ser incluidas en los análisis enunciativos.

2.1.2.a- Inteligencia Artificial (IA)

La Inteligencia Artificial es la rama de las ciencias de la computación encargada de intentar reproducir las capacidades cognitivas humanas para que máquinas sean capaces de realizar tareas que requerirían ese tipo de inteligencia. Hasta el momento, el desarrollo de la disciplina ha intentado replicar y automatizar las capacidades humanas de manera fragmentaria, con avances en distintos niveles para tareas como: el reconocimiento de imágenes o de voces, la identificación de patrones en diagnóstico

médico, la comprensión del lenguaje, o la extracción de opiniones en redes sociales. En este último grupo es que entran las tareas de detección de toxicidad.

Para generar (o como se dice en la jerga de IA, “entrenar”) un modelo de Inteligencia Artificial se parte de la base de una estructura algorítmica -un conjunto de pasos predefinidos- que “aprende” de los datos. Se dice que los modelos “aprenden” cuando adaptan su comportamiento a los datos presentados y mejoran su desempeño en la tarea considerada. Un entrenamiento se considera exitoso cuando una computadora logra aprender de la experiencia (los datos) y puede replicar (aun parcialmente) la cognición humana para la tarea en cuestión.

Una vez que el programa repasa los datos en base a los que fue entrenado, las suficientes (miles o millones de) veces, los modelos se vuelven más finos en la replicación de los resultados que se le presentaron. Por eso se usan las metáforas “entrenamiento” y “aprendizaje”. En los casos más complejos podemos estar frente a cientos de millones de repeticiones y retroalimentaciones de un modelo que se compara contra grandes masas de datos en la búsqueda de predecir comportamientos de la realidad.

Al final, la mayoría de las veces las diferentes formas de Inteligencia Artificial buscan detectar patrones. Christopher M. Bishop (2006) ejemplifica este proceso con la manera en que un modelo de aprendizaje automático puede descifrar dígitos escritos a mano:

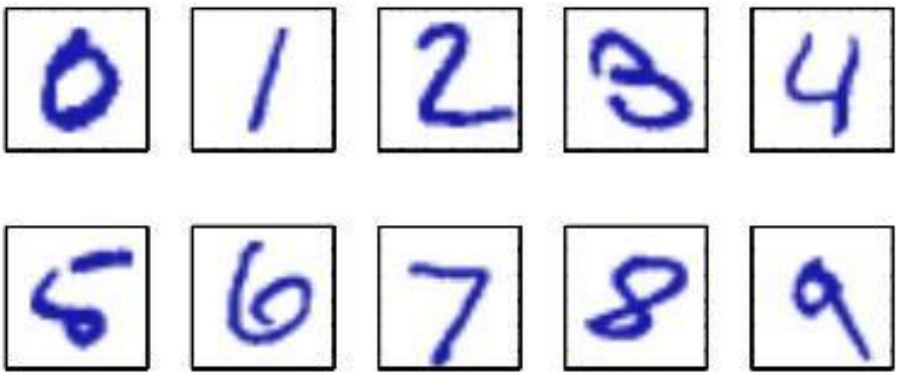


Fig. 1: Ejemplos de dígitos escritos a mano. Fuente: (Bishop, ibidem).

A partir de un conjunto amplio de N dígitos {x1,..., xN} llamado set de entrenamiento, se adaptan los parámetros de un modelo. Estos datos de entrenamiento deben contener categorías para cada dígito que normalmente se completan a mano. Un algoritmo de aprendizaje supervisado se puede expresar como una función y(x) que toma un dígito nunca antes visto, x, como dato de entrada, y genera un vector y, como dato de salida. La forma de esa función y(x) se determinará durante la etapa de

entrenamiento. Si al encontrarse por primera vez con los ejemplos de la Figura 1 el modelo identificara apropiadamente que el primer recuadro se trata de un cero, el segundo de un uno, y así sucesivamente, en base a otros ejemplos que se le fueron enseñando durante su entrenamiento, diríamos que el modelo generalizó correctamente.

Cuando el modelo fue entrenado y se comprobó su performance, estamos ante un modelo que "entendió" la forma en la que se comporta un fenómeno del mundo exterior y sus predicciones pueden aplicarse a casos futuros en los que se desconocen los resultados a priori. En el caso de este trabajo, un modelo entrenado sobre millones de comentarios en castellano -tóxicos o no-, buscará predecir sobre datos que nunca ha visto, es decir, aquellos que recolectamos sobre los 12 youtubers de nuestra selección.

Grosso modo, el campo de la Inteligencia Artificial puede dividirse entre el aprendizaje supervisado y el no supervisado, dependiendo de si el algoritmo se ajusta a partir de una variable objetivo o no. En este trabajo optaremos por técnicas de aprendizaje supervisado, donde determinadas características de un conjunto de datos se asocian a la variación de una variable objetivo. En nuestro caso, esa variable Y será el nivel de toxicidad de cada comentario.

Dentro de los problemas de aprendizaje supervisado encontramos los de regresión y los de clasificación. Los primeros apuntan a dilucidar una variable continua, numérica, mientras que estamos ante un clasificador cuando las posibles respuestas al interrogante son finitas. Si bien nuestra herramienta de análisis funciona de manera regresiva, asignando un valor numérico del 0 al 1 a cada comentario, nosotros colocaremos un umbral para dividir de manera binaria aquellos comentarios, forzando los resultados de una regresión a terminar clasificando comentarios. Por esto, en adelante nos referimos indistintamente a "el modelo" o "el clasificador".

Muchos de los modelos de NLP -dentro del que se cuenta el elegido para este trabajo-, entran dentro de la categorización de aprendizaje supervisado: reciben anotaciones humanas de textos sobre las que aprenden estructuras del lenguaje para predecir sobre datos que no habían observado con anterioridad.

En el caso de los estudios enfocados en fenómenos lingüísticos, el uso de técnicas de IA permite dar con muestras más numerosas y asegura procesos más escalables, eficientes y replicables. Mientras que aquellos trabajos que dependen de la revisión manual y por unidad de comentarios a cargo de analistas humanos pueden asegurarse un mayor nivel de precisión en el análisis de fenómenos a pequeña escala, propuestas como la de este trabajo sólo son posibles escalando herramientas.

Muchos de los modelos lingüísticos actuales están basados en redes neuronales. Las redes neuronales de la informática imitan la estructura de procesamiento de la información de las neuronas biológicas: reciben estímulos externos, realizan cálculos y, a partir de una función específica, generan un valor de salida. Se las llama "redes" porque las neuronas no se activan de forma individual sino en conjunto a través de capas que se computan en secuencias.

A través de estas capas, la red puede aprender con jerarquías, de manera similar a como lo hacemos los humanos, y, además, autoajustar las ponderaciones cada vez que una de ellas sea incorrecta. Con la repetición de este proceso, las predicciones del modelo aumentan en performance. En NLP, muchas veces se usan redes recurrentes porque permiten computar secuencias, este fue uno de los puntos que las arquitecturas siguientes superarían.

En computación se habla de "inteligencia" de las máquinas en vistas a que estos sistemas son porosos a nuevas informaciones y se retroalimentan de sus errores de diagnóstico para aumentar las chances de éxito en las tareas para las que fueron programadas (Poole, 2018).

Aunque la inteligencia a la que puede aspirar una computadora hoy es muy diferente a la de miembros del reino animal (Kaplan, 2022), en tareas cognitivas como el cálculo y el procesamiento hace algunas décadas que la mente humana fue superada por las computadoras. Como una más en la extensísima lista de pruebas al respecto, ejemplifiquemos con los tiempos de análisis de nuestro propio trabajo.

El proceso de recolección de 1.562.810 comentarios tomó nueve días:
comenzó el 14/9/2021 con los comentarios del video "El PERRO del Tío de CJ" de la cuenta AleVera Oficial y acabó el 23/9/2021 con los comentarios de un video publicado por el usuario Tincho Ruiz: "AMIGO MALA INFLUENCIA 10 🚘🤦‍♂️😂 #Cuarentena #CoronaVirus". ¿Cuánto tiempo humano consumiría la tarea de recolectar y clasificar más de un millón y medio de comentarios?

En cuanto al proceso de análisis de los comentarios, las observaciones son muy similares: dimos el puntapié el 24/2/2022 y el último comentario fue interpretado por Perspective.API el 3/3/2022. En siete días el sistema de NLP asignó seis valencias diferentes sobre nuestro corpus de comentarios, lo que totaliza 9.376.860 análisis textuales en una semana. Imaginemos por un momento la cantidad de horas humanas que una tarea de esta naturaleza requeriría.

En los apartados siguientes, detallaremos con mayor profundidad en qué se basan los modelos de lenguaje y en qué medida podemos confiar en los resultados derivados de su actividad.

2.1.2.b- Procesamiento de lenguaje natural (NLP)

El Procesamiento del Lenguaje Natural (PLN o NLP en inglés) es una disciplina que se ubica en la intersección de las Ciencias de la Computación, la Inteligencia Artificial y la Lingüística, encargada de intentar proveer a las computadoras la capacidad de entender y generar el lenguaje escrito. Otros nombres para esta disciplina son minería de texto, o bien lingüística computacional.

En la última década, el campo del NLP ha experimentado avances a pasos agigantados que pueden verificarse en nuestra propia experiencia como usuarios: traductores automáticos, *chatbots*, asistentes inteligentes como Siri o Alexa, entre otros. Gran parte de este avance se explica a partir de la consolidación de las redes neuronales como matriz técnica, que inicia en 2013 con la consolidación de *word-embeddings* (Mikolov, et al., 2013).

La introducción de los Transformers representó un salto de magnitud en el campo del NLP (Vaswani & al, 2017). En muy resumidas cuentas, los Transformers introdujeron un mecanismo de "atención" que permite el cálculo paralelo, más veloz que los cálculos secuenciales o convolucionales de las redes neuronales.

Inicialmente, los Transformers estaban dedicados a tareas de traducción automática, pero en los últimos años la práctica totalidad de los nuevos modelos los ha incluido, perfeccionado y adaptado a diferentes fines.

En BERT (Bidirectional Encoder Representations from Transformers) encontramos una de las aplicaciones más notorias de los Transformers en el último tiempo. En resumen, BERT incorpora de técnicas anteriores el preentrenamiento sobre inmensos corpus sin etiquetas humanas -unos 340 millones de parámetros- seguido de procesos *fine-tuning*, adaptaciones a cada tarea que trae consigo algunos parámetros específicos.

Mientras modelos anteriores consideraban al lenguaje de izquierda a derecha, bajo una lógica causal, BERT opera sobre la premisa de un lenguaje enmascarado: se "ocultan" un número de palabras de una frase sobre las que se intenta predecir. Según Devlin et al. (2019), este método bidireccional abre las puertas a representaciones lingüísticas más fidedignas. Así, en la siguiente frase:

En casa de [MÁSCARA], cuchillo de [MÁSCARA].

BERT intentará predecir las palabras "herrero" y "palo" en función del contexto. Y aunque existan modelos multilingües -diseñados para traducir oraciones al inglés y luego realizar las tareas de NLP- con buenas performances, son los monolingües -

entrenados y puestos en operación en una misma lengua- los que suelen tener mejores resultados.

Esta asunción, que podría parecer sólo teórica, encuentra respaldo experimental en los avances de Canete, et al. (2020) en comparación con modelos multilingües. De hecho, una frase popular como “en casa de herrero, cuchillo de palo” muy muy difícilmente podría ser comprendida cabalmente por un modelo que traduzca la frase al inglés, despojándola de su sentido metafórico inherente al castellano.

En el apartado metodológico profundizaremos más respecto a nuestra herramienta de análisis, pero, solo a modo introductorio, señalaremos que de acuerdo a Pellat y Georgiou (2018), Perspective se entrena con comentarios en castellano al menos desde su alianza con el periódico El País en 2018, y, además, lo más seguro[2] es que Perspective incorpore los Transformers y alguna técnica similar a BERT al interior de su arquitectura.

Antes de que estas novedades fueran una realidad, soluciones a problemas como la identificación de discursos tóxicos obtenían resultados muy pobres comparadas con las de analistas humanos, debido al alto nivel de ruido que caracteriza a los comentarios (errores de ortografía y puntuación, uso de jerga altamente variable y de difícil inclusión en los conjuntos de datos de entrenamiento).

Gracias a la llegada de los modelos basados en redes neuronales y a la novedad de los Transformers, florecieron modelos de lenguaje que sí logran adaptarse a las complejidades de, por ejemplo, los comentarios en Internet.

Entre la semiótica y la IA

Por lo menos dos investigaciones han emprendido un camino similar al nuestro, combinando técnicas predictivas de análisis con instrumentos de la semiótica.

En el cruce de técnicas predictivas de análisis con instrumentos de la semiótica, dos investigaciones que han emprendido un camino similar al nuestro sirvieron de referencia para imaginar un camino aún poco explorado.

Kearney y Octon (2019) buscaron cubrir la tríada peirceana con Funciones Generales de Valor, mientras que Shackell y Sitbon (2019) automatizaron análisis oposicionales aprovechando “la multidimensionalidad, precisión y novedades en la visualización” que brinda la computación. En ambos casos, al uso de técnicas computacionales se le suma una búsqueda de significado, que la Inteligencia Artificial por sí sola no puede brindar:

[2] Decimos “lo más seguro” porque Perspective.API es propiedad de Google y su código no ha sido abierto. En el apartado 3.2 nos centraremos en la herramienta, lo que sabemos de su diseño, sus alcances y limitaciones.

> It remains a major challenge to anticipate which future arguments will win attention or "resonate" with a certain community of consumers, influencers or voters. One reason is that such tasks are not only procedural but also creative. In such endeavors, the past represents a status quo to be disrupted rather than imitated (or even much extrapolated). (…) While computation currently aids in areas such as sentiment analysis and argument mining, which can be matched with argument schemes for argument invention, we assert that some broader aspects of persuasion remain unrecognized. These cross into the domain of semiotics.[3]
>
> Shackell y Sitbon (2019, p. 302).

2.2. Estado de la cuestión.

2.2.1. La sección comentarios: ¿quiénes comentan? ¿qué dicen?

En general, el estereotipo de qué puede encontrarse en la sección de comentarios de YouTube es bastante pobre. No obstante, millones de comentarios nuevos se publican cada día, por lo que es de suponer que las motivaciones detrás de estos mensajes merecen una observación más profunda.

En Schultes, et al. (2013) se señala que, aunque no haya un tipo dominante de comentario y que la sección de comentarios parezca reflejar conductas de la comunicación en el mundo no virtual, estadísticamente la proporción más alta se lo llevan "posteos ofensivos".

El público adolescente conforma una parte considerable de las audiencias de contenidos risibles en Internet. Pires, et al. (2019) realizaron un relevamiento multi-método sobre adolescentes de 8 países en el que se revela que, entre las diferentes dinámicas de comentarios se cuentan: opiniones sobre los videos, contribuciones y recomendaciones para futuros videos del canal y pedidos de ayuda. En resumen, la sección de comentarios puede ser vista por gran parte de estos jóvenes dispersos por el mundo como una herramienta de retroalimentación: un lugar donde se aprende pero también se puede enseñar al productor del video cómo hacerlo mejor.

Para 2011, el arquetipo de usuario que comentaba en YouTube era el de un hombre de 29 años que, en promedio, enviaba textos de no más de 60 caracteres (Thelwall & Sud,

[3] En castellano: "Sigue siendo un gran reto anticipar qué argumentos *futuros* ganarán la atención o 'resonarán' en una determinada comunidad de consumidores, personas influyentes o votantes. Una de las razones es que tales tareas no son sólo de procedimiento, sino también creativas. En estas tareas, el pasado representa un *statu quo* que hay que perturbar en lugar de imitar (o incluso extrapolar mucho). (...) Si bien la computación actualmente ayuda en áreas como el análisis de sentimientos y la minería de argumentos, que pueden emparejarse con esquemas argumentales para la invención de argumentos, afirmamos que algunos aspectos más amplios de la persuasión siguen sin reconocerse. Estos aspectos entran en el ámbito de la semiótica."

2011). Y, aunque pasados 12 años, cabría considerar diferencias en los perfiles de quienes comentan, algunas características parecen permanecer intactas: ya entonces los autores señalaban que las respuestas entre comentarios aumentaban desproporcionadamente en relación a textos negativos que a comentarios positivos. Los datos provistos por Statista (2022) refuerzan la idea de que son mayormente hombres de entre 18 y 44 años quienes más participan en la plataforma.

En la misma línea, Möller, et al. (2018) apuntan que el tipo de video influencia la cantidad y la valencia de los comentarios. Comparativamente, los videos de entretenimiento reciben mayor número de comentarios y más toxicidad que los relativos a la política. Además, los comentarios negativos son peor recibidos por el conjunto de la comunidad, recibiendo menos "me gusta" en promedio que aquellos positivos o de valencia más bien neutra.

En consonancia con Möller et al., Khan (2017) señaló que la principal gratificación que reciben los usuarios de YouTube es a través de los videos de entretenimiento. Según su investigación, el predictor más poderoso para los comentarios fueron las escenas donde se presentan interacciones sociales.

Hay suficientes razones para suponer que, en muchos casos, los usuarios no incurren en comentarios tóxicos sólo como individuos, sino que muchas veces lo hacen inscriptos en colectivos de mayor o menor grado de adhesión identitaria. En la etapa de análisis repasaremos una cierta cantidad de comentarios identificados como "ataques identitarios" por vulnerar la filiación de grupos minoritarios. A este respecto, Murthy y Sharma (2018) relevaron evidencia de acciones en red en los videos que atraen mayor cantidad de comentarios con hostilidad racial.

Con "acciones en red" los autores califican tanto a ataques coordinados a través de diferentes videos de un mismo usuario, como a una economía de compromisos afectivos entre los atacantes por el que una vez que suficientes sujetos están comprometidos en estas prácticas una "formación de rueda" se despliega donde grandes cantidades de usuarios apuntan contra una víctima una y otra vez.

Resultados más recientes parecerían indicar que las observaciones de Murthy y Sharma no son aisladas. Röchert, et al. (2022) también combinaron técnicas de NLP con análisis de redes para demostrar que las personas que expresaron una posición favorable hacia teorías conspirativas tendieron a responder a contenidos o interactuar con usuarios que compartían la misma opinión.

En un estudio de 2020, este último grupo de investigadores usó una matriz analítica idéntica pero orientada a la observación de comentarios relativos a tres temas de

debate social: cambio climático, prohibición de velos islámicos y adopción en parejas del mismo género. Así, encontraron que es más probable que los usuarios crucen comentarios con aquellos que no piensan como ellos que con quienes sí. La tendencia apunta a que la reacción a comentarios disímiles a opiniones propias es más poderosa que aquella que es fruto del acuerdo (Röchert, et al., 2020).

Aunque sin ceñirse a YouTube, Mall, et al. (2020) propusieron una clasificación de los usuarios que vuelcan contenidos tóxicos en la sección de comentarios -esta vez, del foro Reddit-, también con Perspective como herramienta de análisis. Además, definieron los tamaños relativos de los grupos, dato que, entendemos, solo debe ser leído en su contexto de estudio. Según sus observaciones, los sujetos podrían dividirse en:

a- Caprichosos (31,2%): que oscilan entre comentarios con mayor y menor nivel de toxicidad.
b- Pacíficos (25,8%): sus niveles de toxicidad son relativamente bajos y no se alteran durante el tiempo.
c- Radicalizados (25,4%): sus primeros comentarios tienen baja valencia pero gradualmente van aumentando en toxicidad.
d- Estables (17,6%): sus niveles de toxicidad son relativamente altos y no se alteran durante el tiempo.

2.2.2. Los youtubers y sus tácticas frente al algoritmo

Hacemos propia la definición de táctica de Michel De Certeau como “operaciones cuasi microbianas que operan al interior de las estructuras tecnocráticas y de modificar su funcionamiento (…) prácticas a través de las cuales los usuarios se reapropian del espacio organizado por los técnicos de la producción sociocultural” (De Certeau, 1996 [1990], p. 44).

Mientras que de parte de los generadores de contenidos lo más usual sería hablar de “estrategias” de crecimiento, la idea de táctica pone al sujeto a merced de un poder superior con la potencialidad de definir parcialmente su destino. Este poder ayer se llamó algoritmos, hoy redes neuronales y quizás en el futuro vuelva a cambiar de nombre, las que difícilmente cambien serán las posiciones de poder relativas y la adaptación de quien enuncia con fines de visibilidad a las reglas impuestas por las plataformas.

Internet (y YouTube mismo) está repleto de publicaciones de gurúes de redes sociales, consultores, expertos SEO y especialistas en crecimiento que detallan los métodos y métricas para aumentar la visibilidad y las reproducciones en la plataforma.

En todos los casos, estos gurúes señalan caminos para escalar posiciones en el ranking de los buscadores y así ganar mayor visibilidad. Como creemos que es más probable que los youtubers presten atención a estas fuentes que a los papers que citaremos en los párrafos subsiguientes, elegimos tomarnos un instante para analizar los consejos de algunos de estos blogs.

Dean (2017) señala en el número de comentarios la principal variable de correlación con una mayor visibilidad YouTube, mientras que un elevado porcentaje de retención, una duración media de 14 minutos, el número de "compartidos", nuevos suscriptores y "me gustas" y el formato HD en los videos también parecerían ser factores favorables para los rankings.

En marzo de 2022, el primer resultado de la búsqueda "cómo aumentar visitas en YouTube" se lo llevaba un video de la cuenta Romuald Fons (2020) en el que revela cómo traccionando seguidores de Instagram a comentar masivamente un emoji de corazón durante las primeras horas de la subida de un nuevo video, este obtenía mayor visibilidad que los anteriores producidos bajo condiciones similares. Además, para Fons un aumento en la actividad de la pestaña "Comunidad" parecería ser otro aliciente para el buen desempeño de un video.

Si volvemos a las palabras de De Certeau, casos como el de los comentarios masivos de corazones se pueden leer desde la perspectiva de las tácticas. Esta mirada incluye resistencias y subterfugios como los de Fons frente a la rigidez del mandamás algorítmico, pero las victorias no son sino parciales -y los resultados de la acción de Fons no durarán más allá de ese único video-: "Debido a su no lugar, la táctica depende del tiempo, atenta a 'tomar al vuelo' las posibilidades de provecho. Lo que gana no lo conserva, el débil debe sacar provecho de fuerzas que le resultan ajenas. Lo hace en momentos oportunos en que combina elementos heterogéneos para 'aprovechar una ocasión'" (De Certeau, *ibidem*, p. 50).

Las observaciones de este gurú coinciden parcialmente con lo descrito por académicos como Pinto, et al. (2013) y su predicción de la popularidad futura de un video en base a su número de visionados en las primeras horas luego de ser publicado, y con la identificación del número de comentarios como primer predictor de la popularidad (Chatzopoulou, et al., 2010).

Para comprender cómo se actúa en base a estas recomendaciones generales y conclusiones estadísticas, es útil una mirada más minuciosa y contextual de la evolución de cuentas en particular.

Si nos detenemos en Zekiel79, una cuenta de videos paródicos sobre fútbol que experimentó grandes movimientos desde 2015, se observa que las variables responsables de los saltos de escala fueron la aparición de los contenidos en medios de comunicación masiva y "la sinergia producida entre las instancias de producción y reconocimiento" (Fraticelli, 2019, p. 14). El autor analiza factores intradiscursivos que afectarían positivamente al crecimiento de la cuenta en número de suscriptores, uno de los cuales fue la variación en el objeto de la burla que abrió la puerta a un mayor número de compartidos por no apuntar contra la identidad de ningún colectivo en particular.

Entre las prácticas concretas de crecimiento, pueden encontrarse líneas transversales y comunes. Son muy pocos los casos de enunciadores que no incitan explícitamente a sus seguidores a dar "me gusta", compartir, comentar o suscribirse, o todas las anteriores.
De acuerdo a Tur-Viñes y González-Río (2019), todos intentan evitar que sus fans se vuelvan meros espectadores para, en cambio, establecer un diálogo, una comunicación bidireccional que mueva al creador de contenido lo más cerca posible de su audiencia. En su selección de los 10 youtubers más visibles de España, los autores identifican patrones de comportamiento comunes:

a- Reconocen la lealtad de sus seguidores al ofrecer alguna clase de recompensa.
b- Crean expectación y aumentan el interés de sus seguidores al adelantar el contenido de sus próximos videos.
c- La inserción de experiencias personales en los videos, como aspectos de su vida cotidiana, viajes o problemas asociados a su faceta creativa.
d- El uso de una jerga común, solo entendible por la audiencia que mira sus videos de forma sostenida.
e- Concursos y sorteos de merchandising que sirven tanto para sumar una interacción que mantenga viva la comunidad como para promover el canal y los productos del youtuber.

