available at http://ejournal.unp.ac.id/index.php/psikologi/

PRINTED ISSN 2087-8699
ELECTRONIC ISSN 2622-6626

Published by *Jurnal RAP (Riset Aktual Psikologi)*
Universitas Negeri Padang, Indonesia

Vol. 17 No. 1, 2026
Page 1-18

SELF-CONTROL LIMITATIONS IN MITIGATING HATE SPEECH: ROLE TOXIC ONLINE DISINHIBITION AMONG GENERATION Z ON SOCIAL MEDIA

Siti Nabila Safitri[1], Indriyani Santoso[2]
Department of Psychology, Universitas Negeri Padang, Padang, Indonesia[1, 2]
e-mail: sitinabilasafitri@gmail.com[1], indriyani@fpk.unp.ac.id[2]

Submitted: 2026-04-17 Published: 2026-05-01 DOI: https://doi.org/10.24036/rap.v17.i1.85
Accepted: 2026-04-28

Abstract: Self-control Limitations in Mitigating Hate speech: Role Toxic Online Disinhibition Among Generation Z on Social Media. The purpose of the research is to analyze the influence of toxic online disinhibition on hate speech behavior and to test self-control as a moderator in the influence of toxic online disinhibition on hate speech behavior among Generation Z social media users, using a quantitative approach with SEM-PLS analysis and collecting data from 215 Generation Z respondents. The sampling method used is purposive sampling. The research instrument uses a hate speech scale with a Cronbach alpha of 0.923, a toxic online disinhibition scale with a Cronbach alpha of 0.820, and a self-control scale with a Cronbach alpha of 0.81. The results of this study indicate a positive and significant influence of toxic online disinhibition on hate speech behavior with a significance value of 0.000 (p-value <0.05). The higher the level of disinhibition of individuals on social media, the greater the tendency for individuals to express aggressive communication such as insults, provocation, or hate speech on social media. Furthermore, self-control has also been proven to play a moderating role in the influence of toxic online disinhibition on hate speech, with a significance value of 0.036 (p-value <0.05) with an R^2 contribution of 0.45 (45%). This study found that high self-control actually strengthens the influence of toxic online disinhibition and hate speech. The study results contribute to understanding that self-control does not always function as a deterrent to aggressive behavior on social media.

Keywords: *Hate Speech, Toxic Online Disinhibition, Self-Control*

Abstrak: Keterbatasan *Self-Control* dalam Mengurangi *Hate Speech*: Peran *Toxic Online Disinhibiton* pada Generasi Z di Media Sosial. Tujuan penelitian untuk menganalisis pengaruh *toxic online disinhibition* terhadap perilaku *hate speech* serta

menguji *self-control* sebagai moderasi dalam pengaruh *toxic online disinhibition* terhadap perilaku *hate speech* pada pengguna media sosial dari Generasi Z dengan pendekatan kuantitatif yang menggunakan analisis SEM-PLS dan mengumpulkan data dari 215 responden Generasi Z. Pengambilan sampel menggunakan *purposive sampling*. Instrument penelitian menggunakan skala *hate speech* dengan Cronbach alpha 0.923, *toxic online disinhibition* dengan Cronbach alpha 0.820 skala *self-control* dengan Cronbach alpha 0.81. Hasil penelitian ini menunjukkan terdapat pengaruh positif serta signifikan pengaruh *toxic online disinhibition* terhadap perilaku *hate speech* dengan nilai signifikansi 0.000 (p-value <0.05). Semakin tinggi tingkat disinhibisi individu di media sosial, maka semakin besar kecenderungan individu untuk mengekspresikan komunikasi agresif seperti penghinaan, provokasi, maupun *hate speech* di media sosial. Selanjutnya, *self-control* juga terbukti berperan sebagai moderasi dalam pengaruh *toxic online disinhibition* terhadap *hate speech* dengan nilai signifikansi 0.036 (p-value <0.05) dengan R^2 kontribusi pengaruh sebesar 0.45 (45%). Dari penelitian ini menghasilkan bahwa *self-control* yang tinggi justru memperkuat pengaruh *toxic online disinhibition* dan *hate speech*. Hasil studi memberikan kontribusi dalam memahami bahwa *self-control* tidak selalu berfungsi sebagai penghambat perilaku agresif di media sosial.

Keywords: *Hate Speech, Toxic Online Disinhibition, Self-Control*

INTRODUCTION

Menurut Parekh (2012) *hate speech* adalah komunikasi yang ditujukan untuk menghina serta merendahkan kelompok orang tertentu berdasarkan identitas mereka, seperti ras, agama, etnis, orientasi seksual, atau jenis kelamin. *Hate speech* menimbulkan perpecahan sosial serta mengganggu harmoni hubungan antar kelompok dalam masyarakat. Data Kementerian Komunikasi dan Informatika mencatat bahwa sepanjang 2018–2023 terdapat 3.640 kasus *hate speech* terkait isu SARA (KOMINFO, 2021), yang menegaskan bahwa fenomena ini bukan sekadar kasus sporadis, melainkan persoalan yang cukup serius di ruang digital. Hal ini sejalan dengan survei Wendratama (2023) kepada 1.500 pengg una media sosial di 38 provinsi di indoesia yang menunjukkan bahwa *hate speech* merupakan konten negatif yang paling sering ditemui (67%), melampaui disinformasi (66%), penipuan digital (60%), pencemaran nama baik (43%), dan pornografi (40%). Secara konseptual, *hate speech* merujuk pada bentuk komunikasi

yang mengandung provokasi, penghinaan, atau penghasutan berbasis identitas SARA seperti ras, etnis, agama, jenis kelamin, dan kewarganegaraan (Parekh, 2012), berbeda dari *cyberbullying* yang cenderung berulang dan personal, *hate speech* dapat muncul dalam satu kejadian dan menyasar kelompok tertentu (Castellanos et al., 2023), dengan dampak yang tidak hanya memengaruhi kesehatan mental korban tetapi juga berpotensi merenggangkan relasi sosial antar kelompok (Cramer et al., 2020; Nishizawa, 2019).

