Universidad del Desarrollo
Facultad de Ingeniería

LEYES DE ESCALA Y NORMAS DINÁMICAS DE MORALIZACIÓN EN TEXTO

POR: DIEGO JAVIER LEÓN GONZÁLEZ

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

CRISTIAN CANDIA VALLEJOS

DICIEMBRE 2025

SANTIAGO

A mi padre, espero que si eres consciente en algún sitio te alegres de este nuevo logro académico

AGRADECIMIENTO

Segunda vez en mi vida redactando un trabajo final de grado y espero que no sea la última, ya que si de algo me he dado cuenta en los últimos años es que nunca se deja de aprender y uno debe estar actualizándose constantemente. Agradezco a todos los profesores del programa, ya que darse el tiempo de hacer clases en horarios complicados como viernes y sábado sé que no es fácil. Ser alumno de este programa tampoco es fácil, muchas veces quita un fin de semana completo luego de agotadoras semanas laborales, entonces uno nunca descansa.

Agradecimiento especial al Dr. Cristian Candia, que se dio el tiempo durante los meses que dura el Capstone en reunirse a discutir resultados y guiar el trabajo en semanas donde se notaba que no sobraba el tiempo.

TABLA DE CONTENIDO

Resumen **1**

1. Introducción **3**

2. Trabajo Relacionado **6**

3. Hipótesis y Objetivos **8**

4. Datos y Metodología **9**

4.1. Datos 9

4.2. Metodología 11

4.2.1 Pre-procesamiento de texto 11

4.2.2 Medidas de moralización 12

4.2.3 Resumen metodológico 19

5. Resultados **21**

5.1 Distribuciones de medidas de moralización 21

5.2 Leyes de escala en moralización y longitud de texto 22

5.3 Suficiencia de coeficientes en predicción de replies 29

6. Conclusiones **38**

Bibliografía **41**

Resumen

El presente trabajo aborda el estudio de leyes de escala para comprender cómo depende la longitud del texto de una publicación en redes sociales en función de una medida de carga moral o *Moral Loading*. Lo anteriormente descrito corresponde a tendencias y efectos no lineales que se pueden reducir a un ajuste lineal en escala logarítmica. Esto es de utilidad para comprender si los textos más largos son más o menos moralizados que los cortos y permite extraer desviaciones o residuos a partir de dicho ajuste, lo cual se puede utilizar como un predictor adicional para inferir qué textos tienen más o menos interacción o *replies*. Los residuos no son una mera herramienta estadística, sino que reflejan un proceso psicológico complejo de aumento del desarrollo de contenido explicativo cuando se desarrolla un discurso moral. Los residuos tienden a ser nulos cuando el desarrollo de texto está dentro de una tendencia general (norma dinámica), positivos cuando el desarrollo de texto es excesivo debido a la introducción de lenguaje moral y negativos cuando está por debajo de la tendencia general. La metodología de este trabajo busca utilizar estos residuos en modelos regresivos con los *replies* como variable objetivo y estudiar la suficiencia y significancia estadística de los coeficientes asociados a cada predictor. Adicionalmente, los datasets utilizados corresponden a publicaciones en Twitter y Reddit los cuales se utilizan para entrenar los modelos descritos.

Los resultados muestran que las publicaciones en redes sociales con los textos más largos tienen una moralización más alta, mientras que los textos más cortos tienden a tener menor moralización en general y con una varianza mayor. Este comportamiento se observa para dos medidas diferentes de moralización, una con un conteo de palabras morales y otra con medición de similitud de texto mediante *Word Embeddings*.

Las principales conclusiones luego del estudio de leyes de escala, son que los residuos obtenidos a partir de estas leyes definen una norma dinámica que es comparable o similar en comportamiento a una densidad moral o *moral ratio*. Mediante modelos regresivos de coeficientes con los *replies* como variable objetivo, se obtuvo que en la mayoría de los casos los coeficientes asociados a la moralización o *Moral Loading* son estadísticamente significativos y positivos (va en la misma dirección que los *replies*). En contraste, los coeficientes asociados a los residuos son estadísticamente significativos y negativos (va en la dirección opuesta que los *replies*).

1. Introducción

En las últimas décadas las redes sociales han predominado en la forma en que las personas expresan sus ideas e interactúan entre sí. Temas raciales, étnicos y políticos son discutidos y difundidos en publicaciones dentro de redes sociales, las cuales reciben interacciones de usuarios a través de reacciones y comentarios. El comportamiento de estas interacciones en redes sociales ha sido objeto de estudio, teniendo como objetivo entender qué factores son los que tienen mayor influencia en las publicaciones con mayor o menor interacción.

Desde una perspectiva e interpretación psicológica, las publicaciones con contenido altamente moral tienden a propagarse más en redes sociales porque generan fuertes emociones como por ejemplo: indignación, solidaridad, repulsión, entre otros. Más allá del contenido explícito del texto, se trata de un proceso psicológico que hace que las personas elijan compartir frecuentemente este tipo de publicaciones. La moralización actúa como un amplificador del valor social de una publicaciones, lo cual describe un proceso de *Moral Contagion* estudiado por Brady et al. [1, 2, 4]. Adicionalmente, debido a los profundos procesos cognitivos que despiertan estas publicaciones el emisor del mensaje tiende a elaborar más o explayarse más, porque despierta y captura atención [3]. Complementariamente, en [22, 23] se describe un efecto *Moral Reframing*, donde un interlocutor que utiliza técnicas con valor moral en su discurso requiere introducir un mayor contenido explicativo, contextualizar y construir

argumentos. Los conceptos anteriores motivan al estudio de cómo escala la longitud de un texto con el contenido moral de un texto, reflejando estos procesos psicológicos.

Bettencourt et al. [12] evidenció que efectos sociales como productividad o creatividad escalan de forma superlineal con el tamaño urbano, lo cual motiva a continuar modelando efectos sociales y cognitivos como leyes de escala. Adicionalmente, Piantadosi et al. [13] muestra que el lenguaje sigue una estructura matemática, sin diseñarse, es decir, de forma involuntaria. Esto último se denomina ley de Zipf y específicamente plantea que la frecuencia de una palabra está inversamente relacionada con su rango. Esta ley toma relevancia en el contexto de moralización, debido a que a mayor longitud del texto, ciertas categorías crecen en frecuencia, por lo cual intuitivamente lleva a inferir que la moralización puede comportarse como una categoría más ya sea sub o superlineal.

Este trabajo se enfoca particularmente en la presencia de palabras morales en publicaciones en Twitter y Reddit. Si una publicación en algunas de estas redes sociales tiene más o menos presencia de moralización, entonces desde la intuición se debe reflejar en que dicha publicación tenga más o menos interacción. Esta relación que existe entre moralización y cantidad de interacciones es el foco de atención de este trabajo, buscando llegar a esta relación a través del estudio de leyes de escala y normas dinámicas entre longitud del texto y moralización.

Se comienza en este trabajo con la relación que existe entre la longitud de cada texto y la moralización, sin incluir la interacción de usuarios. Esta relación se lleva a cabo utilizando leyes de escala o ley de potencia en escala logarítmica motivado por el trabajo de Bettencourt et al [12]. Lo anterior permite establecer primero qué relación tiene un texto largo o corto con más o menos moralización y segundo estableciendo una norma dinámica a través de desviaciones o residuos obtenidos de este tipo de leyes. Estos residuos se pueden interpretar como una saturación de contenido. Los residuos positivos se asocian a sobre-producción ligado al efecto de *moral reframing* descrito con anterioridad, donde el contenido moral implica un desarrollo argumentativo excesivo. Esta norma dinámica extraída de los residuos establece qué tan alejada está una publicación de una tendencia, lo cual puede ser utilizado posteriormente como un predictor, particularmente como un predictor de interacción de usuarios.

La importancia de este trabajo radica en replicar y extender parte del trabajo previo, donde se han estudiado relaciones con leyes de escala entre la interacción de usuarios y variables asociadas a la moralización. En este trabajo se busca añadir nuevas variables inferidas a través de las normas dinámicas extraídas de la relación entre la longitud del texto y la moralización.

2. Trabajo Relacionado

Dada la complejidad de la interacción de usuarios en redes sociales, ha existido una vasta cantidad de trabajo previo con el objetivo de comprender de qué depende que ciertas publicaciones en redes sociales reciban más o menos interacción. Dada la dependencia no lineal que tiene la interacción de usuarios (como pueden ser re-tweets en el caso de *Twitter*), toma importancia el estudio de leyes de escala a través de leyes de potencia que reducen el problema a uno lineal en escala logarítmica. Bettencourt et al. [12] plantea en su trabajo el estudio de leyes de potencia en el contexto de poblaciones en ciudades y muestra cómo en escala logarítmica se marca una tendencia lineal entre la población y una medida económica GMP por sus siglas en inglés. Adicionalmente, el trabajo de Bettencourt et al. muestra cómo el medir desviaciones de cada observación respecto a la tendencia lineal observada tiene una cierta capacidad predictiva, lo cual utiliza para agrupaciones de ciudades con ciertas características similares. Lo anterior, motiva en el presente trabajo no solo a la utilización de leyes de escala, sino a obtener estas mismas desviaciones o residuos y utilizarlos como predictor de interacción de usuarios.

Particularmente en el contexto de moralización, se ha estudiado la moralización como un conteo de palabras morales presentes en cada texto, lo cual se denomina un enfoque *Linguistic Inquiry Word Counting* (LIWC) [6]. Destaca el trabajo de Brady et al.

[1], donde se pudo establecer la dependencia de los re-tweets como función de la moralización medida con un enfoque LIWC.