Lo que Fraticelli llama sinergia Hidalgo-Marí y Segarra-Saavedra (2017) lo llamarán la creación de un movimiento social de millones en torno al canal. Lo cierto es que la coincidencia en torno al carácter colectivo del crecimiento de un canal es casi absoluta entre académicos y profesionales del mercado, las plataformas en general y YouTube en particular optimizan sus algoritmos por interacciones. La importancia de la faceta comunitaria en Youtube vuelve central a la sección comentarios que, lejos de estar en las sombras, se revelan como uno de los principales recursos de crecimiento de los enunciadores hipermediáticos emergentes.

2.2.3. Taxonomía de la violencia en redes

Si bien existen numerosos estudios sobre toxicidad y violencia en redes, es preciso establecer la separación entre las diferentes categorías que se miden en cada caso y definir cuál será la adoptada a lo largo de nuestra investigación. No sin antes destacar que no existen acuerdos universales ni entre académicos de diferentes campos, ni entre instituciones de la sociedad civil, ni siquiera en el marco del derecho internacional.

Como ya adelantamos en la introducción, la definición de discursos tóxicos que adoptaremos es la de Jigsaw (2020): "lenguaje rudo, irrespetuoso o poco razonable que tiende a expulsar sujetos de la conversación".

Puede que "discursos de odio" sea la categoría que más miradas atrae por parte de la opinión pública, sin embargo, creemos que muchas veces puede ser malentendida. El punto en común entre diferentes definiciones de discursos de odio sería el ataque a un grupo protegido (por su etnia, género, orientación sexual, nacionalidad, entre otras). De acuerdo a Waseem & Hovy (2016, p. 89), la categoría discursos de odio se aplica al:

> Use of a sexist or racial slur, attack a minority, promotes hate speech or violent crime, blatantly misrepresents truth, shows support of problematic hashtags, defends xenophobia or sexism, or contains a screen name that is offensive.[4]

Si bien el foco de nuestro trabajo aprovecha una categoría que no reduce la violencia hacia grupos protegidos sino que la entiende en un marco más abarcativo, también incorporaremos la variable "ataque identitario" para capturar aquellas formas de violencia dirigidas contra colectivos protegidos. Por un lado, desagregaremos el fenómeno en mensajes de diferente naturaleza, y, por el otro, al captar formas diferentes de violencia en redes, nos acercaremos más a determinar si podrían existir incentivos para que youtubers fomenten o eviten moderar la toxicidad al interior de sus canales.

En resumen, diferentes tipos de discursos violentos podrían separarse de la siguiente manera:

[4] "Uso de un insulto sexista o racista, el ataque a una minoría, que promueva discursos de odio o los crímenes violentos, la tergiversación flagrante de la verdad, muestras de apoyo a hashtags problemáticos, la defensa de la xenofobia o el sexismo, o el uso de un nombre de usuario ofensivo."

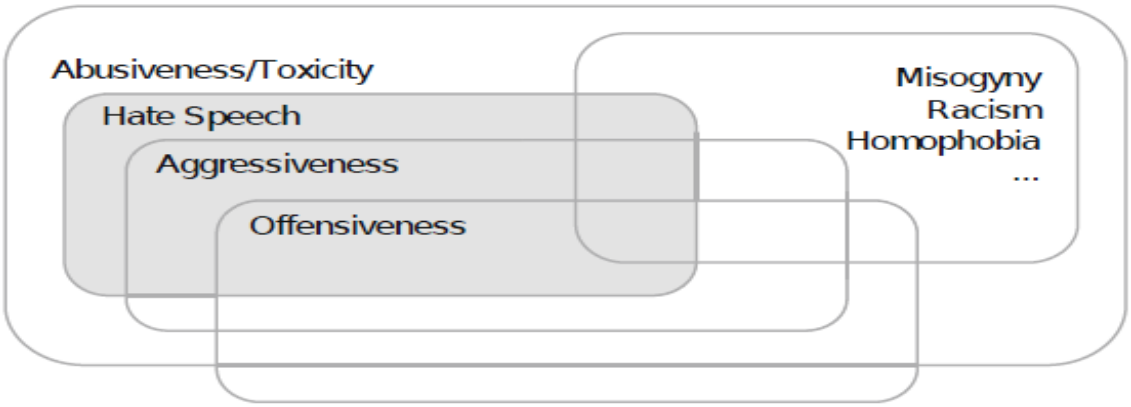


Fig. 2: Diferentes formas de discursos violentos. Fuente: Poletto, et al. (2021)[5]

Como ya detallamos en los apartados sobre Aprendizaje Supervisado y NLP, los modelos de Inteligencia Artificial se entrenan en base a anotaciones de textos llevadas a cabo por humanos. A pesar de que el campo se desarrolla con gran velocidad, un problema inherente a las anotaciones humanas son los desacuerdos sobre qué debe ser definido, como "toxicidad" o "discursos de odio".

Sang y Stanton (2021) comprobaron que la edad y la personalidad son factores que generan divergencias entre anotadores conduciendo a variaciones significativas sobre la tipificación del odio impactando negativamente en las generalizaciones de los modelos. Investigaciones anteriores señalan también en el género (McClelland & Hunter, 1992), la etnia (Cowan & Hodge, 1996) o el peso asignado a la idea de libertad de expresión (Cowan, et al., 2002) influencias significativas en la valencia asignada por anotadores humanos.

En términos comparativos, una de nuestras ventajas al medir toxicidad en lugar de discursos del odio es que, según informó Jigsaw (2020), el acuerdo entre anotadores sobre qué mensajes "expulsan sujetos de una conversación" es mayor que el que puede existir alrededor de otras variables más abiertas a interpretación como "abuso" o "discursos de odio". Este punto redunda en resultados más homogéneos y fidedignos, es decir, mejores performances para los modelos.

2.2.4. Cómo funciona la moderación de comentarios en YouTube

Para el caso de YouTube, es especialmente importante que separemos moderación de comentarios de moderación de contenidos, la relativa a los videos en sí. En lo que respecta a la moderación de contenidos, al igual que la mayoría de otras plataformas, YouTube utiliza un sistema mixto entre sistemas de IA y moderadores humanos para anular contenidos -videos y comentarios en conjunto- que no cumplan con sus

[5] Los puntos señalados por el gráfico de Poletto et al. son, del lado izquierdo (de arriba hacia abajo): Abusos/Toxicidad, Discursos de odio, agresividad y ofensividad. Del lado derecho: misoginia, racismo y homofobia.

lineamientos de comunidad (YouTube, 2020). Sin embargo, en el caso de los comentarios depende de los propios usuarios reportar los comentarios que consideren ofensivos, para que luego un moderador humano tome una decisión ad hoc.

En relación con lo señalado en la introducción, las tasas de remoción de comentarios y de retroalimentación a los infractores ubican a YouTube en el quinto y sexto puesto respectivamente sobre seis plataformas relevadas (Comisión Europea, 2021). Mientras, para con los chats en vivo, la plataforma ofrece un abanico amplio de posibilidades de edición (borrado de mensajes, bloqueos temporales, ocultar a ciertos usuarios del chat o moderación en vivo), en el caso de los contenidos asincrónicos, que hoy en día significan la mayor parte del tráfico de la plataforma, la posibilidad es solo una: remover comentarios (YouTube, 2021).

De acuerdo con el beneficio algorítmico que un gran volumen de comentarios conlleva para cada video es de suponer que muchos enunciadores encuentren más incentivos para evitar moderaciones que para dedicar mayores esfuerzos a la tarea. Dada la naturaleza de las operaciones de moderación, las investigaciones con base empírica sobre comentarios censurados son una difícil empresa para la comunidad científica, puesto que es preciso repetir el *crawling* a través del tiempo para determinar qué comentarios fueron dados de baja. Aun así, en los últimos años algunos investigadores han experimentado con algunas mecánicas al interior del opaco proceso de moderación.

A partir de la presidencia de Donald J. Trump se instaló en algunos sectores la idea de que los grupos de derechas eran más censurados en redes sociales que los de otras ideologías (Kamisar, 2018). Sin embargo, investigaciones recientes sugieren imparcialidad política en los actos de moderación al interior de YouTube. De acuerdo con el relevamiento, sobre 258 videos de diferentes tendencias políticas Jiang, et al. (2019) encontraron que, a pesar de que los videos con contenido de derechas reciben más revisión que los relativos a otras ideologías, ninguna evidencia sustenta la idea de un sesgo ideológico en la moderación de contenidos.

En cambio, sí es más probable que un video reciba “moderación pesada” cuando:

a- el canal tiene una “adherencia partidaria extrema”,
b- el contenido del video es probado falso,
c- se verifica presencia de discursos de odio en los comentarios.

Por fuera de YouTube, Ruiz Caballero et al. (2011) sugieren que la moderación de la participación de los usuarios en sitios de noticias conduce a un diseño de dos modelos de discusión pública claramente delimitados por parte de los encargados de la moderación. Por un lado, las “comunidades de debates” donde ciudadanos con ideas a veces opuestas confrontan en términos respetuosos, por el otro, “comunidades

homogéneas", que también podrían ser definidas como diálogos de sordos, donde predominan los contenidos sentimentales y el debate argumentativo es menor.

Un resultado común a los dos modelos de comunidad es que una vez instaurados sistemas de moderación, la abrumadora mayoría de los comentarios acaban coincidiendo con la línea editorial del medio donde fueron publicados. De replicarse esta conclusión a través de otras plataformas, las políticas de moderación podrían tener un efecto no deseado: la uniformidad de los discursos.

3. Acercamiento al objeto

3.1. Hipótesis y preguntas

Nuestra hipótesis inicial es que aumentos relativos en la toxicidad en los comentarios tienen su contraparte en un aumento en las vistas de los videos o las métricas de interacción (comentarios, "me gustas" e interacciones entre comentarios).

Hablamos de aumentos "relativos" al canal refiriéndonos tanto a la comunidad que se genera en torno a un enunciador como a una ubicación temporal en el que cada video está en línea con el anterior y el siguiente. Nuestra hipótesis se apoya en dos afirmaciones: una del sentido común y otra académica.

La primera es una idea harto conocida sin la suficiente base experimental: "No existe la mala publicidad, solo la publicidad". Si esa afirmación fuese cierta en el contexto estudiado, entonces un aumento en el número relativo de toxicidad y discursos del odio redundaría en mayor visibilidad para la mayoría de los enunciadores y ninguna cantidad de toxicidad severa vertida en los comentarios constituiría una "mala publicidad" para los youtubers. Incluso podría impulsarlos en la empresa por aumentar la propia visibilidad en la plataforma, constituyendo un posible incentivo para la permanencia y profundización de discursividades violentas en los comentarios.

La segunda fuente actuará como puente entre dos de nuestras hipótesis. En "Hegemonía y estrategia socialista" Ernesto Laclau y Chantal Mouffe (1987 [1985]) señalan que toda forma de sentido, remite, se le reconozca o no, a una dimensión antagónica.

El texto hereda de dos autores sobre los que nos gustaría puntualizar antes de continuar. Primero, de Saussure (1991 [1915]) las ideas sobre la naturaleza sistémica del lenguaje y el modelo de significación diferencial: el sentido sólo se daría al interior de un sistema de diferencias, y el signo lingüístico sería arbitrario en la relación entre sus partes. Al final, el sentido de un elemento solo dependería de la relación que entable con todos los otros elementos dentro del mismo sistema de significación, y por

eso es que estudiaremos los videos con más activación tóxica en el marco de las cuentas, en la búsqueda de captar aquellos elementos diferenciales. En segundo lugar, de Wittgenstein (1988 [1953]) toman la idea de los juegos del lenguaje y el carácter performativo de los discursos, que, además, separa a la realidad de un mero estatus mental.

Para esta ontología política posestructuralista que coloca al lenguaje en el centro de su argumentación, el antagonismo opera como condición de posibilidad de todo sentido y toda acción social. Lo que importa a fines de nuestra investigación es la lógica de un "nosotros" opuesto a un "ellos" sin la que sería imposible aprehender la naturaleza de "lo político". En un texto más contemporáneo, Mouffe junto a Íñigo Errejón retoman el tema de la forma adversarial del sentido con una definición más cercana a nuestro problema:

> …la toma de posición, de lealtades colectivas, incluye siempre esa sensación de solidaridad que se crea entre gentes que no necesitan conocerse personalmente, o ser amigos personalmente, para sentirse parte de la misma cosa. Eso tiene que ver siempre con afectos que si se niegan en la política escapan a otros lugares. Por ejemplo, el fenómeno de los equipos de fútbol; la pasión que se tiene, incluso entre la gente a quien no le gusta demasiado el fútbol, es la pasión amigo-enemigo, es el sentirte parte de un mismo colectivo.
> (Mouffe y Errejón, 2015, p. 54-55).

¿Podrían estarse creando colectivos de fans gracias a una mayor visibilidad que tendría entre sus causas a estos intercambios tóxicos? ¿Será que esta lógica antagónica favorece la cohesión al interior de los colectivos como señalan Laclau y Mouffe? ¿Es aplicable esta matriz posmarxista ideada para identificar grupos políticos en pugna para interpretar el vínculo entre youtubers que comen empanadas de shampoo y sus seguidores?

Las referencias a Saussure, Wittgenstein abren la puerta a otra de nuestras preguntas: ¿está condicionada la correlación entre toxicidad severa y mayor visibilidad por juegos del lenguaje que diferencien a nuestros 12 youtubers?

Ya repasamos la importancia de la creación de comunidades activas para la trascendencia de un enunciador en YouTube. Al igual que las hinchadas de los equipos llegan a matarse en la forma agonista de lucha entre adversarios, queríamos indagar si los seguidores de algunos de los youtubers más populares del país podían encontrar en la violencia verbal una forma similar de concentrar y aumentar los alcances de sus comunidades de pertenencia.

En la pregunta acerca de la relación entre el contenido de los videos y sus niveles de toxicidad nos basaremos en la categorización de enunciadores risibles que Fraticelli (2021) elaboró a partir de aportes de Genette (2005 [2002]), Eco (1996) y Steimberg, (2001).

En resumen, los enunciadores risibles se dividirían en cuatro:

- **Cómico**: el enunciador como blanco de la burla tiene una relación inferior con respecto a quien ríe, el enunciatario. Quien es degradado despierta la risa en tanto rompe con un orden social preestablecido por su desconocimiento. Un ejemplo serían Los 3 Chiflados, nos reímos *de* ellos.

- **Humorístico**: enunciador y enunciatario establecen una relación de simetría burlándose de sí mismos en lo que suele ser un acto de reflexión. Para que se mediatice, este enunciador debe ser más que individual, así, Woody Allen invocaba a reírse *con* él a otros miembros de clases medias altas, urbanas, filo-progresistas y sobre escolarizadas.

- **Bufón**: degrada figuras poderosas mostrando sus vicios y debilidades. Por arriesgarse a decir lo que los demás no se atreven, el enunciador carga una capa de heroicidad. Una buena parte del humor feminista (desde Malena Pichot hasta Ali Wong) se vale de la bufonería a sabiendas de las reglas trasgredidas.

- **Bromista**: Aquí no hay heroicidad alguna, y las reglas que se quiebran no ponen a esta figura en riesgos serios; lo que hay es una invitación al enunciatario a ser cómplice de una travesura en la que el objeto de lo risible son normas o comportamientos usualmente aceptados. Dentro de esta forma enunciativa podríamos contar a buena parte del humor “Apto para Todo Público”.

Nos valdremos de estas tipologías en los niveles de la enunciación hipermediática y mediática para identificar si la toxicidad severa afecta de forma diferente a los diferentes registros humorísticos, y, si existiese, cómo de diferentes son las respuestas tóxicas en función del contenido. A este fin, aplicaremos el modelo de análisis enunciativo y del humor de Fraticelli en la búsqueda de respuestas sobre las alteraciones en vistas e interacciones.

Creemos que este cruce entre miradas cuantitativas y cualitativas es el mejor método para estudiar saltos de escala, en la línea de Carlón (2016) entendemos que es en la circulación donde operan los saltos de escala, y que los aportes de la semiótica de redes son decisivos para caracterizar apropiadamente fenómenos que inicialmente podemos identificar a través de la estadística y la visualización computarizada.

La última de nuestras preguntas es: ¿reciben las enunciadoras femeninas la misma cantidad de toxicidad que los masculinos?

Esta cuestión se sitúa en la línea de lo descubierto por Wotanis y McMillan (2014) en el estudio de batallas de rap norteamericanas y por Döring y Mohseni (2020), para una amplia selección de marcos metacomunicativos. En ambos casos, una selección de enunciadoras femeninas recibió una cantidad desproporcionadamente mayor de discursos de odio que sus pares masculinos.

Nuestro trabajo es complementario al de las investigaciones citadas, a la luz de tres diferencias fundamentales:

a- la medición de una variable diferente, toxicidad en lugar de discursos de odio.
b- una muestra mucho más numerosa que los 20 videos de Wotanis y McMillan y los 8.000 comentarios de Döring y Mohseni.
c- el recorte sobre contenidos producidos al interior de una cultura latina, con características muy diferentes a las realidades norteamericanas y alemanas, respectivamente.

Ceñimos nuestra pregunta a enunciadoras femeninas y masculinos en vistas a que ninguno de los 12 youtubers se identifica con una tercera opción de género. Sin embargo, rechazamos las posiciones binarias y creemos en la importancia de incorporar preguntas relacionadas a toda la diversidad de géneros en futuros estudios sobre toxicidad y violencia en redes.

3.2. Metodología

3.2.1. Perspective

Perspective es un algoritmo entrenado para identificar toxicidad en comentarios de redes sociales y otros foros de comentarios, desarrollado por Jigsaw, empresa subsidiaria del grupo Alphabet (dueño de Google), y accesible de manera gratuita para investigadores y otros desarrolladores. Dado un comentario de un usuario en una red social o foro, el algoritmo devuelve un puntaje de entre 0 y 1, siendo 0 la ausencia de toxicidad y 1 la prevalencia absoluta. En base a esos valores, y en la línea de la mayoría de las investigaciones del campo, establecimos un umbral en 0.5 a partir del cual un comentario sería considerado tóxico o no.

Perspective ha probado su robustez en estudios sobre una variedad de fenómenos: en la detección de toxicidad sobre tweets de influencers (Sprejer & al, 2021), en la

clusterización de usuarios tóxicos (Mall & al., 2020), en identificación de relaciones entre seguidores de partidos políticos opuestos (Wu & Resnick, 2021).

Al ser propiedad de un privado, el código de Perspective no es abierto, y, aunque podamos inferir cómo funciona, es imposible asegurarlo al 100%. Aun así, dado que modelos basados en Transformers como BERT son hoy en día el estado del arte, lo más probable es que se basen en estos esquemas que ya repasamos en el marco teórico.

Aunque en este momento no exista otra herramienta más avanzada para tratar un problema de las características del nuestro, también debemos señalar sus limitaciones.

A pesar de los variadísimos orígenes étnico-lingüísticos de los investigadores citados en el tercer párrafo de este apartado, todos y todas han decidido estudiar casos en inglés. Lejos de ser una casualidad, esta tendencia se enmarca al interior de la concentración general del NLP en torno a la lengua más usada y con más recursos en Internet, reforzada en el caso de Perspective por ofrecer análisis más finos y variados. Un ejemplo: las métricas que en castellano y otros idiomas son “experimentales” en inglés ya pasaron esa etapa y se encuentran en pleno funcionamiento.

Luego, en el análisis minucioso pueden encontrarse otras falencias más independientes de la lengua, cuya causa podríamos ubicar en la difícil tarea de señalar palabras clave que pueden ser muy azarosas en comentarios en redes sociales. Un caso con el que nos topamos durante nuestro análisis fue el de Melina Vallejos, una youtuber que apunta a un segmento infanto-juvenil especialmente femenino, con una enunciación que podría ser calificada de inocente; por eso, fue una sorpresa encontrar que el video en el que presentaba a su nueva perrita tenía altísimos niveles de toxicidad relativa. Era la palabra “perrita” la que hacía saltar las alarmas de Perspective –ubicando la valencia por encima del 0.7 en casi todos los casos. Las observaciones fueron descartadas.

En resumen, nuestra tecnología de medición es la más avanzada hasta el momento, pero incluso más allá de Perspective, todavía queda mucho campo por recorrer en la forma en que los humanos buscamos entender el lenguaje por vía computacional.

En principio, recogimos la valencia para seis medidas de toxicidad diferentes. Las siguientes definiciones son traducciones de la web de Jigsaw (2020):

> **Toxicidad**: Lenguaje rudo, irrespetuoso o poco razonable que tiende a expulsar sujetos de la conversación.
> **Toxicidad severa**: Un comentario muy odioso, agresivo, irrespetuoso o de otro modo muy probable que haga que un usuario abandone una discusión o

renuncie a compartir su perspectiva. Este atributo es mucho menos sensible a formas más leves de toxicidad, como comentarios que incluyen usos positivos de malas palabras.
Ataque identitario (experimental): Comentarios negativos o de odio dirigidos a alguien por su identidad [de género, sexual, étnica, etc.].
Vulgaridad (exp.): Malas palabras u otro lenguaje obsceno o profano.
Amenaza (exp.): Describe la intención de infligir dolor, lesiones o violencia contra un individuo o grupo.
Insulto (exp.): Comentario insultante, incendiario o negativo hacia una persona o un grupo de personas.

¿Qué significa que las últimas cuatro métricas se encuentren en etapa experimental? Mientras que las primeras dos se encuentran en funcionamiento pleno, cuatro están aún en etapa experimental para el lenguaje castellano, por lo que sus resultados no deben suponerse precisos de antemano, y deben ser leídos con precauciones extra.

Teniendo en cuenta que la medida estándar, "toxicidad" a veces podía incurrir en interpretaciones casi victorianas, marcando como tóxicos comentarios solo por contar con adjetivos calificativos o insultos menores, decidimos quedarnos con la medida de toxicidad severa como principal a nuestros fines. Por otra parte, teníamos un alto interés en identificar qué porción de los discursos tóxicos atacaban a grupos protegidos. A ese fin, también consideramos la métrica de ataque identitario.

3.2.2. Análisis de correlación de variables

Durante la etapa de análisis cuantitativo, nos valdremos del Coeficiente de Correlación de Pearson para determinar la relación asociativa entre variables. El análisis de correlación busca indagar acerca del grado de cercanía entre dos variables.

Un ejemplo de dos variables correlacionadas sería el del consumo de helado y los cambios de temperatura. En general, a medida que sube la temperatura en una región geográfica, también aumenta el consumo de helado en esa misma área; aquí decimos que estamos ante una correlación positiva. Al pasar el verano, las temperaturas bajan y el consumo de helado también tenderá a descender, aquí la correlación puede ser igual de fuerte que en el final de la primavera, solo que en este segundo caso hablaremos de correlaciones negativas.

El riesgo ante el que no debemos caer como investigadores es la falacia de correlación, es decir, la lectura en clave de causación donde los instrumentos sólo nos indican correlación. Si así fuera, para nuestro caso del helado donde solo tuvimos en cuenta dos variables, tendríamos posibilidades de afirmar que es el aumento del consumo de

helado el que hace subir las temperaturas en una región determinada. Por más ridículo que pueda parecer este planteo, confundir correlación con causación sigue siendo muy común, especialmente en ciencias sociales. En este estudio nos contentamos con probar o refutar que dos variables evolucionan en cercanía.

Dado un conjunto de observaciones (X_1, Y_1), (X_2, Y_2), … , (X_N, Y_N) para las dos variables a estudiar, la fórmula de la correlación de Pearson es:

$$r = \frac{N \sum xy - \sum (x)(y)}{\sqrt{N \sum x^2 - \sum (x^2)][N \sum y^2 - \sum (y^2)]}}$$

Fig. 3: Fórmula de la correlación de Pearson.

Siendo:

r: el coeficiente de correlación
N: el número de pares de puntajes.
∑xy: la sumatoria de los productos de puntajes pares.
∑x: la sumatoria de los puntos en x (variable independiente).
∑y: la sumatoria de los puntos en y (variable dependiente).
$\sum x^2$:la sumatoria del cuadrado de los puntos en x.
$\sum y^2$:la sumatoria del cuadrado de los puntos en y.

La centralidad de nuestros análisis rondará en torno a dos variables: en primer lugar, la "toxicidad severa" determinada bajo un umbral de 0.5. Segundo, el puntaje Z de las vistas e interacciones por video, calculado tomando cada youtuber como unidad.

El puntaje Z es una medida que determina la cantidad de desviaciones estándar que un dato se aleja de la media. Esta fórmula nos permite determinar qué tan usual es un dato en torno a la distribución general.

$$z = \frac{x - \mu}{\sigma}$$

Fig. 4: Fórmula del puntaje Z.

Siendo:

σ: desviación estándar de la población.
μ: la media de la población.