Penyebaran *hate speech* di platform digital memperkuat polarisasi sosial dan menjadi isu serius secara global (Rahmi, 2024), terutama di kalangan Generasi Z yang lahir 1997–2012 (Dimock, 2019) dan tumbuh sebagai *digital natives* dengan intensitas tinggi dimana gen Z menjadikan sosial media sebagai sumber informasi utama dan aktivitas sehari-hari (Evita et al., 2023). Di Indonesia, kelompok ini mencakup sekitar 33% populasi (Christiani & Ikasari, 2020) dan menunjukkan kecenderungan lebih tinggi terpapar sekaligus terlibat dalam *hate speech* akibat keterbukaan informasi digital (Ningrum et al., 2019; Nobata et al., 2016). Faktor anonimitas dan kemudahan manipulasi identitas turut memperkuat *online disinhibition effect* kondisi ketika individu bebas dalam berekspresi tanpa mempertimbangkan norma sosial (Istiani & Islamy, 2020) yang secara empiris berkorelasi dengan meningkatnya kecenderungan melakukan *hate speech* (Wachs & Wright, 2018).

Salah satu faktor yang memicu perilaku *hate speech* di media sosial adalah adanya kebebasan berekspresi serta anonimitas dalam ruang digital (Istiani & Islamy, 2020; Silva et al., 2021). Kondisi ini berkaitan dengan konsep *online disinhibition effect*. Pada peneltian Suler (2004) mendefinisikan *online disinhibition effect* kondisi dimana seseorang menjadi lebih sulit mengendalikan perilaku, pikiran, dan perasaannya saat berinteraksi di dunia maya dibandingkan berinteraksi secara langsung di dunia nyata. *Online disinhibition effect* memiliki 2 dimensi yang berlawanan yakni *benign disinhibition* yang bersifat positif dan *toxic online disinhibition* yang mendorong perilaku negatif seperti penggunaan bahasa kasar, kemarahan, serta ancaman terhadap orang lain (Cheung et al., 2016). Penelitian Shafira & Ardelia (2025) menyebutkan bahwa *Online disinhibition effect* membuat orang yang menggunakan internet merasa lebih leluasa dalam menyampaikan pendapat mereka. Hal ini membuat mereka

lebih terbuka/ bebas menyampaikan pikiran dan perasaan, termasuk cara yang tidak langsung untuk menunjukkan perasaan benci. Interaksi di media sosial juga menciptakan lingkungan yang mendukung penggunaan bahasa yang merendahkan dan provokatif.

Online disinhibition effect menggambarkan kecenderungan seseorang untuk berperilaku lebih bebas dan kurang terbatasi dalam media sosial dibandingkan saat berinteraksi secara langsung. Kebebasan berekspresi harus diikuti oleh tanggung jawab moral dan kemampuan untuk berbagi pengetahuan dengan orang lain. Etika komunikasi di media sosial sangat penting dengan mempertimbangkan tiga hal: etika dalam hal waktu, konten, dan komunikasi. (Badjatiya et al., 2017). Untuk mengendalikan etika berkomunikasi di media sosial dibutuhkan *self-control* yang tinggi untuk mengurangi perilaku *hate speech* di lingkungan virtual yang terjadi di platform media sosial(Ramadhani & Merida, 2024).

Self-control merupakan kemampuan seseorang untuk menahan, mengganti, atau mengubah respons yang biasanya muncul, dan mengelola perilaku, cara berpikir, dan perasaan mereka (De Ridder et al., 2012). Penelitian Irmayanti & Chusniyah (2024) memaparkan bahwa peningkatan *self-control* berperan penting dalam mengurangi potensi perilaku agresif online, termasuk *hate speech*, karena mampu memperlambat respon impulsif dan membantu individu mempertimbangkan dampak sosial dari pesan yang mereka unggah. Hal tersebut didukung dalam penelitian (Wachs et al., 2019) bahwa lingkungan daring seperti media sosial yang ditandai anonimitas, ketidakjelasan dan tidak tatap muka dapat mengurangi kemampuan untuk mengendalikan diri. Selaras dengan penelitian(Lapidot-Lefler & Barak, 2012) *toxic online disinhibiton effec*t terjadi karena hilangnya inhibisi ditunjukkan dalam perilaku agresif yang tidak terlihat di dunia nyata. Sehingga *Self-control* berfungsi sebagai variabel moderasi yang memperlemah efek *toxic online disinhibition effect* terhadap munculnya *hate speech* dan perilaku agresif di platform digital, sehingga kebebasan berekspresi tetap berada dalam batas etika komunikasi yang bertanggung jawab.

Penelitian yang membahas peran *self-control* sebagai moderasi pengaruh *toxic online disinhibition* dan *hate speech* masih terbatas. Penelitian sebelumnya hanya mempelajari hubungan antara dua variabel secara terpisah, seperti hubungan *self-*

control dan *Online disinhibition effect* (Khairani & Guspa, 2025) serta pengaruh *self-control* terhadap *hate speech* (Jung, 2023) . Penelitian sebelumnya menjelaskan bahwa *online disinhibition effect* berkaitan dengan perilaku *hate speech*, sementara *self-control* berperan sebagai faktor protektif dalam perilaku negatif di media sosial. Namun, penelitian di Indonesia masih meneliti variabel tersebut secara terpisah, serta studi global belum ada yang menguji *self-control* sebagai variabel moderasi.