C. Candia et al. [9, 10] abordó el problema desde una perspectiva diferente a LIWC. Se propuso una forma alternativa de medir moralización basada en trabajo previo en *Distributed Dictionary Representations* (DDR) [8, 21], donde se muestra como se pueden representar y comparar textos y frases utilizando representaciones vectoriales como *Word Embeddings*. El trabajo de C. Candia et al. utilizó este enfoque DDR para entregar una medida de moralización basada en similitud de un texto con un diccionario moral también representado como un vector. De esta manera, qué tan moral es un texto o frase ya no está limitado a sí todas sus palabras están o no están en un diccionario moral, sino qué similitud semántica tiene con una diccionario o base moral, sin ser necesariamente igual. En este último trabajo se mostró un efecto de *Moral Penalty*, donde para un nivel fijo de moralización, la cantidad de *replies* de tweets y posts en Reddit decae conforme la densidad moral aumenta.

3. Hipótesis y Objetivos

En este trabajo se otorga énfasis a estudiar la relación entre la longitud de cada mensaje con alguna medida de moralización, donde particularmente se estudian dos formas de medir moralización: una a través de conteo de palabras morales en un texto y otra a través de similitud del texto con un diccionario moral a través de *word embeddings*. La razón de estudiar previamente esta relación es para establecer alguna norma en el espacio de longitud del texto y la moralización, la cual se plantea como hipótesis que tendrá una capacidad como predictor sobre la interacción de usuarios y ser en parte un reflejo de efectos psicológicos como la producción de texto excesivo debido al desarrollo del contenido moral.

El objetivo general en base a lo anterior se desprende que es evaluar, mediante análisis de leyes de escala, cómo la moralización y la longitud de un texto influyen en mensajes con mayor o menor interacción de usuarios.

Los objetivos específicos son:

1. Medir moralización en texto según diferentes metodologías disponibles en la literatura.
2. Analizar la dependencia de la moralización con la longitud de cada mensaje.
3. Entrenar modelos regresivos para predecir mensajes con más interacción utilizando desviaciones o residuos en torno a una norma moral.

4. Datos y Metodología

4.1. Datos

Los datos objeto de estudio de este trabajo corresponden a publicaciones en Twitter y Reddit. Ambos datasets rondan 1 millón de publicaciones divididas por tópico (canales en el caso de Reddit). Cada una de estas publicaciones, contiene texto crudo típico de redes sociales, con hashtags, enlaces (url), menciones, entre otros. La Tabla 1. muestra el detalle de cantidad de datos y tópicos/canales. Se puede observar una cantidad de 10 tópicos en Twitter, mientras que en Reddit son solo 4 canales.

Dataset	Cantidad de datos	Cantidad de tópicos/canales
Twitter	986,685	10
Reddit	1,067,832	4

Tabla 1. Datasets de Twitter y Reddit.

En la Figura 1 y 2 se puede observar la distribución de los datos de Twitter y Reddit respectivamente. Se puede observar que en el caso de Twitter, pese a ser 10 tópicos solo 3 concentran la mayor parte de los datos: MeToo, Climate Change y Gun Control. En el caso de Reddit que son solo 4 canales, más del 80% de los datos son del canal The_Donald.

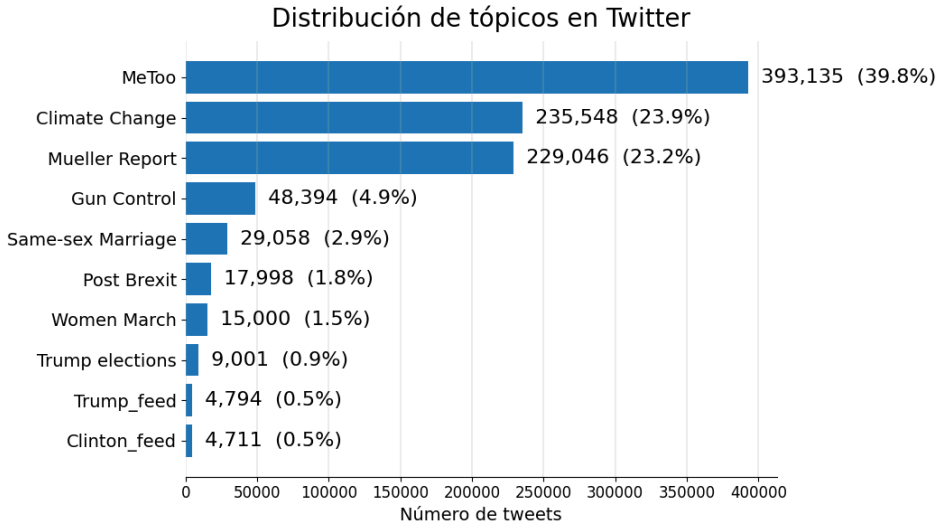

Figura 1. Distribución de tópicos en dataset de Twitter. Fuente: elaboración propia.

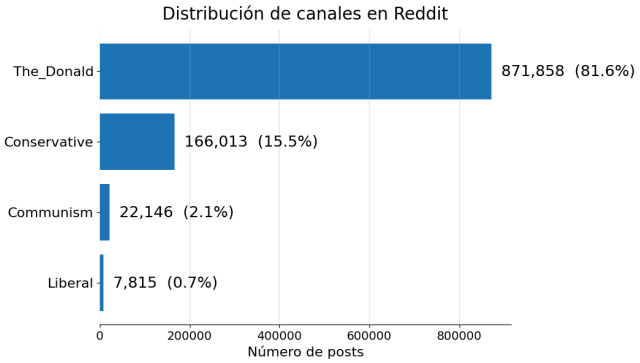

Figura 2. Distribución de tópicos en dataset de Reddit. Fuente: elaboración propia.

4.2. Metodología

La metodología planteada para este trabajo se encuentra dividida en 3 aspectos principales: pre-procesamiento de texto, definición de medidas de moralización a utilizar, extracción de curvas en escala logarítmica (log-log) relacionando la longitud de cada texto con la moralización del texto y finalmente la predicción de interacción de usuarios. A continuación se explican los 2 primeros aspectos y el resto se integra en un resumen al final de la sección. La subsección de medidas de moralización es fundamental, ya que se definen explícitamente los indicadores que son utilizados en los resultados de este trabajo.

4.2.1 Pre-procesamiento de texto

Cada texto presente en los datasets de Twitter y Reddit son datos crudos de texto libre, lo cual significa que tienen una cantidad importante de información que no es útil semánticamente. A continuación se muestra un tweet de ejemplo:

"Over 1,100,000 people killed by guns in the USA since @JohnLennon was shot and killed on Dec 8 1980\r\n#StopGunViolence https://t<U+2026>"

Se pueden realizar las siguientes observaciones de información semántica que no es de utilidad:

- Números: 1,100,000 por sí solo no aporta información alguna.

- Menciones: Una mención a John Lennon, donde el nombre sin el @ y separado tampoco aporta información.
- Fechas: tanto mes, como día y año.
- URL: una dirección web sin información útil.
- Hashtags: acá se observa que el hashtag #StopGunViolence de forma literal no es útil, pero si puede llegar a serlo si se procesa correctamente. En este caso concreto bastaría eliminar el # y separarlo en sus palabras por separado.

Realizando un proceso de limpieza y extrayendo los tokens a cada uno de los mensajes, el texto de ejemplo queda de la siguiente forma:

["Over", "people", "killed", "by", "guns", "in", "the", "usa", "since", "was", "shot", "and", "killed", "on", "dec", "stop", "gun", "violence"]

donde en el mensaje fueron eliminadas todas observaciones hechas anteriormente sobre números, fechas, enlaces y menciones. En el caso del hashtag, este se mantuvo, pero fue procesado, de forma de mantener sus palabras por separado.

4.2.2 Medidas de moralización

Para poder realizar todo el análisis propuesto, es primordial definir cómo será

medida la moralización en un texto dado. En este contexto, se definen 2 enfoques diferentes: conteo de palabras (LIWC) y otro a través de word embeddings, lo cual en la literatura es denominado Distributed Dictionary Representation (DDR).

Dado el texto utilizado en el pre-procesamiento utilizado de ejemplos:

["Over", "people", "killed", "by", "guns", "in", "the", "usa", "since", "was", "shot", "and", "killed", "on", "dec", "stop", "gun", "violence"]

Se define la medida *Moral Loading* según la ecuación (1)

$$Moral\ Loading\ = \ N\acute{u}mero\ de\ palabras\ morales \quad (1)$$

Utilizando la ecuación (1) en el ejemplo, Moral Loading = 3, ya que "killed" (que aparece 2 veces) y "violence" son consideradas palabras morales. Este indicador mide qué tan cargado está un texto de palabras morales, sin importar si dicha carga moral es alta o baja en relación al texto completo.

Se define *Moral Density* según la ecuación (2)

$$Moral\ Density\ = Moral\ Ratio\ = \frac{N\acute{u}mero\ de\ palabras\ morales}{N\acute{u}mero\ total\ de\ palabras} \quad (2)$$

Moral Density en este caso de enfoque LIWC es idéntico a un *Moral Ratio* según (2).

Siguiendo el texto de ejemplo, son 3 palabras morales dividido en la longitud del texto (18 tokens en este caso), por lo cual tiene un Moral Density = 3/18 = 0.16. Esta densidad moral se fundamenta intuitivamente en que una carga moral de 3 palabras no da contexto de cuánto representan esas 3 palabras en el texto completo. Por esta razón, con este enfoque LIWC es directo compararlo con la cantidad total de palabras del texto, definido como un *ratio*. No es igual esa carga moral de 3 palabras si el texto tiene 10 palabras que si tiene 20, por lo cual da cuenta qué tan "denso" es el texto en términos de moralización.

Para decidir si una palabra es o no moral se utiliza un diccionario moral. Se utiliza el *Moral Foundations Dictionary 2.0* [18, 19, 20]. Este diccionario contiene 2041 palabras morales. De esta manera, para definir *Moral Loading* se examina cada mensaje palabra a palabra y anotando cuántas de ellas están presentes en el diccionario moral.