Deberíamos entender por Z "la desviación de la media por autor de…" cada variable sobre la que calculamos este puntaje. Al comparar niveles de toxicidad contra el

puntaje Z de vistas, podemos centrarnos en las variaciones de las visualizaciones antes que en el número “en crudo”, y cuánto afecta la toxicidad severa a esas variaciones.

En la etapa de análisis veremos valores del puntaje Z de -1.6, 2 o 3, entre otros. ¿Pero qué significan esos puntajes? Para entenderlo, observemos y desarrollemos el siguiente gráfico.

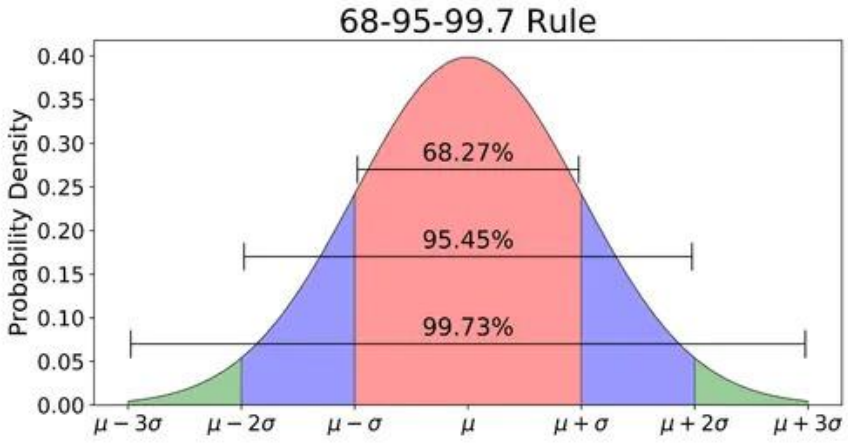


Fig. 5: Interpretación del puntaje Z sobre una distribución normal. Fuente: McLeod (2019).

Sobre la presunción de una distribución normal de los datos, cuando el puntaje Z de una variable supera el -1 o el 1 ese dato se encuentra en el 32% menos habitual de la distribución. Es decir que si la variable vistas de un video tuviera un puntaje Z igual a 1, ese video estaría en el tercio más visto entre los del youtuber correspondiente. Si Z superase 2 o -2, entonces estaríamos ante el 5% de las observaciones menos frecuentes, un valor muy atípico. Incluso más, para valores mayores de entre -3 y 3, sobre una distribución normal estaríamos viendo una observación con el 0.03% de probabilidades, muy difícilmente repetible.

¿Para qué nos sirve este indicador? Para medir alteraciones en las variables al interior de una cuenta (o comunidad). Si un video tuviese, por ejemplo, un puntaje Z en comentarios de 2 puntos, la probabilidad de ocurrencia de ese evento sería muy poco habitual, de apenas un 5%. Esta medida de probabilidad es útil para contextualizar los datos absolutos en torno a lo que sería normal para cada cuenta. Y decimos para cada cuenta porque los puntajes Z se calcularon por enunciador; 100 mil vistas de diferencia pueden ser una variación enorme para los canales menos visitados, pero no así para aquellos donde la media de vistas suele ser más elevada.

Mientras que muchas veces los problemas de ciencias naturales se adaptan mejor a los modelos, los problemas sociales suelen revestir una complejidad diferente y a menudo están influidos por muchas más variables no mensurables -o que no se pensó en medir- y que para la estadística serían consideradas “ruido”. Por ello, es preciso no

replicar ciegamente los modelos de un conjunto de ciencias al otro sin la debida pregunta por las adaptaciones necesarias.

De acuerdo con lo recomendado por Cohen (1988), los umbrales de qué consideraremos correlaciones estadísticamente significativas deberán estar adaptados al estado del arte del tratamiento de problemas sociales. Ahora: ¿cuándo estamos ante una correlación estadísticamente significativa en ciencias sociales? Cohen zanjó las discusiones de la siguiente manera: un R en torno a 0.1 indica una correlación baja, alrededor de 0.3 vemos una relación mediana entre variables y por encima de 0.5 estaríamos ante correlaciones de fuerte intensidad.

Fig. 6: Guía para leer una matriz de correlaciones en base a Cohen (1988).

Si bien en los últimos años han surgido cuestionamientos a esta clasificación, como el de Bakker, et al. (2012) que apuntan contra los sesgos de publicación de las revistas científicas, la categorización de Cohen no parece encontrar aún competencia que la destrone.

Además, del R, seguimos la línea más usual en investigaciones de este tipo, considerando válidas las correlaciones que cuenten con un *p*-valor menor a 0.05. El *p*-valor es la probabilidad, bajo un modelo estadístico determinado, de observar las características actuales en los datos o aún más extremas.

Un ejemplo: si en un experimento en el cual arrojamos un dado (que suponemos no adulterado) se observa que sacamos diez veces seguidas un 5 o un 6, la probabilidad de que esto ocurra es muy baja (y esta baja probabilidad bien podría llevarnos a cambiar nuestra creencia de que el dado no esté adulterado). Si la característica que observamos es la media del valor de los dados (en este caso, por encima de 5 para diez observaciones) obtendremos un p-valor bastante bajo. Una práctica común es utilizar un umbral previamente determinado y, cuando el resultado de la ecuación arroja un valor menor a éste, decimos que estamos ante resultados “estadísticamente significativos”. El nivel de significancia representa la probabilidad de que la hipótesis nula (el dado no está adulterado) sea cierta, cuando *p* es menor a 0.05, esa probabilidad es menor al 5%.

Entonces, si para nuestro caso una hipótesis nula fuera “no hay correlación entre la desviación estándar de toxicidad y X variable” (vistas o interacción). Un *p* menor a 0.05 permitiría rechazar la hipótesis nula con, al menos, 95% de seguridad. La fórmula para el *p*-valor es la siguiente:

$$z = \frac{\hat{p} - p0}{\sqrt{\frac{p0\,(1 - p0)}{n}}}$$

Fig. 7: Fórmula del *p*-valor.

Siendo:

p̂: la proporción de la muestra.
p0: la población asumida de la proporción en la hipótesis nula.
n: el tamaño de la muestra.

Para terminar con este apartado, nos gustaría dar cuenta de un fenómeno que observaremos en la etapa de análisis.

A partir de los trabajos de Edward P. Simpson, el matemático norteamericano Colin Blyth (1972) bautizó como “la paradoja de Simpson” al tipo de distribución de los datos en el que la tendencia al interior de un grupo desaparece o se modifica al combinar esos datos con los de otros grupos. La paradoja de Simpson, especialmente usual en ciencias sociales de acuerdo a lo descrito por Wagner (1982), es una causa común de confusiones por desafiar la intuición de los observadores.

Por ejemplo, en Rogier, et al. (2013) se describió cómo dos relaciones negativas entre las dosis de un medicamento y las posibilidades de recuperaciones entre pacientes podrían leerse como relaciones positivas si no se separasen los grupos por género. Este ejemplo traído de la psiquiatría revela la importancia de ciertas divisiones categóricas a la hora de analizar correlaciones.

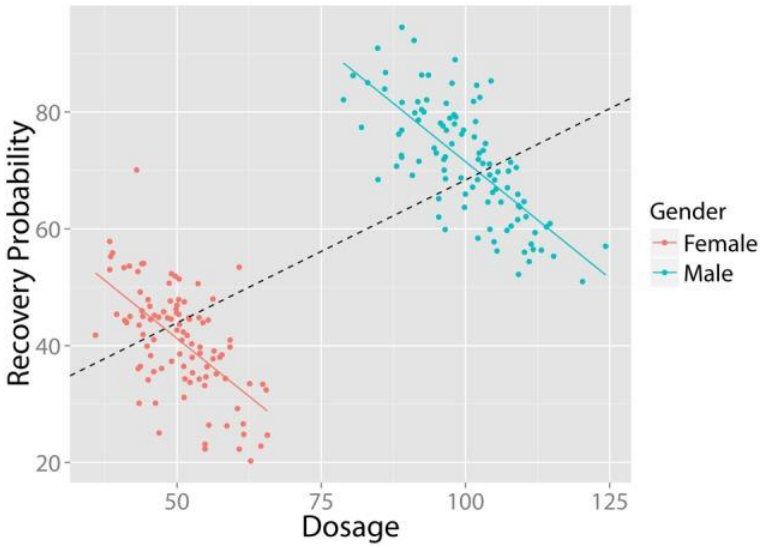


Fig. 8: Ejemplo de la paradoja de Simpson extraído de Rogier, et al. (ibidem, p. 2).[6]

3.2.3. Modelo de análisis enunciativo y humor.

El modelo de análisis del que nos valdremos es heredero de nociones veronianas. Verón toma de Charles S. Peirce (1994 [1901]) la tríada interpretante-signo-objeto y la adapta en operaciones-discurso-representaciones. Además, incorpora el concepto de "semiosis" que Peirce definía como relación necesariamente triádica, en la que el signo o representamen, el objeto y el interpretante, son manifestaciones de las categorías faneroscópicas de la primeridad, segundidad y terceridad. El interpretante, a su vez, es un signo y conforma otra tríada, lo que da lugar a una semiosis infinita.

Verón adapta la semiosis peirceana y llama semiosis social a la dimensión significante de los fenómenos sociales, una red interdiscursiva infinita que descansa sobre dos hipótesis que se contrapesan: toda producción de sentido es necesariamente social, pero, a su vez, todo fenómeno social es un proceso de construcción de sentido.

El esquema de circulación de Verón presenta un eslabón de la cadena de la semiosis al que entramos a través de un producto específico, que es nuestro discurso de referencia D_i. El D_i es la configuración empírica, material, de signos que puede ser sometida a análisis (por ejemplo, un video de YouTube). Es entre las instancias de producción y reconocimiento, cualitativamente diferentes, que circulará el sentido en las sociedades.

[6] El eje de las Y se lee como "probabilidad de recuperación", el de las X como "dosis", mientras que los matices de la derecha indican que los puntos celestes son para el sexo masculino, mientras que los rojos para el femenino.

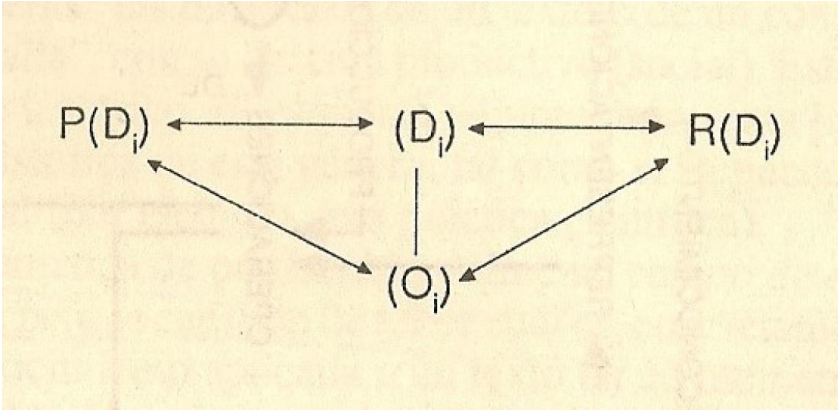


Fig. 9: La doble triada veroniana. Fuente: Verón (1987, p. 132)

Siendo:

(D_i): El discurso de referencia.
P(D_i): Condiciones discursivas de la producción de (Di).
R(D_i): Condiciones discursivas del reconocimiento de (Di).
(O_i): Objeto del discurso de (Di).

En el marco teórico explicamos cómo la circulación ha devenido hipermediática en nuestras sociedades contemporáneas y los contenidos han rebasado las fronteras de las redes sociales o los medios masivos en movimientos que llamamos intersistémicos. Nuevos dispositivos son atravesados por el sentido y la adaptación de la teoría veroniana a los fines del análisis hipermediático representa un desafío. Para dar cuenta de la forma en que esa materialidad significante diseña el vínculo entre las instancias de producción y reconocimiento es que haremos uso del modelo de análisis enunciativo y del humor de Fraticelli (2021).

El autor asume que los análisis internos -como el que arroja Perspective- no son suficientes para explicar la producción de sentido ni lo risible. No solo las condiciones de producción y reconocimiento deberán ser tenidas en cuenta a la hora del análisis: el diseño, que, lejos de ser neutral, habilita (o no) diferentes formas de interacción constituirá una de las capas analíticas del modelo. Veamos una esquematización:

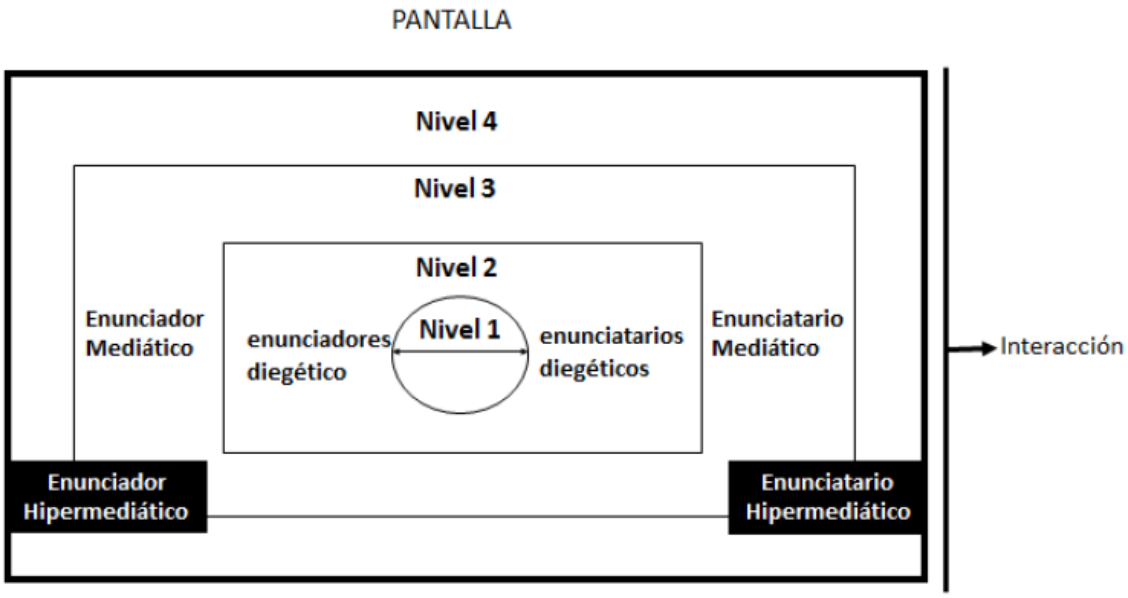


Fig. 10: Esquema del dispositivo enunciativo de las redes sociales hipermediáticas. Fuente: Fraticelli (2021, p. 122).

El nivel 4 corresponde a la escena enunciativa compartida por Enunciador y Enunciatario Hipermediáticos. El primero se define como la condensación de la figura del enunciador y el propietario de la cuenta, el autor "no es una persona sino una figura de contacto entre el texto y lo extratextual" (Fraticelli, *ibidem*, p. 123). Incorporar cualidades del enunciador risible como su nombre, foto de perfil, descripción de cuenta, entre otros, contextualiza los mensajes.

El nivel 3 es el de la escena de los Enunciadores y Enunciatarios Mediáticos de cada publicación. Sus propiedades "dependerán de los lenguajes, dispositivos técnicos, modalidades y clasificaciones discursivas (tipos, géneros, estilos, etc.) puestos en juego" (Fraticelli, *ibidem*, p. 123). Por ejemplo, un video puede asemejarse al lenguaje cinematográfico o podría haber una transmisión en directo como sucede con la televisión. La diferencia entre los niveles 4 y 3 es sobre todo temporal. En el primer nivel se construye un contrato de lectura de largo plazo, mientras que el segundo puede variar de publicación en publicación.

El nivel 2 consiste en las escenas enunciativas dadas en cada publicación. Allí veremos personajes, música, efectos de edición que, dentro de la diégesis, funcionan como múltiples enunciadores y enunciatarios.

Por último, el nivel 1 estará constituido por "las escenas enunciativas de los enunciados producidos por los enunciadores diegéticos" (Fraticelli, *ibidem*, p. 124). El autor da el ejemplo de una cita al interior de un video, allí, la persona citada sería un enunciador de primer nivel.

Este será el nivel enunciativo de nuestro análisis, donde prestaremos especial atención a los marcos metaenunciativos risibles: lo cómico, el humor, la bufonería y la broma.

Durante nuestro análisis, distinguiremos caracterizaciones hipermediáticas, las que se producen a nivel cuenta, de las mediáticas o diegéticas (niveles 3 y 2, respectivamente). Así, veremos que un youtuber puede tener en general una enunciación hipermediática bromista, generando risa desde la ruptura de parámetros socialmente aceptados, pero en el nivel diegético podemos estar ante una escena cómica si para con sujetos mediáticos ajenos a esas reglas sociales se establece una risa despreciativa con mirada descendente. Como señalara Verón (2004), estos análisis sólo pueden ser comparativos, No trabajan con un solo soporte sino en un universo de competencia dentro del cual buscamos identificar qué diferencia a las enunciaciones, a las maneras del decir.

Para terminar, incluimos un análisis del sistema de cuenta de Fraticelli (2019). Si durante toda la historia de la comunicación los enunciadores se retroalimentaron de las instancias de reconocimiento -por ejemplo, a través del rating o los focus groups- la novedad de la era virtual es la falta de mediación entre los colectivos de receptores y quien emite el mensaje.

Solo mediados por la plataforma, "los comentarios llegan 'directamente' a los propietarios de las cuentas y operan como condicionamiento productivo" (Fraticelli, *ibidem*, p. 56). La relación de los colectivos y las cuentas puede leerse como un sistema de producción discursiva y graficarse de la siguiente manera:

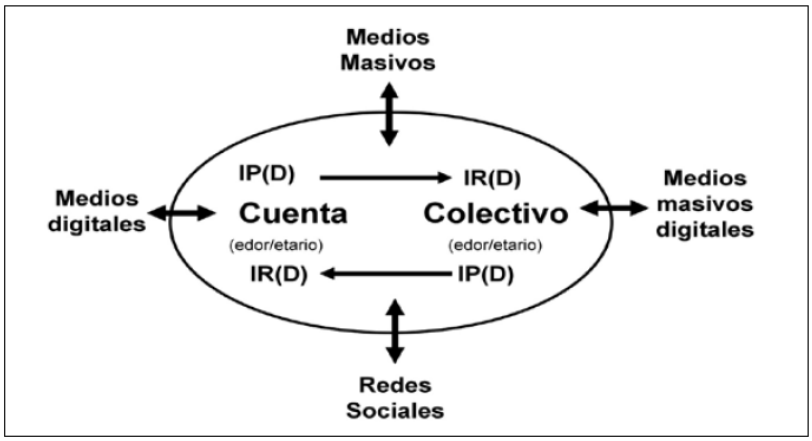


Fig. 11: Esquema del sistema cuenta. Fuente: Fraticelli (ibidem, p. 57).

En palabras de Fraticelli:

> El sistema de cuenta se compone de la cuenta y de su colectivo individuado. Ambos ocupan alternativamente la instancia de producción IP(D) y reconocimiento IR (D). Y, a

> su vez, son enunciadores y enunciatarios debido a que los discursos intercambiados los construyen de esa manera. El entorno del sistema lo forman otros sistemas de las redes sociales y las instituciones mediáticas (medios masivos, medios masivos digitales y medios digitales). Con ellos mantiene relaciones de interpenetración.
> (Fraticelli, *ibidem*, p. 57)

¿A qué prestaremos atención cuando observemos el sistema cuenta? Especialmente a movimientos en conjunto que las variables atadas a los receptores verifiquen en períodos temporales coincidentes.

3.2.4. Corpus

Como señalamos en la introducción, pusimos el foco sobre la sección Comedia en YouTube Argentina. Al tratarse de un vastísimo universo, seleccionamos las 12 cuentas más populares según la plataforma Noxinfluencer al 17 de marzo de 2021, descartando aquellas que no contaran con una mayoría de contenido risible a pesar de estar incorporadas en la sección.

De esta manera, nos aseguramos una uniformidad en cuanto al castellano en su variedad argentina como lengua común y una muestra de un tamaño suficientemente extenso como para generalizar sobre poblaciones mayores.

Las cuentas seleccionadas fueron: AleVeraOficial, FFran Gomez, Guille Aquino, Hecatombe Produccion, Marito Baracus, Nico Villa, PassThor, Rodriguez Galati, Tincho Ruiz, Melina Vallejos, Nimu y Romi. Coincidentemente con lo observado por Tur-Viñes y González-Río (2019) en relación al mayor peso relativo de enunciadores masculinos sobre femeninos en las categorías de entretenimiento en YouTube, solo los últimos 3 son canales llevados a cabo por enunciadoras femeninas, mientras que en el caso de Hecatombe Producción se trata de un colectivo mayormente conformado por varones aunque con una mujer en un rol protagónico.

¿Tiene sentido recortar por países en una red donde se supone que las fronteras corresponden a los idiomas en lugar de a los estados-nación? Según Brodersen, et al. (2012) sí. Los investigadores concluyeron que la popularidad de los videos de YouTube está ceñida a intereses determinados por fronteras geográficas precisas, por lo que la promesa de una circulación libre de fronteras en Internet parece más un slogan que una realidad. Dado el desarrollo extendido y autónomo que han tenido en Argentina los productos culturales nacionales es de esperar que gran parte de la circulación ocurra entre habitantes del suelo argentino.

En cuanto al recorte temporal, definimos usar los equinoccios de otoño (20 de marzo) de 2020 y 2021 con principio y fin de nuestro análisis. Nuestro recorte elude tanto los cambios generados por la pandemia de COVID-19, como las arbitrariedades del

calendario gregoriano, sin dejar de atender variables estacionales al prolongarse durante 365 días.

La recolección de datos se llevó a cabo a través de la adaptación de YouTube Data Tools el crawler de la iniciativa Digital Methods de la Universidad de Ámsterdam (Rieder, 2015). El problema ante el que nos encontrábamos era que mientras la API de YouTube obligaba a descargar los datos video por video con una cuota muy baja como para cumplir nuestras expectativas, el sistema de Rieder requería una descarga por ID de video y sortear un Captcha en cada operación. Con 876 videos en nuestro recorte, la indexación video por video parecía titánica.

Así fue como automatizamos el proceso de entrada de ids, dividiendo los resultados del crawler en carpetas por autor. El resultado: un dataset que inicialmente constaba de 1.562.810 comentarios (filas) y 19 columnas con información sobre los comentarios y los videos donde habían sido vertidos. Luego de limpiar comentarios duplicados y errores, el conjunto de datos final resultó en 1.280.492 entradas.

Inicialmente, para cada comentario contábamos con columnas tan variadas como: id del video, título del video, canal, fecha de publicación del video, número de vistas del video, likes al video, dislikes al video, comentarios totales, réplicas por comentario, likes por comentario, fecha de publicación del comentario, nombre del autor, texto, id del canal del autor, url del canal del autor, opción binaria por si es respuesta o no, nombre de a quien se le respondió. Además, creamos la variable del género del youtuber, y adjuntamos al set de datos producto del análisis de Perspective.

Nuestro corpus es lo suficientemente amplio y heterogéneo como para ser objeto de muchas investigaciones, entre las que imaginamos: estudios de sistemas de red entre usuarios que comentan, nuevas pruebas a Perspective sobre un mismo conjunto de datos, miradas microhistóricas del desarrollo de una sola de las 12 cuentas durante el primer año de la pandemia. El set de datos queda disponible en GitHub para su aprovechamiento por la comunidad científica.

Para evitar entrar en más detalles técnicos, en el anexo de este trabajo ponemos a disponibilidad de la comunidad nuestro código en lenguaje PHP, a la espera de que futures investigadores no tengan que pasar las muchas semanas de elaboración que tomó en nuestro caso y puedan reutilizarlo en próximas investigaciones sobre comentarios de YouTube. Por lo demás, detallaremos algunos datos poblacionales del conjunto recolectado.

Youtuber	Suscriptores al 20/03/2021	Cantidad de videos	Cantidad de comentarios
Romi	2.680.000	140	212.981
Nico Villa	2.110.000	125	313.198
maritobaracus	1.680.000	54	73.837
Hecatombe Producciones	1.560.000	50	41.661
Melina Vallejos	1.220.000	9	18.171
Tincho Ruiz	1.030.000	91	90.509
PassThor	727.000	106	209.330
FFran Gomez	599.000	85	43.708
Guille Aquino	544.000	12	32.719
Rodriguez Galati	536.000	111	31.112
AleVera Oficial	476.000	27	37.155
Nimu	442.000	66	175.425

Tabla 1: Tabla de datos poblacionales totales de los 12 youtubers con más suscriptores en YouTube Argentina, sección Comedia.