Penelitian ini bertujuan untuk mengisi kesenjangan dari penelitian sebelumnya. Hipotesis dalam penelitian ini terdiri dari dua hipotesis utama, hipotesis pertama terdiri dari hipotesis awal yaitu tidak terdapat pengaruh *toxic online disinhibition* terhadap *hate speech* pada gen Z di media sosial dan hipotesis alternatif yaitu terdapat pengaruh *toxic online disinhibition* terhadap *hate speech* pada gen Z di media sosial. Hipotesis kedua terdiri dari yaitu hipotesis awal yaitu *self-control* tidak berperan secara signifikan sebagai moderasi dalam pengaruh *toxic online disinhibition* terhadap *hate speech* pada gen Z di media sosial dan hipotesis alternatif *self-control* berperan sebagai moderasi dalam pengaruh *toxic online disinhibition* terhadap *hate speech* pada gen Z di media sosial.

RESEARCH METHODS

Penelitian ini merupakan penelitian kuantitatif eksplanatif yang dilakukan menggunakan survey serta data yang diperoleh langsung (Sugeng, 2022). Penelitian menggunakan subjek gen Z dengan rentang usia 16-29 tahun sebanyak 215 responden dengan kriteria; (1) aktif menggunakan sosial media, (2) gen Z yang lahir pada tahun 1997- 2010, (3) pernah/aktif berpartisipasi dalam percakapan atau debat di platform media sosial terkait isu-isu yang sensitive atau kontroversi pada individu/ kelompok dalam hal agama, ras, identitas gender dll. Pengambilan sampel menggunakan *non-probability sample* yaitu *purposive sampling* dimana pengambilan responden bedasarkan kriteria tertertu yang telah ditentukan oleh peneliti (Sugiyono, 2022). Penelitian ini menggunakan 215 sample Generasi Z seluruh indonesia yang pernah/aktif berpartisipasi dalam percakapan atau debat di platform media sosial terkait isu-isu yang sensitif atau kontroversi pada individu/ kelompok dalam hal agama, ras, identitas gender dll.

Data untuk penelitian ini dikumpulkan melalui penyebaran kuesioner secara online

menggunakan platform *Google Forms*.Kuesioner disebarkan melalui berbagai platform media sosial seperti WhatsApp, Instagram, dan Twitter. Selain itu, responden diberikan penjelasan tentang tujuan penelitian dan instruksi untuk menjaga kerahasiaan data mereka sebelum mengisi kuesioner, sehingga data yang dikumpulkan bersifat anonim serta hanya digunakan untuk kepentingan penelitian.

Instrumen *hate speech* menggunakan skala yang dikembangkan oleh (Putri et al., 2024) yang terdiri dari 48 item lalu peneliti melakukan reduksi item terhadap instrumen. Proses ini dilakukan dengan menggugurkan item yang memiliki *corrected item-total correlation* yang rendah, namun tetap mempertahankan integritas konstrak dan redaksi bahasa dari alat ukur orisional yang menghasilkan 23 item final dengan skala Likert 5 poin tanpa modifikasi redaksi, serta diuji validitas dan reliabilitasnya melalui SEM-PLS (outer model) menggunakan SmartPLS yang mendapatkan nilai cronbach alpha sebesar 0.923. Instrumen *toxic online disinhibition* diadopsi dari (Mantara et al., 2023) berdasarkan dimensi *toxic disinhibition* dalam *online disinhibition scale*, terdiri dari 4 item dengan skala Likert 5 poin, hasil uji validitas menunjukkan nilai $\chi^2(6, 252)=0.48$, $p=0.49$ dan reliabilitas sebesar 0.820. Selanjutnya, Skala *self-control* yang digunakan dalam penelitian ini merupakan adaptasi dari *Brief Self-Control Scale* yang dikembangkan oleh De Ridder et al. (2012) dan telah disesuaikan dalam konteks bahasa Indonesia oleh (Arifin & Milla, 2020). Intstrumen terdiri dari 10 item dengan skala Likert 7 poin, serta memiliki reliabilitas *Cronbach's Alpha* sebesar 0.81 (inhibition = 0.68; initiation = 0.69).

Structural Equation Modeling berbasis *Partial Least Squares* (SEM-PLS) dengan menguji dua tahap outer model dan inner model. Analisis data menggunakan perangkat lunak SmartPLS 4.0 Metode ini dipilih karena tidak membutuhkan data berdistribusi normal multivariat dan memungkinkan untuk mengestimasi variabel laten melalui kombinasi linier dari variabel manifes (Hair et al., 2021).

