

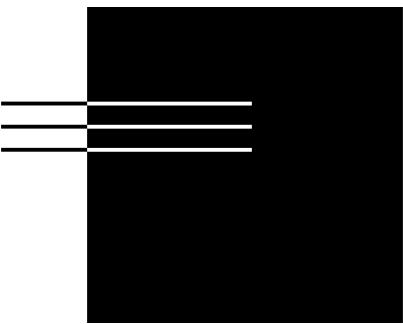
ANALISIS SPAM PADA KOMENTAR YOUTUBE MUSIC VIDEO UNTUK OPTIMALISASI MODERASI KONTEN

Pemanfaatan IBM Granite LLM untuk Deteksi Spam



Content

01	Project Overview
02	Tools and Data Source
03	Analysis Process
04	Insight & Findings
05	AI Support Explanation
06	Conclusion & Recommendation

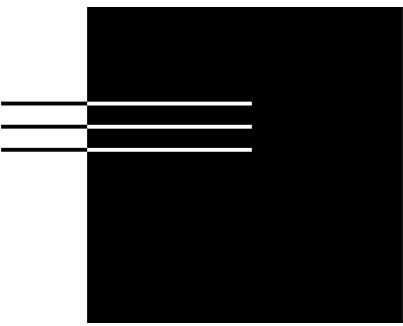


Project Overview

YouTube merupakan salah satu platform video terbesar dengan jutaan pengguna yang mengakses video dari berbagai kategori termasuk music video (MV). MV sering menjadi video dengan jumlah penonton dan komentar yang tinggi dan menjadikannya rentan terhadap komentar spam yang dapat mengganggu kenyamanan penonton dan berpotensi mengarah pada penipuan atau phishing.

Besarnya jumlah komentar membuat proses penanganan spam secara manual menjadi sulit dilakukan oleh pemilik channel. Oleh karena itu, **proyek ini bertujuan untuk menganalisis komentar spam pada YouTube Music Video sebagai langkah awal dalam mengoptimalkan moderasi konten menggunakan Large Language Model (LLM).**

Dengan memahami pola dan distribusi spam, pemilik channel dan platform dapat mengidentifikasi waktu rawan spam serta kata kunci yang sering muncul pada spam, sehingga proses deteksi dan penanganan komentar spam dapat dioptimasi menggunakan LLM. Hal ini akan membantu menjaga kualitas diskusi pada kolom komentar dan meningkatkan pengalaman penonton.



Tools and Data Source

TOOLS

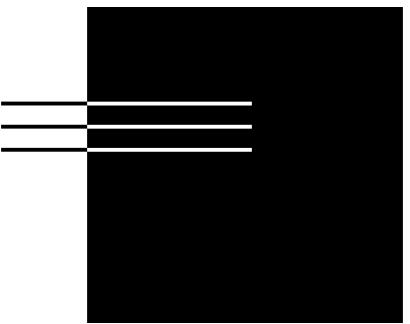


- Bahasa Pemrograman: Python
- Platform Analisis: Google Colab
- Model AI: IBM Granite LLM
[IBM Granite - Hugging Face](#)

DATA SOURCE



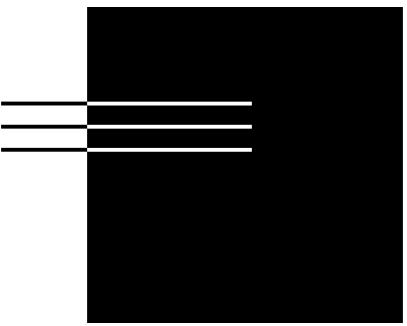
- Sumber: Kaggle
[YouTube Comments Spam Dataset](#)
- Jumlah:
1,953 entries,
6 atribut.



Tools and Data Source

DATASET DESCRIPTION

- comment_id: ID unik komentar.
- author: Nama pengguna yang mengirim komentar.
- date: Tanggal komentar dibuat.
- content: Isi komentar.
- video_name: Nama video tempat komentar dibuat.
- class: Label target dengan 1 = spam dan 0 = non-spam.



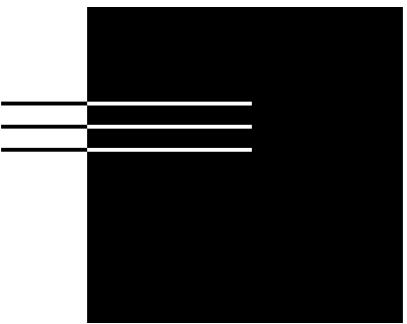
Analysis Process

PREPARATION

- Menyiapkan library yang dibutuhkan.
- Mengimpor dataset CSV dan membuat DataFrame.

DATA UNDERSTANDING

- Melihat isi dataset menggunakan `df.shape`, `df.info()`, `df.head()` untuk eksplorasi awal.
- Melihat karakteristik dan statistik dataset.
- Memeriksa data duplikat dan missing value.



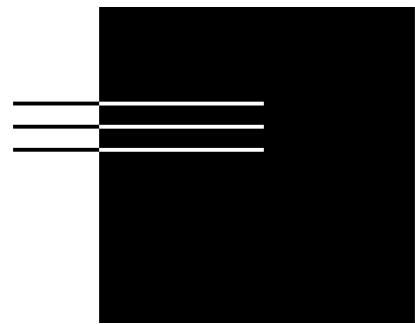
Analysis Process

DATA CLEANING

- Membersihkan data berdasarkan temuan pada tahap understanding.
- Menyesuaikan tipe data setiap kolom.
- Menambahkan kolom YEAR_MONTH untuk memudahkan visualisasi tren.

EXPLORATORY DATA ANALYSIS (EDA)

- Melihat jumlah unik pada setiap kolom.
- Melakukan EDA untuk melihat distribusi label, waktu, video, kata, dan author.
- Visualisasi untuk membantu menemukan pola.