Perspective puede devolver errores ante ciertos escenarios como comentarios vacíos, demasiado largos, determinados caracteres especiales o repetidos. Para sortear este obstáculo, a las herramientas tradicionales de limpieza de datos de Python sumamos pysentimiento, una librería diseñada por Pérez, et al. (2021) para facilitar procesos de minado de opiniones.

Las métricas[7] que consideramos -junto a sus etiquetados en el set de datos- fueron las siguientes:

[7] Para más información sobre qué y cómo mide YouTube en las métricas de vistas e interacción, recomendamos consultar la documentación oficial en https://developers.google.com/youtube/analytics/metrics

Métrica	Versión corta[8]	Indicador	Definición
Toxicidad severa	*sev_toxicity*	Toxicidad - Perspective	Un comentario muy odioso, agresivo, irrespetuoso o de otro modo muy probable que haga que un usuario abandone una discusión o renuncie a compartir su perspectiva.
Ataques identitarios	*ident_attack*	Toxicidad - Perspective	Comentarios negativos o de odio dirigidos a alguien por su identidad.
Vistas al video	*vistas*	Vistas - YouTube	El número de veces que se ha visto un video.
"Me gusta" el video	*likes*	Interacción - YouTube	El número de veces que los usuarios indicaron que les gustaba un video dándole una calificación positiva.
"No me gusta" el video	*dislikes*	Interacción - YouTube	El número de veces que los usuarios indicaron que no les gustaba un video dándole una calificación negativa.
Comentarios al video	*comentarios*	Interacción - YouTube	Los mensajes que los usuarios comentaron en un video.
"Me gusta" a comentarios	*likeCount*	Interacción - YouTube	Número de veces que los usuarios indicaron que les gustaba un comentario con una calificación positiva.
Réplicas a comentarios	*replyCount*	Interacción - YouTube	Los mensajes que los usuarios enviaron en respuesta a otro comentario.
Género del o la youtuber	*genero_yt*	Enunciador - El. Propia	Indicamos las enunciadoras femeninas con un "1" y los masculinos con un "0".

Tabla 2: Principales métricas para el análisis de correlaciones cruzadas.

Para tamizar la calidad de los análisis, nos valimos de la Precisión o Tasa de Verdaderos Positivos. Esta métrica divide los verdaderos positivos por sobre el total de positivos para dar cuenta de la exactitud de las clasificaciones. Así, comprobamos la tasa de verdaderos positivos para la métrica de ataques identitarios y, a pesar de encontrarse en etapa experimental, los resultados fueron muy alentadores: 71% de precisión en promedio.

[8] Acortamos los nombres de las variables a fines de una visualización más clara.

4. Análisis

Dividimos a los enunciadores hipermediáticos de acuerdo con los regímenes enunciativos que establece Fraticelli (2021). En vistas a que algunos youtubers publicaron cientos de videos en el plazo estudiado -y, de nuevo, analizar individualmente las 826 publicaciones sería tarea para otra tesis entera-, optamos por una caracterización por canal.

Aun cuando durante un año de publicaciones algunos youtubers hayan variado las formas de la enunciación entre video y video, nuestra generalización busca abarcar a las cuentas como conjunto. En resumen, los 12 youtubers más consumidos de la sección Comedia de Argentina se dividen en:

Cómicos (quienes se ríen de arriba hacia abajo): Nimu.

Humorísticos (quienes se ríen de forma horizontal): Rodriguez Galati.

Bufones (quienes se ríen de abajo hacia arriba, con heroicidad): Guille Aquino.

Bromistas (quienes se ríen por el placer de romper una regla, sin heroicidad): AleVera, Maritobaracus, PassThor, Romi, Nico Villa, FFran Gomez, Hecatombe Producciones, Melina Vallejos, Tincho Ruiz.

No obstante, aquí estamos solo ante el nivel hipermediático. En la fase cualitativa pondremos la lupa sobre la escena enunciativa mediática y diegética en la búsqueda de coincidencias discursivas entre publicaciones de diferentes autores.

4.1 Etapa cuantitativa

4.1.1. Experimentos con Perspective

Más allá de nuestros puntos en contacto con herramientas de las ciencias de la computación y nuestro interés por dar con procedimientos lo más eficientes posibles, una pregunta que en una investigación de ciencias sociales no puede dejarse de lado es si lo que mide la herramienta automática realmente se ajusta a su propia definición de toxicidad.

Pregunta especialmente necesaria cuando tratamos discursos risibles, que, muy a menudo, habilitan desde producción un juego donde las groserías y los insultos forman parte de la circulación. ¿Estamos verdaderamente ante discursos tóxicos cuando se cita un insulto intradiegético? ¿A quién ofendería o “tendería a alejar de la discusión” un comentario de estas características? Repasemos un conjunto de ejemplos para graficar la situación.

Para la cuenta maritobaracus, seleccionamos como exponente de mayor toxicidad el video “Chupa Pig (Temporada 1 Episodio 1)”. Este es uno de esos casos donde la obscenidad es una puerta abierta desde enunciación: los personajes están dibujados con penes en las caras, hay escenas de sexo en el fondo mientras se desarrolla la acción central, vemos referencias al semen e insultos severos y sexistas durante gran parte de la narración. Ahora, observemos diez de los comentarios clasificados como severamente tóxicos por Perspective:

Índice	**Comentario**
1	Tienen cara de pija xd
2	Jajjajaa Chupa pig
3	Paso el tiempo se ve un poquito deteriorada la chupa pig
4	Marito sos un crack bldo
5	Alguien Vio que El Chupa George xd Se estaba Cogiendo el dinosaurio..F por el dinosaurio
6	Es peppa pig pero chupa pig xd
7	@usuario mira vete al club de solo niñas mecas estúpida!!!
8	@usuario inapropiado para quien? Te mereces un récord al más pelotudo
9	Jajaja un portugués o brasileño sin cultura
10	Nigga what a fack xd

Tabla 3: Selección de diez comentarios marcados como tóxicos por Perspective para “Chupa Pig (Temporada 1 Episodio 1)”.

¿Cuántos de estos diez son realmente tóxicos? A nuestro entender, los últimos cuatro: el segundo, el séptimo, octavo y noveno. Aquí nos enfrentamos ante la dificultad de la clasificación binaria entre tóxico y no tóxico. Aunque quizás los comentarios séptimo, octavo y noveno sean indiscutiblemente tóxicos, por atacar a un individuo en particular o discriminar contra una nacionalidad, podría discutírsenos que en el contexto argentino el término “nigga” no tiene el mismo significado que en culturas angloparlantes y que la intención del usuario no fue publicar un comentario tóxico sino revelar sorpresa. A nuestro criterio, los puntos de contacto con una cultura global como la que se desarrolla en YouTube y el ataque a una minoría étnica marcan la pauta de que el décimo comentario de la Tabla 3 es efectivamente tóxico, pero el ejemplo no deja de señalar las dificultades de estudiar discursos violentos priorizando las formas sobre el sentido.

En pos de comprobar este punto de manera más particular, analizamos los comentarios de nuestra selección de seis videos de manera manual. Para los videos con más de 250 comentarios señalados como tóxicos, revisamos una selección aleatoria de 200, mientras que para el resto estudiamos la totalidad. De acuerdo a nuestro criterio, el nivel de precisión de Perspective para con su propia definición fue el siguiente:

Video	Canal	Comentarios con toxicidad severa detectada	Precisión de la toxicidad severa
Chupa Pig (Temporada 1 Episodio 1)	maritobaracus	943	32%
METAL GEAR SALE MAL PARTE 3	maritobaracus	21	57%
TROLEADA POR LEER COMENTARIOS	Nimu	212	34%
PRIMER STREAM EN YOUTUBE	Nimu	4	75%
Smoke consigue NOVIO y CJ se pone Celoso	AleVera Oficial	273	40,5%
La HISTORIA del Canal - GTA San Andreas	AleVera Oficial	10	70%

Tabla 4: Tasa de Verdaderos Positivos en detección de comentarios tóxicos para la selección de seis publicaciones que estudiaremos en la etapa cualitativa.

A primera vista, a más comentarios detecta el sistema basado en Transformers como tóxicos, más erra. Pero además -y sobre este punto profundizaremos en el análisis individual de las publicaciones-, también a más obscenidad proponga el productor, más falsos positivos recolecta Perspective. Solo a modo introductorio de entre los tres videos más tóxicos del cuadro, y considerando cada publicación sobre los promedios de toxicidad enunciada al interior de cada sistema cuenta, entendemos que el video de maritobaracus es el más obsceno, seguido del de Nimu, y por último, el de AleVera Oficial: el orden de los errores coincide exactamente.

Retomaremos estas observaciones en la discusión final, pero una conclusión exploratoria de esta fase apuntaría a las limitaciones de una mirada únicamente estadística de un fenómeno cuyas aristas contextuales no podemos ignorar si lo que buscamos estudiar es realmente el lenguaje y su desarrollo

Para ponderar los resultados de la herramienta a nivel del set de datos, tomamos 200 muestras de comentarios señalados como tóxicos y comprobamos la precisión en relación a los videos donde se hubieran publicado esos comentarios. Para determinar si un texto era realmente tóxico, contrastamos estos comentarios con los contenidos de los videos, de manera que si, por ejemplo, se citaba un insulto, consideraríamos ese comentario como un falso positivo por estar en línea con lo producido por el enunciador y no "expulsar" a nadie de la conversación.

Nuestra métrica de precisión para el set de datos completo fue de 57%, en la etapa cualitativa profundizaremos sobre algunos videos seleccionados. Además de aquellos que entraban en el juego propuesto por el youtuber, buena parte de los falsos positivos vinieron de mensajes con adjetivos descalificativos, insultos o muchos errores ortográficos, pero que no "tendían a alejar" a individuos o colectivos de la discusión.

Índice	**Comentario**
1	Jajjajaja yo la dejo y me voy ala mierda 😂😂😂
2	Like si sos un vago como yo xddddd
3	Romi es la versión femenina del demente gracias Romi estos cansado de ver los videos viejos del demente gracias mí aptinensia por la tía demente acabo igual no soy hater tullo ni del demente
4	Alto kpo 😇👌
5	Haaa me cage asta la patas con el de la sirena
6	Sube más de Yao porr favorrrrrrrr 😂😂😂😂😂😂😂😂 el mejor video que eh visto tuyo, eres el mejor loco
7	Que cagazo el último vídeo…
8	Jxfkñ hg esa porque hdgñ

Tabla 5: Muestra de falsos positivos con Perspective.

4.1.2. Validación y refutación de hipótesis

Nuestra primera hipótesis, la de que un incremento en la toxicidad severa de los comentarios se correlacionaría con mayor cantidad de visionados y/o reacciones, puede observarse desde el mismo prisma que la segunda: el tipo de enunciación humorística influye en la correlación entre discursos severamente tóxicos y visionados o reacciones.

A continuación veremos matrices de correlación, una forma de visualización que informa acerca de los niveles de relación cruzada entre variables. De aquí en más,

prestemos especial atención a las dos primeras columnas, las de los puntajes Z de toxicidad severa y de ataques identitarios:

	z_sev_toxicity	z_ident_attacks	sev_toxicity	ident_attacks
z_vistas	0.14	0.10	0.10	0.07
z_comentarios	0.11	0.08	0.08	0.06
z_likes	0.13	0.10	0.10	0.07
z_replyCount	0.18	0.16	0.14	0.13
z_likeCount	0.18	0.14	0.13	0.12
dislikes	0.15	0.08	0.09	0.08
likes	0.10	0.09	0.04	0.11
vistas	0.12	0.09	0.08	0.10
comentarios	0.11	0.04	0.02	0.08
replyCount	0.15	0.09	0.07	0.14
likeCount	0.11	0.09	0.13	0.22

Fig. 12: Matriz de correlaciones del set de datos completo. Ninguna correlación entre las métricas de toxicidad y de interacción rompe la barrera del 0.2 para ser calificada como asociación considerable.

Cuando buscamos correlaciones significativas entre formas de toxicidad y medidas de vistas o interacción para los 12 youtubers en conjunto no se observan valores considerables. Ni los valores absolutos ni las desviaciones de la media por autor de ninguna de nuestras variables objetivo parecerían tener siquiera una correlación medianamente atendible por encima del 0.2.

Es decir que no encontramos evidencia en nuestros datos que apoye la primera hipótesis de que aumentos en los comentarios tóxicos a una publicación contribuyen a aumentar su visibilidad, al menos cuando observamos al fenómeno a través del ranking de las 12 cuentas. Ninguno de los datos recolectados permitiría afirmar que los 12 youtubers más populares de la sección “Comedia” en Argentina podrían capitalizar la toxicidad en los comentarios de sus videos.

Pero, ¿es esta tendencia común a todos los youtubers? Volvamos sobre matrices de correlación, pero ahora, nos centraremos en tres cuentas donde sí se verifican correlaciones estadísticamente significativas: maritobaracus, AleVera Oficial y Nimu.

4.1.2.a- maritobaracus - Sistema cuenta

El 13.2% de los comentarios de esta cuenta fueron calificados como severamente tóxicos. En maritobaracus, toxicidad severa y ataques identitarios se correlacionan de diferente manera con vistas y métricas de interacción. Mientras que en el primer caso verificamos un nivel de relación considerable, los ataques a grupos protegidos no parecerían ir de la mano con ningún salto en la visibilidad al interior de este sistema cuenta.

	z_sev_toxicity	z_ident_attacks	sev_toxicity	ident_attacks
z_vistas	0.36	-0.15	0.36	-0.15
z_comentarios	0.38	0.06	0.38	0.06
z_likes	0.51	0.12	0.51	0.12
z_replyCount	0.36	0.01	0.36	0.01
z_likeCount	0.44	0.14	0.44	0.14
dislikes	0.23	-0.13	0.23	-0.13
likes	0.51	0.12	0.51	0.12
vistas	0.36	-0.15	0.36	-0.15
comentarios	0.38	0.06	0.38	0.06
replyCount	0.36	0.01	0.36	0.01
likeCount	0.44	0.14	0.44	0.14

Fig. 13: Matriz de correlaciones de la cuenta maritobaracus. Consideramos medianas las correlaciones por encima de 0.2 y fuertes las que superen el 0.5.

La variación por la media de autor (z_…) de comentarios y réplicas se relacionan en 0.38 y 0.39 respectivamente, y además la correlación con el z de likes entre comentarios también es atendible, con R = 0.33. Los mensajes tóxicos parecerían ir de la mano con un aumento en la participación de usuarios que, sin toxicidad mediante, no interaccionarían de la misma manera. El p-valor más elevado para este conjunto de observaciones fue del 0.007, por lo que sobre un estándar de 0.05 podemos afirmar estar ante conclusiones estadísticamente significativas.

Sin embargo, la suba en las interacciones no redunda en un ascenso relativo de las vistas. La dimensión temporal puede aportarnos algo más de claridad al respecto:

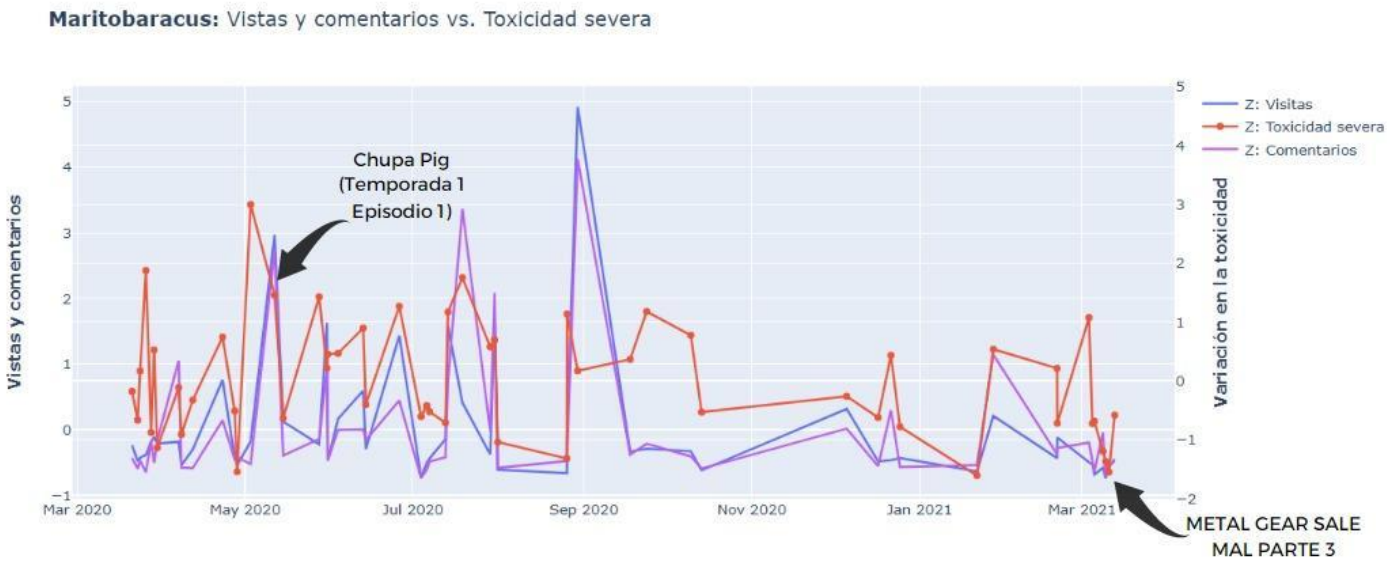


Fig. 14: Evolución de vistas, comentarios y toxicidad en función del tiempo, medidas en puntajes z, para los videos de maritobaracus. Para cada variable, un valor 0 indica la media de dicho enunciador, valores por encima de 0 indican mediciones por sobre la media, y negativos por debajo.

En maritobaracus es especialmente notable como, en la mayoría de los casos, si un video no supera la desviación de la media por autor en toxicidad, tampoco lo hará en cantidad de vistas o de comentarios. Entre julio y septiembre de 2020 dos períodos a la baja para ambas métricas coinciden.

Si seguimos las líneas de comentarios y vistas, veremos que en algunas áreas del gráfico pierden asociación el octavo video -entre marzo y mayo 2020-, dos veces antes de julio, una vez muy fuerte hacia agosto y en dos casos más -uno en diciembre 2020 y otro ya en febrero-. En todos esos casos, el Z toxicidad severa acompaña la pendiente de la Z de comentarios.

Para la etapa cualitativa seleccionamos dos publicaciones: “Chupa Pig (Temporada 1 Episodio 1)”, uno de los videos con una desviación de la media de toxicidad más alta (1.45) y también desviaciones en vistas (2.97) y comentarios (2.74) que colocan a la publicación entre el 0.03% y el 1% menos probable de ocurrencias. Además, interpretamos en este video un ejemplo de la enunciación más típica del canal, mientras que otros con más toxicidad relativa fueron trailers, videos demasiado cortos o aclaraciones a polémicas, que no constituyen el contenido más usual para esta cuenta. En segundo lugar, nos quedamos con “METAL GEAR SALE MAL PARTE 3” por ser el video con menos desviación de la media de toxicidad de este recorte (-1.6) y también métricas de visionado e interacción muy pobres (-0.63 en vistas, -0.54 en comentarios, -0.8 en likes).

Al final, tanto en el caso de maritobaracus como en el resto de nuestras selecciones,

buscamos recortar un video con mucha toxicidad en sus comentarios y otro con muy poca, manteniendo en la consideración de los recortes a las formas enunciativas más propias de cada canal. Si nuestro análisis estadístico aborda el problema de manera panorámica, con métricas que ponen el foco en la desviación de la media, nuestro análisis semiótico apuntará a los extremos, donde creemos que ciertos patrones pueden revelarse con más claridad.

4.1.2.b- AleVera Oficial - Sistema cuenta

Pasamos a AleVera Oficial, donde 5.6% de los comentarios fueron señalados como severamente tóxicos por Perspective. Ahora son los ataques a grupos protegidos los que más fuertemente se correlacionan con aumentos en las vistas y las interacciones, si bien las correlaciones con otras formas de toxicidad severa son también significativas.

	z_sev_toxicity	z_ident_attacks	sev_toxicity	ident_attacks
z_vistas	0.38	0.52	0.38	0.52
z_comentarios	0.14	0.34	0.14	0.34
z_likes	0.28	0.45	0.28	0.45
z_replyCount	0.25	0.52	0.25	0.52
z_likeCount	0.39	0.63	0.39	0.63
dislikes	0.38	0.68	0.38	0.68
likes	0.28	0.45	0.28	0.45
vistas	0.38	0.52	0.38	0.52
comentarios	0.14	0.34	0.14	0.34
replyCount	0.25	0.52	0.25	0.52
likeCount	0.39	0.63	0.39	0.63

Fig. 15: Matriz de correlaciones de la cuenta AleVera Oficial.

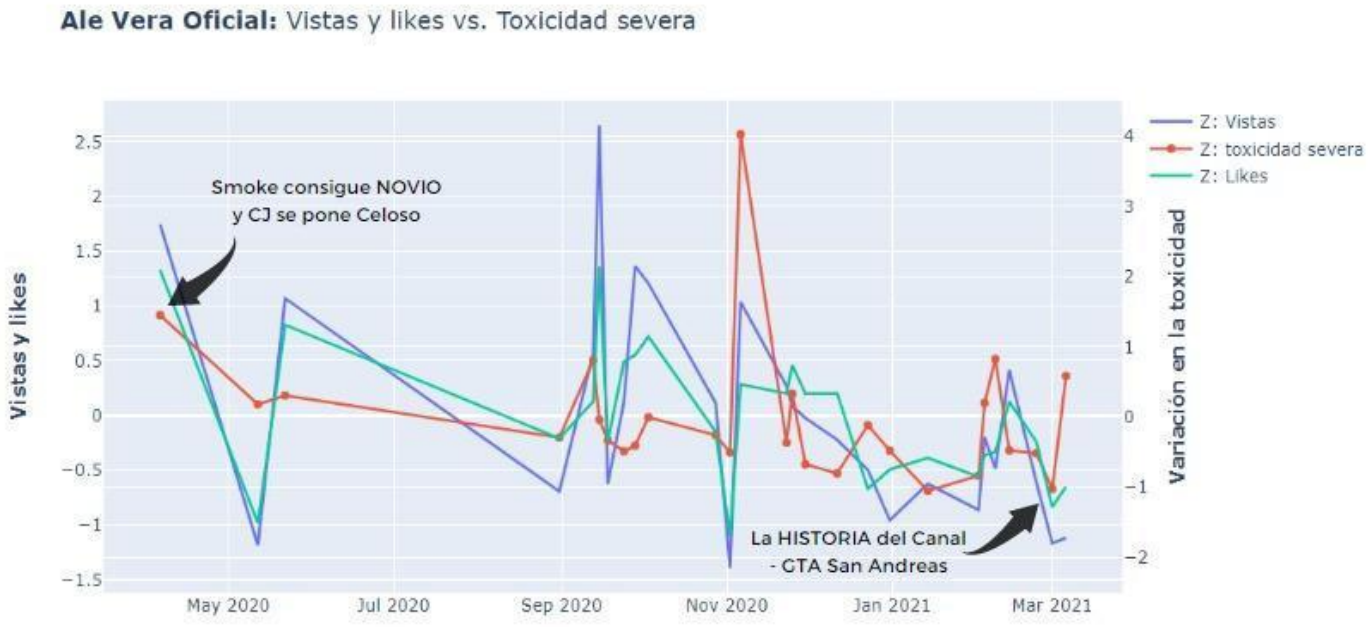
La correlación entre ataques identitarios y la variación en vistas, réplicas y likes a comentarios es fuerte, mientras que con comentarios a secas y likes la correlación es mediana. En la fase siguiente veremos que muchas veces el humor en AleVera pasa por señalar una homosexualidad vivida de forma penosa o la deficiencia mental de un personaje, y estas huellas en producción pueden luego verse en comentarios que, en un marco experimental, solo pueden verse como altamente tóxicos.

A grandes rasgos, las correlaciones que vimos como fuertes para los ataques identitarios, son moderadas para los casos de toxicidad severa: vistas, likes entre

comentarios y likes, en ese orden, son las métricas más relevantes en movimiento lineal junto a las desviaciones en la media de toxicidad por autor.

Como se aprecia en la Figura 15, entre marzo y noviembre de 2020 toxicidad y visitas e interacciones se mueven casi siempre en la misma dirección. Entre diciembre y enero parecería haber desplazamientos dispares, pero al comenzar febrero de 2021 la tendencia vuelve a corroborarse. Mientras que en maritobaracus era el número de comentarios la principal métrica de interacción que correlacionaba con los ascensos en la toxicidad, en AleVera, los likes se relacionan más intensamente con mensajes violentos.

Fig. 16: Relaciones entre vistas y likes y toxicidad en AleVera Oficial.



Para el análisis cualitativo, nos quedamos con el segundo video más tóxico, “Smoke consigue NOVIO y CJ se pone Celoso” (también segundo al rankear variación en las vistas y en likes) y con “La HISTORIA del Canal - GTA San Andreas”, el segundo menos tóxico (-1.03, apenas 0.02 puntos Z de toxicidad severa menos que el peor), con puntajes Z para vistas y likes también en el podio negativo.

