Social Media:
Great Data, Big Problems

Ashley Y Roccapriore

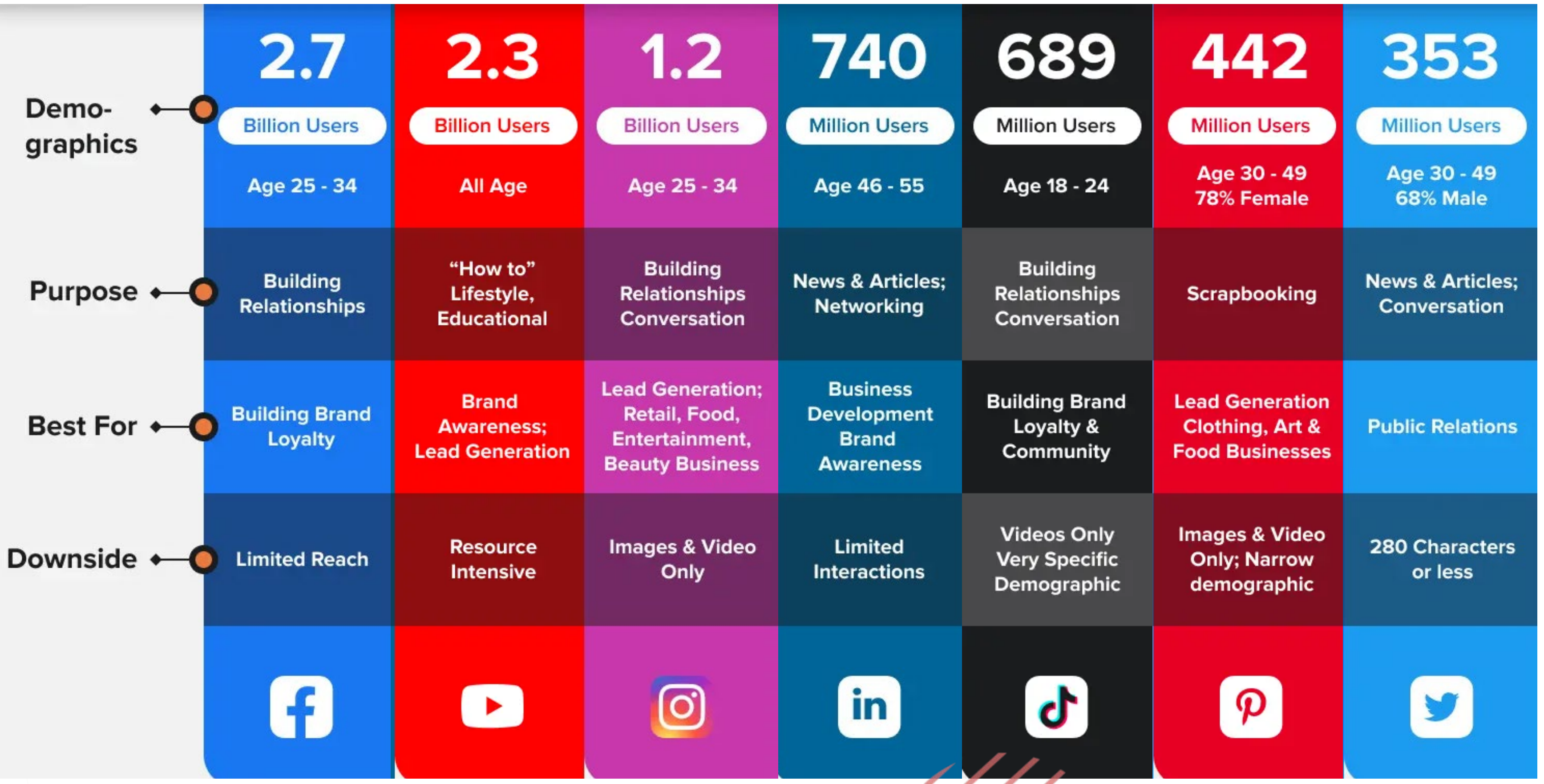
Content and Textual Analysis in Organizational Research PDW: Techniques and Applications (2024)