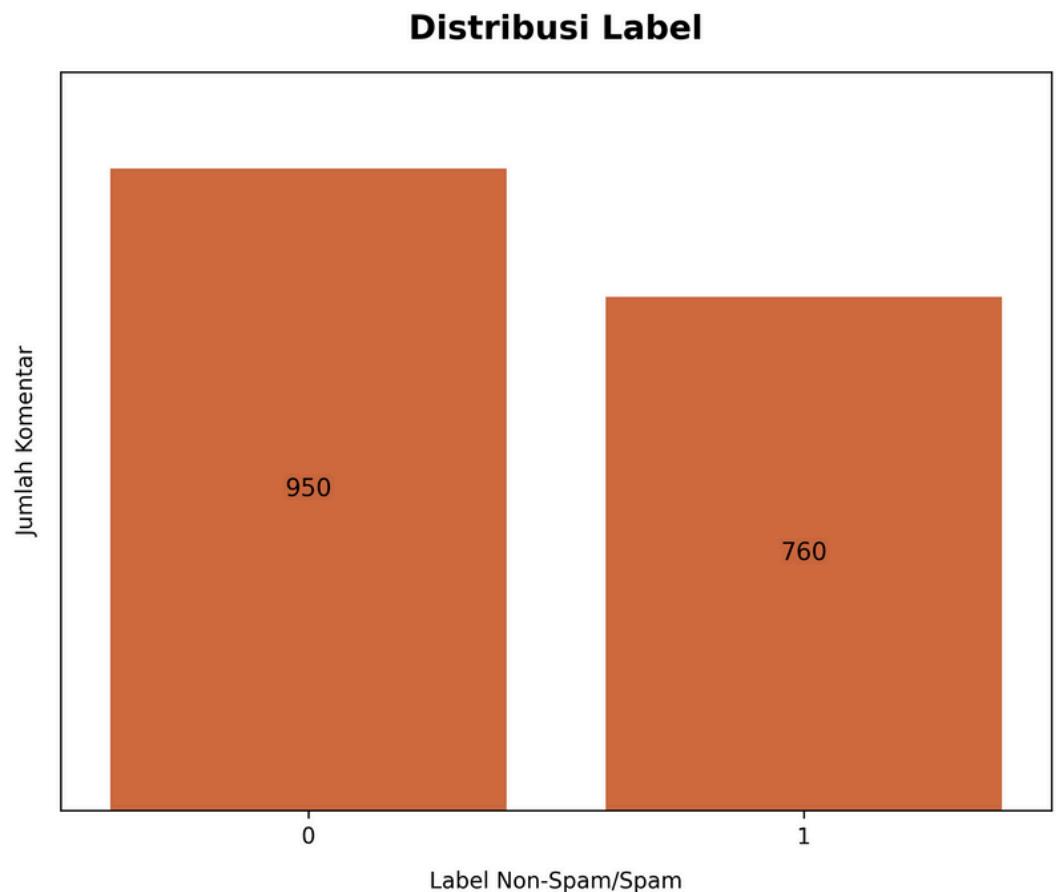


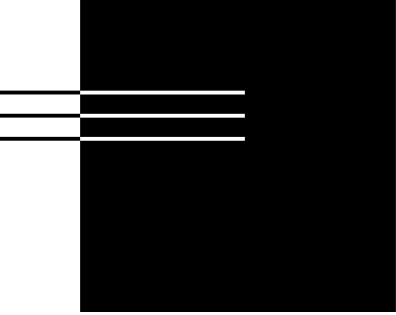
Insight & Findings

Dataset yang telah dibersihkan menyisakan 1.710 komentar dari 1.615 pengguna unik yang diambil dari 5 music video populer selama 19 bulan dalam rentang Juli 2013 – Mei 2015.

Hasilnya **lebih dari separuh komentar tersebut teridentifikasi sebagai spam (950 komentar)**. Temuan ini menunjukkan bahwa masalah spam menjadi tantangan nyata dalam menjaga kualitas interaksi penonton pada video musik di YouTube.