Otro enfoque para definir *Moral Loading* y *Moral Density* es a través de *word embeddings*. Un *word embedding* permite representar cada palabra como un vector U, de este modo, cada texto queda representado mediante un vector global V_{doc} definido de la siguiente forma:

$$V_{doc} = \frac{1}{n}\sum_{i=1}^{n} U_i \quad (3)$$

donde n es el número de tokens totales presentes en cada mensaje. En este trabajo se

utiliza el Word Embedding pre-entrenado **word2vec "*Google News 300"*** motivado por el trabajo de Mikolov et al. [17], donde 300 hace alusión a la dimensión de cada vector. De la misma forma se representa el diccionario moral como un solo vector de la forma siguiente:

$$V_{moral} = \frac{1}{2041} \sum_{i=1}^{2041} M_i \quad (4)$$

donde M_i representa la representación vectorial de cada palabra presente dentro del diccionario moral (de las 2041 palabras del diccionario moral).

Con este enfoque, se define un *Moral Loading*, el cual para distinguirlo de la definición con el otro enfoque (de conteo de palabras o LIWC) se le nombrará como *DDR Moral Loading* como la similitud coseno entre V_{doc} y V_{moral} según la ecuación 5:

$$DDR\ Moral\ Loading = cosine\ similarity\ (V_{doc}, V_{moral}) = \frac{V_{doc} \cdot V_{moral}}{|V_{doc}||V_{moral}|} \quad (5)$$

Esta medida es análoga al *Moral Loading* definido por enfoque LIWC, donde la diferencia radica en que en vez de definir la carga moral en base a si una palabra es o no moral, permite incorporar similitud. Una palabra que según el enfoque LIWC es no moral, el enfoque DDR permite ver que en realidad es "similar" a una palabra moral.

La Figura 3 muestra una ilustración de cómo funciona la similitud coseno en 2D,

a modo de entender cómo matemáticamente cuantifica la similitud a través de un producto punto.

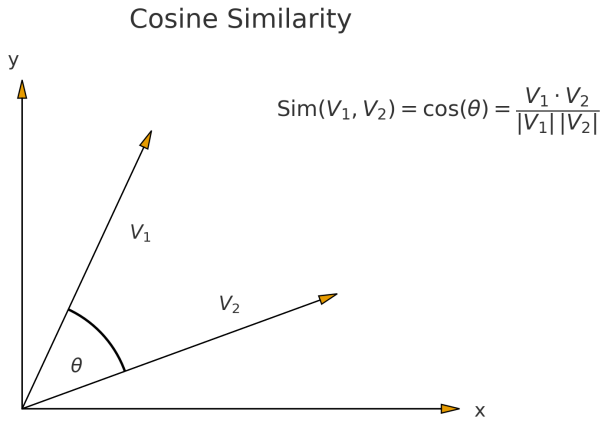

Figura 3. Ilustración de *Cosine similarity*. Fuente: Elaboración propia.

Definida la medida *DDR Moral Loading,* se puede definir análogamente un *DDR Moral Density,* el cual no se puede definir como un *ratio* directamente (como en el caso del enfoque con conteo de palabras), ya que la definición de Moral Loading es una medida completamente diferente y no tiene una relación proporcional a la longitud del texto. Para hacer una conexión con el enfoque LIWC, se necesita definir un ratio de la forma siguiente:

$$DDR\ Moral\ Density\ = \frac{1}{Número\ total\ de\ palabras} \sum_{i=1}^{N} \hat{I}_i \quad (6)$$

donde $\hat{I}_i$ representa un 1 si la palabra i es moral y 0 si no. Sin embargo, DDR Moral

Loading es una medida del texto completo y no de cada palabra, por lo que se define una medida individual para cada palabra de la siguiente forma:

$$w_i = cosine\ similarity(U_i, V_{moral}) \quad (7)$$

donde U_i es la representación vectorial según el *word embedding* para la palabra i y V_{moral} es la representación vectorial del diccionario moral completo. Usando (7) se puede definir si la palabra i es o no moral según (8):

$$\hat{I}_i = 1\ si\ w_i > \tau\ y\ 0\ si\ no \quad (8)$$

donde τ representa un umbral para decidir si cada palabra i es o no moral. De esta forma, se encuentra definido $\hat{I}_i$ y por ende la ecuación (6) queda también completamente definida. Se observa que esta forma de medir la densidad moral es una medida en el rango [0, 1] dependiente de cada w_i y si ese valor es mayor o menor que un umbral τ. Sin embargo, lo anterior es una transición "brusca" o drástica entre 0 y 1, sin utilizar la información de similitud que tiene w_i, ya que lo colapsa inmediatamente a 0 o 1. Con el fin de incorporar lo anterior mencionado, en vez de utilizar dicha transición, se incorpora una función suave entre 0 y 1, pero que siga dependiendo de un umbral τ. Una función que intuitivamente cumple lo anterior es una función sigmoide $\boldsymbol{\sigma}$, tal que se define los valores s_i de la siguiente forma:

$$s_i = \boldsymbol{\sigma}(\tau \, x \, w_i) = \frac{1}{1+exp(\tau \, x \, w_i)} \quad (9)$$

El valor en (9) para la palabra i cuantifica la contribución individual de esta palabra controlada por un umbral τ. De esta forma, promediando los valores s_i se obtiene finalmente la definición de *DDR Moral Density:*

$$DDR\ Moral\ Density = \frac{1}{N}\sum_{i=1}^{N} s_i = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\sigma}(\tau \, x \, w_i) \quad (10)$$

En (10) se muestra un indicador que mide qué tan moral es el texto completo considerando las contribuciones individuales, donde si 1 de 18 palabras por ejemplo (siguiendo mismo ejemplo de LIWC) están sobre el umbral y las restantes 17 por debajo del umbral, entonces *DDR Moral Density* tenderá a cero. Análogamente, si 17 están sobre el umbral y 1 por debajo, entonces tenderá a 1.

No se trata simplemente de la presencia de términos moralizados, sino de la intensidad relativa con la que el contenido moral ocupa el espacio semántico del mensaje. Esto significa que esta definición marca una alta densidad semántica moral cuando una proporción sustantiva de su construcción de texto está dedicada a expresar por ejemplo condenas, obligaciones, derechos, virtudes o transgresiones. Desde esta perspectiva, la densidad moral es una medida local que captura cuánto peso semántico moral se distribuye en relación con la longitud total del texto.

La Figura 4 muestra cómo se comporta la función $\boldsymbol{\sigma}(\tau\, x\, w_i)$ para diferentes valores de τ. Se puede ver que para $\tau = 0.3$ cada valor contribuye aproximadamente proporcional (aproximadamente lineal), mientras que al aumentar τ la función se vuelve más dicotómica. En el límite de τ tendiendo a valores grandes, la función se vuelve un escalón, es decir una transición entre 0 y 1 sin valores intermedios, que es justamente lo que describe la ecuación (6).

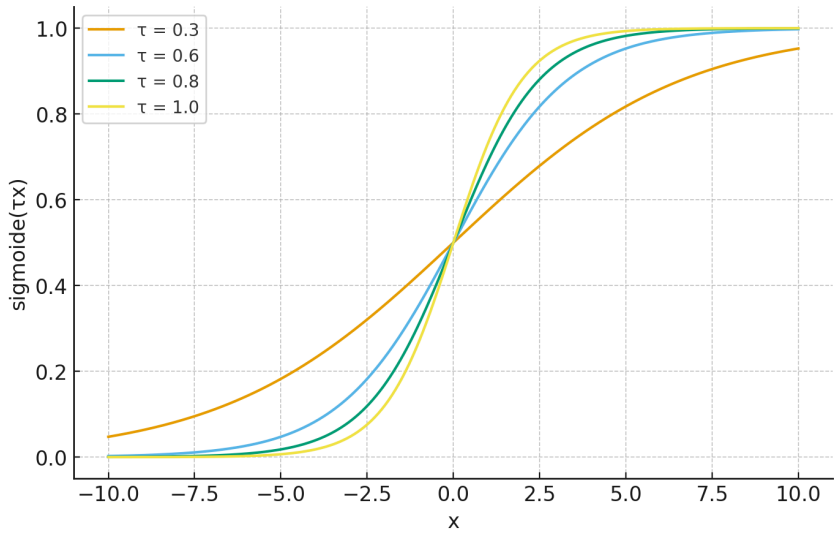

Figura 4. Comportamiento función sigmoide. Fuente: elaboración propia.

4.2.3 Resumen metodológico

A modo de resumen, la metodología se resume en la Figura 5.

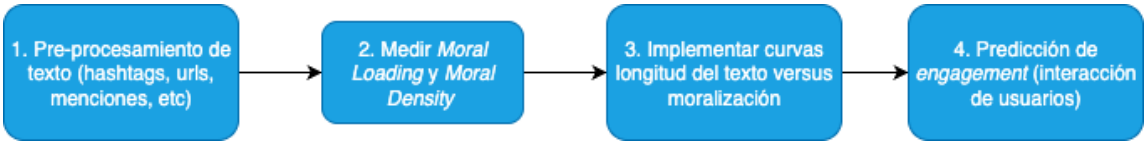

Figura 5. Resumen metodología. Fuente: elaboración propia.

El bloque 1 y 2 fueron explicados en detalle en las subsecciones anteriores, donde en el punto 2 se utilizan ambos de enfoque, es decir, de conteo de palabras y con word embeddings (DDR). El punto 3. implica graficar la medida *Moral Loading* vs Longitud del texto en escala logarítmica, esto permite inferir si los textos más largos son más o menos moralizados que los cortos y además permite extraer coeficientes de leyes de escala a través de regresiones lineales en escala logarítmica. Este punto es crucial, pues aplicando leyes de escala en este espacio *Moral Loading* vs Longitud del texto se puede extraer en esta escala logarítmica qué tan alejado está cada mensaje de la recta definida por leyes de escala y utilizar esto como una "norma moral". Esta norma es utilizada en el punto 4. para predecir el *engagement* o interacción de usuarios (número de replies tanto en Twitter como en Reddit). Esta predicción se vislumbra previamente que tendrá un pobre desempeño en términos regresivos, ya que la interacción o *engagement* depende no solo de la moralización, sino de una gran cantidad de factores no considerados de este trabajo. No obstante, la predicción se realiza con modelos con coeficientes (Regresión Lineal y Negative Binomial), por lo cual se puede realizar un análisis de suficiencia de coeficientes, pese al pobre desempeño predictivo.