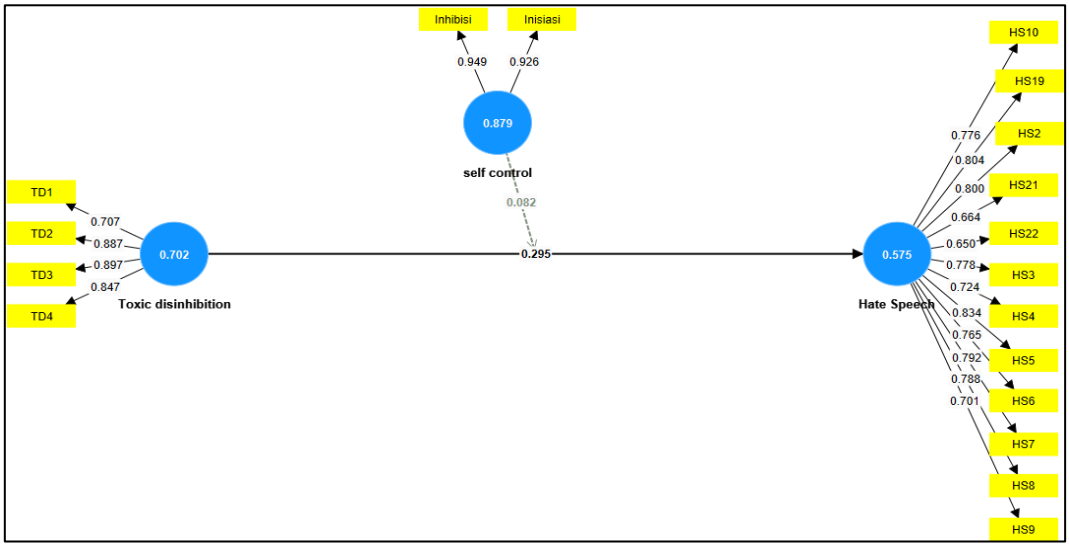
RESULTS AND DISCUSSION

RESULTS

Mayoritas responden pada penelitian ini berusia antara 22- 29 tahun dengan persentase 52,1% dari seluruh responden, diikuti oleh responden dengan rentang umur 16- 21 tahun dengan persentase 47,9%. Jumlah responden dengan jenis kelamin laki-laki sebesar 25,2% dari total responden, dan subjek perempuan sebesar

74,8% dari total responden. Kemudian, isu yang sering diikuti oleh responden didominasi oleh isu gaya hidup sebesar 29,2% kemudian diikuti oleh isu sosial yang sedang viral sebesar 28,8%, politik sebesar 25% dan terakhir SARA sebesar 16,7%.

Gambar 1. Model Penelitian Akhir



Pada gambar 1 Konstruk *toxic online disinhibition* diukur melalui empat indikator (TD1–TD4) dengan nilai factor loading 0,707–0,897 yang representasikan konstruk yang baik. Konstruk *self-control* terdiri dari dua dimensi yang sudah dilakukan *second order*, yaitu inhibisi (0,949) dan inisiasi (0,926), dengan kontribusi yang sangat kuat. Sementara itu, konstruk *hate speech* awalnya diukur dengan 23 item, namun beberapa indikator dieliminasi karena memiliki factor loading di bawah ambang batas, sehingga tersisa 12 item yang memenuhi validitas konvergen. Meskipun demikian, indikator dengan loading 0,50–0,70 masih dipertahankan karena dianggap memadai dalam merepresentasikan konstruk (Hair et al., 2021).

Tabel 1. Hasil Evaluasi Outer Model

Variable	α	ρα	ρc	AVE
Hate speech	0.923	0.939	0.942	0.575
Toxic Online Disinhibition	0.854	0.854	0.903	0.702
Self - Control	0.864	0.882	0.936	0.879

Hasil evaluasi outer model pada tabel 1 memaparkan bahwa seluruh konstruk *hate speech, toxic disinhibition,* dan *self-control* telah memenuhi kriteria reliabilitas internal. Hal ini ditunjukkan oleh nilai *Cronbach's Alpha* (α), rho_A (ρA), dan *Composite Reliability* (ρC) yang seluruhnya di atas batas 0,70. Semua konstruk telah memenuhi *convergent validity* yang diukur melalui koefisien *Average Variance Extracted* (AVE) dimana konstruk dinyatakan memiliki *convergent validity* yang baik jika koefisien AVE >0,50 yang berarti konstruk mampu menjelaskan lebih dari 50% varians indikatornya. Berdasarkan hasil analisis, nilai AVE untuk *hate speech* sebesar 0,575, *toxic online disinhibition* sebesar 0,702 >0,50, dan *self-control* sebesar 0,879 >0,50. Maka dari itu, indikator-indikator pada penelitian ini reliabel dan valid dalam merepresentasikan konstruk yang diukur (Hair et al., 2021).

Tabel 2. HTMT

No	Variable	1	2	3	4	5	6
1	*Hate speech*						
2	*Self-Control*	0.669					
3	*Toxic Disinhibiiton*	0.422	0.257				
4	*Inhibition*	0.704	-	0.266			
5	*Initiation*	0.629	-	0.248	-		
6	*Self-control x Toxic disinhibition*	0.111	0.131	0.203	0.098	0.180	

Pada tabel 2 *Discriminant validity* dilihat dari nilai HTMT. Pada tabel tersebut, seluruh koefisien HTMT <0.90 (valid) sehingga setiap konstruk pada model penelitian ini memuat *discriminant validity* yang memadai. Koefisien HTMT antara *self-control* dan dimensinya (inhibition dan initiation) tidak perlu diinterpretasikan (Sarstedt et al., 2019). Hal ini dikarenakan kedua dimensi tersebut merupakan bagian dari konstruk self-control.