Evitamos seleccionar el punto más alto en toxicidad que se ve hacia noviembre de 2020 porque en ese video, “GTA San Andreas - El HIJO del Tío de CJ”, aparecen dos personajes cuyos nombres hacen saltar de las alarmas de Perspective de manera artificial: “el tío gilipollas” y su hijo “gilipollitas” (recordemos que Perspective incluyó en su entrenamiento datos de comentarios en el diario El País, al considerar el cambio de dominio parece bastante verosímil que cualquier comentario con una variación de la palabra “gilipollas” no se salve del clasificador). En el video los dos personajes parecen mostrar algún tipo de discapacidad mental, por lo que si bien no descartamos

completamente la observación y la toxicidad presente en la publicación, a la hora de mirar con más detenimiento preferimos un caso arquetípico de comentarios violentos antes que un *outlier*.

4.1.2.c- Nimu - Sistema cuenta

	z_sev_toxicity	z_ident_attacks	sev_toxicity	ident_attacks
z_vistas	0.09	-0.04	0.09	-0.04
z_comentarios	0.38	0.16	0.38	0.16
z_likes	0.12	0.02	0.12	0.02
z_replyCount	0.39	0.16	0.39	0.16
z_likeCount	0.33	0.08	0.33	0.08
dislikes	0.32	0.15	0.32	0.15
likes	0.12	0.02	0.12	0.02
vistas	0.09	-0.04	0.09	-0.04
comentarios	0.38	0.16	0.38	0.16
replyCount	0.39	0.16	0.39	0.16
likeCount	0.33	0.08	0.33	0.08

Fig. 17: Matriz de correlaciones de la cuenta Nimu.

En Nimu el 5.3% de los comentarios fueron calificados como severamente tóxicos. Al ver las correlaciones entre la desviación de la media por youtuber de comentarios tóxicos e interacciones, el camino que trazamos en nuestra primera hipótesis parece revelarse ante nosotros: más toxicidad va de la mano con más interacciones, y más interacciones van de la mano con más vistas, el objetivo último de todo youtuber profesional.

En Nimu son los comentarios y réplicas los que se correlacionan con la toxicidad. Cuando veíamos los datos poblacionales de los 12 youtubers en la Tabla 1, Nimu era la cuarta cuenta con más comentarios a la vez que la decimosegunda con más seguidores. Es decir, proporcionalmente los seguidores de Nimu son más activos que la media de los enunciadores humorísticos. Cuando el público se Nimu se “enciende” en actividad, buena parte hace lo propio en toxicidad.

En la comparación entre las correlaciones con toxicidad severa y ataques identitarios, el caso de Nimu se parece más a maritobaracus que a AleVera Oficial, los insultos a grupos protegidos no parecerían ser indicador de un aumento en la visibilidad sino más bien lo contrario. En la etapa semiótica exploraremos algunas posibles causas de esta diferencia.

Cuando optamos por una mirada temporal vemos que, durante la circulación de las 66 publicaciones de Nimu en nuestro período, las correlaciones entre toxicidad y vistas tienen algunos períodos de irregularidad (sobre todo, entre agosto y septiembre de 2020 y entre enero y febrero de 2021) que explicarían por qué la correlación con vistas no fue tan fuerte.

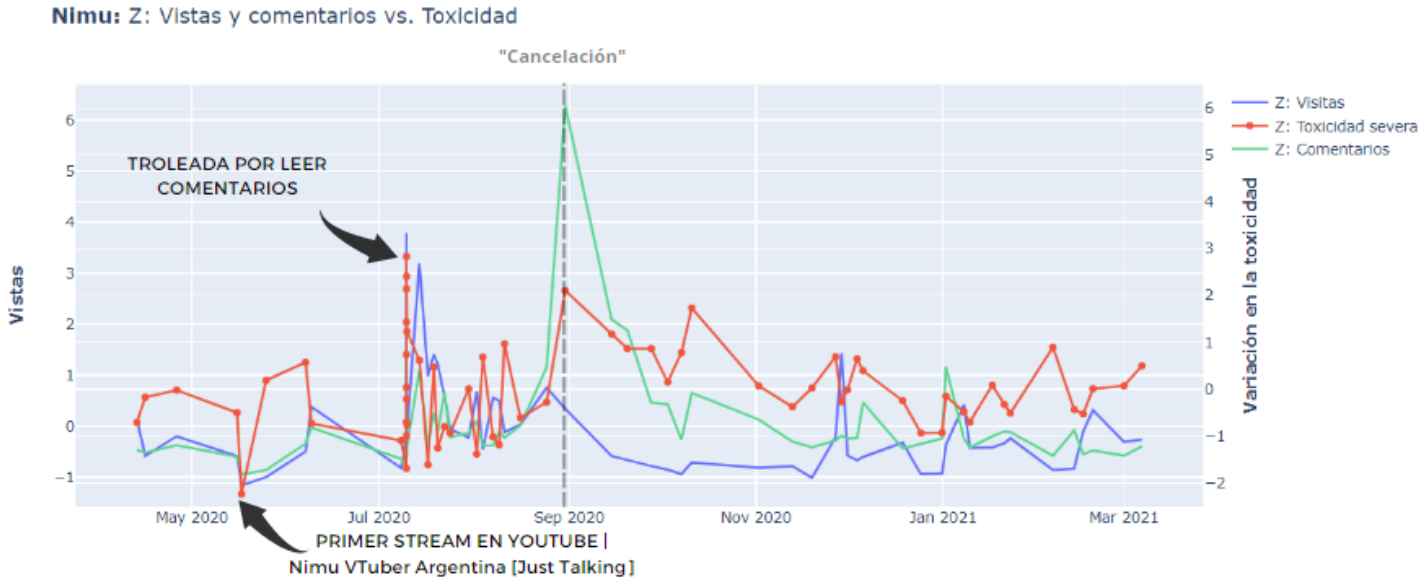


Fig. 18: Relaciones entre vistas y comentarios y toxicidad en Nimu.

Creemos que una perspectiva por períodos de diferente naturaleza es la óptima para el caso de Nimu porque su nivel de actividad fue irregular a lo largo del año recortado. En los tres meses entre mayo y julio de 2020 solo publicó cinco videos, luego desde julio hasta septiembre entró en una época más prolífica con decenas -doce de los cuales se publicaron en una misma semana- de videos para un período de igual duración. En septiembre entró en un período de relativo silencio, luego de haber sido “cancelada” por la comunidad otaku luego de un conflicto con otra vtuber. Pasado el conflicto, en noviembre de 2020 vuelve a publicar, en promedio, casi un video semanal. Así, los períodos de menos publicaciones son más estables en su correlación entre toxicidad y vistas e interacciones, mientras que el período frenético previo a septiembre es más inestable.

Para la etapa cualitativa seleccionamos el video con más comentarios tóxicos “TROLEADA POR LEER COMENTARIOS”, que, asimismo, fue uno de los más vistos del canal y “PRIMER STREAM EN YOUTUBE | Nimu VTuber Argentina [Just Talking]“, una publicación que no solo tuvo el Z de toxicidad más bajo sobre sus 66 videos (-2.23), además fue el de más baja desviación sobre la media en: vistas, likes y comentarios.

4.1.2.d- Una mirada panorámica

Antes de pasar al apartado siguiente, volvamos a nuestras dos hipótesis, la general -para la que no encontramos evidencia- y la que puntualiza sobre diferentes juegos del lenguaje -a priori, comprobada-. ¿Cómo explicaríamos que sobre un recorte en base a juegos del lenguaje específicos la relación entre toxicidad y visitas e interacciones exista, pero se diluye cuando observamos la totalidad del set de datos?

Esta respuesta tiene dos partes. La primera es que creemos estar ante un caso de la Paradoja de Simpson, el fenómeno según el cual una tendencia al interior de un grupo signado por una variable desaparece al ser mezclado sin considerar esa característica que lo distingue.

De manera similar al experimento de Rogier, et al. (2013) en el que un medicamento actuaba de forma diferente sobre pacientes hombres que en mujeres, cuando vemos a los youtubers según un ranking que mezcla los diferentes regímenes enunciativos del humor por estar organizado alrededor del número de suscriptores no vemos ninguna correlación entre toxicidad y vistas o interacciones. Ahora: ¿qué variable categórica explicaría la existencia de correlaciones significativas en algunos casos y no en otros como lo hacía el género en el experimento de Rogier? Responder esta segunda pregunta constituirá nuestra tarea central en la siguiente etapa de análisis cualitativo.

Para terminar, nuestra tercera y última hipótesis tampoco fue respaldada por nuestra evidencia. Dentro de los datos que recolectamos, no encontramos ninguna correlación entre que una enunciadora se identifique con el género femenino y que reciba mayor cantidad de comentarios tóxicos o ataques hacia su identidad. Así lo vemos en la siguiente matriz de correlaciones.

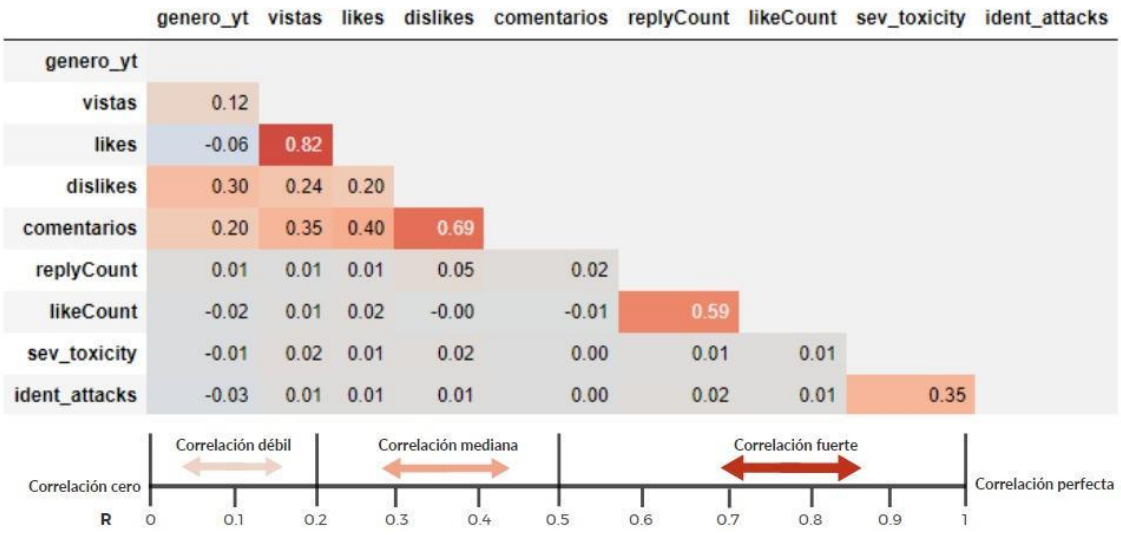


Fig. 19: Matriz de correlación para el género de los youtubers. Solo los dislikes parecen estar levemente asociados al género del propietario de la cuenta.

En este sentido, el único punto relevante que observamos fue una correlación de 0.3 (mediana, pero considerable y estadísticamente significativa) entre autopercibirse como mujer y recibir dislikes en los videos. Si bien la relación no es fortísima, y esta variable no fue una de las centrales en nuestro análisis, el dato podría interpretarse en la línea de los muchos estudios sobre género en redes sociales que señalan que, bajo condiciones similares, youtubers femeninas reciben más rechazo en YouTube en particular.

Fox, et al. (2015) corroboraron un punto que ya habíamos mencionado en la introducción: las consecuencias de la violencia en redes -esta vez ciñéndose a feminidades- pueden rastrearse incluso en entornos offline. Por más verídicas que consideremos sus afirmaciones y pruebas experimentales, creemos que, para nuestro contexto, los resultados recolectados son insuficientes para respaldar a la mayor parte de la literatura del tema.

En cuanto a los demás 9 youtubers, las correlaciones fueron inferiores a 0.2 en prácticamente todos los casos. Un solo enunciador escapa a esta lógica y es Guille Aquino.

	z_sev_toxicity	z_ident_attacks	sev_toxicity	ident_attacks
z_vistas	0.16	0.41	0.16	0.41
z_comentarios	0.09	0.16	0.09	0.16
z_likes	0.07	0.24	0.07	0.24
z_replyCount	0.47	0.51	0.47	0.51
z_likeCount	0.07	0.35	0.07	0.35
dislikes	0.41	0.16	0.41	0.16
likes	0.07	0.24	0.07	0.24
vistas	0.16	0.41	0.16	0.41
comentarios	0.09	0.16	0.09	0.16
replyCount	0.47	0.51	0.47	0.51
likeCount	0.07	0.35	0.07	0.35

Fig. 20: Matriz de correlaciones para Guille Aquino.

En Guille Aquino la toxicidad severa apenas se correlaciona con un aumento en las réplicas entre comentarios. Son los ataques identitarios los que sí están asociados, de forma moderada con un aumento en las vistas y en los likes a comentarios, y fuertemente con las réplicas.

Separamos 100 muestras de ataques identitarios que apuntaban, en el siguiente orden, a: personas de izquierda, feministas, peronistas, “negros”, liberales y judíos. El 59.6% eran réplicas a otros comentarios, que, en muchos casos, se desarrollaron en forma de largas discusiones centradas en unos pocos comentarios.

Consideramos a Aquino un caso intermedio, porque si bien la correlación solo ocurre con ataques identitarios, de acuerdo a Perspective el 58% de los comentarios discriminadores coinciden con los severamente tóxicos. En nuestra muestra, 74% de los comentarios señalados como ataques identitarios también eran severamente tóxicos.

4.2. Etapa cualitativa

4.2.1. maritobaracus - Enunciador Hipermediático

maritobaracus es el nombre de fantasía de Juan Arnone, un ex columnista de programas televisivos de videojuegos que se convirtió en uno de los pioneros del humor en YouTube Argentina. Desde 2011 publica videos en los que dobla otros contenidos de la cultura masiva como películas, series o videos de redes sociales, además de producir series propias como el Combo Loco o Chupa Pig.

Desde sus inicios, sus fotos de perfil fueron diferentes variaciones de payasos malvados que preanuncian un humor satírico que para muchos receptores podría ser considerado violento o de mal gusto. El rebajamiento de los recursos cómicos es en maritobaracus tan intenso como en ningún otro enunciador de nuestra muestra primaria, aunque también sus enunciados mediáticos pueden contener cierta perspectiva social.

Así, temas escatológicos, sexuales, alusiones a discapacidades, drogas o violencia de género pueden compartir video con mensajes fácilmente identificables como progresistas, feministas o anti-corporaciones. Al momento de la recolección de nuestros datos, la cuenta se presentaba en su descripción como “Humor hardcore gay antifascista”. Un punto que viene a complejizar la enunciación es que el propietario de la cuenta ha hecho pública en entrevistas su propia homosexualidad, por lo que la tesis de un enunciador discriminador perdería fuerza frente a una interpretación de una búsqueda antisistema que, ya veremos se repite también en Nimu y AleVera Oficial.

Además de las groserías y los doblajes en muchas ocasiones las voces sugieren características del enunciador diegético (masculinidad para mujeres trans, voces graves y trabadas que sugieren discapacidades mentales para personajes de poca inteligencia o feminidad para hombres homosexuales), el collage es otro de los recursos estilísticos más repetidos en su canal, así, diferentes tonos o tipos de imagen se funden para crear una sensación de incomodidad en el enunciatario.

¿Por qué decimos que en maritobaracus estamos ante una enunciación bromista? Porque para este enunciador hipermediático, lo risible no cae en sí mismo o en sus pares, sino en los enunciadores mediáticos o diegéticos; cuando nos reímos con maritobaracus no nos reímos de él o de quienes son parecidos a él, nos reímos de la forma en la que están diseñados los personajes o de sus parlamentos poco apropiados a una situación. El recurso de lo inadecuado siempre es ignorado por personajes que no pueden percibir que están violentando normas elementales de comportamiento, lo que los vuelve objeto de burla, cómicos.

La de maritobaracus fue la cuenta con más comentarios señalados como tóxico, con 13.2% del total marcados registrados por encima de nuestro umbral de 0.5. Nuestra hipótesis es que la mayor parte de estos comentarios forman parte de una respuesta esperada y fomentada por un enunciador hipermediático que incluye diferentes discursividades tóxicas en sus producciones. Son tóxicos en la mirada desde las formas, no así desde el contexto.

"Chupa Pig (Temporada 1 Episodio 1)" [9]

En tanto parodia de la animación Peppa Pig, el enunciador mediático (nivel 3) de este video es un enunciador partido. Por un lado parecería imitar los recursos estilísticos y verosímiles de Peppa Pig, la serie infantil que propone un entretenimiento por demás inocente, y por el otro se expone, por ejemplo, cuando los personajes aparecen dibujados con caras en forma de pene. Las imágenes de transición, también emulando a las del programa original, revelan una tipografía que incluye penes, bocas y manchas de semen en el fondo. Las reglas de un juego de transparencia se rompen apenas comenzado el video.

En el nivel diegético (el 2) nos enfrentamos ante una enunciación cómica. Ninguno de los personajes parece notar extrañeza ante escenas de fondo que muestran masturbaciones, sexo oral, plantas de marihuana, efectos de sonido incluyen flatulencias, sonidos pegajosos. La definición de Perspective hablaba de "lenguajes rudos, irrespetuosos", y si analizáramos la transcripción del video, oración por oración seguramente la mayoría de los sistemas de detección de toxicidad se activarían a cada paso. Repasemos un diálogo protagonizado por el padre de la protagonista de relato entre los segundos 0:37 y 0:52:

-Hola jefe, chúpeme bien la chota.
-¿Eh? ¿Cómo va a decir eso? Lo voy a tener que despedir con justa causa, sin indemnización.
-No necesito su mugroso dinero: mi hija se hizo influencer, ahora somos multimillonarios -y sigue, en otro escenario en el que aparece lanzando billetes- tomá

[9] El link a "Chupa Pig…" es: https://www.youtube.com/watch?v=L_u88nDn0y4

amor, pedí sushi y depilate la concha que esta noche festejamos.

Además, haciendo gala del espíritu punk de su descripción, los personajes de maritobaracus critican a YouTube como empresa por pagar muy poco a los productores de videos, y señala como “boludas” a las personas que buscan volverse ricas produciendo videos en la plataforma.

Fig. 21: La protagonista baila inocentemente frente a su casa -situada sobre una colina con forma de glúteos durante una penetración- donde su madre practica sexo oral a su padre. Este tipo de imágenes son corrientes en los videos de maritobaracus.

En suma, tenemos un video que revela un alto nivel de toxicidad desde producción. ¿Qué vemos en los comentarios?

En la Tabla 3 accedimos a una muestra aleatoria de comentarios señalados como severamente tóxicos para esta publicación. El 60% (en la línea de los resultados totales) incluían o sugerían groserías, pero como respuesta validante al contenido del video, lejos de ser pruebas de una toxicidad real.

En general, los comentarios verdaderamente tóxicos en esta publicación corresponden a la lógica de construcción desde el antagonismo expuesta por Laclau y Mouffe. Hay dos tipos de comentarios tóxicos: primero, las respuestas a quejas de individuos ajenos al colectivo de fans, así, si alguien crítica al youtuber o a su producto por sus formas incorrectas, las respuestas pueden ser “@usuario tremendo pendejo 😳” o “Más mentirosa, si fueras una madre no escribirías como boluda.”; segundo, otros se unen a la crítica frente a la industria del entretenimiento que parecería representar el video: “AL FIN UNA BUENA SERIE PARA VER CHUPALA NETFLIX” o, a youtubers más *mainstream* “Boludo alta bronca q Yao Cabrera tenga 7 millones de subs y miente y vos q tenés alto contenido solo 2 millones”.

“METAL GEAR SALE MAL PARTE 3” [10]

En este caso, el youtuber se reapropia de las animaciones del Metal Gear, un videojuego de guerra marcado por la retórica del honor castrense y pronorteamericana, que en la reapropiación de baracus acaba contando historias de un amor homosexual ridiculizado entre miembros de las milicias.

Esta vez los recursos con los que se modifica la narrativa original sobre la que se ironiza se reducen a solo tres: edición alterada, para modificar el orden y contenido de la historia original; iteraciones de secuencias, para prolongar diálogos y generar un efecto de ridículo; y el guion que, además de abrir posibilidades impropias de un contexto militar -como que soldados hablen en lenguaje inclusivo-, es leído por voces que buscan encarnar estereotipos sobre la homosexualidad masculina.

Si en el anterior video la enunciación mediática era plenamente cómica, en este caso vemos una apelación humorística a temas como el cambio climático, las acusaciones de violencia de género o las violaciones intracarcelarias. A falta de una transformación de la imagen (como sucedía con los personajes con cara de pene en “Chupa Pig…”) lo risible pasa por un recurso penoso: que el planeta esté en peligro o que algunos reclusos en cárceles sean víctimas de abuso sexual es llevado al frente en diálogos como el siguiente:

-Coronel, necesito unas toallitas.

-No Snake, eso contamina.

-Pero a la conchuda de Mary le está sangrando.

-El mundo también está sangrando, y tenemos que cuidarlo entre todes. Dejame que te muestre un pequeño video explicativo.

Allí entra en escena un recorte ya no de animación sino de grabaciones humanas que presenta una tensión entre las imágenes de lo que parecería un informe televisivo con tono serio y un narrador que declara que “tanto la ONU como la CGT se han expresado a mediados de 2019, haciendo un llamado a las naciones del mundo para que reduzcan un 65% el uso de toallitas de cara a [la copa del mundo de] Qatar 2022”. La equiparación de dos instituciones no equivalentes como la ONU y la CGT, y de las toallitas femeninas con un mundial de fútbol son dos de los recursos de los que se vale la enunciación mediática. Si bien los temas escatológicos vuelven a hacerse presentes, la enunciación mediática del video es antes humorística que cómica o bromista.

En comparación con “Chupa Pig…” las escenas con alta toxicidad son mucho menos crudas. Además, la publicación es la tercera de una serie que va descendiendo en cantidad de vistas (102 mil vistas para el primero, 79 mil para el segundo y 54 mil para

[10] El link a “Metal Gear sale mal…” es: https://www.youtube.com/watch?v=mxCFm1F-R_E

el tercero, al momento de la redacción de nuestra investigación) y cuya primera edición tuvo un puntaje Z en vistas de -1.6.

La presencia de solo 21 comentarios tóxicos permite observarlos en su totalidad. Durante la etapa cuantitativa repasamos cómo, mientras una mayor toxicidad se correlacionaba con un aumento de vistas e interacciones, los ataques identitarios tenían una correlación levemente negativa con las métricas de visibilidad. Por eso, que la mayor parte de los comentarios tóxicos ataquen a colectivos sexuales o raciales podría darnos un indicio de una ruptura en el contrato de lectura que explicaría la pobre performance del video en general.

Índice	**Comentario**
1	- Ah no, ah, tu tu tu tu tu tuf tuf tuff yeeee tuff yeeee tuff tuff tuff yeeeee yeee... - Chau grone, me tenes podrido - Black day matter, yeeee
2	Pero al final son todos gay en el metal gear ??? Jajaja
3	A nadir le importa negro
4	Eeeeh Marito sos de mardel o de aldosivk
5	marito baracus cuando vas a subir la segunda temporada de chupa pig :([subela por favor]
6	El culpable de la contaminación global es Macri😡😡😡
7	Cállate negro
8	La referencia a la Farmacia Violenta de los 3 Cabiados, ajjajajajaj Burifandela, Garbanzo re sacado!
9	¿Para cuándo un Marito Baracus sale mal y Marito Baracus políticamente correcto? (?
10	La famosa BURIFANDELA
11	Aguante Los3Cabiados grodo ferneteroooo
12	@usuario que hace grone? No rompas los huevos otra vez vos no estoy de humor
13	Kpo Marito
14	@usuario pero porque las grandes le rompen el orto jojojooooooo
15	Corré Esnake te van a acusar de femicidio hahahahaha
16	<ahref="https://www.youtube.com/watch?v=mxCFm1F-R_E&t=5m11s">5:11 Yiiiie Agh me resetie las rodillas bich
17	COMANDANTE: NO ESNAKE,EL MUNDO ESTA SANGRANDO..... OTACON: YA TE ACLARE LAS COSAS ESNAKE XD JAJAJA
18	BLACK LIES! BLACK LIES COMO MENTIRAS!!
19	BURIFANDELA JAJAJAJAJAJAJAJAJ

20	Un poco cutre este dobleje me gusta la saga metal gear pero tu la cagaste te tengo que desir muy poco humor y mucha trolleada asco.
21	"Black Lives Matter". Ha Ha Ha Ha !