5. Resultados

5.1 Distribuciones de medidas de moralización

En la Figura 6 se muestra la medida de *Moral Loading* para los datos de Twitter. Se puede apreciar que la moralización o *Moral Loading* con el enfoque LIWC de conteo de palabras se concentra principalmente en el rango de [0, 5] palabras morales. Se desprende por ende que por lo general cada texto tiene menos de 5 palabras morales. En el gráfico de la derecha se muestra la misma medida con el enfoque DDR. En este último caso se puede ver que se distribuye de una manera más amplia centrada en aproximadamente 0.5 de *cosine similarity.* Lo anterior, se debe a que el enfoque DDR permite extraer información moral de ciertos textos en base a similitud y no pertenencia estricta a un diccionario moral.

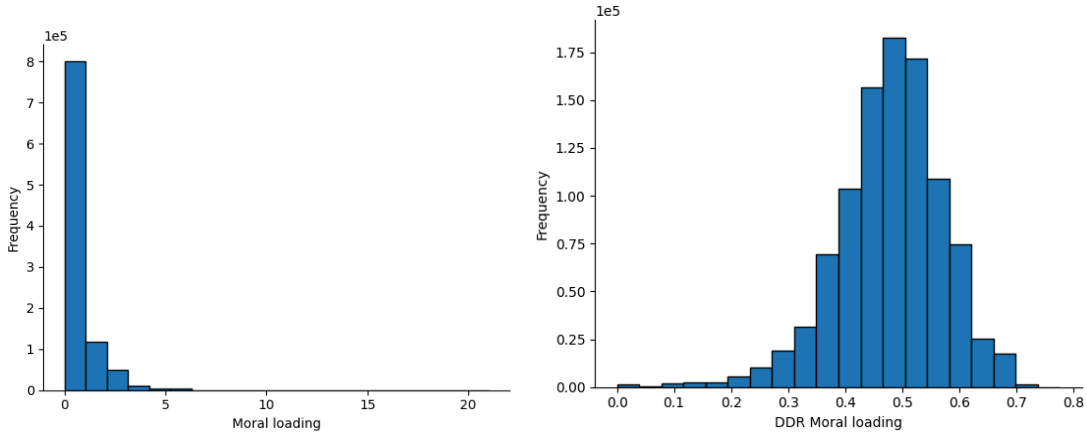

Figura 6. Cambio de distribuciones de *Moral Loading* entre enfoque LIWC (izquierda) y DDR (derecha).

5.2 Leyes de escala en moralización y longitud de texto

En la Figura 7 y Figura 8 se visualizan las dependencias entre la longitud de cada texto y número de palabras morales para los 10 tópicos de Twitter y 4 canales de Reddit respectivamente utilizando el enfoque de conteo de palabras. Se puede observar en todos los casos líneas verticales discretas para cada nivel de palabras morales. Mientras menor es la cantidad de palabras morales presentes, más incierta es la longitud (línea discreta cubre un mayor rango vertical en el gráfico), lo cual significa que los textos menos moralizados tienen por lo general menor longitud, pero con mayor varianza. Mientras más largo es el texto las líneas verticales se hacen más estrechas, evidenciando que los **textos más largos están fuertemente moralizados** con una menor varianza.

Todos los gráficos en las Figura 7 y Figura 8 son en escala logarítmica en ambos ejes y la recta en azul muestra la tendencia lineal en este espacio utilizando leyes de escala. Se reportan los coeficientes encontrados en cada uno de los casos. En todos los casos se observa visualmente que el ajuste con leyes de escala es adecuado, ya que visualmente la tendencia lineal es apreciable y describe el comportamiento descrito anteriormente, es decir, textos largos más moralizados.

En todos los gráficos se muestra una *colorbar* cuantificando la interacción de usuarios a través del número de *replies*. Por lo general no se aprecia tendencia alguna de que existan más *replies* para textos más o menos moralizados. Sin embargo, si se puede apreciar comparando los distintos tópicos y canales que son desiguales. Tópicos como *Same-Sex Marriage* o *Gun Control* concentran en su mayoría tweets de baja cantidad de

replies, mientras que *Trump Feed* por ejemplo tiene una alta cantidad de interacción o *replies*.

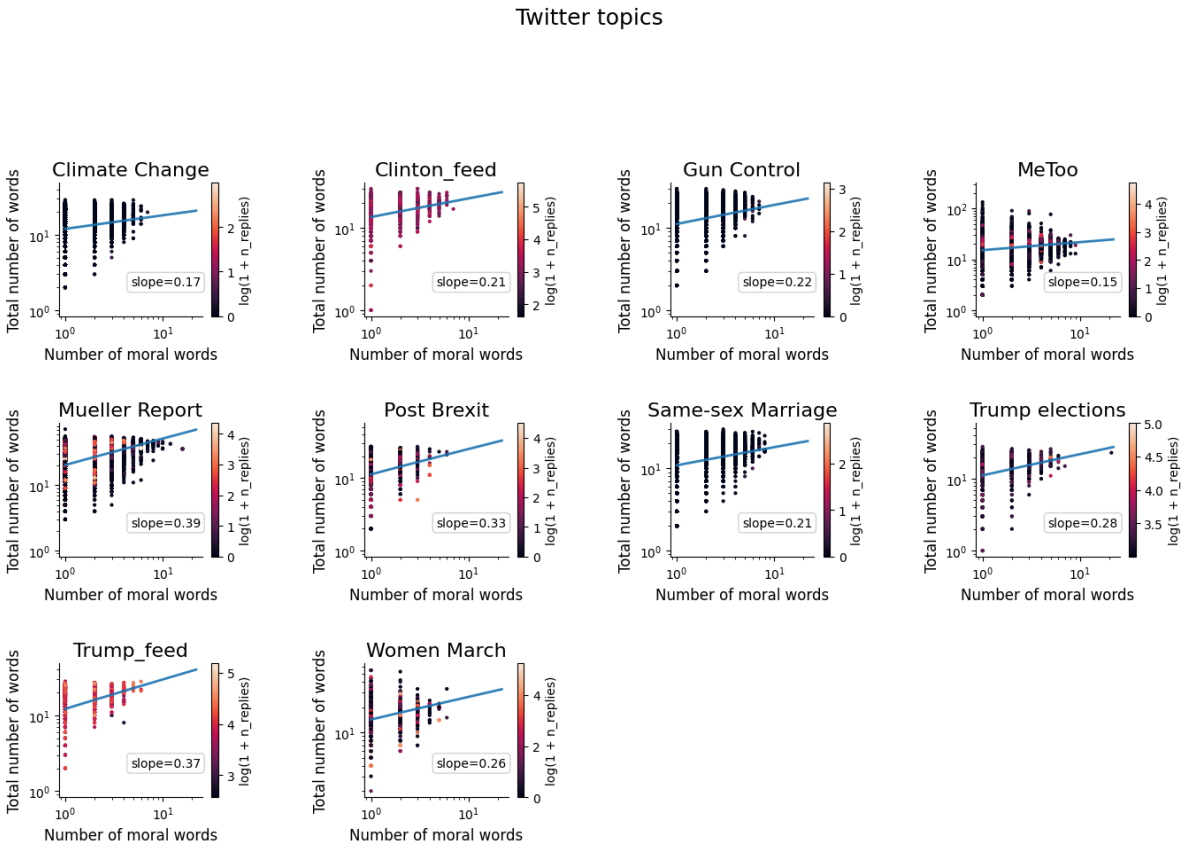

Figura 7. Dependencia entre longitud de texto y moralización usando enfoque de conteo de palabras en Twitter. Se añaden leyes de escala con coeficientes encontrados.

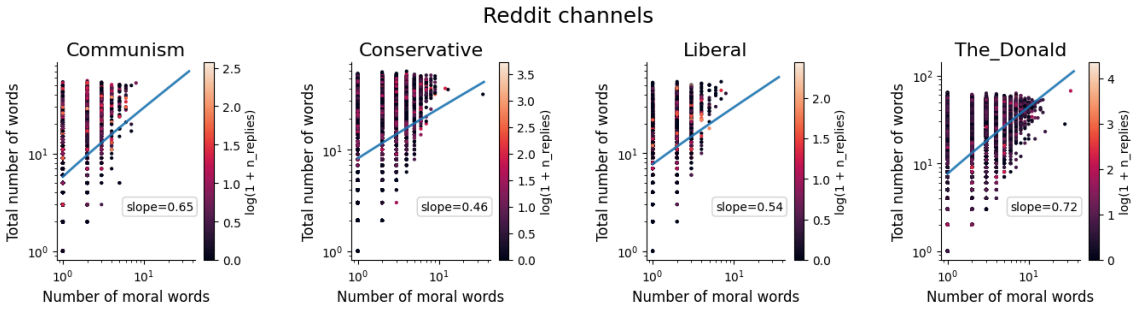

Figura 8. Dependencia entre longitud de texto y moralización usando enfoque de conteo de palabras en Reddit. Se añaden leyes de escala con coeficientes encontrados.

En la Figura 9 y Figura 10 se muestran las tendencias del largo del texto versus moralización, pero esta vez utilizando el enfoque DDR usando *Cosine Similarity*. Las tendencias del largo del texto en función de la moralización no cambia en relación al enfoque de conteo de palabras. Se observa nuevamente la misma tendencia que los textos más largos son más moralizados que los cortos y que los textos menos moralizados tienen una incerteza mayor. Adicionalmente, en este caso no se observan líneas verticales discretas como con el enfoque de conteo de palabras, ya que en este caso la medida de *Moral Loading* es un continuo.