COMMENT_ID	1710
AUTHOR	1615
DATE	1709
CONTENT	1540
VIDEO_NAME	5
CLASS	2
YEAR_MONTH	19

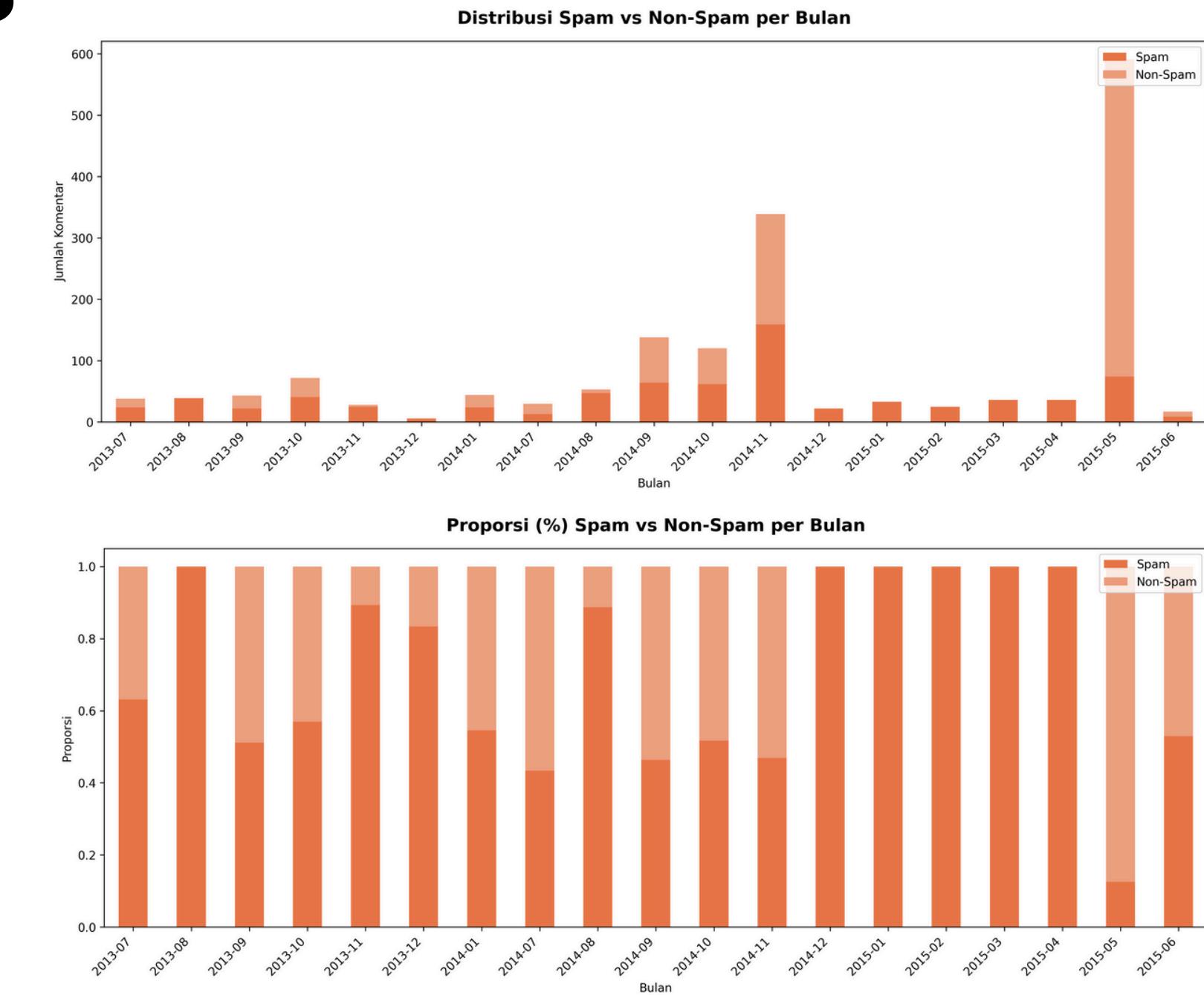


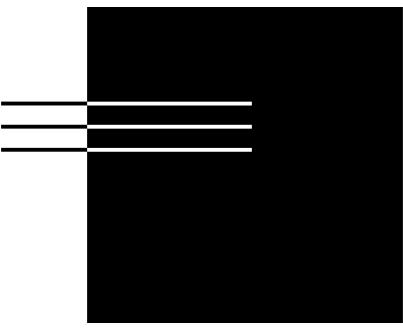


Insight & Findings

Analisis waktu menunjukkan bahwa **spam cenderung muncul mendominasi di bulan-bulan dengan aktivitas komentar yang rendah.** Saat volume komentar ramai seperti Mei 2015, spam justru turun proporsinya karena komentar non-spam meningkat signifikan.

Tercatat selama **lima bulan berturut-turut (Desember 2014 – April 2015)**, semua komentar yang masuk adalah **100% spam** yang menandakan adanya **pola spam wave/pola spam** muncul secara konsisten dalam periode sepi komentar yang perlu diwaspadai pemilik channel.

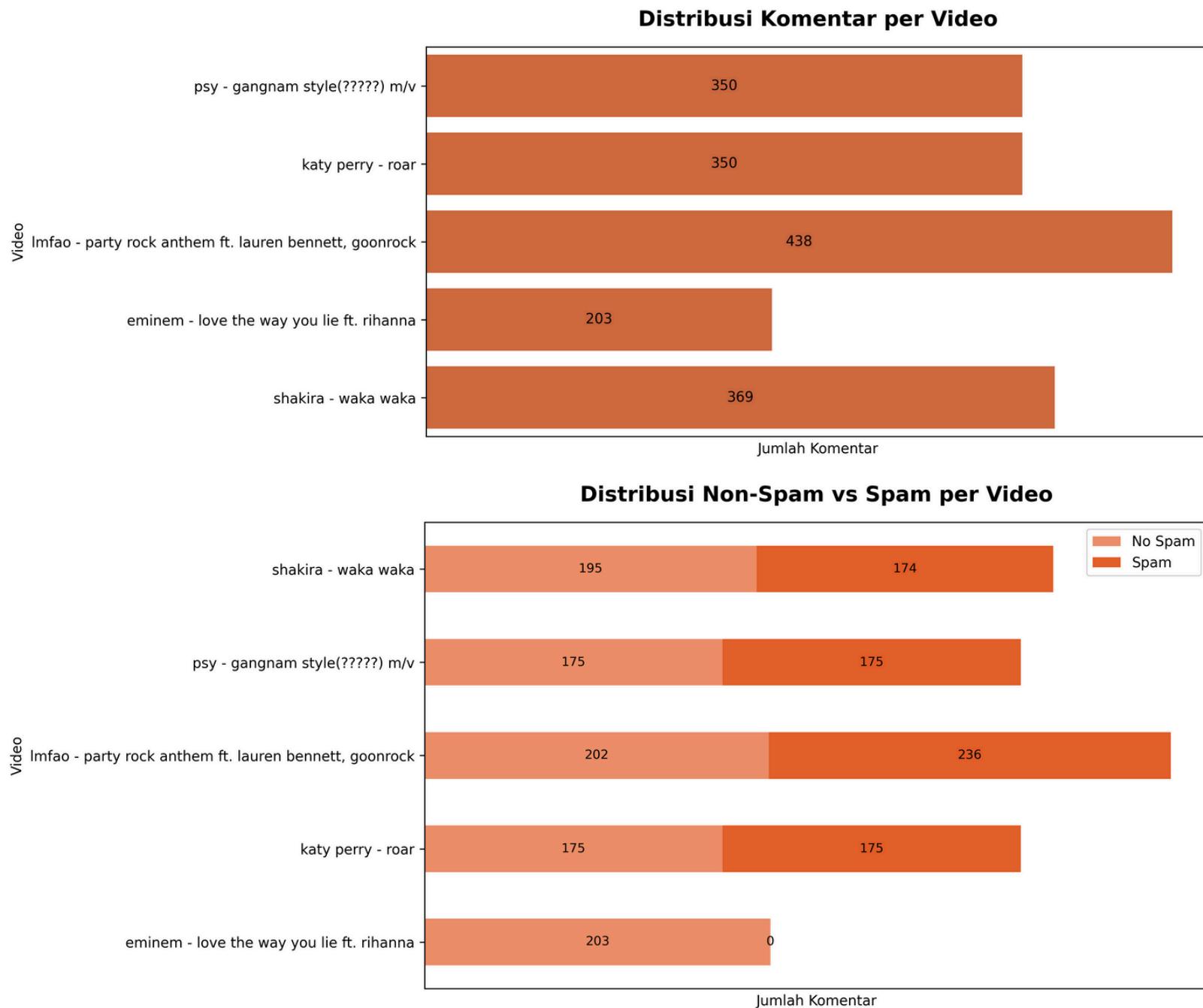


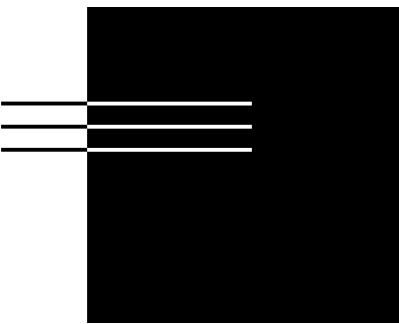


Insight & Findings

LMFAO - Party Rock Anthem menjadi MV dengan jumlah komentar terbanyak sekaligus memiliki distribusi spam tertinggi di antara semua video yang dianalisis. Sebaliknya, Eminem - Love The Way You Lie menjadi satu-satunya MV yang tidak memiliki komentar spam dan memiliki jumlah komentar yang paling sedikit. Sedangkan video lainnya menunjukkan distribusi spam dan non-spam relatif seimbang.