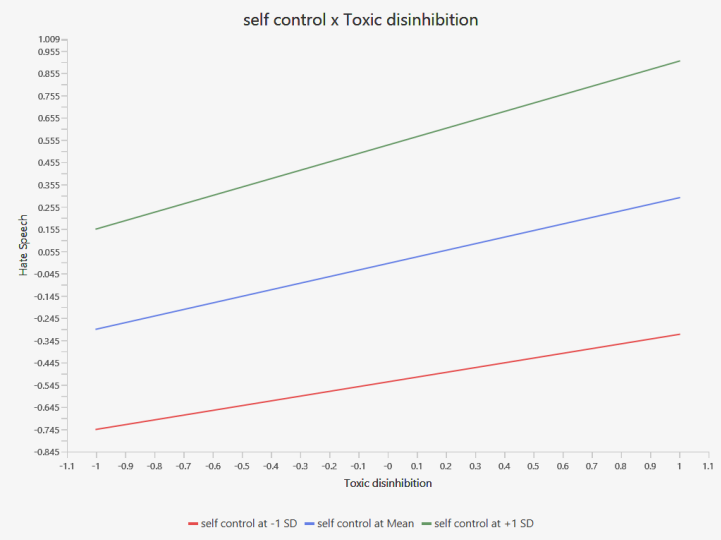
Tabel 3. Hasil Bootstrapping

Hipotesis	Path Coefficient	Sample mean (M)	95% interval kepercayaan path coefficient		p-value	F square
			5%	95%		
Toxic disinhibition -> *Hate speech*	0.295	0.299	0.204	0.394	0.000	0.144
Self-control X Toxic Disinhibition - > *Hate speech*	0.082	0.081	0.004	0.153	0.036	0.014

Pada tabel 3 hasil analisis *boostrapping* menunjukkan bahwa *toxic online disinhibition* terdapat pengaruh positif yang signifikan terhadap *hate speech* dengan nilai signifikan sebesar 0.000 (p-value<0.05). Hal ini mengimplikasikan bahwa semakin tinggi *toxic online disinhibition* pada Generasi Z dimana semakin tinggi kecenderungan melakukan *hate speech*. Pada nilai *effect size* (f^2 = 0,144) menunjukkan pengaruh berada pada kategori kecil hingga sedang.

Selanjutnya, hasil analisis menunjukkan *self-control* berperan sebagai variabel moderasi pengaruh *toxic online disinhibition* dan *hate speech* dengan nilai signifikansi 0,036 (p- value <0,05) juga mengindikasikan pengaruh moderasi tersebut positif signifikan secara statistic. Nilai *effect size* (f^2) sebesar 0,014 > 0,01 menunjukkan kekuatan efek moderasi yang diberikan oleh *self-control* tergolong rendah (Hair et al., 2021).

Gambar 2. *Simple Slope Analysis*



Hasil *simple slope analysis* pada gambar 2 diatas menunjukkan bahwa kemiringan hubungan antara *toxic online disinhibition* dan *hate speech* berbeda pada setiap tingkat *self-control*. Pada kondisi *self-control* rendah (−1 SD), nilai slope sebesar 0,213, yang berarti setiap kenaikan 1 *toxic online disinhibition* akan meningkatkan *hate speech* sebesar 0,213. Pada tingkat *self-control* rata-rata, slope meningkat menjadi 0,295, sedangkan pada self-control tinggi (+1 SD) mencapai 0,377. Perbedaan ini menunjukkan bahwa pengaruh *toxic online disinhibition* terhadap *hate speech* menjadi

semakin kuat seiring meningkatnya *self-control*. Selisih kemiringan antara *self-control* tinggi dan rendah sebesar 0,164 (0,377 – 0,213), yang mengindikasikan adanya efek moderasi dengan arah memperkuat (enhancing effect). Dengan kata lain, individu dengan *self-control* tinggi mengalami peningkatan *hate speech* yang lebih besar dibandingkan individu dengan *self-control* rendah ketika *toxic online disinhibition* meningkat.

Tabel 4. Evaluasi Kecocokan Model

Variable	R Square	SRMR	Q Square
Hate speech	0,451	0.064	0.428

Pada tabel 4 Hasil evaluasi model menunjukkan bahwa model penelitian menjelaskan hubungan antarvariabel dengan baik. Menurut nilai koefisien determinasi (R^2) sebesar 0,451 untuk variabel *hate speech* 45,1% variasi perilaku *hate speech* dapat dijelaskan oleh variabel dalam model. Sisanya faktor-faktor di luar model memengaruhi 54,9%. Nilai tersebut berada di kelas moderat, menunjukkan bahwa model dapat dijelaskan dengan baik. Selain itu, koefisien SRMR 0,064 berada di bawah batas 0,08 yang menunjukkan model memiliki tingkat kesesuaian yang baik dengan data empiris. Koefisien Q^2 0,428 juga menunjukkan bahwa model memiliki kemampuan prediktor yang baik terhadap variabel endogen. Berdasarkan nilai R^2, SRMR, dan Q^2 peneliti dapat melakukan inferensi model penelitian ini memiliki kualitas yang baik serta mampu menjelaskan variable *hate speech*.

DISCUSSION

Hasil penelitian menunjukkan pengaruh *toxic online disinhibition* terhadap *hate speech* memiliki arah positif serta signifikan (0.000 <0.05) pada Gen Z di media sosial, sehingga semakin tinggi *disinhibisi* dalam ruang daring, maka semakin besar kecenderungan individu mengekspresikan komunikasi agresif. Hasil ini sejalan dengan Wachs & Wright (2018) yang menyatakan bahwa individu lebih bebas dan kurang terkontrol dalam interaksi daring dibandingkan tatap muka. Kondisi ini dipengaruhi oleh berkurangnya empati akibat tidak adanya respons emosional langsung, faktor anonimitas, invisibilitas, dan jarak psikologis yang mendorong individu lebih berani mengekspresikan emosi negatif seperti komentar kasar, penghinaan, dan *hate speech* (Suler, 2004). Agresi verbal di media sosial merupakan fenomena yang cukup umum terjadi pada

pengguna internet, khususnya kelompok usia muda yang memiliki intensitas penggunaan media sosial yang tinggi (Silva et al., 2021). Gen Z sebagai kelompok yang tumbuh bersama perkembangan teknologi digital memiliki keterlibatan yang sangat tinggi dalam aktivitas media sosial. Hal ini membuat mereka lebih sering terlibat dalam diskusi maupun perdebatan online, terutama pada isu-isu sosial yang sensitif seperti politik, agama, maupun identitas kelompok yang mempertinggi prevalensi terjadinya konflik komunikasi digital yang memicu munculnya *hate speech*.