Tabla 6: Comentarios tóxicos en "METAL GEAR SALE MAL PARTE 3".

4.2.2. AleVera Oficial - Enunciador Hipermediático

A diferencia de todos los demás enunciadores de nuestra selección inicial, AleVera Oficial establece una narración ficcional única que transcurre a lo largo de capítulos. Forma parte de un subgénero marginal pero muy popular en YouTube en castellano, los "loquendos" o "loquenderos", productores de contenido que usan el software de animación Loquendo para intervenir a personajes de videojuegos que se vuelven protagonistas de sus propios fanzines. Dentro de este subgénero, los videos más populares -dentro de los se cuentan los producidos por AleVera Oficial- se realizan sobre la base de las animaciones del videojuego Grand Theft Auto (GTA): San Andreas.

El GTA San Andreas es un videojuego lanzado en 2004 que gira en torno a la vida de un delincuente apodado CJ. La historia tiene referencia a las guerras de pandillas de los noventa en la costa oeste de los Estados Unidos y gran parte de las misiones que debe cumplir el jugador constan de robos, asesinatos, venta de drogas, proxenetismo o atentados terroristas. El juego ha recibido acusaciones de estereotipos raciales, de género e incluso llegó a ser el único juego de venta masiva clasificado como "Solo para adultos" luego de que un usuario activase un mod en el que personajes tenían sexo explícito. Los insultos raciales y el lenguaje soez son moneda corriente al interior del verosímil de un juego que acabó por ser el más vendido en la historia de la PlayStation 2.

Sobre la base de este videojuego y de sus personajes es que AleVera Oficial montó una historia alternativa, en la que lo risible pasa por una modificación entre las expectativas de un videojuego violento y una historia mucho más telenovelesca. La caracterización de este cuarto nivel girará entonces mucho más alrededor de los personajes del GTA que cubren espacios de la cuenta que de la foto de perfil (una animación de un joven, castaño de tez blanca, sin demasiadas particularidades) o el nombre del propietario (bastante genérico para un varón joven argentino).

Así, el peligroso pandillero CJ se encuentra en los videos de esta cuenta con situaciones mucho más corrientes que las aventuras del videojuego: el amor, los conflictos familiares o los gustos gastronómicos generan risa en cierto público en tanto no serían propios del personaje que conocieron. Con el mismo recurso, un personaje secundario del GTA San Andreas como el capo de la droga Big Smoke es, en las

historias de AleVera Oficial, un hombre homosexual supuestamente afeminado que intenta sin éxito concretar un vínculo con CJ.

Si bien a diferencia de maritobaracus este enunciador no se plantea explícitamente como antisistema, consideramos que organizar toda una narración alrededor de un juego que idealiza la violencia y una vida de delincuencia se posiciona frente a los valores de la convivencia pacífica y la meritocracia, tradicionales en nuestras sociedades occidentales. Basta con googlear "GTA" para verificar la resistencia que la saga genera hasta el día de hoy en medios tradicionales.

La violencia y los insultos están en la génesis de las historias del canal, por lo que gran parte de lo que Perspective identificó como toxicidad severa en realidad correspondería a una respuesta positiva por parte de individuos que aceptan y participan de la propuesta del youtuber. Esta característica que no solo se encuentra en los videos, sino también en los comentarios de una comunidad que, a pesar de mostrarse muy fiel al enunciador, muchas veces se dirige a él con expresiones como "gordo fan del nutella" u "homúnculo". Además, en muchos videos vemos una interpenetración entre producción y reconocimiento cuando el youtuber incluye los comentarios de sus seguidores en el último minuto de sus videos.

Igual que en el caso de maritobaracus, el enunciador hipermediático es bromista en tanto la apropiación de un discurso (en este caso, de un videojuego) figura al objeto risible por fuera de la órbita del yo que enuncia y las narraciones plagadas de ridículos no tienen una función de crítica sino que provocan la risa por la sola ruptura de las normas de lo esperable.

En el total de esta cuenta, sólo 5.6% de los comentarios fueron señalados como tóxicos por Perspective. Comparativamente, sus resultados tuvieron la precisión más alta de las relevadas. En las próximas páginas veremos que en gran parte esto se debe al elevado número de violencia dirigida a hombres homosexuales que circula en este canal.

"Smoke consigue NOVIO y CJ se pone Celoso" [11]

El conflicto de esta historia es que Smoke, un mafioso homosexual que en capítulos anteriores fallaba en seducir a CJ, anuncia que consiguió pareja, lo que despierta sorpresivos celos en el protagonista. CJ intenta recuperar el interés de Smoke a través de diferentes hazañas (roba un camión de helados, toca un solo de bajo en la calle, busca besarlo repentinamente), pero falla. Ocurre un encuentro entre CJ y el novio de Smoke donde se golpean hasta dejarse inconscientes mutuamente, y el tercer acto termina con los dos personajes centrales hacen las paces y acuerdan que es mejor continuar su amistad como de costumbre. Luego de un par de chanzas sobre la

[11] El link a "Smoke consigue NOVIO…" es: https://www.youtube.com/watch?v=YClxtHqBOmU

violencia entre hombres gay, y la continuación de una línea argumental de episodios pasados sobre “Smoke Malvado”, el capítulo finaliza.

A pesar de que la diégesis no adopta una posición expresamente homofóbica, sí hace referencia a que en épocas pasadas el protagonista “mandaba al hospital” a Smoke por sus acercamientos. Lejos de la literalidad, la lectura de estos diálogos debe considerar los verosímiles del universo GTA, donde la violencia es moneda corriente; aun así, que estas situaciones provocan risas en cierto público es indicativo, al menos, de una alta aceptación a tópicas discriminatorias.

Que gran parte de la narración transcurra en torno a acciones ridículas que serían producto de la homosexualidad de personajes ubica a la enunciación mediática en el rango de lo cómico. El enunciatario se ríe junto al enunciador de hombres gays que no pueden resistir sus impulsos sexuales.

La conversión del protagonista de la serie no parece haber convencido al colectivo de fans de AleVera Oficial, que parecían preferir la historia en sus términos heteronormados. Salvo por los errores groseros, en la Tabla 7 los textos sexto y séptimo, los comentarios marcados como severamente tóxicos por Perspective se dividen en tres grupos principales:

Índice	**Comentario**
1	negro pandillero te extrañe
2	Joder este video es la polla con cebolla
3	Jajaja acabo de mirar esto yo realmente no tengo nada en contra de los homosexuales ni nada de eso, pero ver a CJ tipo homosexual me da mucha pena ajena ya que siempre eh tenido la imagen de ser pandillero gánster y por eso se me hace raro verlo tipo así jajaja pero igual genial video men tienes una linda fan que ama tus vídeos.
4	No se si la cagaron, Cj no es así mierda 🙄😠
5	asi que....cj es gay
6	Si ases a CJ GAY TE DENUNCIO EL CANAL Y DISLIKE Y UN MENOS SUBCRIBTOR MENOS NO ES BROMA LO ARE
7	Por que hiciste a cj celoso el no era gay
8	El no es cj pinche pendejo
9	@usuario mierda es tu comentario aparte de ti
10	Like si te violo esmoke xd

Tabla 7: Muestra aleatoria de 10 comentarios marcados como severamente tóxicos por Perspective para “Smoke consigue NOVIO y CJ se pone Celoso”.

Primero, aquellos que celebran la vuelta a la actividad del canal luego de algunos meses sin publicaciones con groserías amistosas como “maldito vago adicto ala nutella”; si bien marcamos este tipo de toxicidad amistosa como falsos positivos excepto en casos que ataquen a un colectivo protegido (como el primero de la Tabla 7), creemos que aportan información sobre las características de la comunicación al interior de la comunidad de fans. Segundo, otros reaccionan con enojo a la conversión homosexual del personaje protagónico y se lo hacen saber al creador de la serie; este fue el grupo más abundante en el total de los mensajes verdaderamente tóxicos. Tercero, un grupo continuó situaciones a las que hacía referencia la narración, en chistes que incluían un nivel elevado de toxicidad, por ejemplo, el comentario con índice 10: “Like si te violo esmoke xd”. La equiparación de las imágenes del hombre gay y el hombre violador está sugerida en el relato del video, y explicitada en este grupo de comentarios que también retoman tópicas machistas o de violencia de género.

“La HISTORIA del Canal - GTA San Andreas” [12]

Esta publicación rompe con la lógica narrativa del canal, escapa al relato de los personajes del GTA San Andreas para centrarse en una autobiografía del canal y su propietario. Así, no es tan sorprendente que este video haya sido el menos visto del canal para el período estudiado.

A pesar de su tono de documental biográfico, la enunciación mediática de este video corresponde a la categoría humorística. El enunciador mediático se reconoce a sí mismo y a sus creaciones como objetos risibles: la narración es transparente en relación a que tal o cual personaje es un estereotipo o a las diferencias entre las creaciones de AleVera Oficial y las personalidades originales del GTA San Andreas, y hasta el locutor confiesa que, cuando un personaje muere “va a la papelera de reciclaje antes de ir al paraíso” o ser restaurado a la historia sin más, e incluso que “los personajes principales pueden morir todas las veces que sea necesario y no mueren”.

Además de los recursos estilísticos habituales de la cuenta, varios metatextos estructuran la supuesta historia del canal: referencias bíblicas, a Star Wars o Matrix. La contraposición de diferentes estructuras y el objetivo antes narrativo que risible hacen de este un video bastante atípico a los patrones del canal, y aventuramos en ese punto la explicación de su pobre performance.

Perspective marcó sólo 10 comentarios como severamente tóxicos. Algunos de los señalados con precisión forman parte de discusiones entre usuarios o son insultos a personajes de la saga de AleVera Oficial, mientras otros tienen poco que ver con los contenidos del video y parecerían groserías casi aleatorias.

[12] El link a “La HISTORIA del Canal…” es: https://www.youtube.com/watch?v=xQhp8tVeAI4

Índice	Comentario
1	DEJA DE ASER ESO, Y MEJOR MEJORA TU GTA SA ANTIGUO CON MAL GRAFICO, Y HAS MÁS VIDEOS DE COMEDIA.
2	Entonces smoke no es gay solo actúa
3	Es gay solo que en los vídeos es obsesivo hacia cj pero detrás de camaras no es irritante
4	Yo quisiera poder hir al universo de gta sanandreas como alevevera 😔
5	Fuck clases de historia
6	Solo aún subnormal no le gusta esta joya de video
7	@usuario :O por un momento pense que me llamarias gordo mamon
8	Smoke 😠
9	un fan de german, que asco
10	Mierda ojalá tu fueras mi profesor de matemáticas para que nadie reprobara

Tabla 8: Los 10 comentarios marcados como severamente tóxicos por Perspective para "La HISTORIA del Canal - GTA San Andreas".

4.2.3. Nimu - Enunciadora Hipermediática

Nimu forma parte de otro subgénero nacido en los últimos años de desarrollo de las plataformas audiovisuales: los vtubers. Un vtuber es un enunciador hipermediático personalizado en la figura de un avatar virtual. Estos personajes de animación funcionan a través de modelos de Inteligencia Artificial que dan lugar a una gran independencia en los movimientos y las acciones de una gráfica que vendría a reemplazar a la figura del youtuber humano tradicional.

En casi todos los casos, las animaciones de los vtubers replican figuras del animé o la cultura otaku. A pesar de que los fenómenos relativos a esta cultura filo-japonesa puedan parecer invisibilizados en ciertos escenarios de la cultura masiva, redes como YouTube permiten dimensionar el fenómeno cuantitativamente: Las dos principales exponentes globales de los vtubers son Kizuna A.I. y Gawr Gura, con 4.5 millones y 4 millones de suscriptores respectivamente. La cuenta global de Disney tenía, al momento de la redacción de este trabajo, 5.4 millones de suscriptores, la de la cadena ESPN, 8.4 millones. No estamos ante fenómenos de dimensiones muy diferentes a los gigantes de las industrias culturales tradicionales. En el plano local, Nimu contaba con 741 mil suscriptores al momento de la redacción, un tercio de los 2.7 millones de la cuenta de la Televisión Pública Argentina.

Las figuras del propietario de la cuenta (muchas veces, un individuo en las sombras) y un enunciador que protagoniza la acción, enmascarado en la virtualidad, se disocian en lo que es un movimiento clave para la identificación de colectivos que, en muchos casos, inician su relación con la cuenta partiendo de un alto nivel de compromiso para con representaciones de lo japonés y otaku. En el caso de Nimu, estamos ante una adolescente que se muestra como un personaje naíf al trastabillar en la práctica de videojuegos de moda o reacciona con cierta sobreactuación infantilizada.

Posiblemente el acercamiento más acabado desde las ciencias sociales a la intensidad identitaria de los colectivos otakus al interior de las plataformas en castellano se encuentre en el trabajo de Álvarez Gandolfi (2016):

> El otakismo es un recurso identitario, una base a partir de la cual puede construirse una identidad alternativa y duradera, tanto individual como colectiva, cuya puesta en escena en el ciberespacio implica una autonominación que resignifica los semas negativos presentes en las nominaciones externas –sociales y mediáticas– de los fans otakus como "anormales" o "inmaduros". Así, estos sujetos pueden procesar grupalmente sus experiencias compartidas de subordinación simbólica, identificándose con los "exóticos" y subalternizados objetos de la cultura de masas japonesa y concibiendo a su consumo como algo "diferente", dentro de una economía cultural hegemonizada por Estados Unidos.
> (Álvarez Gandolfi, *ibidem*, p. 48)

Y es posible que en esta construcción de una identidad colectiva desde la diferencia y la subalternidad encontremos la explicación a por qué la diferencia de correlaciones entre toxicidad severa y ataques identitarios era tan notoria en Nimu.

Para que hablemos de un pleno enunciador hipermediático cómico, tenemos que estar ante una cuenta que, al menos, coquetee en la propiedad con la noción de *fake*. En la línea de Álvarez Gandolfi, el enmascaramiento de los vtubers puede leerse como parte de un rechazo a una cultura del entretenimiento que demanda ciertos estándares estéticos a quien enuncia y su contexto: ahora ya no hace falta ser bello, construir un culto a la personalidad, tener los mejores equipos de grabación o un fondo diseñado de acuerdo a los cánones de YouTube. Basta con un software.

Cuando un vtuber como Nimu hace humor con sus pequeñas desgracias de la cotidianeidad virtual, no estamos ante escenas demasiado diferentes a las de Charles Chaplin en El Gran Dictador u Homero Simpson: quien ríe, lo hace desde la satisfacción de superioridad para con un sujeto que ha obrado contra las reglas de la sociedad. También en Nimu lo risible cae en la figura de la enunciadora: sus errores durante los juegos, intercambios poco afortunados con sus fans o situaciones incómodas la ubican en una triple frontera entre lo ridículo, lo penoso y lo tierno o dulce que habilita la risa ansiosa.

Sobre el total de los comentarios recolectados, un 5.3% del total fue señalado por Perspective como severamente tóxico.

“TROLEADA POR LEER COMENTARIOS | Nimu” [13]

Si en el nivel 4 señalamos que Nimu es antes que nada cómica, en este YouTube Short se muestra con una enunciación humorística. La pieza dura solo unos segundos y consta del recorte de un stream en el que el avatar lee comentarios sobre un fondo de Mario Bros. Mientras la vtuber lee los agradecimientos de sus fans se cuela un mensaje desafortunado:

-...muchas gracias por compartir: Emanuel Villasaña, Javier Lisana, Elmano Seadordeculos…, ¿qué? ¡Espérense!

Luego el video continúa con Nimu aparentemente fingiendo un desconocimiento de la situación, hasta que, al final, cae en la cuenta de lo ocurrido. Al final, la enunciadora se expone en un momento de equivocación penosa: cayó en una trampa que cualquiera hubiera podido pasar por alto y se reconoce como objeto risible. La identificación de los enunciatarios se verifica en gran parte de los comentarios, que se ríen junto a ella y de una chanza que, al final, solo provoca un daño superficial y momentáneo. El video es cómico en tanto el error es cometido por la enunciadora que se expone en su yerro e incluso insiste en degradarse ante el enunciatario, que ríe desde fuera de la escena.

Índice	**Comentario**
1	Elmano Seador de Culos: Saliste Trolliadisima
2	Tu nombre: A L F R E D O P A R E D E S
3	@usuario youruber virtual, virtualtuber vtuber
4	@usuario JAJAJAJA te dolió? Hasta editado pone XD
5	detrás del avatar de anime seguro hay un gordo rancio con voz de loli
6	@usuario trolling Dancing
7	aaa :O el manoseador de culos, hijo de.
8	Ah mamada de ninu
9	eres una mentirosa desagradable
10	Con el manoseador de culos
11	Elmanos Eadord Eculos buen nombre para rey demonio esta casi a la par de Anos Voldigoad
12	Entrega el orto

[13] El link a “troleada por leer comentarios…” es: https://www.youtube.com/shorts/gDTQV7l2AMc

13	JJAJAJAAAJJAAJAJAJJAAJAJAJAJAJAJA amo su puta inosenciaaaaaaaaaa
14	@usuario C H A M P I O N

Tabla 9: Muestra aleatoria de 14 comentarios marcados como severamente tóxicos por Perspective en "TROLEADA POR LEER COMENTARIOS | Nimu".

Entre los comentarios señalados como severamente tóxicos con precisión, vemos que algunos enunciatarios insultan a Nimu y hasta comentan con una violencia de género que hasta ahora no habíamos observado en nuestra investigación; otros parecen responder indignados ante la exageración o fabulación del suceso por parte de la vtuber. Del "un saludo de Elmano Seadordeculos" inicial al "entrega el orto" como respuesta hay un larguísimo trecho que sólo puede ser explicado como una forma de violencia de género.

Si bien nuestros resultados estadísticos son categóricos respecto a la no existencia de una tendencia general a mayor toxicidad para con enunciadoras femeninas, la observación cualitativa parecería probar que estos individuos se enfrentan a tipos de violencia muy diferente a sus pares masculinos.

"PRIMER STREAM EN YOUTUBE | Nimu VTuber Argentina [Just Talking]" [14]

Esta publicación es cualitativamente única. Normalmente, la vtuber produce streams en Twitch de los que sube a YouTube un resumen o edición de los momentos más graciosos. En este caso, Nimu transmitió en vivo a través de YouTube sus reacciones a otras publicaciones de la plataforma (más que nada compilaciones de videos virales o de bromas pesadas, publicaciones sobre cultura japonesa y hasta videos que formaron parte de nuestro recorte inicial como "AJEDREZ BAJO EL AGUA" de FFran Gomez), un video mucho más largo y con transiciones más lentas de lo usual para su público de YouTube.

La enunciación mediática de esta pieza es bromista, Nimu produce un stream donde el objeto risible recae en las escenas enunciativas de los videos a lo que reacciona con risas superficiales contando con (y registrando en vivo) la complicidad del enunciatario hipermediático.

El video tuvo menos comentarios que la media de la enunciadora, 226 en total, de los que solo cuatro fueron marcados como tóxicos por nuestro modelo. Sobre estos cuatro, vemos dos ataques identitarios (uno a chilenos, otro a otakus), algo parecido a una amenaza de muerte y un comentario con espacios entre cada letra que provoca un error en el modelo. El colectivo de seguidores de Nimu, normalmente muy activo y fiel, en este caso participó más en el stream que en la sección de comentarios de YouTube.

4.2.4. Una mirada panorámica II

[14] El link a "Primer stream en YouTube…" es: https://www.youtube.com/watch?v=q_7bnUtqHQ4

¿Y los otros 9 youtubers? ¿Qué diferencia los contratos de lectura de los enunciadores restantes de maritobaracus, AleVera Oficial y Nimu, donde la toxicidad sí se asocia con más vistas e interacción? ¿La correlación entre toxicidad y visibilidad es uniforme a través de este grupo tan heterogéneo?

El punto de contacto entre todos y todas es transmitir discursos avalados por los valores mayoritarios de época. Rodriguez Galati, FFran Gomez, Tincho Ruiz y Hecatombe Producciones crean sketches, que despiertan la risa por lo penoso o lo ridículo, pero sus contenidos están casi exclusivamente determinados por escenas de la vida cotidiana. Romi, PassThor, Nico Villa reaccionan a videos virales, la mayoría de las veces figurando situaciones ridículas, a las que casi siempre responden desde posturas de sentido común, invalidando el discurso o la acción presentada en el video original. Por último, el humor infanto-juvenil de Melina Vallejos que repitió sobre todo dos temáticas: sus perros y la "la Abuela Rita", un personaje que encarna conflictos intergeneracionales con un tono complaciente y familiar.

Así, con diferencias, en todos estos contratos de lectura hay, de una manera u otra, una cláusula de entretenimiento pasajero. Ni los mensajes ni los escenarios que se presentan ante el receptor cuestionan la realidad como es, sino que operan sobre ella sin más.

Solo Guille Aquino escapa a esta norma. Aquino enuncia desde una minoría mucho más numerosa que los antisistemas, los otakus o los fans de GTA: el progresismo urbano. En todos sus sketches Aquino pone en cuestión temáticas sociales y políticas. Allí, vimos cómo los ataques identitarios se asociaban especialmente con un aumento en las réplicas entre comentarios y las visualizaciones totales.

Trazando un paralelismo parlamentario, la diferencia entre Aquino y nuestra selección de tres es la diferencia entre una primera minoría y la representación marginal de un puñado. Estamos ante un caso intermedio desde lo cuantitativo, pero también desde lo cualitativo.

5. Discusión

En un contexto con tantas variables como el analizado, que una sola de ellas, la toxicidad en los comentarios, aislada, se correlacione con un aumento en vistas de forma moderada o fuerte es un resultado notorio e indicativo de que existe cierta asociación entre los comentarios abusivos y la visibilidad que adquieren ciertos videos risibles con enunciación mediática cómica.

Si bien encontramos una correlación moderada, estos resultados están matizados por la ponderación de un clasificador que tuvo una precisión de casi 60% para el total del

set de datos, pero para casos de alta toxicidad bajó a números cercanos al 35%. Además, -y a pesar de haber recolectado los datos apenas semanas después del cierre del nuestro recorte temporal- no podemos descartar que cierto número de comentarios de una toxicidad muy severa hayan sido moderados y eliminados antes de nuestro *crawling*, alterando la naturaleza del fenómeno que finalmente analizamos.

¿Deberíamos entonces arribar a la conclusión de que Perspective no es una herramienta eficaz en la detección de toxicidad en redes sociales? Creemos que no, sino que la particularidad de los discursos risibles suma una cantidad de ruido enorme, como los insultos usados de forma afectiva en el contexto argentino y las referencias o citas a groserías del video.

En la actualidad no existe ningún modelo de NLP que incluya los discursos en producción a modo contextual, tanto nuestro diseño investigativo como cualquier otro que estudiase toxicidad a gran escala se encontraría con el mismo escollo. Creemos que un desafío de estas características podría ser subsanado en futuros trabajos con un sistema que incluya la transcripción de los videos a priori del análisis.

Durante el trabajo señalamos en más de una ocasión que la diseminación de discursos tóxicos en la virtualidad tenía probadas consecuencias en el mundo exterior. ¿Cuál sería entonces la consecuencia de que ciertos grupos, como los observados, crecieran alrededor de mecánicas tóxicas?

El reciente trabajo de Chen, et al. (2022) clarifica en gran medida a este respecto. Este grupo de investigadores de universidades norteamericanas siguió el comportamiento de usuarios que interactuaron con videos de alta toxicidad en YouTube (mayormente sobre teorías conspirativas, supremacismo blanco o noticias falsas de canales de extrema derecha) y concluyeron en que las interacciones no amplifican tanto los contenidos como sí radicalizan al grupo que buscó esos contenidos intencionalmente. La intención de los usuarios se operacionalizó a través del trackeo de links que estos sujetos de experimentación clickeaban voluntariamente, los canales a los que se suscribían y los clicks a las recomendaciones de YouTube.

De acuerdo a sus datos, cerca del 80% de las interacciones con estos canales de alta producción de discursos tóxicos provinieron del 0.6% de los sujetos de su experimentación, y, luego de una interacción positiva, esa minoría de sujetos ávidos de contenidos tóxicos tenía en 1 de cada 4 recomendaciones de videos, contenido similar al que ya habían apoyado con interacciones pasadas. En conclusión: la toxicidad en YouTube no provoca tanto que quien vea videos de pastelería pase a ver videos de extrema derecha, sino que radicaliza a grupos con tendencias “extremistas” -en lo que podría ser calificado como el complemento algorítmico a la teoría antagonística de Laclau y Mouffe-. YouTube no amplifica la demanda tanto como facilita la oferta de discursos tóxicos.