AUBURN
Harbert College of Business

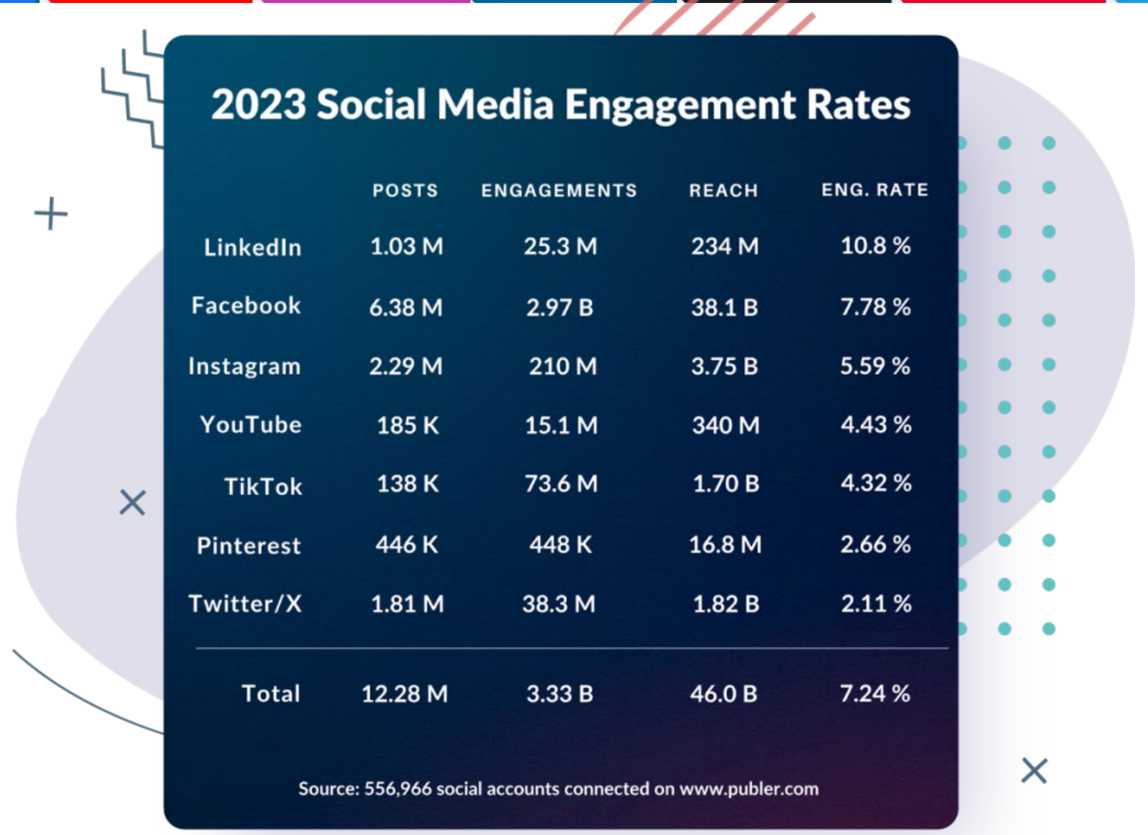
AGENDA

- Social Media Data Types
- How to Get Social Media Data
- Social Media Data Problems
- New Solutions to Data Problems
- Using CATA for Social Media Data

SOCIALMEDIA
DATA TYPES



Demo-graphics	2.7 Billion Users Age 25 - 34	2.3 Billion Users All Age	1.2 Billion Users Age 25 - 34	740 Million Users Age 46 - 55	689 Million Users Age 18 - 24	442 Million Users Age 30 - 49 78% Female	353 Million Users Age 30 - 49 68% Male
Purpose	Building Relationships	"How to" Lifestyle, Educational	Building Relationships Conversation	News & Articles; Networking	Building Relationships Conversation	Scrapbooking	News & Articles; Conversation
Best For	Building Brand Loyalty	Brand Awareness; Lead Generation	Lead Generation; Retail, Food, Entertainment, Beauty Business	Business Development Brand Awareness	Building Brand Loyalty & Community	Lead Generation Clothing, Art & Food Businesses	Public Relations
Downside	Limited Reach	Resource Intensive	Images & Video Only	Limited Interactions	Videos Only Very Specific Demographic	Images & Video Only; Narrow demographic	280 Characters or less



2023 Social Media Engagement Rates

	POSTS	ENGAGEMENTS	REACH	ENG. RATE
LinkedIn	1.03 M	25.3 M	234 M	10.8 %
Facebook	6.38 M	2.97 B	38.1 B	7.78 %
Instagram	2.29 M	210 M	3.75 B	5.59 %
YouTube	185 K	15.1 M	340 M	4.43 %
TikTok	138 K	73.6 M	1.70 B	4.32 %
Pinterest	446 K	448 K	16.8 M	2.66 %
Twitter/X	1.81 M	38.3 M	1.82 B	2.11 %
Total	12.28 M	3.33 B	46.0 B	7.24 %