Temuan ini menunjukkan bahwa **video dengan popularitas tinggi cenderung memiliki potensi spam yang lebih besar** sehingga memerlukan perhatian lebih dalam moderasi komentar.

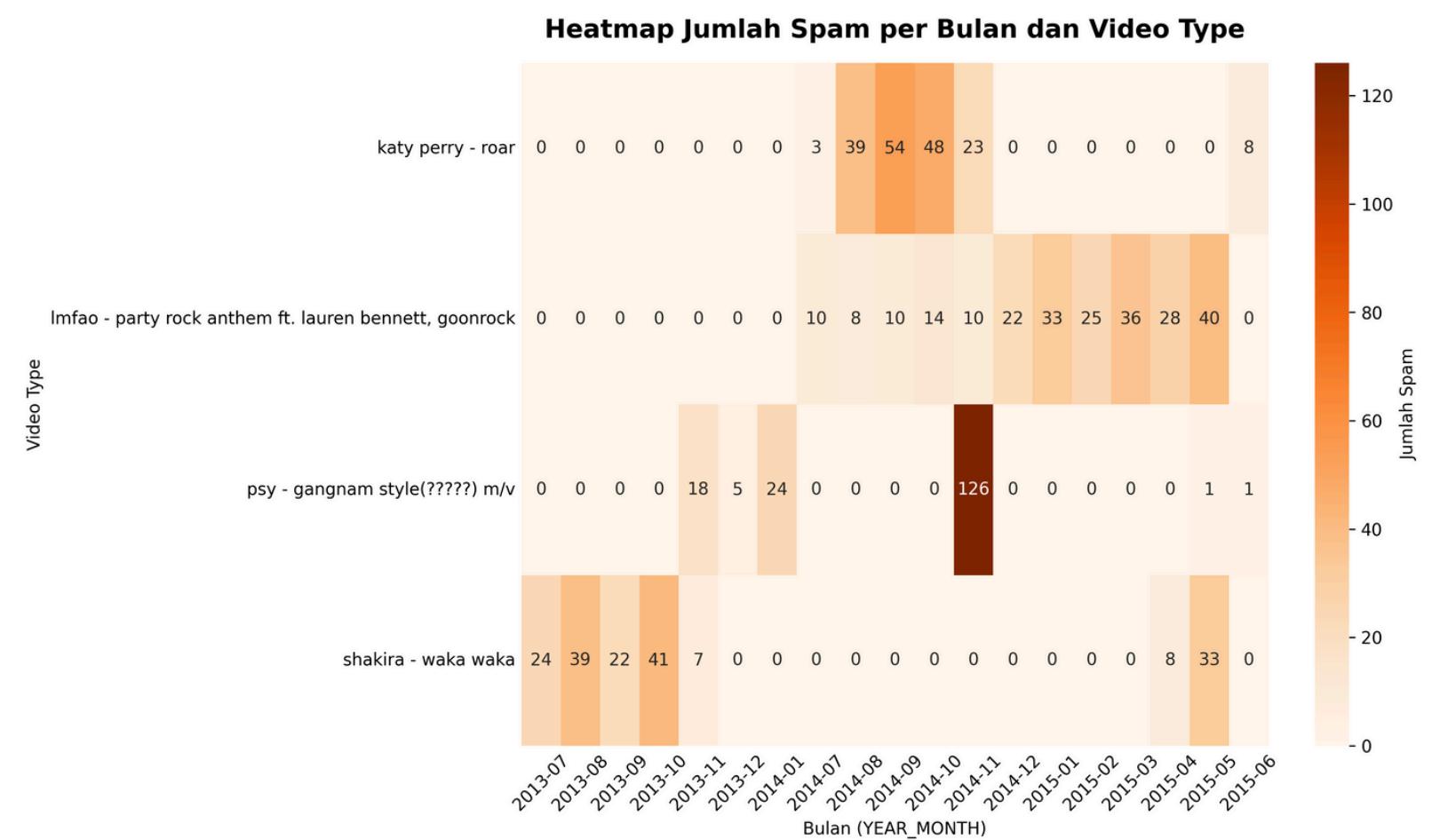


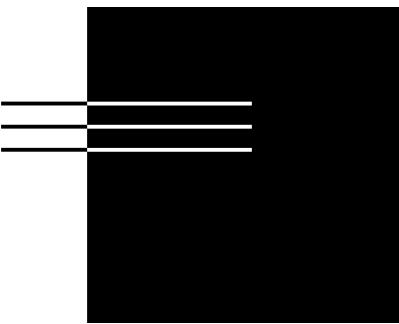


Insight & Findings

Hasil analisis heatmap mengungkap bahwa MV LMFAO - Party Rock Anthem memiliki spam secara beruntun selama 11 bulan tanpa jeda. Hal tersebut menunjukkan **adanya aktivitas spammer yang terfokus pada satu video dalam jangka panjang**. Sementara itu **puncak aktivitas spam justru terjadi pada MV PSY - Gangnam Style yang memiliki distribusi komentar seimbang di November 2014 dengan 126 spam, bukan pada video dengan komentar terbanyak**.

Temuan ini menegaskan bahwa spam dapat hadir secara terus-menerus pada video tertentu dan dapat memuncak pada periode tertentu di video lainnya. Hal ini penting bagi pemilik kanal untuk memantau dan melakukan moderasi secara rutin untuk menjaga kualitas interaksi di kolom komentar.