Tal y como repasamos en el apartado sobre las tácticas de adaptación al algoritmo, a pesar de requerir un esfuerzo y compromiso mínimos, los likes influyen decisivamente en la visibilidad de un video. Sobre el total de nuestro set de datos, la correlación entre los likes y la desviación de la media por autor de vistas fue de 0.72, muy fuerte. Similar es el caso entre el puntaje Z de comentarios y las vistas, con una correlación entre moderada y fuerte con R=0.41. De una manera o la otra, las variaciones en las métricas de interacción al final se asocian con variaciones en las vistas, el objetivo primario de todo youtuber.

¿Qué pasaría entonces si, a través de una intuición alimentada por las analíticas de la plataforma, un enunciador descubriese que sus videos con más toxicidad son también los de más likes? ¿O más comentarios, que también lo favorecen? Sabiendo que más interacciones suelen correlacionarse con más vistas ¿Repetiría -o profundizaría- estrategias y formas enunciativas de un video anterior donde haya verificado alta toxicidad, en la búsqueda de replicar esos resultados positivos? ¿Podrían encontrar youtubers como Nimu, maritobaracus o AleVera Oficial incentivos para expandir la toxicidad en la plataforma?

Respuestas plenas a estas preguntas escapan a los alcances de nuestra experimentación y quedan pendientes para futuros trabajos que tengan antes a los enunciadores que a los comentarios de los enunciatarios en el centro de las hipótesis. Lo que es seguro es que, de acuerdo a nuestra muestra, para el caso de las enunciaciones mediáticas cómicas con cierto contenido antisistema, todas las condiciones e incentivos para una capitalización de la toxicidad parecerían estar dadas, y no hay ningún equilibrio institucional o algorítmico que se oponga a ello.

Nimu, maritobaracus, AleVera Oficial y, en menor medida, Guille Aquino, son lo que Howard Becker llamó *outsiders*[15]. Sujetos que infringen, en plena conciencia, normas acordadas por los grupos mayoritarios. “Es desviado quien ha sido exitosamente etiquetado como tal y el comportamiento desviado es aquel que la gente etiqueta como tal” (Becker, 2009 [1963], p. 28).

Becker no considera a la desviación inherente a la acción, sino a la reacción: la mirada que apunta sobre el grupo *outsider* lo define desde la perspectiva de otros sectores, que fueron exitosos al establecer sus normas como las válidas. El grupo *outsider* nace entonces en la interacción social. La desviación no es una cualidad del acto que una persona comete sino una aplicación de reglas y sanciones sobre el individuo infractor a manos de terceros.

Si bien se descalifica como *outsider* a aquello que no es funcional al sistema de valores preponderantes, lejos de posiciones funcionalistas, Becker pone sobre la mesa el

[15] En la versión citada, *outsider* se tradujo como “marginal”.

componente de poder que implica el concepto de lo funcional en sí. Para él, la función de un grupo es producto de una confrontación política antes que algo intrínseco a la naturaleza del grupo.

Repasemos: Nimu es una personalidad virtual que forma parte del mundo otaku, maritobaracus construye su enunciación sobre groserías severas y apelaciones a temáticas explícitamente sexuales que lo ponen en conflicto con la cultura tradicional y AleVera Oficial produce a través de un videojuego que fue objeto de incontables críticas por su naturaleza violenta. Guille Aquino enuncia desde una progresía porteña que, aunque minoritaria y etiquetada, merece un matiz al diferenciarse de los 3 casos puros en la escala de aprobación social. Los títulos de otakus, "loquenderos" o "progres" son las etiquetas que diferencian a estos *outsiders* de la aparente normalidad.

La enunciación de estos youtubers se funda en la contestación a los valores tradicionales, es en el choque con esas normas o sus representantes que se verificaron buena parte de los episodios de alta toxicidad. Por otro lado, el humor acorde a las normas de época, representado en los otros 8 youtubers de la muestra y su colectivo de fans, no lleva en su gen el desmarcarse con violencia de un afuera. Ellos son ese afuera de los *outsiders*, o al menos representan los valores de la mayoría.

En cuanto a los significados y efectos de los insultos y otro tipo de discursos tóxicos en el contexto argentino, la literatura no abunda -y, si recortáramos sólo en discursos online, los cardos rodantes pasarían ante nuestros ojos como en un western-. Desde una perspectiva antropológico-lingüística, Gabriel Hernández (2014) fue el que más se aproximó a esta temática con su estudio de jóvenes de la Provincia de Buenos Aires y llegó a la conclusión de que el contexto determina el significado anti-cortés de un insulto.

Entre amigos, un insulto puede significar confianza "ya que una relación afectiva cercana posibilita y habilita una resignificación de las 'malas palabras' como marcas de identidad común y de amistad" (Hernández, *ibidem*, p.45). En entornos de desconocidos -como YouTube-, las "malas palabras" tendrían para Hernández la función de construir una identidad positiva para un grupo determinado o de dañar la imagen de un tercero.

Respecto a los ataques identitarios en Argentina, Alderete, et al. (2012) estudiaron grupos de jóvenes en el Norte argentino y concluyeron que, dentro de su muestra, la discriminación racial/étnica es un lugar común entre los jóvenes latinoamericanos y, que en el análisis bivariado, un porcentaje mayor de jóvenes con síntomas depresivos reportó haber estado expuesto a insultos raciales. Las consecuencias de los discursos discriminatorios sobre la autopercepción en jóvenes no parecerían ser muy distintos en el contexto argentino al que se registran en muchas otras latitudes influidas por

culturas tecnológicas con características compartidas (para más profundidad al respecto, remitirse a la ya citada investigación de David Lester, et al. sobre el suicidio de la adolescente canadiense Amanda Todd).

Si, además, tomáramos en consideración que al momento de la recolección de nuestros datos los youtubers de la muestra oscilan entre los 22 (Nimu y Melina Vallejos), y los 36 años (Guille Aquino) y sumáramos todas las marcas en la enunciación que dan cuenta de un enunciatario joven (desde las formas, en los verosímiles o lenguajes establecidos, hasta los contenidos, con una demanda de conocimientos de animé o videojuegos que raramente se encuentren en generaciones más avanzadas), entonces podríamos asegurar que la mayoría de los receptores y productores de los comentarios sobre los que se centra esta investigación son también jóvenes y argentinos.

En total, la evidencia hace pensar que las consecuencias de la toxicidad severa detectada en nuestro recorte podrían rebasar la virtualidad y, aunque buena parte de la literatura apunta a impactos negativos en la salud mental de aquellos que son blanco de odio, el interrogante acerca de la dirección y magnitud de estas consecuencias también queda pendiente de próximas indagaciones. Al final, la suma de inferencias conduce rápidamente al debate sobre el alcance de la moderación de contenidos y comentarios en redes sociales, debate vastísimo y multidisciplinario en el que solo nos sentimos capaces de colaborar desde las conclusiones de nuestro experimento.

Por un lado, YouTube controla férreamente qué videos se publican para no herir las susceptibilidades de los titanes de la industria del entretenimiento -no por nada, señala Van Dijck (2013), el slogan de YouTube pasó de “Your online video repository”[16] a “Broadcast Yourself”[17] poco después de que Google adquiriera la compañía en 2006 para sacarla de las tinieblas del contenido sin copyright-. Pero por el otro, la sección de comentarios, que se presupone satelital o residual, está mucho más desatendida en términos de moderación. Acaso por la dificultad en monetizarla en el corto plazo, acaso por una audiencia de demandas de los usuarios al respecto.

Así explicaba Van Dijck los cambios en las prioridades de la plataforma:

> Poniendo el énfasis en sus cualidades de red social por encima de sus propósitos como sitio de contenido generado por los usuarios, el primer YouTube promocionaba el material de video como un medio para la conformación de comunidades y la actividad grupal. Si se mira con atención la interfaz de YouTube de 2008, todavía se advierte en ella el lugar central del usuario: los botones que permiten comentar los videos de otros miembros y establecer comunidades son muy visibles en la página de inicio. Sin

[16] Se traduce como: “Tu repositorio online de videos”.
[17] Podría traducirse como “Transmitite a vos mismo” o “Transmití por tu cuenta”.

> embargo, las distintas características de la interfaz que promueven la conectividad poco a poco complementaron o sustituyeron a aquellas que alentaban la creatividad: la inclusión de videos en miniatura y botones destacados para acceder a los rankings de los "más vistos" y "los favoritos" potenciaron las rutinas de los consumidores.
> (Van Dijck, ibidem, p. 121)

Aun cuando el diseño de la interfaz no los coloque en un lugar tan central como antaño, en algunos tipos de enunciadores, la toxicidad tiene cierta relación lineal con las vistas y las interacciones; en estos registros, los cómicos en nuestro recorte, todas las condiciones e incentivos parecerían dados para que se extiendan los videos que más comentarios con toxicidad cosechan.

Los resultados de Perspective, principalmente entrenado a partir de comentarios en medios online, se resienten ante nuestro desafío de naturaleza doble: a los cambios de dominio, de sitios de noticias a videos risibles en YouTube, se le suman los cambios propios de la variedad lingüística rioplatense.

De seguir la línea de Bender y Koller (2020), las imprecisiones más llamativas que detectamos a lo largo de nuestro trabajo podrían explicarse en el acercamiento al lenguaje desde las formas -antes que los contenidos y el contexto- que hacen las técnicas actuales de NLP. En un medio virtual lleno de formas, las tareas de modelado de lenguaje solo se valen de este recurso en una búsqueda de significado destinada a fallar por no contar con una comprensión holística de un fenómeno como el sentido:

> …in contrast to some current hype, meaning cannot be learned from form alone. This means that even large language models such as BERT do not learn "meaning"; they learn some reflection of meaning into the linguistic form which is very useful in applications.[18]
> (Bender y Koller, *ibidem*, p. 5193).

Nuestro estudio de método mixto, sobre una muestra de más de 1.2 millones de comentarios, da cuenta de que el entusiasmo a partir de las posibilidades que abren las nuevas técnicas de la llamada Inteligencia Artificial debe ser acompañado con una sana dosis de escepticismo y con la cogobernanza de aquellos que serían impactados por sus decisiones.

De lo contrario, aplicadas a tareas de moderación, las diferentes formas de Inteligencia Artificial que venían a liberar a los humanos de tareas repetitivas y mecánicas para que los trabajadores puedan volcarse a tareas más creativas (Lee, 2019), acabarán limitando relaciones afectivas reales y la creatividad. La producción de contenidos

[18] En castellano: "En contraste con el entusiasmo actual, el sentido no puede aprenderse sólo a partir de la forma. Esto significa que incluso los grandes modelos lingüísticos como BERT no aprenden el 'sentido', sino que aprenden algún reflejo del sentido en la forma lingüística que es muy útil en las aplicaciones."

risibles, uno de los géneros más populares y abundantes en Internet, nos parece un buen ejemplo de esta mecánica. Es más una observación que un vaticinio, hoy las críticas “al algoritmo” son especialmente comunes entre enunciadores hipermediáticos de géneros risibles.

Van Dijck, et al. (2018) señalan que muchas plataformas se han vuelto sorprendentemente influyentes antes de que pudiera comenzar un debate real sobre bien común y valores públicos. Los placeres tecnológicos no solo han hecho que tales debates pasen casi desapercibidos para la opinión pública, sino que la velocidad de adopción ha contribuido a un mito de la neutralidad técnica que, a pesar de sus numerosas refutaciones desde las más diversas perspectivas (Heidegger, 1997 [1954]; Haraway, 1995 [1984] Feenberg 2012 [1991]; Castoriadis, 2004) sigue formando parte de cierto sentido común.

La solución a problemas de naturaleza social como la moderación de contenidos tóxicos en redes sociales difícilmente se encuentre en una Inteligencia Artificial autónoma entrenada a partir de datos más apropiados o arquitecturas algorítmicas más eficientes. Asimismo, una resolución que deje de lado herramientas técnicas parece inevitablemente destinada al fracaso.

En cambio, mecanismos de regulación y supervisión humana, transparentes con las partes involucradas, podrían mitigar algunos de los daños que sistemas automáticos pueden propagar con máxima eficacia. Creemos que los resultados de nuestro estudio dan soporte parcial a ambos argumentos: tanto normas de moderación estrictas como sistemas automáticos pueden infligir más daño del que curan sin añadir una mirada holística y contextual que comprenda a la circulación de sentido en su naturaleza indivisible.

Diez años antes de que ELIZA, uno de los primeros modelos de NLP se alzara a los tropiezos en los laboratorios del MIT, Martin Heidegger advertía respecto a que, mientras en el peligro reside lo salvador, extremado, este peligro puede recluir a la humanidad en lo ya establecido. Puede que en el presente sus palabras sean de ayuda para transitar la delgada frontera entre la automatización y la humanización de procesos necesarios en el diseño de normas para sociedades más prósperas, justas y democráticas:

> Lo esente de la técnica amenaza al desocultar, amenaza con la posibilidad de que todo desocultar vaya a parar al establecer y que todo se conciba únicamente en el desvelamiento de lo constante. El hacer humano jamás puede enfrentar este peligro inmediatamente. El esfuerzo humano no puede por sí solo conjurar el peligro. Sin embargo, la reflexión humana puede meditar que todo lo salvador tiene que ser una esencia más elevada, aunque emparentada al mismo tiempo con lo amenazado por el

peligro.
(Heidegger, 1997 [1954], p. 146)

6. Conclusiones

Durante nuestra investigación analizamos más de 1.2 millones de comentarios de 826 videos humorísticos argentinos con la herramienta más avanzada para detección automática de toxicidad de la actualidad. Nuestra búsqueda de correlaciones entre toxicidad severa y vistas e interacciones se guio, primero, por la pista de que colectivos homogéneos se unirían bajo lógicas antagónicas, segundo, por la creencia de que enunciadoras femeninas podrían recibir más hostilidad que los masculinos y, por último, de que diferentes juegos en el lenguaje modificarían el grado de respuesta tóxica de los receptores.

La prueba experimental no fue suficiente para validar dos de nuestras hipótesis: la correlación entre vistas e interacción y toxicidad a la escala de los primeros 12 puestos del ranking de la sección “Comedia” en YouTube Argentina no fue tal, así como tampoco la tesis de que enunciadoras femeninas recibirían más toxicidad que los masculinos.

En el caso de esta segunda hipótesis, no pensamos que nuestros resultados sean suficientes para descartar una prevalencia de la violencia en redes hacia personas femeninas. Más bien, consideramos a este como un primer experimento, a ser replicado sobre sets de datos balanceados y con recortes que no se enfoquen en el número de suscriptores sino en la coincidencia estilística entre youtubers femeninos, masculinos y de otros géneros.

En cuanto a la hipótesis primaria, consideramos que parte de la explicación de esta correlación entre leve y nula podría encontrarse en el viraje contemporáneo hacia contenidos más acordes a los valores de época, posiblemente como consecuencia de la diseminación de una cultura de la cancelación que despertó en muchos enunciadores la pregunta acerca de “los límites del humor” y, en las plataformas, un afán por no amplificar contenidos “extremos”. Progresivamente, buena parte de la oferta y la demanda *mainstream* se adaptó, minimizando un factor violento que, no obstante, es intrínseco al humor. En términos generales, la robustez de nuestros resultados nos permite afirmar -en contra de cierto sentido común- que en la etapa contemporánea sí existe la publicidad negativa y que los enunciadores hipermediáticos más populares parecerían esquivar las situaciones de alta toxicidad.

Entre aquellos que no solo no adaptaron su enunciación a los valores de época, sino que, en general, construyeron discursos contrarios a algunas de esas normas, encontramos como exponentes más puros a maritobaracus, Nimu y AleVera Oficial. En estos tres canales, las correlaciones estadísticamente significativas entre toxicidad

severa y vistas o diferentes métricas de interacción fueron mayormente moderadas y, en algunos casos, fuertes. Asimismo, en Guille Aquino se verifica un fenómeno similar con la variable de ataques identitarios.

Más allá de su gran visibilidad, los youtubers donde las correlaciones son suficientemente fuertes representan valores marginados: la vida virtual y el otakismo, la violencia explícita, la progresía. Allí, podríamos afirmar que lo tóxico garpa. Mientras que el resto de la muestra, más propensa a contenidos más integrados, no tuvo correlaciones significativas.

El clasificador captó un nivel de ruido, propio de la instancia de producción risible, que afectó nuestras posibilidades de dar con una conclusión definitiva en cuanto a la relación entre toxicidad en los comentarios y visibilidad de un video humorístico.

Sin embargo, este precedente no solo nos abre la puerta a nuevas líneas investigativas. También pone luz sobre las limitaciones del uso autónomo de herramientas estadísticas en los estudios sobre toxicidad en el campo de lo risible, y de la necesidad de incorporar el contexto en los sistemas de NLP para una comprensión completa de la circulación discursiva. El de la necesidad de contexto en la identificación de discursos del odio y toxicidad es uno de los interrogantes que el campo del NLP busca definir en los últimos años, y esperamos que nuestra prueba experimental aporte a la posición por la positiva.

Solo pudimos arribar a estas conclusiones desde el cruce de métodos computacionales de estudio del lenguaje con la teoría veroniana que enlaza producción y reconocimiento, en cadenas cuya ruptura coincide con la ruptura del sentido pleno. Nos esperanzan los aportes que la adaptación contemporánea de la perspectiva veroniana pueden hacer en el desafío de que algún día las computadoras puedan aproximarse a la cognición humana del lenguaje y consideramos los resultados de este experimento como el primer paso de un camino interdisciplinar que solo podrá hacerse al andar.

7. Bibliografía

Alakrota, A., Murray, L., & Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science 142*, 174-181.

Alderete, E., Monteban, M., Gregorich, S., Kaplan, C. P., Mejıa, R., & Pérez-Stable, E. J. (2012). Smoking and exposure to racial insults among multiethnic youth in Jujuy, Argentina. *Cancer Causes Control 23*, 37-44.

Álvarez Gandolfi, F. (2016). Cibercultura otaku: un análisis interdiscursivo de identidades fan puestas en escena en grupos de Facebook. Perspectivas de la Comunicación 9, 31-57.

Article 19. (2015). Hate speech explained: A toolkit. *Technical report, Article 19*.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*, 543-554.

Becker, H. (2009 [1963]). Outsiders: hacia una sociología de la desviación. Siglo XXI Editores. Buenos Aires.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (págs. 5185-5198). Online: Association for Computational Linguistics.

Benesch, S. (2003). Vile Crime or Inalienable Right, Defining Incitement to Genocide. Virginia Journal of International Law, 48(3), pp. 485-528.

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer. Singapur.

Bleich, E. (2013). Freedom of Expression versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the. *Journal of Ethnic and Migration Studies*, 283-300.

Blout, E., & Burkart, P. (2020). White supremacist terrorism in charlottesville: Reconstructing 'unite the right'. *Studies in Conflict & Terrorism*, 1-22.

Blyth, Colin R. (1972). On Simpson's Paradox and the Sure-Thing Principle. Journal of the American Statistical Association.

Brodersen, A., Scellato, S., & Wattenhofer, M. (2012). YouTube around the world: geographic popularity of videos. *Proceedings of the 21st international conference on World Wide Web*, (págs. 241-250).

Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Perez, J. (2020). Spanish pre-trained BERT model and evaluation data. En *PML4DC at ICLR 2020.*

Carlón, M. (2015). Público, privado e íntimo: el caso Chicas Bondi y el conflicto entre derecho a la imagen y libertad de expresion en la circulacion contemporánea. *Editora da Universidade Federal de Alagoas*, 211-232.

Carlón, M. (2016). Apropiación contemporánea de la teoría de la comunicación de Eliseo Verón. En E. Vizer, & C. Vidales, *Comunicación, campo(s),*

teorías y problemas. Una perspectiva internacional. Salamanca: Comunicación Social Ediciones.

Carlón, M. (2020). *Circulación del sentido y construcción de colectivos: en una sociedad hipermediatizada.* San Luis: Nueva Editorial Universitaria.

Castoriadis, C. (2014 [1978]). Técnica. Buenos Aires. Revista Artefacto, 5. pp. 50-66.

Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010). A First Step Towards Understanding Popularity in YouTube. *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, (págs. 1-6).

Chen, A., Nyhan, B. Reifler, J., Robertson, R., Wilson, C. (2022). Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. Obtenido de: https://cpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/5/2293/files/2022/04/YouTube.pdf

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York: Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* . Hillside, New Jersey: Lawrence Erlbaum Associates.

Comisión Europea. (2021). *6th evaluation of de Code of conduct.* Unión Europea.

Cowan, G., & Hodge, C. (1996). Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, 355-374.

Cowan, G., Resendez, M., Marshall, E., & Quist, R. (2002). Hate speech and constitutional protection: Priming values of equality and freedom. *Journal of Social Issues 58*, 247-263.

De Certeau, M. (1996 [1990]). *La invención de lo cotidiano. I. Artes de hacer.* México: Universidad Iberoamericana.

Dean, B. (2017). *Backlinko*. Obtenido de https://backlinko.com/youtube-ranking-factors

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Döring, N., & Mohseni, M. R. (2020). Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. *Studies in Communication and Media*, 62-88.

Eco, U. (1996). *La estrategia de la ilusión.* Buenos Aires: Editorial Lumen/Ediciones de la Flor.

Feenberg, A. (2012 [1991]) Transformar la tecnología, una nueva visita a la teoría crítica. Bernal, Universidad Nacional de Quilmes. 21-67.

Fons, R. (2020). *YouTube*. Obtenido de ¿NO CRECES EN YOUTUBE? (Haz Esto!) - Cómo Conseguir Más Visitas y Suscriptores: https://www.youtube.com/watch?v=muSG1nPsbpw

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *Association for Computing Machinery, 51*, 1-30.

Fox, J., Cruz, C., & Young Lee, J. (2015). Participating in online sexist behavior increases levels of hostile sexism and has offline impacts in the workplace. *Computers in Human Behavior 52*, 436-442.

Fraticelli, D. (2019). Los colectivos mediáticos de las redes: Algunas observaciones desde el humor ¿y más allá? *Mediaciones*, 47-63.

Fraticelli, D. (2019). Zekiel79: youtubers y medios masivos. La circulación de una. En M. Carlón, *La (in)comunicación contemporánea: de las redes sociales a los medios masivos y de los medios masivos a las redes sociales.*

Fraticelli, D. (2021). Enunciación y humor en las redes (o cómo estudiar memes sin perder el chiste) . *La Trama de la Comunicación, 25*, 115-129.

Genette, G. (2005 [2002]). *Muertos de risa.* Buenos Aires: Siglo XXI editores.

Goodfellow, I., Bengio, Y., Couville, A. (2016). Deep Learning. MIT Press.

Haraway, D. (1995 [1984]). Manifiesto ciborg. En Ciencia, cyborgs y mujeres. La reinvención de la naturaleza. Madrid. Cátedra.

Heidegger, M. (1997 [1954]). "La pregunta por la técnica", en Ciencia y técnica. Santiago de Chile, Editorial Universitaria, 3ra edición.

Hernández, G. (2014). Manifestación de la descortesía y anticortesía en jóvenes de la Provincia de Buenos Aires, Argentina: usos y representaciones de "malas palabras" e insultos. *Signo y Seña, 26*, 23-47.

Hidalgo-Marí, T., & Segarra-Saavedra, J. (2017). El fenómeno youtuber y su expansión transmedia. Análisis del empoderamiento juvenil en redes sociales. *Journal of Communication*, 43-56.

Jiang, S., Robertson, R. E., & Wilson, C. (2019). Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media.*

Jigsaw. (2020). *Why we use the term "toxicity"*. Obtenido de https://jigsaw.google.com/the-current/toxicity/#:~:text=Throughout%20this%20report%20we%20use,make%20someone%20leave%20a%20discussion.

Kamisar, B. (3 de 8 de 2018). Conservatives cry foul over controversial group's role in YouTube moderation. Washington D.C.

Kaplan, A. (2022). *Artificial Intelligence, Business and Civilization: Our Fate Made in Machines.* Routledge.

Kearney, A., & Octon, O. (2019). Making Meaning: Semiotics Within Predictive Knowledge Architectures.

Khan, M. L. (2017). Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior*, 236-247.

Kievit, Rogier; Frankenhuis, Willem; Waldorp, Lourenz; Borsboom, Denni. (2013). Simpson's paradox in psychological science: a practical guide. Frontiers in Psychology (4). pp 1-14.

Koehler, D. (2019). Violence and Terrorism from the Far-Right: Policy Options to Counter an Elusive Threat. *Technical report, International Centre for Counter-Terrorism, La Haya*.

Laclau, E., & Mouffe, C. (1987 [1985]). *Hegemonía y estrategia socialista Hacia una radicalización de la democracia.* Madrid: Siglo XXI.

Lee, K. (2019). AI's real impact? Freeing us from the tyranny of repetitive tasks. Obtenido de https://www.wired.co.uk/article/artificial-intelligence-repetitive-tasks

Lees, A., & al, e. (2022). A New Generation of Perspective API: Efficient Multilingual Character-level Transformers.