Hasil penelitian ini juga menjelaskan bahwa *self-control* berperan sebagai variabel moderasi dalam pengaruh *toxic online disinhibition* terhadap *hate speech* (0.036 <0.05). Namun, yang menarik dalam penelitian ini *self-control* yang tinggi tidak menurunkan kecenderungan *hate speech*, tetapi justru berkaitan dengan tingkat *hate speech* yang lebih tinggi ketika *toxic online disinhibition* meningkat. Hasil analisis menunjukkan bahwa terdapat pengaruh yang kompleks *self-control* dengan *hate speech* dalam komunikasi digital.

Self-control dikaitkan dengan penurunan perilaku agresif atau impulsif. Namun, *self-control* tidak selalu menekan perilaku agresif**.** Dalam penelitian Rai (2019) ditemukan bahwa seseorang dengan *self-control* yang tinggi justru lebih mungkin melakukan agresi ketika mereka memandang tindakan tersebut sebagai sesuatu yang benar secara moral (moralistic aggression) yang terdapat peran evaluasi moral dalam keputusan perilaku agresi tersebut. Dalam konteks *hate speech*, pelaku kemungkinan menggunakan *self-control* untuk mengatur perilakunya agar selaras dengan keyakinan moral, sehingga penyampaian pesan kebencian dianggap sah oleh pelaku. Maka dari itu, perilaku *hate speech* tidak selalu muncul sebagai tindakan impulsif, tapi dapat merupakan bentuk ekspresi yang disengaja di media sosial.

Penelitian menunjukkan bahwa agresivitas verbal di media sosial dipengaruhi faktor emosional, kognitif dan ideologis (Wachs et al., 2021). Individu dengan keyakinan kuat terhadap nilai atau identitas kelompok tertentu cenderung merasa memiliki legitimasi moral untuk menyerang kelompok lain, sehingga dalam konteks ini *self-control* tidak selalu menekan agresivitas, melainkan dapat membentuk ekspresi komunikasi yang lebih terstruktur. Hasil ini sejalan dengan karakteristik Generasi Z sebagai *digital natives* yang mudah mengekspresikan opini secara

terbuka di media sosial (Christiani & Ikasari, 2020), terutama ketika mereka merasa memiliki keyakinan moral yang kuat terhadap suatu isu. Selain itu, kemampuan mengenali *hate speech* yang lebih tinggi pada individu berpendidikan juga dapat mendorong keberanian individu mengekspresikan pandangan ekstrem (Costello et al., 2016), sehingga dalam kondisi tertentu, bahkan individu yang memiliki *self-control* yang baik tetap berpotensi terlibat dalam *hate speech*.

Self-control juga berperan sebagai regulasi yang memungkinkan individu memilih kapan dan bagaimana mereka mengekspresikan agresi (Voggeser et al., 2018). Ini berarti bahwa perilaku *hate speech* pada Gen Z adalah fenomena yang kompleks dan dipengaruhi oleh interaksi faktor situasional dan faktor psikologis (Jiménez-Díaz et al., 2025). *Toxic online disinhibition* sebagai faktor situasional memberikan ruang bagi individu untuk mengekspresikan perilaku agresif dalam komunikasi digital. Sementara itu, *self-control* sebagai faktor psikologis dapat mempengaruhi bagaimana individu mengelola dan mengekspresikan emosi maupun opini mereka dalam ruang digital.

Hasil ini dapat dijelaskan lebih dalam melalui *Moral Disengagement Theory* (Bandura, 1999). Bandura menjelaskan bahwa individu pada dasarnya memiliki standar moral internal yang berfungsi sebagai kendali perilaku. Namun, melalui berbagai mekanisme kognitif seperti *moral justification*, *diffusion of responsibility*, dehumanisasi korban, dan *advantageous comparison* kontrol moral internal tersebut dapat dilepaskan secara selektif sehingga perilaku antisosial seperti *hate speech* dipandang sebagai tindakan yang sah, bahkan mulia (Bandura, 1999). Dalam penelitian ini, kondisi *toxic online disinhibition* yang tinggi memperkuat mekanisme pelepasan moral ini memudahkan individu Gen Z untuk mendehumanisasi target ujaran kebenciannya, sehingga kendali moral yang seharusnya menghambat agresi justru menjadi tidak aktif. Selanjutnya , ketika individu memiliki *self-control* yang tinggi, kemampuan kognitif tersebut justru dimanfaatkan untuk memformulasikan *hate speech* secara lebih strategis dan terencana bukan menghambatnya karena pelaku telah melalui proses pembenaran moral yang membuat tindakannya terasa benar secara ideologis.