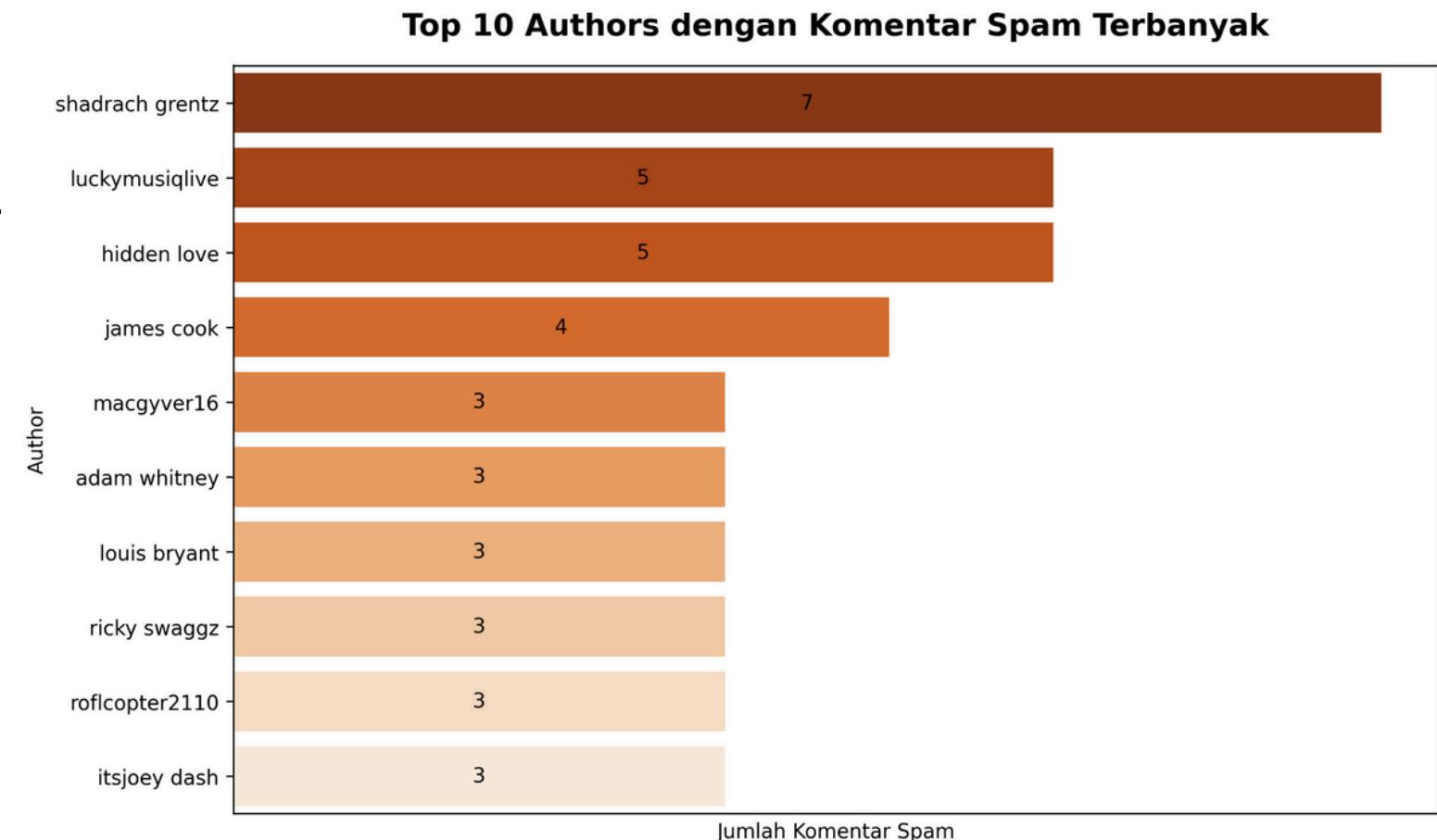


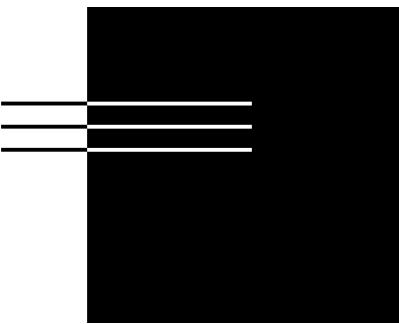


Insight & Findings

Berdasarkan top 10 author dengan komentar spam terbanyak, terlihat bahwa tidak ada satu pun akun spammer yang mendominasi secara signifikan. Author dengan spam terbanyak hanya memiliki 7 komentar spam dan author lainnya rata-rata memiliki 3-5 komentar spam.

Pola ini menunjukkan bahwa **aktivitas spam cenderung tersebar tipis pada banyak akun berbeda**. Maka dari itu, upaya moderasi tidak bisa hanya menargetkan satu akun besar tetapi perlu memantau banyak akun dengan intensitas rendah.