Lester, D., McSwain, S., & Gunn III, J. F. (2013). Suicide and the internet: The case of Amanda Todd. *International Journal of Emergency Mental Health and Human Resilience, 15*, 179-180.

Mall, R., & al., e. (2020). Four Types of Toxic People: Characterizing Online Users' Toxicity over Time. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (págs. 1-11). New York: Association for Computing Machinery.

McClelland, K., & Hunter, C. (1992). The perceived seriousness of racial harassment. *Social Problems 39*, 92-107.

McGraw, P., & Warren, C. (2010). *Benign Violations: Making Immoral Behavior Funny.* Washington D.C.: Psychological Science.

McLeod, S. A. (2019). Z-score: definition, calculation and interpretation. Simply Psychology. www.simplypsychology.org/z-score.html

Mikolov, T., Ilya, S., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

Möller, A. M., Kühne, R., Baumgartner, S. E., & Peter, J. (2018). Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube. *Social Science Computer Review*, 510-528.

Mouffe, C., & Errejón, Í. (2015). *Construir pueblo. Hegemonía y radicalización de la democracia.* Barcelona: Icaria.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on arabic social media. *Proceedings of the First Workshop on Abusive Language Online*, (págs. 52-56).

Murthy, D., & Sharma, S. (2018). Visualizing YouTube's comment space: online hostility as a networked phenomena. *New Media & Society*, 191-213.

New York's Department of Financial Services and Department of Health. (2019). To Optum. New York. Obtenido de https://dfs.ny.gov/system/files/documents/2019/10/20191025160637.pdf

Nockleby, J. T. (2000). Hate speech. Encyclopedia of the American Constitution, 3:1277–79.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*, 447-453.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis. *LREC, 10*, 1320-1326.

Peirce, C. S. (1994 [1901]). *Collected Papers.* Cambridge: Harvard University Press.

Pellar, Marie & Georgiou, Patricia. (2018). Perspective Launches in Spanish with El Pais. Obtenido de https://medium.com/jigsaw/perspective-launches-in-spanish-with-el-pa%C3%ADs-dc2385d734b2

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *arXiv:2106.09462*.

Pinto, H., Almeida, J. M., & Gonçalves, M. A. (2013). Using early view patterns to predict the popularity of youtube videos. *Proceedings of the sixth ACM international conference on Web search and data mining*, (págs. 365-373).

Pires, F., Masanet, M.-J., & Scolari, C. A. (2019). What are teens doing with YouTube? Practices, uses and metaphors of the most popular audio-visual platform. *Information, Communication & Society*, 1175-1191.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 477-523.

Poole, D. (2018). Computational Intelligence: A Logical Approach. *Oxford University Press*, 1.

Poruthiyil, P. V. (2021). Case study: A dangerous speech that triggered lynchings in. *Dangerous Speech Project.*

Rieder, B. (2015). YouTube Data Tools (Version 1.22) [Software].

Röchert, D., Neubaum, G., Ross, B., & Brachten, F. (2020). Opinion-based Homogeneity on YouTube: Combining Sentiment and Social. *Computational Communication Research*, 81-108.

Röchert, D., Neubaum, G., Ross, B., & Stieglitz, S. (2022). Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube. *Information Systems, 103*.

Ruiz Caballero, C., & al, e. (2011). Public Sphere 2.0? The Democratic Qualities of Citizen Debates in Online Newspapers. *The International Journal of Press/Politics, 16*, 463-487.

Sang, Y., & Stanton, J. (2021). The Origin and Value of Disagreement Among Data Labelers: A Case Study.

Saussure, F. d. (1991 [1915]). *Curso de lingüística general.* Madrid: Alianza.

Schultes, P., Dorner, V., & Lehner, F. (2013). Leave a Comment! An In-Depth Analysis of User Comments on YouTube. *WIRTSCHAFTSINFORMATIK PROCEEDINGS*.

Shackell, C., & Sitbon, L. (2019). Computational Opposition Analysis Using Word Embeddings: A Method for Strategising Resonant Informal Argument. *Argument & Computation*, 301-317.

SimilarWeb. (2021). Obtenido de Top Websites Ranking: https://www.visualcapitalist.com/the-50-most-visited-websites-in-the-world/

Sprejer, L., & al, e. (2021). An influencer-based approach to understanding radical right viral tweets.

Statista. (2021). *Most popular websites worldwide as of June 2021, by total visits.* Obtenido de https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/

Statista. (2022). *Distribution of YouTube users worldwide as of January 2022, by age group and gender.* Obtenido de https://www.statista.com/statistics/1287137/youtube-global-users-age-gender-distribution/

Steimberg, O. (2001). *Sobre algunos temas y problemas del análisis del humor gráfico en Signo y seña.* Buenos Aires: Instituto de Lingüística, Facultad de Filosofía y Letras, UBA.

Thelwall, M., & Sud, P. (2011). Commenting on YouTube Videos: From Guatemalan. *Information Science and Technology, 63*, 616–629.

Tur-Viñes, V., & González-Río, M. (2019). Youtuber and Community Management strategies. *Revista Latina de Comunicación Social, 74*, 1291-1307.

Van Dijck, J. (2013). The culture of connectivity. A critical history of social media. New York, Oxford University Press.

Van Dijck, J., Poell, T., & De Wall, M. (2018). *The Platform Society.* New York: Oxford University Press.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia. (2017). Attention Is All You Need.

Verón, E. (1985). El análisis del “Contrato de Lectura”, un nuevo método para los estudios de posicionamiento de los soportes de los media. *"Les Medias: Experiences, recherches, actualles, aplications*".

Verón, E. (1987). El sentido como producción discursiva. En *La semiosis social. Fragmentos de una teoría de la discursividad* (pág. 127). Barcelona: Gedisa.

Verón, E. (1993). *La semiosis social. Fragmentos de una teoría de la discursividad.* Barcelona: Gedisa.

Verón, E. (1999). Esquema para el análisis de la mediatización. *Diálogos de la comunicación*, 9-17.

Verón, E. (2001). El living y sus dobles. Arquitecturas de la pantalla chica. En *El cuerpo de las imágenes.* Buenos Aires: Norma.

Verón, E. (2004). Cuando leer es hacer: la enunciación en el discurso de la prensa gráfica. En *Fragmentos de un tejido.* Barcelona: Gedisa.

Verón, E. (2013). La semiosis social, 2. Ideas, momentos, interpretantes. Buenos Aires: Paidós.

Wagner, Clifford H. (1982). Simpson's Paradox in Real Life. The American Statistician 36 (1), 46-48.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop* (págs. 88-93). San Diego, California: Association for Computational Linguistics.

Wittgenstein, L. (1988 [1953]). *Investigaciones filosóficas.* México: Crítica-UNAM.

Wotanis, L., & McMillan, L. (2014). Performing Gender on YouTube. *Feminist Media Studies, 14*, 912-928.

Wu, S., & Resnick, P. (2021). Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don’t Talk to Conservatives.

YouTube. (2016). *Report inappropriate content*. Obtenido de https://support.google.com/youtube/answer/2802027?hl=en&co=GENIE.Platform%3DDesktop

YouTube. (2020). Obtenido de Community guidelines: https://www.youtube.com/intl/ALL_in/howyoutubeworks/policies/community-guidelines/?utm_campaign=1008960&utm_source=paidsearch&yt_product=ytgen&yt_goal=eng&utm_medium=googlesearch&utm_content=txt&yt_campaign_id=hyw&yt_creative_id=&utm_keyword=%2Byoutube%20

YouTube. (2021). *How to use moderation tools*. Obtenido de https://support.google.com/youtube/answer/10888907?hl=en&ref_topic=9257792

8. Anexos

8.1. Anexo I: crawler de YouTube en PHP

Pasos para ejecutar:

1- Guardar el código PHP en una carpeta.
2- En esa carpeta, creamos otra carpeta bajo el nombre “crondata”.
3- Crear archivos .txt con un id de video por línea.
4- Ejecutar el código en una terminal con el siguiente comando (actualizando el “NOMBREDEARCHIVO”). : php mod_video_info.php videolist=./NOMBREDEARCHIVO.txt
5- El crawler ya debería haber comenzado. Cuando la cuota llegue a su fin, debemos volver a la terminal y comprobar el último id que se recolectó. Esto será indicado con el siguiente mensaje: BAJANDO COMENTARIOS PARA EL VIDEO [id].
6- Debemos volver al .txt del youtuber, borrar los ids anteriores al indicado, y repetir volver a ejecutar el código hasta que todos los videos del youtuber pasen por el crawler.
7- Una vez recolectados los comentarios para todos los videos de ese youtuber, entramos a la carpeta crondata y los resultados de nuestro crawling ya están listos. Sería recomendable guardar ese conjunto de archivos en una carpeta por youtuber, antes de continuar con más recolecciones.

```
<?php include("html_head.php"); ?>


    <div class="rowTab">

        <div class="sectionTab">

            <h1>Video Info and Comments Module</h1>

        </div>

    </div>


    <div class="rowTab">

        <div class="fullTab">
```

```
            <p>This module starts from a video id and retrieves
basic info for the video in question and provides a number of
analyses of the comment section.

            Comments are retrieved via the <a
href="https://developers.google.com/youtube/v3/docs/commentThreads/
list" target="_blank">commentThreads/list</a> API endpoint.</p>


            <p>The number of comments the script is able to
retrieve can vary wildly. In some cases, only a relatively small
percentage is made available, while in others well over

            100.000 comments have been successfully retrieved. This
seems to be mainly related to the age of the video in question.</p>


            <p>The module creates the following outputs:

                <ul>

                    <li>a tabular file containing basic info and
statistics about the video;</li>

                    <li>a tabular file containing all retrievable
comments, both top level and replies;</li>

                    <li>a tabular file containing comment authors
and their comment count;</li>

                    <li>a network file (gdf format) that maps
interactions between users in the comment section;</li>

                </ul>

            </p>


            <p>The first three elements can be shown directly in
the browser by enabling HTML output.</p>

        </div>

    </div>


    <div class="rowTab">

        <div class="sectionTab"><h1>Parameters</h1></div>
```

```
    </div>


    <form action="mod_video_info.php" method="get">


    <div class="rowTab">

        <div class="leftTab">Video id:</div>

        <div class="rightTab">

            <input type="text" name="videohash" value="<?php
if(isset($_GET["videohash"])) { echo $_GET["videohash"]; } ?>" />
(video ids can be found in URLs, e.g. <span
class="grey">https://www.youtube.com/watch?v=</span><b>aXnaHh40xnM<
/b>)

        </div>

    </div>


    <div class="rowTab">

        <div class="leftTab">HTML output:</div>

        <div class="rightTab">

            <input type="checkbox" name="htmloutput" <?php
if($_GET["htmloutput"] == "on") { echo "checked"; } ?> /> (displays
HTML result tables in addition to file exports)

        </div>

    </div>


    <div class="g-recaptcha" data-
sitekey="6Lf093MUAAAAAIRLVzHqfIq9oZcOnX66Dju7e8sr"></div>


    <div class="rowTab">

        <div class="leftTab"></div>

        <div class="rightTab">

            <input type="submit" />
```

```
        </div>
    </div>


    </form>


<?php


// blocked video example:
https://www.youtube.com/watch?v=pLN59ZOweUE


$folder = $datafolder;


// allow for direct URL parameters and command line for cron
// e.g. php mod_video_info.php videohash=aXnaHh40xnM or php
mod_video_info.php videolist=videolist_xy.tab (file must be in
cronfolder)
// don't forget to set $cronfolder in config.php
if(isset($argv)) {
    parse_str(implode('&', array_slice($argv, 1)), $_GET);
    $folder = $cronfolder;
}


$feed = array();
$feed["comments"] = array();


$video = array();




if(isset($_GET["videolist"])) {
```

```
    $filename = "./".$folder."/".$_GET["videolist"];


    $header = ["url"];
    $videolist = array();
    if(($handle = fopen($filename, 'r')) !== FALSE) {
        while (($row = fgetcsv($handle,0,"\t",chr(8))) !== FALSE) {
            if(!$header)
                $header = $row;
            else
                $videolist[] = array_combine($header, $row);
        }
        fclose($handle);
    } else {
        echo "no list file found"; exit;
    }
    foreach($videolist as $video) {
        $filename =
"videoinfo_".$video["channelId"]."_".$video["videoId"]."_".date("Y_
m_d-H_i_s");
        getComments($video["url"]);
    }



} else if(isset($_GET["videohash"])) {


    echo '<div class="rowTab">
            <div class="sectionTab"><h1>Results</h1></div>
         </div>
         <div class="rowTab">';
```

```
    if(RECAPTCHA) {
        if($_GET["g-recaptcha-response"] == "") {
            echo "Recaptcha missing.";
            exit;
        }
        testcaptcha($_GET["g-recaptcha-response"]);
    }

    if($_GET["videohash"] == "") {
        echo "Missing video id.";
        exit;
    }

    echo 'Processing:';

    $videohash = $_GET["videohash"];
    $html = $_GET["htmloutput"];
    $commentonly = $_GET["commentonly"];
    $filename = "videoinfo_".$videohash."_".date("Y_m_d-H_i_s");

    $video = getInfo($videohash);
    $nodecomments = getComments($videohash);
    $commenters = getCommenters($nodecomments);
    makeNetwork($nodecomments);

    echo '<br /><br />The following files have been generated:<br />';
    echo '<a href="./data/'.$filename.'_basicinfo.tab" download>'.$filename.'_basicinfo.tab</a><br />';
```

```
    echo '<a href="./data/'.$filename.'_comments.tab"
download>'.$filename.'_comments.tab</a><br />';

    echo '<a href="./data/'.$filename.'_authors.tab"
download>'.$filename.'_authors.tab</a><br />';

    echo '<a href="./data/'.$filename.'_commentnetwork.gdf"
download>'.$filename.'_commentnetwork.gdf</a><br />';

    echo '<br />';



    if($html == "on") {


        echo "<hr /><br />";


        // output basic video info table

        echo '<table class="resulttable">';

        foreach($video as $key => $data) {

            echo '<tr class="resulttable">';

            echo '<td class="resulttableHi"><b>'.$key.'</b></td>';

            echo '<td class="resulttable">'.$data.'</td>';

            echo '</tr>';

        }

        echo '</table>';


        echo "<br /><br />";



        // output author list

        echo '<table class="resulttable">';

        foreach($commenters as $username => $count) {

            echo '<tr class="resulttable">';
```

```
            echo '<td
class="resulttableHi"><b>'.$username.'</b></td>';
            echo '<td class="resulttable">'.$count.'</td>';
            echo '</tr>';
        }
        echo '</table>';

        echo "<br /><br />";



        // output full comment table
        echo '<table class="resulttable">';
        echo '<tr class="resulttable">';
        foreach(array_keys($nodecomments[0]) as $key) {
            echo '<td class="resulttableHi"><b>'.$key.'</b></td>';
        }
        echo '</tr>';
        foreach($nodecomments as $comment) {
            $style = ($comment["isReply"] == 0) ?
"resulttable":"resulttableHi";

            echo '<tr class="resulttable">';
            foreach($comment as $element) {
                echo '<td class="'.$style.'">'.$element.'</td>';
            }
            echo '</tr>';
        }
        echo '</tr>';
        echo '</table>';
```

```
    }


    echo '</div>';
}



function getInfo($videohash) {


    global $html,$filename,$folder;


    // forbidden: fileDetails,processingDetails,suggestions
    $restquery =
"https://www.googleapis.com/youtube/v3/videos?part=statistics,conte
ntDetails,snippet,status,topicDetails&id=".$videohash;


    $reply = doAPIRequest($restquery);
    if(count($reply->items) == 0) {
        echo "<br /><br />No results found. You are probably not
using a valid video id."; exit;
    }
    $reply = $reply->items[0];
    $video = array();


    $video["id"] = $reply->id;


    $video["published"] = date("Y-m-d H:i:s", strtotime($reply-
>snippet->publishedAt));
    $video["published_unix"] = strtotime($reply->snippet-
>publishedAt);
    $video["title"] = preg_replace("/\s+/", " ",$reply->snippet-
>title);
```

```
    $video["description"] = preg_replace("/\s+/", " ",$reply-
>snippet->description);

    $video["channelId"] = $reply->snippet->channelId;

    $video["channelTitle"] = $reply->snippet->channelTitle;



    $video["duration"] = $reply->contentDetails->duration;

    $video["dimension"] = $reply->contentDetails->dimension;

    $video["definition"] = $reply->contentDetails->definition;

    $video["caption"] = $reply->contentDetails->caption;

    $video["allowedIn"] = (isset($reply->contentDetails-
>regionRestriction->allowed)) ? implode(",",$reply->contentDetails-
>regionRestriction->allowed):"";

    $video["blockedIn"] = (isset($reply->contentDetails-
>regionRestriction->blocked)) ? implode(",",$reply->contentDetails-
>regionRestriction->blocked):"";



    $video["licensedContent"] = $reply->contentDetails-
>licensedContent;

    $video["viewCount"] = $reply->statistics->viewCount;

    $video["likeCount"] = $reply->statistics->likeCount;

    $video["dislikeCount"] = $reply->statistics->dislikeCount;

    $video["favoriteCount"] = $reply->statistics->favoriteCount;

    $video["commentCount"] = $reply->statistics->commentCount;



    $video["uploadStatus"] = $reply->status->uploadStatus;

    $video["privacyStatus"] = $reply->status->privacyStatus;

    $video["license"] = $reply->status->license;

    $video["embeddable"] = $reply->status->embeddable;

    $video["publicStatsViewable"] = $reply->status-
>publicStatsViewable;
```

```
    $content = "";
    foreach($video as $key => $data) {
        $content .= $key."\t".$data."\n";
    }
    writefile("./".$folder.$filename."_basicinfo.tab",$content);

    return $video;
}



function getComments($videohash) {

    global $html,$filename,$folder;



    // get toplevel comments first

    $nextpagetoken = null;
    $run = true;
    $comments = array();

    echo "<br /><br />Getting comments: "; flush(); ob_flush();

    while($run == true) {

        $restquery =
"https://www.googleapis.com/youtube/v3/commentThreads?part=snippet&
maxResults=100&videoId=".$videohash;
```

```
        if($nextpagetoken != null) {
            $restquery .= "&pageToken=".$nextpagetoken;
        }

        $reply = doAPIRequest($restquery);

        foreach($reply->items as $item) {
            $comments[] = $item;
        }

        echo " " . count($comments); flush(); ob_flush();

        if(isset($reply->nextPageToken) && $reply->nextPageToken !=
"") {
            $nextpagetoken = $reply->nextPageToken;
        } else {
            $run = false;
        }
    }


    // work through top level comments and get replies

    $nodecomments = array();
    $counter = 0;

    echo "<br /><br/>Digging into thread structure: "; flush();
ob_flush();

    foreach($comments as $comment) {
```

```
        echo " " . $counter; flush(); ob_flush();

        $counter++;


        $tmp = array();

        $tmp["id"] = $comment->id;

        $tmp["replyCount"] = $comment->snippet->totalReplyCount;

        $tmp["likeCount"] = $comment->snippet->topLevelComment-
>snippet->likeCount;

        $tmp["publishedAt"] = date("Y-m-d H:i:s",
strtotime($comment->snippet->topLevelComment->snippet-
>publishedAt));

        $tmp["authorName"] = preg_replace("/\s+/", " ",$comment-
>snippet->topLevelComment->snippet->authorDisplayName);

        $tmp["text"] = preg_replace("/\s+/", " ",$comment->snippet-
>topLevelComment->snippet->textDisplay);

        $tmp["authorChannelId"] = $comment->snippet-
>topLevelComment->snippet->authorChannelId->value;

        $tmp["authorChannelUrl"] = $comment->snippet-
>topLevelComment->snippet->authorChannelUrl;

        $tmp["isReply"] = 0;

        $tmp["isReplyTo"] = "";

        $tmp["isReplyToName"] = "";

        $tmp["videoId"] = $videohash;



        //print_r($tmp);


        $nodecomments[] = $tmp;


        if($tmp["replyCount"] > 0) {
```

```
            $replies = array();
            $nextpagetoken = null;
            $run = true;

            while($run == true) {

                $restquery =
"https://www.googleapis.com/youtube/v3/comments?part=snippet&textFo
rmat=plainText&maxResults=100&parentId=".$tmp["id"];

                if($nextpagetoken != null) {
                    $restquery .= "&pageToken=".$nextpagetoken;
                }

                $reply = doAPIRequest($restquery);

                foreach($reply->items as $item) {
                    $replies[] = $item;
                }

                if(isset($reply->nextPageToken) && $reply-
>nextPageToken != "") {
                    $nextpagetoken = $reply->nextPageToken;
                } else {
                    $run = false;
                }
            }

            foreach($replies as $reply) {
```

```
                $tmp2 = array();
                $tmp2["id"] = $reply->id;
                $tmp2["replyCount"] = "";
                $tmp2["likeCount"] = $reply->snippet->likeCount;
                $tmp2["publishedAt"] = date("Y-m-d H:i:s",
strtotime($reply->snippet->publishedAt));
                $tmp2["authorName"] = preg_replace("/\s+/", "
",$reply->snippet->authorDisplayName);
                $tmp2["text"] = preg_replace("/\s+/", " ",$reply-
>snippet->textDisplay);
                $tmp2["authorChannelId"] = $reply->snippet-
>authorChannelId->value;
                $tmp2["authorChannelUrl"] = $reply->snippet-
>authorChannelUrl;
                $tmp2["isReply"] = 1;
                $tmp2["isReplyToId"] = $tmp["id"];
                $tmp2["isReplyToName"] = $tmp["authorName"];

                $nodecomments[] = $tmp2;
            }
        }
    }

    echo '<br /><br/>The script retrieved '.count($nodecomments).'
comments from '.count($comments).' top level comments.';

    $content = implode("\t",array_keys($nodecomments[0])) . "\n";
    foreach($nodecomments as $comment) {
        $content .= implode("\t",$comment) . "\n";
```

```
    }
    writefile("./".$folder.$filename."_comments.tab",$content);



    return $nodecomments;
}



function getCommenters($nodecomments) {

    global $filename,$folder;

    $authors = array();

    foreach($nodecomments as $comment) {
        if(!isset($authors[$comment["authorName"]])) {
            $authors[$comment["authorName"]] = 0;
        }
        $authors[$comment["authorName"]]++;
    }

    arsort($authors);

    $content = "";
    foreach($authors as $key => $data) {
        $content .= $key."\t".$data."\n";
    }
    writefile("./".$folder.$filename."_authors.tab",$content);
```

```
    return $authors;
}



function makeNetwork($nodecomments) {

    global $filename,$folder;

    $nodes = array();
    $edges = array();

    foreach($nodecomments as $nodecomment) {

        if(!isset($nodes[$nodecomment["authorName"]])) {
            $nodes[$nodecomment["authorName"]] = 0;
        }
        $nodes[$nodecomment["authorName"]]++;

        $tmp =
preg_match_all("/oid=\"\d+\">(.*)<\/a>/U",$nodecomment["text"],$out
);

        if(count($out[1]) > 0) {

            foreach($out[1] as $ref) {
                if(!isset($nodes[$ref])) {
                    $nodes[$ref] = 0;
                }
```

```
                $edgeid = $nodecomment["authorName"] . "_|_|X|_|_"
. $ref;

                if(!isset($edges[$edgeid])) {

                    $edges[$edgeid] = 0;

                }

                $edges[$edgeid]++;

            }


        } else if ($nodecomment["isReply"] == 1) {


            if(!isset($nodes[$nodecomment["isReplyToName"]])) {

                $nodes[$nodecomment["isReplyToName"]] = 0;

            }


            $edgeid = $nodecomment["authorName"] . "_|_|X|_|_" .
$nodecomment["isReplyToName"];

            if(!isset($edges[$edgeid])) {

                $edges[$edgeid] = 0;

            }

            $edges[$edgeid]++;

        }

    }




    $nodegdf = "nodedef>name VARCHAR,label VARCHAR,commentCount
INT\n";

    foreach($nodes as $nodeid => $nodedata) {

        $nodeid = preg_replace("/,/", " ", $nodeid);

        $nodegdf .= $nodeid . "," . $nodeid  . "," . $nodedata .
"\n";
```

```
    }


    $edgegdf = "edgedef>node1 VARCHAR,node2 VARCHAR,weight
INT,directed BOOLEAN\n";
    foreach($edges as $edgeid => $edgedata) {
        $tmp = explode("_|_|X|_|_",$edgeid);


        $edgegdf .= preg_replace("/,/", " ", $tmp[0]) . "," .
preg_replace("/,/", " ", $tmp[1]) . "," . $edgedata . ",true\n";
    }


    $gdf = $nodegdf . $edgegdf;


    writefile("./".$folder.$filename."_commentnetwork.gdf",$gdf);
}


?>


</body>
</html>
```