La interacción de usuarios tanto para Twitter como para Reddit nuevamente no se localiza en ningún sector particular, es decir, no se observa que textos más largos ni más cortos tengan mayor interacción. Los textos más o menos moralizados tampoco se observa en general que tenga más o menos interacción. Nuevamente la única observación que se extrae al respecto es que ciertos tópicos de Twitter o canales de Reddit tienen menos variabilidad de interacción de usuarios.

Se desprende que pese a la diferencia de enfoque, el *Moral Loading* no cambia su comportamiento frente a la longitud de cada texto. Lo anterior, permite concluir que a nivel de resultados se encontró prácticamente una equivalencia entre el enfoque de conteo de palabras y el enfoque DDR que emplea *Cosine Similarity*.

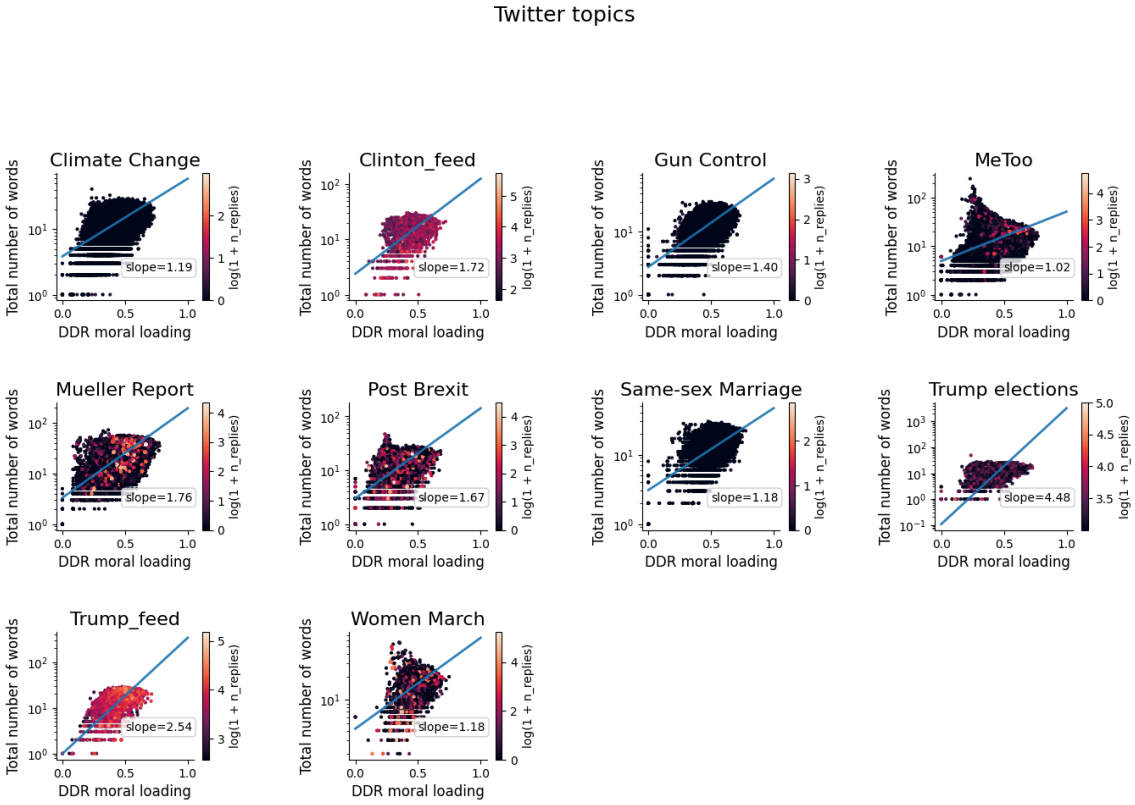

Figura 9. Dependencia entre longitud de texto y moralización usando enfoque DDR en Twitter. Se añaden leyes de escala con coeficientes encontrados.

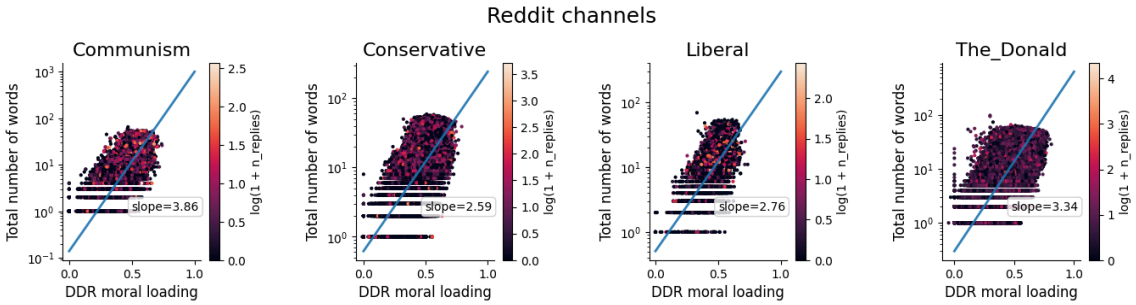

Figura 10. Dependencia entre longitud de texto y moralización usando enfoque DDR en Reddit. Se añaden leyes de escala con coeficientes encontrados.

En la Figura 11 y Figura 12 se muestran los residuos en base a los ajustes encontrados anteriormente mediante leyes de escala para los tópicos de Twitter y canales de Reddit respectivamente. Un residuo corresponde a la diferencia entre el valor real y el valor encontrado mediante el ajuste lineal en espacio logarítmico. La importancia de graficar la distribución de residuos es que cuantifica qué tan cerca o qué tan lejos se encuentra cada texto de una "tendencia general". En todos los tópicos de Twitter y en todos los canales de Reddit se puede observar que son distribuciones aproximadamente normales visualmente centrada en cero. Los residuos de ahora en adelante en este trabajo se utilizan como una variable predictora más, con el objetivo de analizar y cuantificar qué tan influyente es en la interacción de usuarios. De forma intuitiva, se vislumbra que los residuos visto como un predictor tenga una relación similar en comparación al *Moral Density*.

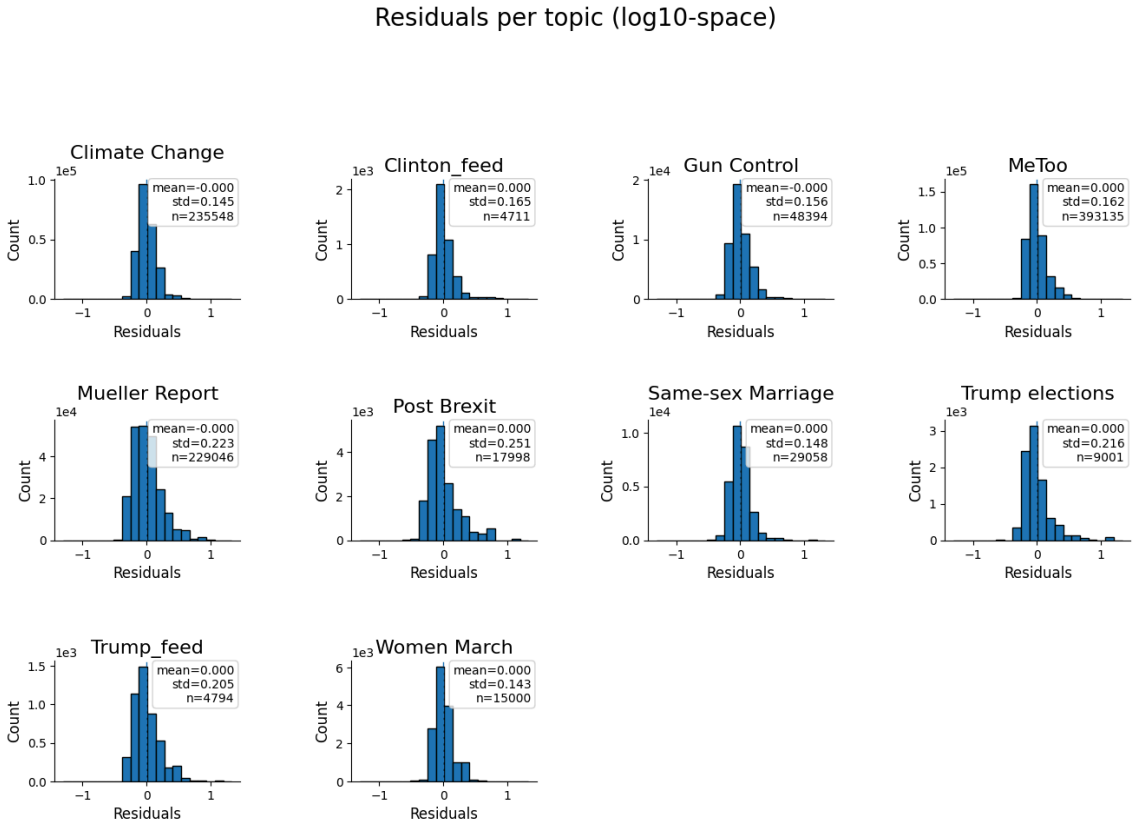

Figura 11. Distribución de residuos dado por leyes de escala en el caso de Twitter.

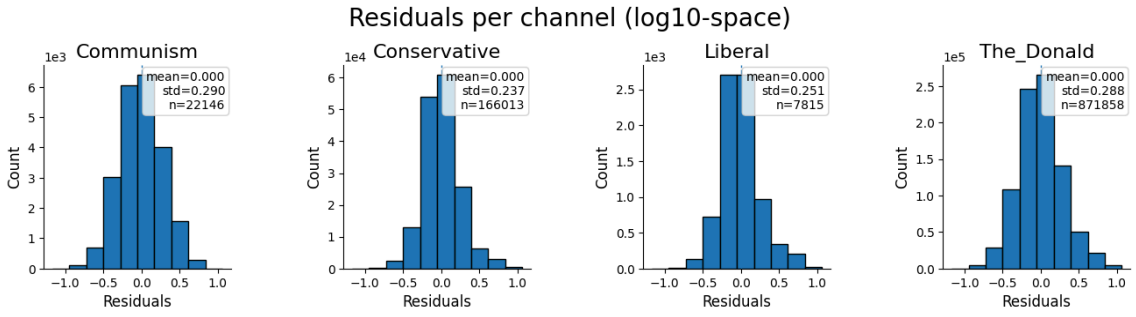

Figura 11. Distribución de residuos dado por leyes de escala en el caso de Reddit.