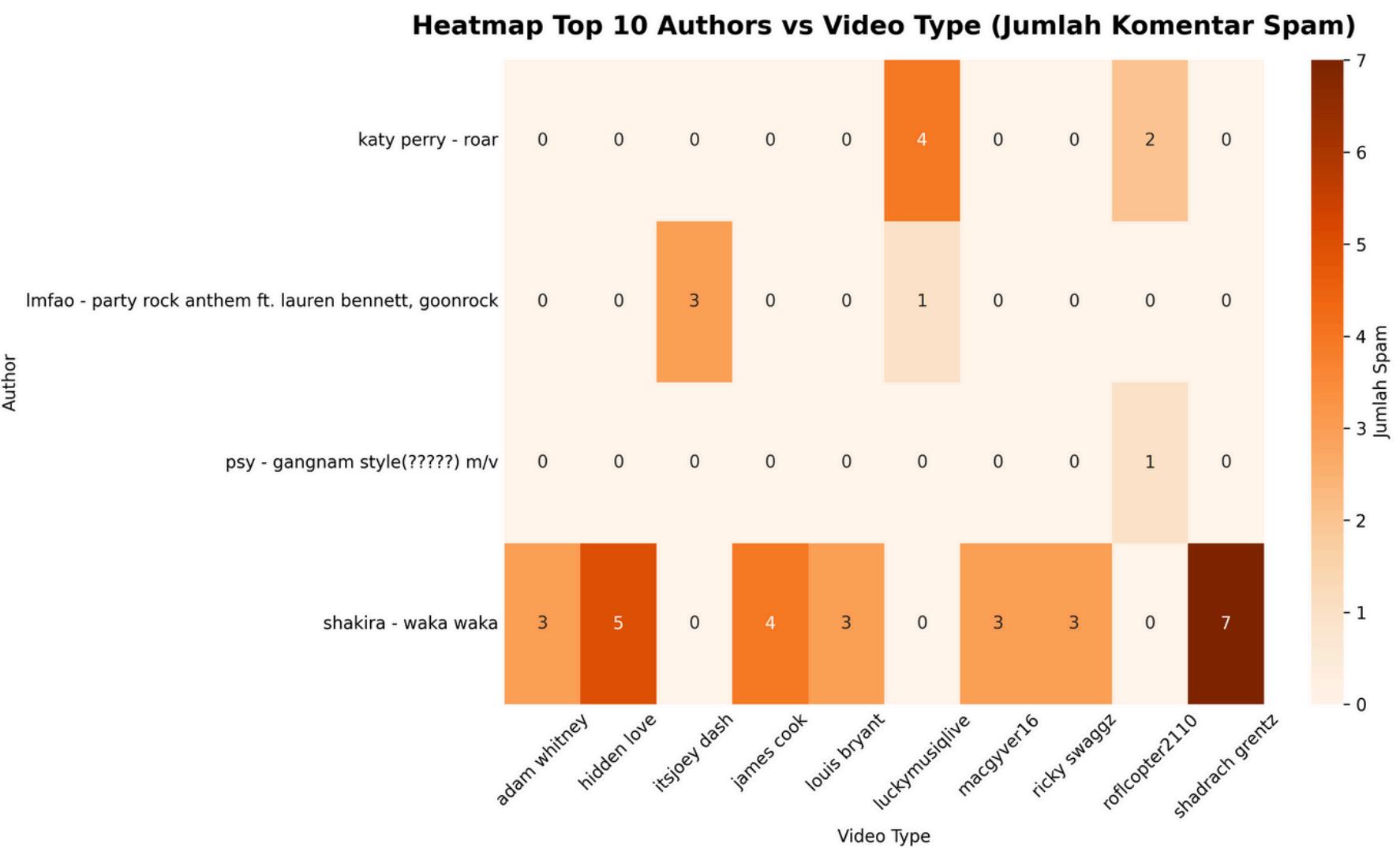


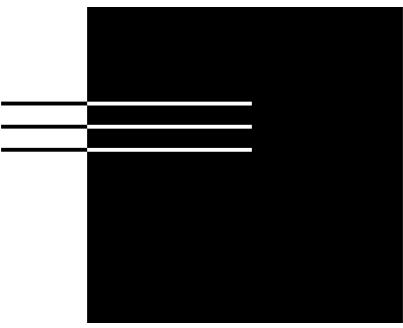


Insight & Findings

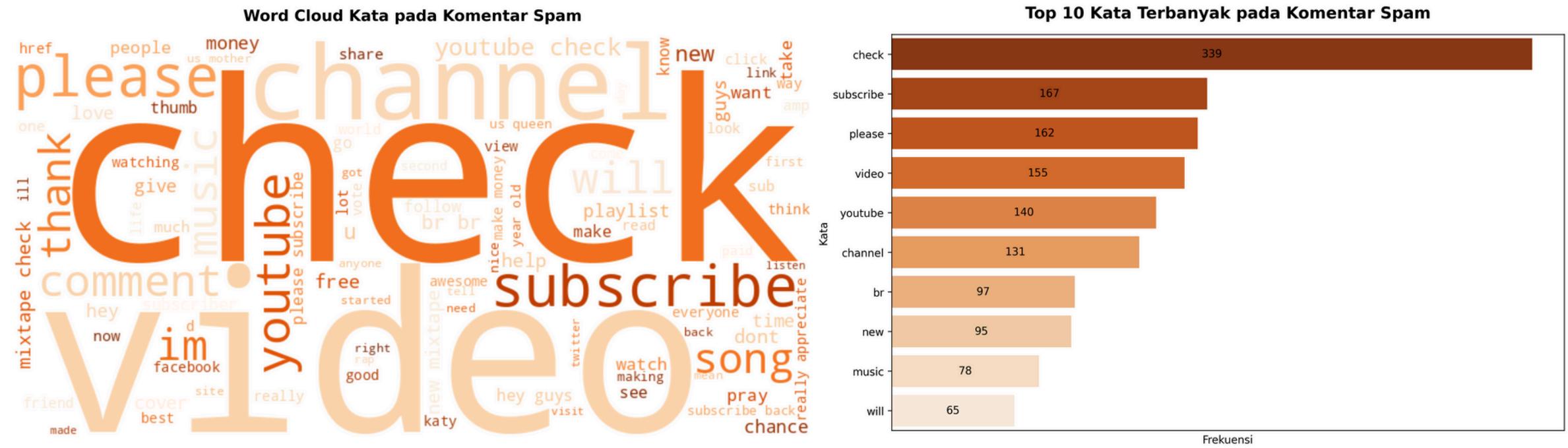
Dari analisis heatmap top 10 spammer, ditemukan bahwa **pola spam intensif dengan jumlah komentar spam tinggi per author lebih banyak ditujukan pada video tertentu** yaitu MV Shakira. Hal ini menunjukkan adanya spammer yang fokus menyerang satu video secara berulang.

Sebaliknya **spam yang tinggi pada MV populer seperti LMFAO dan PSY justru tidak berasal dari top spammer, melainkan dari banyak author berbeda dengan intensitas spam rendah** sehingga tidak terlihat dominan pada heatmap ini.



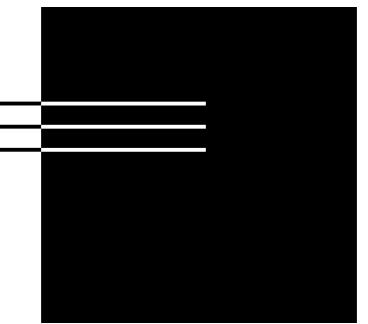


Insight & Findings



Hasil wordcloud menunjukkan bahwa **konten spam di komentar MV YouTube didominasi kata seperti *check*, *subscribe* dan *please* yang mengindikasikan pola ajakan untuk mengunjungi tautan atau channel lain.** Selain itu, kata *video* juga sering muncul dan memperkuat indikasi promosi terselubung dalam komentar spam.