Gottfredson & Hirschi (1990) dalam *General Theory of Crime*-nya memandang *self-control* sebagai sifat yang stabil dan

terbentuk sejak masa kanak-kanak, di mana individu dengan *self-control* rendah lebih rentan terhadap perilaku menyimpang, termasuk agresi, karena kecenderungan impulsif dan orientasi pada kepuasan segera. *Self-control* yang tinggi seharusnya berfungsi protektif terhadap *hate speech*. Namun, hasil penelitian ini justru berbeda dengan prediksi teori tersebut, yang menjelaskan bahwa model Gottfredson dan Hirschi tidak memadai untuk menjelaskan agresi daring yang bersifat ideologis dan terencana. Sebaliknya, model trait kontemporer memandang *self-control* sebagai kapasitas regulasi yang bersifat dinamis, dapat bervariasi antarindividu maupun antarsituasi, dan tidak selalu diarahkan untuk menekan perilaku negatif (De Ridder et al., 2012). Apabila tujuan yang dipegang bersifat agresif dan telah melalui pembenaran moral (Bandura, 1999), maka *self-control* yang tinggi justru akan memperkuat ekspresi agresi tersebut, bukan melemahkannya.

Hasil penelitian ini berkontribusi memahami dinamika perilaku komunikasi digital Generasi Z. Hal ini dapat menunjukkan upaya untuk mengurangi *hate speech* di media sosial, yang perlu disertai dengan penguatan literasi digital, etika komunikasi online, dan kesadaran akan efek sosial dari *hate speech*.

CONCLUSION AND SUGGESTIONS

Conclusion

Self-control yang tinggi tidak selalu menekan *hate speech* justru dalam kondisi *toxic online disinhibition* tinggi, individu dapat menggunakan kemampuan regulasi dirinya untuk mengekspresikan agresi secara disengaja dan terstruktur, bukan impulsif semata. Penelitian ini memiliki beberapa keterbatasan, yakni: jumlah responden yang belum mempresentasikan bagaimana keadaan di lapangan seluruhnya, desain *cross-sectional* yang membatasi inferensi kausal, dan instrumen *self-report* yang menyebabkan bias subjektivitas. Strategi yang lebih efektif perlu diarahkan pada pengelolaan konteks digital itu sendiri, seperti mengurangi faktor pemicu disinhibisi, serta memperkuat literasi digital yang menekankan empati, refleksi moral, dan konsekuensi sosial dari komunikasi online. Selain itu, program edukasi perlu mengarahkan individu agar tidak hanya mampu mengendalikan diri, tetapi juga mampu mengevaluasi yang mendasari perilaku agresif mereka.

Suggestion

Penelitian selanjutnya disarankan untuk memperluas jumlah sampel, serta

menambahkan variabel lain seperti kontrol sosial, *moral disengangement, self-regulation* untuk memperoleh pemahaman yang lebih menyeluruh mengenai dinamika *hate speech* di media sosial.

REFERENCES

Arifin, H. H., & Milla, M. N. (2020). Adaptasi dan properti psikometrik skala kontrol diri ringkas versi Indonesia. *Jurnal Psikologi Sosial*, *18*(2), 179–195. https://doi.org/10.7454/jps.2020.18

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 759–760. https://doi.org/10.1145/3041021.3054223

Bandura, A. (1999). Moral Disengagement in the Perpetration of Inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., & Bilz, L. (2023). Hate speech in adolescents: A binational study on prevalence and demographic differences. *Frontiers in Education*, *8*, 1076249. https://doi.org/10.3389/feduc.2023.1076249

Cheung, C. M. K., Wong, R. Y. M., & Chan, T. K. H. (2016). Online Disinhibition: Conceptualization, Measurement, and Relation to Aggressive Behaviors. *Proceedings of International Conference on Information Systems*, 1–10.

Christiani, L. C., & Ikasari, P. N. (2020). *Generasi Z dan Pemeliharaan Relasi Antar Generasi dalam Perspektif Budaya Jawa* (Pt. 2). *2*, 84–105.

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*, 311–320. https://doi.org/10.1016/j.chb.2016.05.033

Cramer, Robert. J., Fording, R. C., Gerstenfeld, P., Kehn, A., Marsden, J., Deitle, C., King, A., Smart, S., &

Nobles, M. R. (2020). *Hate-Motivated Behavior: Impacts, Risk Factors, And Interventions*. Project HOPE. https://doi.org/10.1377/hpb20200929.601434

De Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking Stock of Self-Control: A Meta-Analysis of How Trait Self-Control Relates to a Wide Range of Behaviors. *Personality and Social Psychology Review*, *16*(1), 76–99. https://doi.org/10.1177/1088868311418749

Dimock, M. (2019, January 17). Where Millennials end and Generation Z begins. *Pew Research Center*. https://www.pewresearch.org/short-reads/2019/01/17/where-millennials-end-and-generation-z-begins/

Evita, N., Prestianta, A. M., & Asmarantika, R. A. (2023). Patterns of media and social media use in generation z in Indonesia. *Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies)*, *7*(1), 195–214. https://doi.org/10.25139/jsk.v7i1.5230

Gottfredson, M., & Hirschi, T. (1990). *A General Theory of Crime. Stanford*. CA: Stanf. Univ. Press.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*. Springer Nature.

Irmayanti, N., & Chusniyah, T. (2024). Empathy in the Digital Age: The Role of Self-Control and Social Control in Addressing Cyberviolence. *Bisma The Journal of Counseling*, *8*(2). https://doi.org/10.23887/bisma.v8i2.86154

Istiani, N., & Islamy, A. (2020). Fikih Media Sosial Di Indonesia. *ASY SYAR'IYYAH: JURNAL ILMU SYARI'AH DAN PERBANKAN ISLAM*, *5*(2), 202–225. https://doi.org/10.32923/asy.v5i2.1586

Jiménez-Díaz, O., Wachs, S., Del Rey, R., & Mora-Merchán, J. A. (2025). Associations Between Searching and Sending Cyberhate: The Moderating Role of the Need of Online Popularity and Toxic Online Disinhibition. *Cyberpsychology, Behavior, and Social Networking*, *28*(1), 37–43.