La Tabla 2 muestra la correlación de Spearman entre los residuos encontrados anteriormente por tópico/canal y la DDR *Moral Density*. Se puede observar que en todos los casos es positiva superando umbrales de 0.7 y 0.8 que son valores apreciablemente altos, indicando que **los residuos y densidad moral van en la misma dirección en todos los casos.** Se vislumbra que los residuos como variable predictora debe tener un comportamiento similar a la densidad moral.

Tópico/canal	Correlación de Spearman
Climate Change	0.38
Clinton_feed	0.47
Gun Control	0.34
MeToo	0.36
Mueller Report	0.14
Post Brexit	0.25
Same-sex Marriage	0.46
Trump elections	0.76
Trump_feed	0.66
Women March	0.18
Communism	0.83
Conservative	0.72
Liberal	0.74
The_Donald	0.74

Tabla 2. Correlación de Spearman entre residuos y densidad moral para cada tópico/canal.

5.3 Suficiencia de coeficientes en predicción de replies

Con los residuos encontrados anteriormente, se entrenan modelos regresivos *Negative Binomial* entregando los residuos y *DDR Moral Loading* como predictores y el número de *replies* como variable a predecir. El problema de predecir interacción de usuarios es un problema complejo que puede depender de una gran cantidad de factores, por lo que en este trabajo no se busca obtener métricas regresivas altas como Error Cuadrático Medio (MSE), sino estudiar y analizar la suficiencia de coeficientes, con el fin de analizar la influencia de cada predictor en la variable a predecir o variable objetivo.

Las Figuras 13 muestra los valores de los coeficientes asociados a *DDR Moral Loading* (derecha) y residuos (izquierda) encontrados por cada tópico y canal de Twitter y Reddit respectivamente usando un modelo *Negative Binomial*. Dado que no se encontró diferencia significativa en el análisis entre el enfoque conteo de palabras y DDR, se adopta este último por simplicidad y evitar duplicar información. Los coeficientes que son más positivos, es decir, están más hacia la derecha, marcan una tendencia a que mientras mayor es la variable mayor es la interacción. En contraparte, los coeficientes que más negativos son marcan una tendencia inversa, es decir, mientras mayor es la variable menor es la cantidad de interacción o *replies*.

En el caso de *DDR Moral Loading* (derecha) se observa prácticamente en la mayoría de tópicos y canales una tendencia directa, es decir, los textos más moralizados tienden a tener mayor cantidad de *replies*. En el caso del tópico de Twitter MeToo por ejemplo, se puede ver que el coeficiente es positivo y significativamente mayor al resto

con un valor de 10.48 ± 0.01. En el caso del tópico Women March y Clinton Feed se puede ver que son casos en los cuales se marca una tendencia inversa.

En el caso de los residuos (izquierda) se puede observar en la mayoría de los casos que la tendencia es inversa (coeficientes negativos), es decir, mientras mayor es el valor del residuo menor es la cantidad de *replies*. Los tópicos Climate Change, Gun Control, MeToo y Gun Control tienen esta tendencia acentuada con un valor estadísticamente significativo (p-valor muy cercano a 0). El coeficiente de MeToo por ejemplo tiene un valor de -2.95 ± 0.05, lo cual es apreciablemente negativo. La excepción en este caso nuevamente es el tópico Women March y Clinton Feed. Pese a ser excepciones, se observa que tiene consistentemente una tendencia opuesta en relación a *DDR Moral Loading*, es decir, en ningún caso se observa que los coeficientes de *DDR Moral Loading* y Residuos tengan el mismo signo.

Se desprende que por lo general los textos más moralizados tienen más interacción para niveles fijos de residuos y los textos con los mayores residuos bajo moralización fija (más lejanos a la tendencia general) tienden a tener menor interacción. Se concluye que **se observa un efecto de saturación, similar al efecto *Moral Penalty* descrito en [9]**. Este último efecto se realizó estudiando *DDR Moral Loading* en conjunto a *DDR Moral Density*, por lo cual se desprende que existe una cierta similitud (como fue previsto con anterioridad) entre *DDR Moral Density* y los residuos obtenidos mediantes leyes de escala. Si un texto tiene una alta moralización o *DDR Moral Loading* fijo y por ende una cantidad de *replies* apreciable, lo que se encontró es que en general

los residuos actúan similar a un *Moral Ratio*, es decir, si se satura el texto de palabras morales, entonces el número de *replies* por lo general tenderá a disminuir. Esto significa una sobre-producción de texto, es decir, son casos **donde el emisor tiende a exagerar su discurso argumentativo por sobre la tendencia o norma moral**, explayándose y derivando a un largo texto que repele los *replies* en vez de atraerlos.

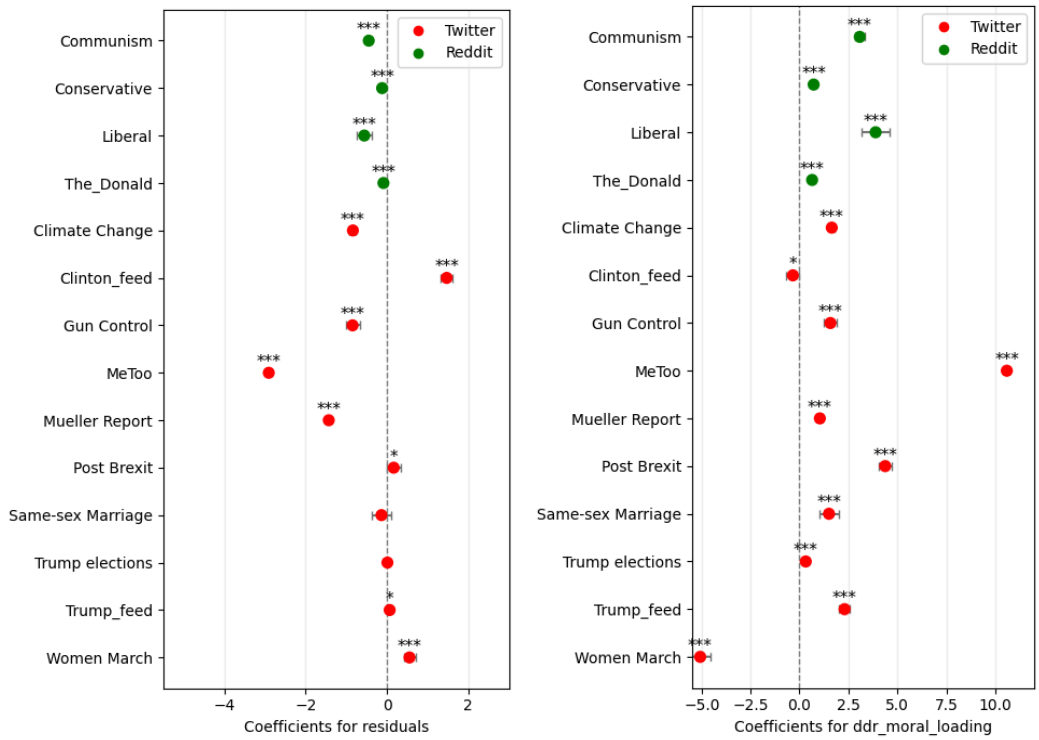

Figura 13. Valores de coeficientes e intervalos de confianza (95%) asociado en un modelo *Negative Binomial* entrenado para predecir *replies* o interacción de usuario. Se marca con *, ** y *** los p-valor asociados para 0.1, 0.01 y 0.05.

La Figura 14 muestra análogamente los mismos resultados que en la Figura 13, pero esta vez se utiliza un modelo *Zero inflated Negative Binomial*, dado que varios de los tópicos tienen una tendencia a tener una cantidad alta de ceros. Se puede ver que el comportamiento en general no cambia, donde los tópicos y canales con coeficientes más significativos en la Figura 13 se mantienen similares en este caso, salvo que en Reddit los canales *Communism* y *Liberal* tienden a hacerse algo más negativos.

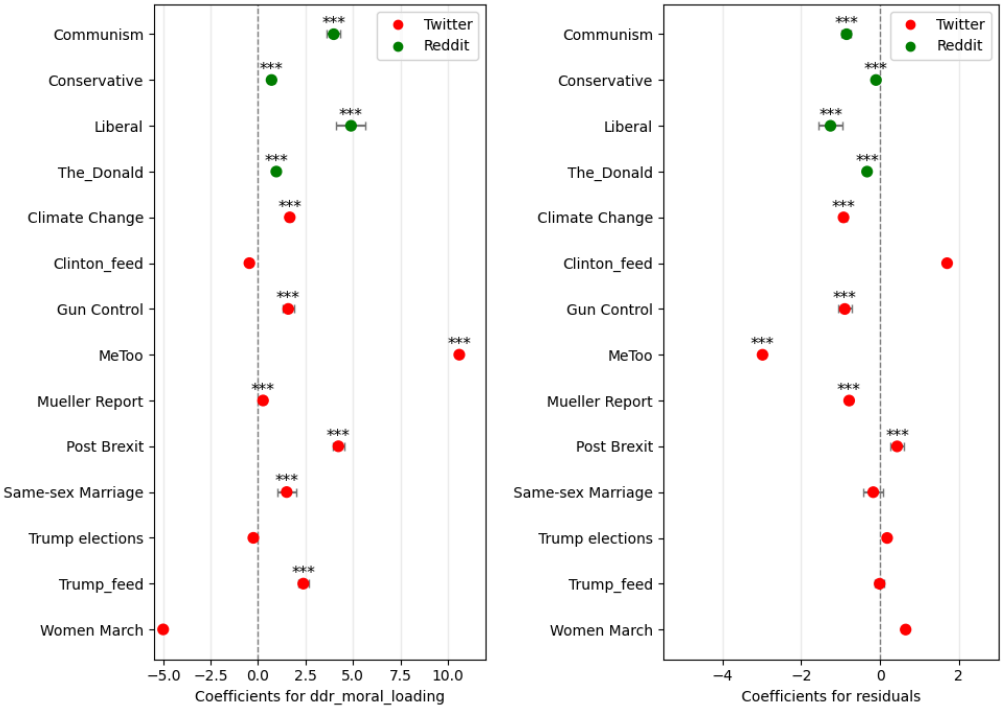

Figura 14. Valores de coeficientes e intervalos de confianza (95%) asociado en un modelo *Zero Inflated Negative Binomial* entrenado para predecir *replies* o interacción de usuario. Se marca con *, ** y *** los p-valor asociados para 0.1, 0.01 y 0.05.