AI Support Explanation

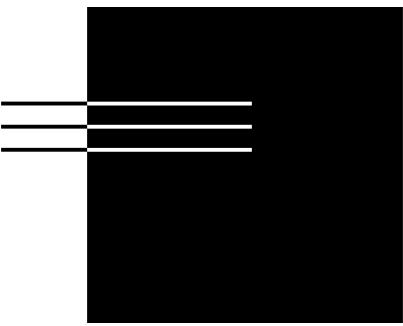


The screenshot shows the IBM WatsonX Prompt Lab interface. At the top, there's a navigation bar with 'IBM watsonx' and various dropdowns like 'Upgrade', 'UNJ', 'Sydney', and 'AR'. Below the navigation is a toolbar with 'AI guardrails off', 'Unsaved', 'New prompt', and a 'Deploy' button. The main area is titled 'Projects / Capstone_YoutubeSpam_Granite8B / Prompt Lab'. It features a 'Chat' tab selected, followed by 'Structured' and 'Freeform'. A central panel displays four sample questions generated by the AI model 'granite-3-2b-instruct':

- What are more efficient alternatives to a 'for loop' in Python?
- What is the Transformers architecture?
- Create a chart of the top NLP use-cases for foundation models.
- Describe generative AI using emojis.

At the bottom, there's a text input field with 'Type something...' placeholder text and a URL at the bottom left: 'au-syd.dai.cloud.ibm.com/projects/9d536ec8-9898-43bc-8783-3b9a82a75a0...'. A watermark 'HACKTIV8' is visible in the bottom right corner of the screenshot.

Dalam proyek ini, **IBM WatsonX Granite 3-2b-Instruct** digunakan sebagai AI support untuk menguji apakah AI mampu mendeteksi komentar spam YouTube secara otomatis berdasarkan pola yang ditemukan saat analisis manual.



AI Support Explanation

PROMPT

You are a spam classification assistant for YouTube comments.

Your task:

Classify each YouTube comment below strictly as either:

- "SPAM"
- "NOT SPAM"

Classification criteria:

Label as "SPAM" if the comment contains:

- Unsolicited promotion
- Links or attempts to redirect users
- Repeated phrases
- Scams or misleading information
- Requests to subscribe or check channels deceptively

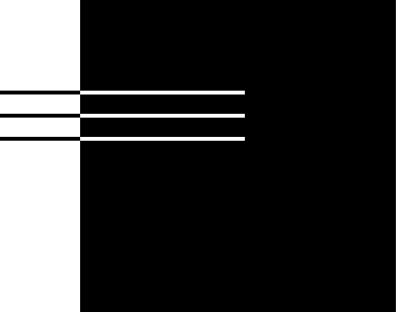
Otherwise, label as "NOT SPAM".

Important: Only output the labels "SPAM" or "NOT SPAM" for each comment, in the order given separated by commas.

Comment: {30 comments is here}

Prompt disusun berdasarkan prinsip clarity dengan instruksi jelas, relevance dengan kriteria spam yang sesuai hasil analisis, dan accuracy dengan memaksa output hanya dua label yang dapat langsung diverifikasi dengan dataset.

Hal ini membantu WatsonX Granite dalam memahami tugas secara fokus dan konsisten dalam klasifikasi.



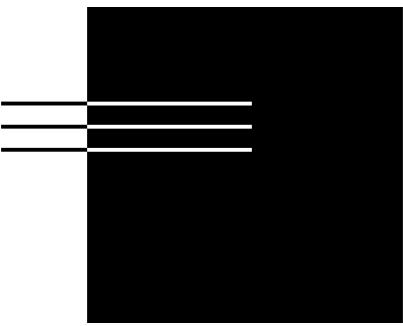
AI Support Explanation

No	Komentar yang Salah Diklasifikasikan oleh AI	Label Sebenarnya	Prediksi Granite
11	"check out this video on youtube:"	SPAM	NOT SPAM
13	"check out this playlist on youtube:"	SPAM	NOT SPAM
14	"check out this video on youtube:"	SPAM	NOT SPAM
17	"check out this video on youtube:"	SPAM	NOT SPAM
20	"huh, anyway check out this youtube channel:	SPAM	NOT SPAM
22	"katy you are a shit go die!!!! roar is a	NOT SPAM	SPAM
26	"check out partyman318 for good tunes!! :d"	SPAM	NOT SPAM
27	"i love that you subscribed"	SPAM	NOT SPAM
28	"check out this video on youtube:"	SPAM	NOT SPAM
29	"like if you came here to see how many views this	SPAM	NOT SPAM
30	"i am from brazil please subscribe to my channel	SPAM	NOT SPAM

Tabel Kesalahan Klasifikasi Granite WatsonX pada Sampel 30 Komentar

RESULT

- 30 komentar diuji menggunakan Granite.
- **Granite berhasil mengklasifikasikan 19/30 komentar dengan benar dan mendapatkan akurasi 63% akurasi.**
- Terjadi 11 salah klasifikasi pada komentar ambigu seperti ajakan halus atau tanpa link. Komentar “check out this video on youtube:” seharusnya SPAM namun terdeteksi NOT SPAM oleh Granite.
- Kasus unik pada komentar No. 22 terdeteksi SPAM padahal seharusnya NOT SPAM karena mungkin Granite menangkap kata kasar sebagai spam.

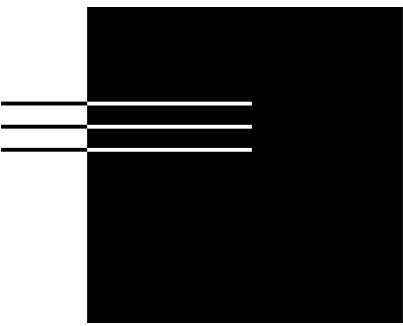


AI Support Explanation

INSIGHT

- Granite efektif mendeteksi spam terbuka (link/ promosi eksplisit).
- Namun, model kesulitan menangkap promosi halus tanpa link.
- **Prompt perlu diperbaiki agar lebih eksplisit untuk meningkatkan akurasi.**

IBM WatsonX Granite 3-2b-Instruct terbukti dapat mempercepat proses labeling komentar spam pada dataset baru untuk efisiensi analisis ke depannya.