https://doi.org/10.1089/cyber.2024.0305

Jung, C. W. (2023). Role of Informal Social Control in Predicting Racist Hate Speech on Online Platforms: Collective Efficacy and the Theory of Planned Behavior. *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 507–518. https://doi.org/10.1089/cyber.2022.0107

Kenny, D. A. (2018, September 15). *SEM: Moderation*. https://davidakenny.net/cm/moderation.htm

Khairani, A., & Guspa, A. (2025). Pengaruh Self Control terhadap Toxic Online Disinhibition Effect pada Generasi Z Pengguna Aplikasi X. *CAUSALITA : Journal of Psychology*, *3*(1), 243–253. https://doi.org/10.62260/causalita.v3i1.525

KOMINFO. (2021). https://www.komdigi.go.id/berita/pengumuman/detail/siaran-pers-no-143-hm-kominfo-04-2021-tentang-sejak-2018-kominfo-tangani-3-640-ujaran-kebencian-berbasis-sara-di-ruang-digital

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

Mantara, A. Y., Sa'id, M., Zahra, G. A., Rizkina, A. T., Febriyanti, L., & Prastika, S. B. (2023). Adaptation of the Online Disinhibition Effect Scale. *KnE Social Sciences*. https://doi.org/10.18502/kss.v8i19.14381

Ningrum, D. J., Suryadi, S., & Chandra Wardhana, D. E. (2019). Kajian Ujaran Kebencian Di Media Sosial. *Jurnal Ilmiah KORPUS*, *2*(3), 241–252. https://doi.org/10.33369/jik.v2i3.6779

Nishizawa, Y. (2019). *A Study on the Mental Foundations and Evolving Legal Norms Regarding Hate Speech in Japan: Bandwagon Effect or Social Desirability Bias*. https://ynishiza.doshisha.ac.jp/ynishiza2014/downloadables/Nishizawa2019%20jpsa19%20Mental%20Foundations%20Regarding%20HS%20Regulation.pdf

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*, 145–153. https://doi.org/10.1145/2872427.2883062

Parekh, B. (2012). Is There a Case for Banning Hate Speech? In M. Herz & P. Molnar (Eds.), *The Content and Context of Hate Speech* (1st ed., pp. 37–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.006

Putri, A., Kurniawan, R., Mardianto, & Utami, R. H. (2024). Hubungan *Cyberwellness* dengan Perilaku *Hatespeech* pada Remaja di Media Sosial. *CAUSALITA : Journal of Psychology*, *2*(2), 244–253. https://doi.org/10.62260/causalita.v2i2.329

Rahmi. (2024). Empathy and Hate Speech in Social Media: The Case of Indonesia. *International Journal of Social Science and Human Research*, *07*(03). https://doi.org/10.47191/ijsshr/v7-i03-29

Rai, T. S. (2019). Higher self-control predicts engagement in undesirable moralistic aggression. *Personality and Individual Differences*, *149*, 152–156. https://doi.org/10.1016/j.paid.2019.05.046

Ramadhani, A. Z., & Merida, S. C. (2024). Self-Control and the Phenomenon of Toxic Online Disinhibition in Teenagers Who Have Twitter. *Nusantara Journal of Behavioral and Social Sciences*, *3*(2), 45–52. https://doi.org/10.47679/202454

Sarstedt, M., Hair, J. F., Cheah, J.-H., Becker, J.-M., & Ringle, C. M. (2019). How to Specify, Estimate, and Validate Higher-Order Constructs in PLS-SEM. *Australasian Marketing Journal*, *27*(3), 197–211. https://doi.org/10.1016/j.ausmj.2019.05.003

Shafira, M. T. I., & Ardelia, V. (2025). Hubungan Self-Esteem dengan Online Disinhibition pada Emerging Adult Pengguna Media Sosial X. . . *Character*, *12*.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2021). Analyzing the Targets of Hate in

Online Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 687–690. https://doi.org/10.1609/icwsm.v10i1.14811

Sugeng, B. (2022). *Fundamental Metodologi Penelitian Kuantitatif (Eksplanatif)*. Deepublish.

Sugiyono. (2022). Metode Penelitian Kuantitatif. In *Metode Penelitian Kuantitatif*. Alfabeta.

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Voggeser, B. J., Singh, R. K., & Göritz, A. S. (2018). Self-control in Online Discussions: Disinhibited Online Behavior as a Failure to Recognize Social Cues. *Frontiers in Psychology*, *8*, 2372. https://doi.org/10.3389/fpsyg.2017.02372

Wachs, S., Mazzone, A., Milosevic, T., Wright, M. F., Blaya, C., Gámez-Guadix, M., & O'Higgins Norman, J. (2021). Online correlates of cyberhate involvement among young people from ten European countries: An application of the Routine Activity and Problem Behaviour Theory. *Computers in Human Behavior*, *123*, 106872. https://doi.org/10.1016/j.chb.2021.106872

Wachs, S., & Wright, M. F. (2018). Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition. *International Journal of Environmental Research and Public Health*, *15*(9), 2030. https://doi.org/10.3390/ijerph15092030

Wachs, S., Wright, M. F., & Vazsonyi, A. T. (2019). Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition. *Criminal Behaviour and Mental Health*, *29*(3), 179–188. https://doi.org/10.1002/cbm.2116

Wendratama. (2023, May 27). Riset Pengaturan Konten Ilegal dan Berbahaya di Media Sosial. *Engelbertus Wendratama*. https://wendratama.com/2023/05/27/riset-pengaturan-konten-ilegal-dan-berbahaya-di-media-sosial/