Con el fin de expandir visualmente el análisis extraído anteriormente, de la Figura 15 a la 19 se muestran los efectos marginales para los diferentes tópicos de Twitter y canales de Reddit. Se le denomina efecto marginal a dejar fijo en diferentes niveles una de las variables (residuos o *DDR Moral Loading*) y hacer un barrido sobre la otra variable. En todos los casos se puede observar como la cantidad de *replies* decae conforme los residuos aumentan para niveles fijos de *DDR Moral Loading*. Similarmente, para niveles fijos de residuos, la cantidad de *replies* aumenta conforme el *DDR Moral Loading* aumenta. En todos los casos se sombrea el intervalo de confianza construido en base a los coeficientes expuestos anteriormente, donde para MeToo por ejemplo se puede ver que se tiene una incerteza prácticamente imperceptible, lo cual es coincidente con que es el tópico más predominante en todo el dataset de Twitter. En el caso de Mueller Report en la Figura 16 si bien se observa el mismo efecto, la incerteza es mayor observándose intervalos de confianza más anchos.

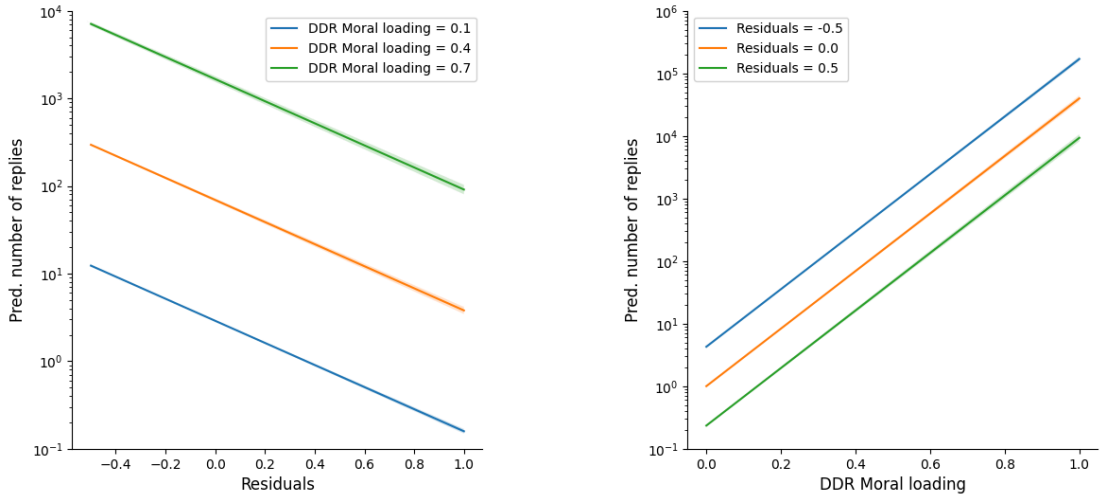

Figura 15. Efectos marginales para el tópico MeToo en Twitter.

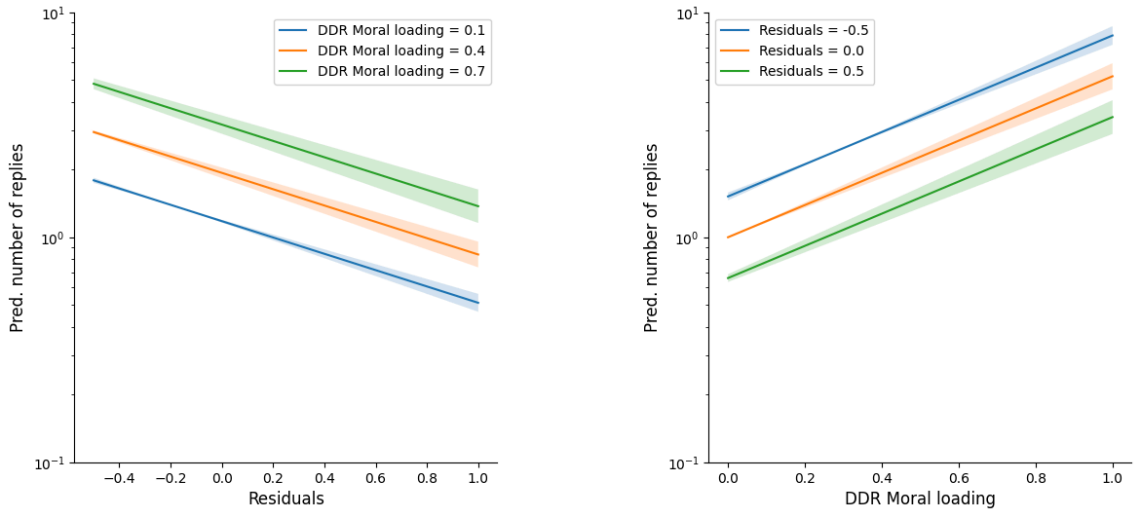

Figura 16. Efectos marginales para el tópico Climate Change en Twitter.

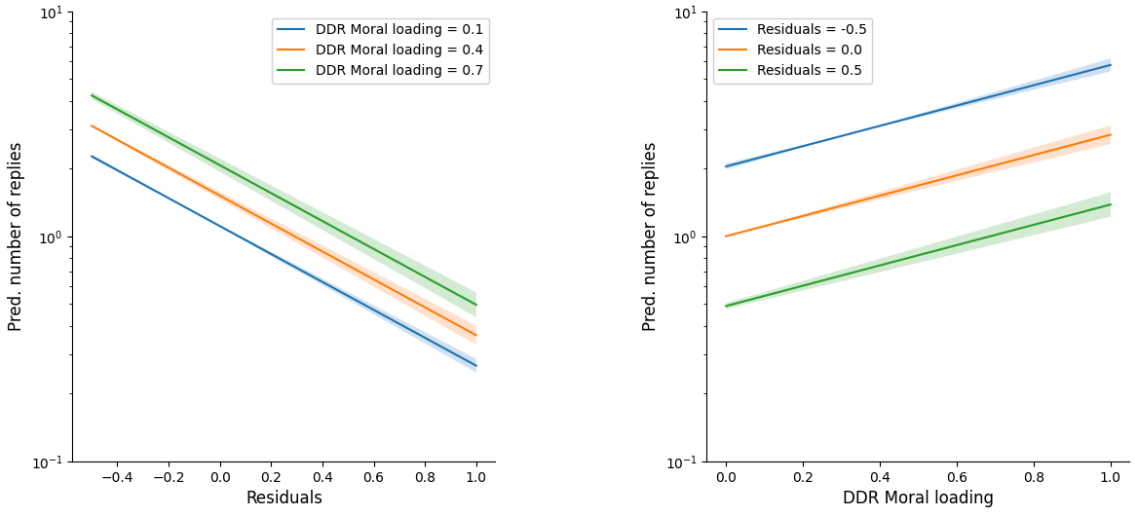

Figura 17. Efectos marginales para el tópico MeToo en Twitter.

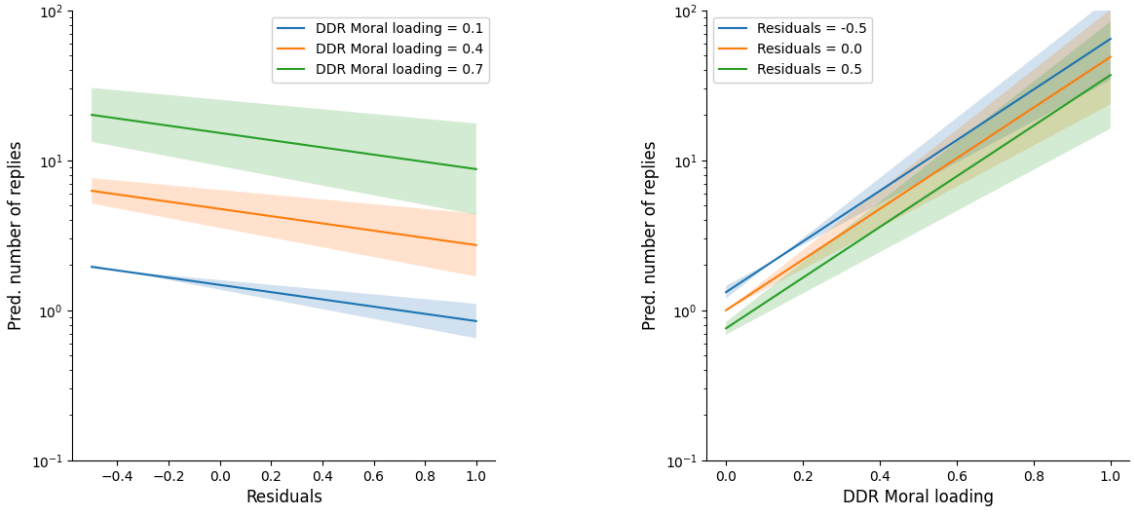

Figura 18. Efectos marginales para el tópico Mueller Report en Twitter.