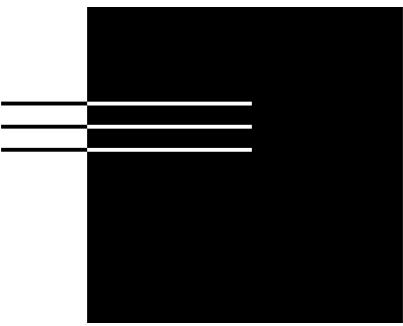


Conclusion & Recommendation

CONCLUSION

Dari hasil analisis ditemukan bahwa:

- Spam sering muncul mendominasi pada bulan dengan aktivitas komentar rendah yang mengindikasikan spammer memanfaatkan periode sepi untuk beraksi.
- Spam dapat hadir secara terus-menerus dalam periode panjang pada video tertentu/spam wave, seperti pada MV LMFAO - Party Rock Anthem yang mengalami spam 11 bulan berturut-turut.
- Puncak spam dapat terjadi secara mendadak pada periode tertentu pada video populer, seperti PSY - Gangnam Style pada November 2014.
- Spam tidak terpusat pada satu akun dominan, melainkan terdistribusi secara luas ke banyak akun dengan intensitas rendah.
- Konten spam umumnya berisi ajakan untuk mengunjungi tautan atau channel lain dengan kata kunci seperti check, subscribe, please, dan video.

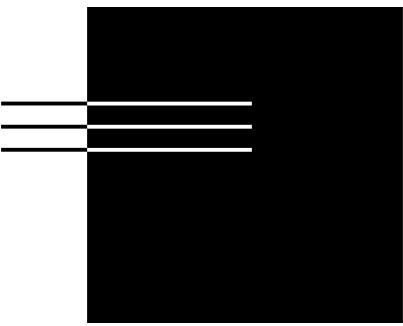


Conclusion & Recommendation

CONCLUSION

Analisis komentar pada 5 YouTube Music Video selama 19 bulan menemukan bahwa spam merupakan masalah nyata dalam menjaga kualitas interaksi di kolom komentar YouTube dengan lebih dari separuh komentar teridentifikasi sebagai spam.

Temuan dan hasil analisis menegaskan bahwa spam pada komentar YouTube MV terjadi secara konsisten, terdistribusi pada banyak akun, dan sering memanfaatkan momen terpusat pada satu akun dominan.



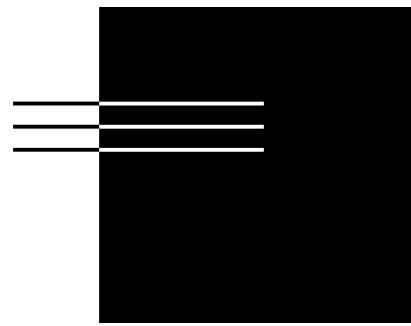
Conclusion & Recommendation

RECOMMENDATION

01 Gunakan LLM untuk Deteksi Otomatis Komentar Spam

Large Language Model (LLM) dapat memahami konteks komentar dengan baik, sehingga mampu membedakan spam dan non-spam secara akurat, termasuk pada komentar dengan kalimat yang samar. Hal ini membantu mengurangi kesalahan deteksi saat memfilter komentar.

Dalam hal ini, IBM WatsonX Granite 3-2b-Instruct dapat menjadi pilihan tepat karena mampu memproses komentar dalam skala besar secara cepat dan konsisten dengan pemahaman kontekstual yang baik. Dengan demikian, IBM WatsonX Granite 3-2b-Instruct dapat menangani masalah spam YouTube secara otomatis yang dapat membantu menjaga kualitas diskusi di kolom komentar tanpa membebani moderasi manual.



Conclusion & Recommendation

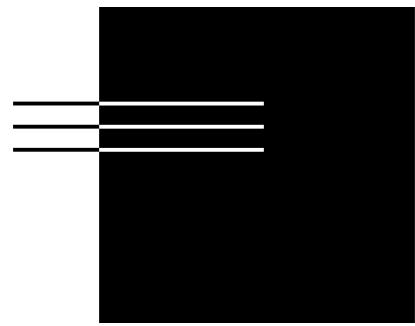
RECOMMENDATION

02 Prioritaskan Moderasi pada Video Populer dan Periode Traffic Rendah

Video populer lebih sering menjadi target spam karena jumlah penontonnya tinggi sementara pada periode traffic rendah, komentar spam cenderung mendominasi. Maka dari itu dengan memprioritaskan moderasi pada video-video ini dan saat periode sepi, upaya pembersihan spam akan lebih efektif dan efisien.

03 Pantau Akun-Akun dengan Intensitas Spam Rendah secara Kolektif

Spam pada YouTube sering tersebar tipis di banyak akun berbeda, bukan terpusat pada satu akun besar. Oleh karena itu, pemantauan perlu dilakukan secara kolektif dengan mendeteksi pola aktivitas spam lintas akun untuk mencegah penyebaran spam secara masif.



Conclusion & Recommendation

RECOMMENDATION

04 Gunakan Kata Kunci Spam sebagai Filter Awal

Berdasarkan hasil analisis, kata seperti “check”, “subscribe”, dan “please” mendominasi konten komentar spam. Kata-kata tersebut dapat digunakan sebagai filter awal untuk menandai komentar mencurigakan sebelum diproses lebih lanjut secara otomatis oleh LLM.

05 Bangun Dashboard Monitoring Tren Spam per Bulan dan Video

Dengan dashboard, pemilik channel dapat memantau tren spam berdasarkan bulan dan video secara visual. Hal ini membantu mendekripsi pola lonjakan komentar spam/periode spam wave dengan cepat sehingga dapat tindakan moderasi dapat segera dilakukan untuk menjaga kualitas diskusi di kolom komentar.

THANK YOU!

FIND ME:

E-mail

azzahmstudio@gmail.com

LinkedIn

Azzah M

PRESENTED BY:

Azzah M