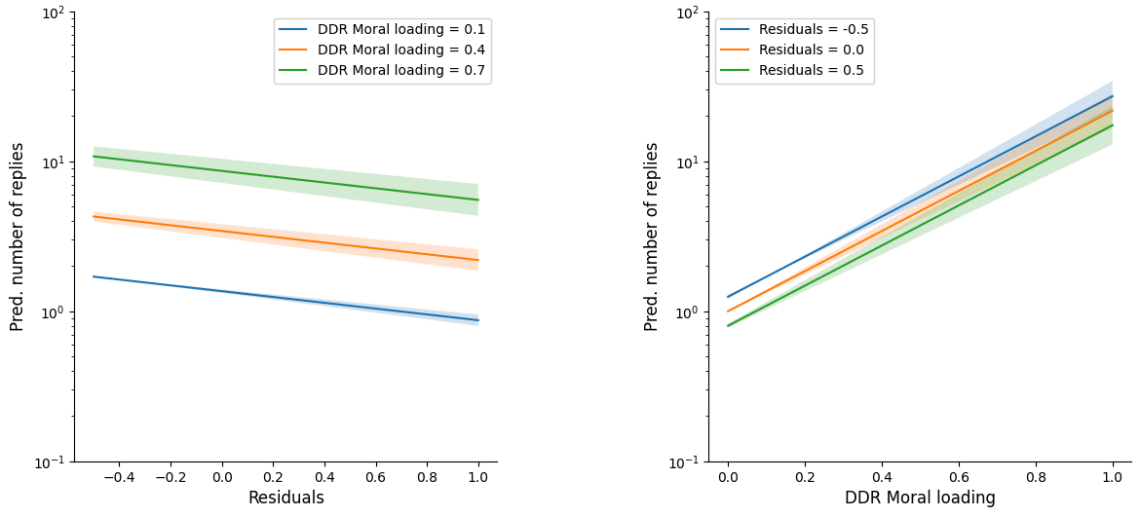

Figura 19. Efectos marginales para el canal Communism en Reddit.

Finalmente, es importante mencionar que si bien los modelos de regresión *Negative Binomial* permitieron extraer la similitud entre residuos y *DDR Moral Density* con coeficientes significativos en la mayoría de casos y con efectos marginales analizados anteriormente, existen **limitaciones derivadas de factores externos** no considerados que también afectan en la cantidad de *replies* y pueden introducir sesgos. En redes sociales en general, la probabilidad de recibir alto impacto o alta interacción depende no solo del contenido moral o normativo del mensaje, sino también de condiciones estructurales y contextuales que no fueron incorporadas en este análisis: la hora y el día de publicación (que determinan ciclos de actividad), características del autor como identidad, reputación y número de seguidores, así como atributos formales del mensaje (presencia de imágenes, hashtags o enlaces) que modulan la visibilidad. La omisión de estos factores introduce la posibilidad de sesgo en las estimaciones, **inflando**

o atenuando los efectos atribuidos al contenido moral. Por lo tanto, los resultados deben entenderse como asociaciones condicionadas a las variables disponibles y no como efectos causales. Futuros estudios deberían incorporar controles temporales, posibles metadatos del autor y modelos multinivel para aislar con mayor precisión la contribución del contenido moral a la cantidad de *replies*.

6. Conclusiones

Se puede concluir un cumplimiento de los objetivos planteados para este trabajo, ya que se logró estudiar en profundidad cómo la moralización y la longitud de un texto influyen en que un texto tenga más o menos interacción de usuarios. Se encontró dicha influencia mediante una medida de moralización o *Moral Loading* y residuos o desviaciones obtenidos mediante leyes de escala, las cuales a su vez se obtuvieron en el espacio de longitud de texto versus moralización, definiendo una norma dinámica en este espacio. Los residuos encontrados se concluye que en general tienen un comportamiento similar a una densidad moral. Lo anterior es debido a que se observaron efectos marginales que describen que para la mayoría de tópicos y canales en Twitter y Reddit para niveles fijos de moralización o *Moral Loading*, la interacción o *replies* de las publicaciones tienden a disminuir, donde en la mayoría de tópicos se encontraron coeficientes estadísticamente significativos y negativos que respaldan este comportamiento. Los residuos se concluye empíricamente que en general no solo se comportan similar a la densidad moral, sino que cómo fue previsto, marcan una tendencia general de cuánto se expande el contenido explicativo para un nivel dado de moralización. Lo anterior permite inferir residuos muy positivos que significa una sobre-producción de texto, es decir, el emisor exagera su construcción argumentativa. Residuos negativos en contraparte se interpretan como falta de esta construcción argumentativa.

Se concluye que a través de las leyes de escala estudiadas en el espacio de longitud del texto versus moralización que en todos los tópicos y canales sin excepción los textos más largos son más moralizados que los cortos. Se concluye que los textos más cortos tienen una incerteza mayor, concentrando una mayor varianza de moralización. En contraste, los textos más largos disminuyen radicalmente su varianza de moralización.

Se desprende que entre los dos enfoques de medidas de moralización no se observaron diferencias apreciables entre moralización por conteo de palabras morales y la moralización obtenida mediante el uso de similitud con representaciones vectoriales de palabras. El análisis realizado muestra que se observa el mismo comportamiento, es decir, en ambos casos se evidencia que la longitud del texto escala con la carga moral, por lo cual se desprende la utilidad del análisis mediante leyes de escala.

Como trabajo futuro, se propone principalmente extender este análisis a datasets de otros contextos y periodos de tiempo. Particularmente, se propone extender el análisis realizado en este trabajo con un control más detallado de factores externos como hora del día y cantidad de seguidores. Lo anterior permite evitar sesgos y aislar con mayor precisión la contribución del contenido moral a la cantidad de *replies*. Adicionalmente, en este trabajo no se utilizó *MoralBert*, el cual es un tercer enfoque de medir moralización mediante un modelo con aprendizaje profundo o *Deep Learning*. Es

esperable que las conclusiones no cambien al extender el trabajo de esta forma, pese a que cambie la forma de medir moralización.

Bibliografía

[1] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. V. Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, vol. 114, p. 201618923, Jun. 2017, doi: 10.1073/pnas.1618923114.

[2] Brady, W. J., Crockett, M. J., & van Bavel, J. J. (2020). "The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online". *Perspectives on Psychological Science, 15(4), 978–1010*.

[3] Brady, W. J., Gantman, A. P., & van Bavel, J. J. (2019). "Attentional capture helps explain why moral and emotional content go viral". *Journal of Experimental Psychology: General.*

[4] Brady, W. J., Rathje, S., Globig, L., & Van Bavel, J. J. (2025). "Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis". https://doi.org/10.31219/osf.io/s4w2x_v2

[5] Brady, W. J., & Van Bavel, J. J. (2025). "Social identity shapes antecedents and functional outcomes of moral emotion expression". *Journal of Experimental Psychology: General.*

[6] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.

[7] Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: *LIWC 2001. Lawrence Erlbaum Associates.*

[8] Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). "Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis". *Behavior Research Methods*, 50(1), 344–361.

[9] C. Candia, M. Atari, N. Kteily, and B. Uzzi, "The Moral Penalty Effect: Overuse of Moral Language Dampens Content Engagement on Social Media,", 2023.

[10] C. Candia, M. Atari, N. Kteily, and B. Uzzi, "*Supplementary Material — The Moral Penalty Effect: Overuse of Moral Language Dampens Content Engagement on Social Media*," Data Science Institute, Universidad del Desarrollo; Santiago, Chile, Sept. 22, 2025.

[11] Bettencourt LMA, Lobo J, Strumsky D, West GB (2010), "Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities". PLoS ONE 5(11): e13541. doi:10.1371/journal.pone.0013541

[12] L. Bettencourt, “The Origins of Scaling in Cities,” *Science*, vol. 340, pp. 1438–1441, Jun. 2013, doi: 10.1126/science.1235823.

[13] Piantadosi, S.T., Zipf’s word frequency law in natural language: A critical review and future directions. *Psychon Bull Rev 21, 1112–1130* (2014). https://doi.org/10.3758/s13423-014-0585-6

[14] V. Preniqi, I. Ghinassi, J. Ive, C. Saitis, and K. Kalimeri, “MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions,” in *Proceedings of the 2024 ACM Conference on Information and Knowledge Management (CIKM)*, Sept. 2024, pp. 433–442, doi: 10.1145/3677525.3678694.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Advances in Neural Information Processing Systems*, vol. 26, Oct. 2013.

[16] Garten, J., Hoover, J., Johnson, K.M. *et al.* Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behav Res* 50, 344–361 (2018). https://doi.org/10.3758/s13428-017-0875-9

[17] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, vol. 2013, Jan. 2013.

[18] Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehghani, M. (2019). “Moral foundations dictionary for linguistic analyses 2.0”.

[19] Kennedy, B., Atari, M., Davani, A. M., Hoover, J., Omrani, A., Graham, J., & Dehghani, M. (2021). “Moral concerns are differentially observable in language”. Cognition, 212.

[20] Hopp, F.R., Fisher, J.T., Cornell, D. *et al.* “The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text”. *Behav Res* 53, 232–246 (2021). https://doi.org/10.3758/s13428-020-01433-0

[21] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). “Distributed representations of words and phrases and their compositionality”. *Advances in Neural Information Processing Systems*, 26, 3111–3119.

[22] M. Feinberg and R. Willer, “From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence?,” *Personality & Social Psychology Bulletin*, vol. 41, Oct. 2015, doi: 10.1177/0146167215607842.

[23] C. Wolsko, H. Ariceaga, and J. Seiden, “Red, White, and Blue Enough to Be Green: Effects of Moral Framing on Climate Change Attitudes and Conservation Behaviors,” *Journal of Experimental Social Psychology*, vol. 65, Feb. 2016, doi: 10.1016/j.jesp.2016.02.005.