Source: 556,966 social accounts connected on www.publer.com

GETTING DATA

Data Providers

Provides data after inputting platform credentials

- Free: SocialBlade , Dataverse, Social Media Archive
- Paid: Brightdata , Oxylabs , Sociality

3rd Party APIs

Requires application through platform or 3rd party

- Internal: specific to social media platform
- Providers: CrowdTangle , PhantomBuster

Platform Data Scraping

Typically uses an API, but can also use HTML

- Python: Beautiful Soup, Scrapy, Selenium
- R: rvest, Rselenium

Manual Collection

Most labor intensive and may introduce bias

- Copy/pasting data from websites
- Manually coding specific data (e.g., images)

SOCIALMEDIA DATA PROBLEMS

Hashtags

Duplicated Data

Sponsor Posts

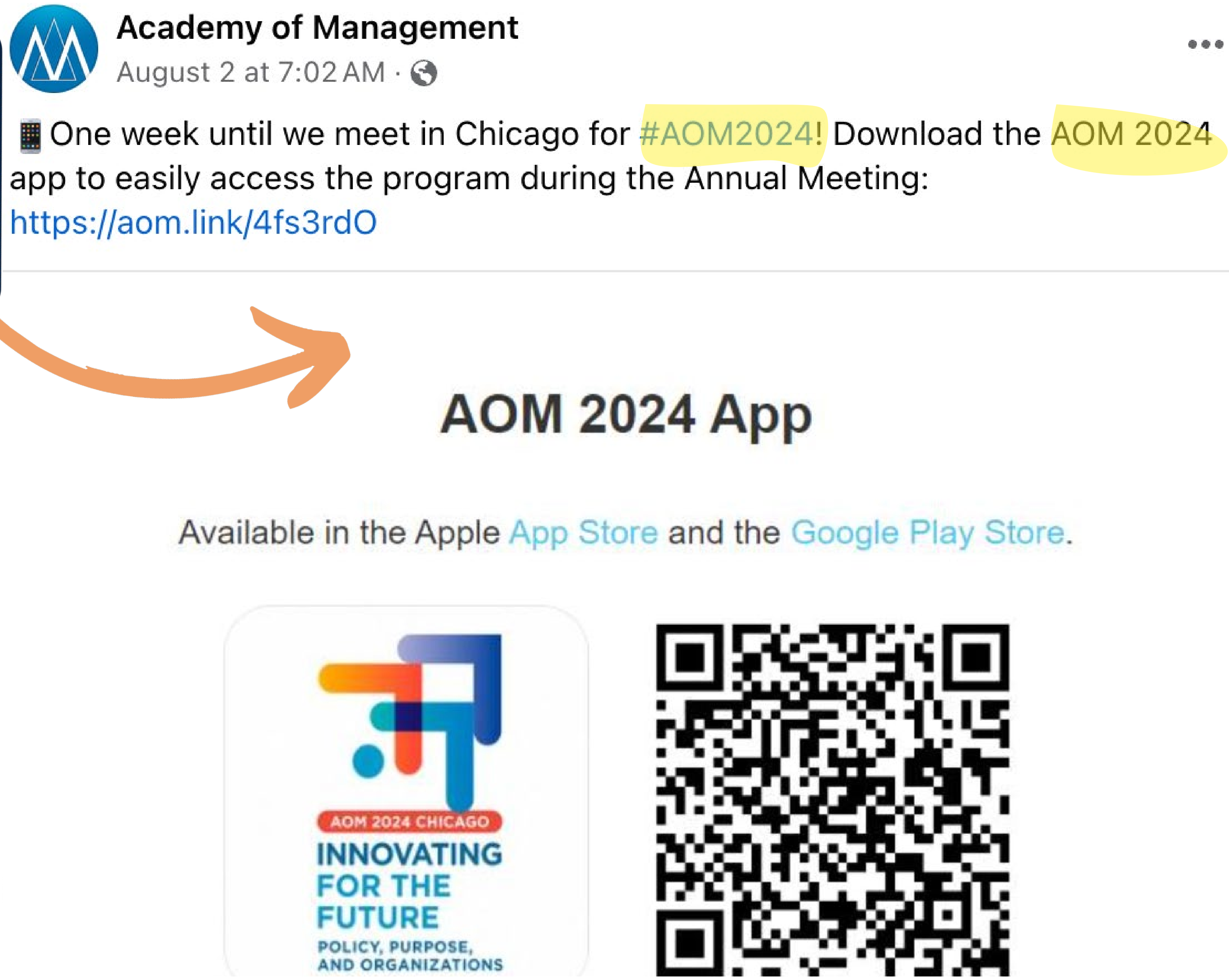
Trolling & Sarcasm

Reposting

Emojis

PROBLEM 1:
Cleaning Data - Hashtags

Hashtags
Emojis
Reposting



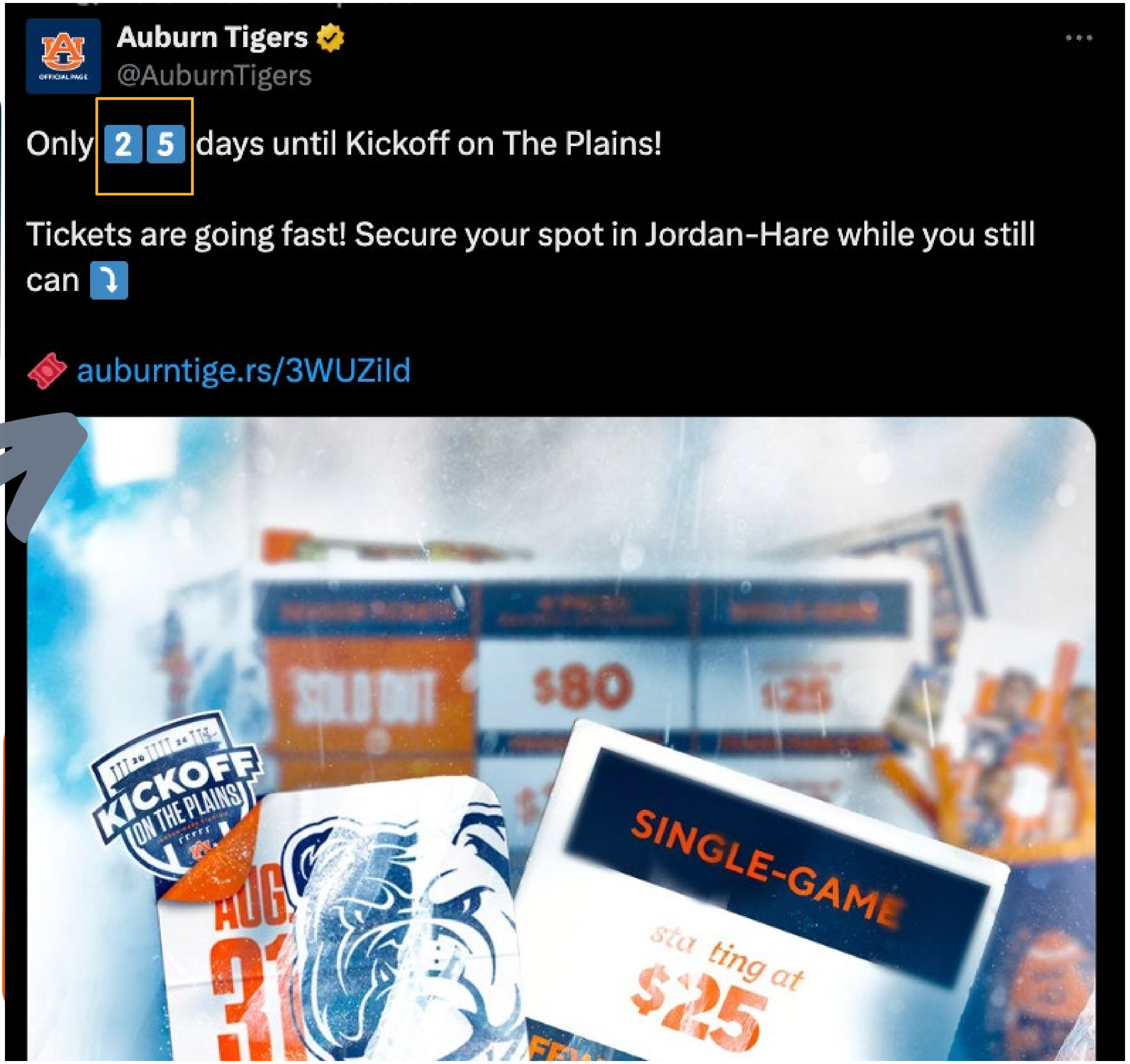
Academy of Management
August 2 at 7:02 AM ·

📱One week until we meet in Chicago for #AOM2024! Download the AOM 2024 app to easily access the program during the Annual Meeting: https://aom.link/4fs3rdO

AOM 2024 App

Available in the Apple App Store and the Google Play Store.

PROBLEM 2:
Cleaning Data - Emojis


Auburn Tigers
@AuburnTigers

Only 2️⃣5️⃣ days until Kickoff on The Plains!

Tickets are going fast! Secure your spot in Jordan-Hare while you still can ⤵️

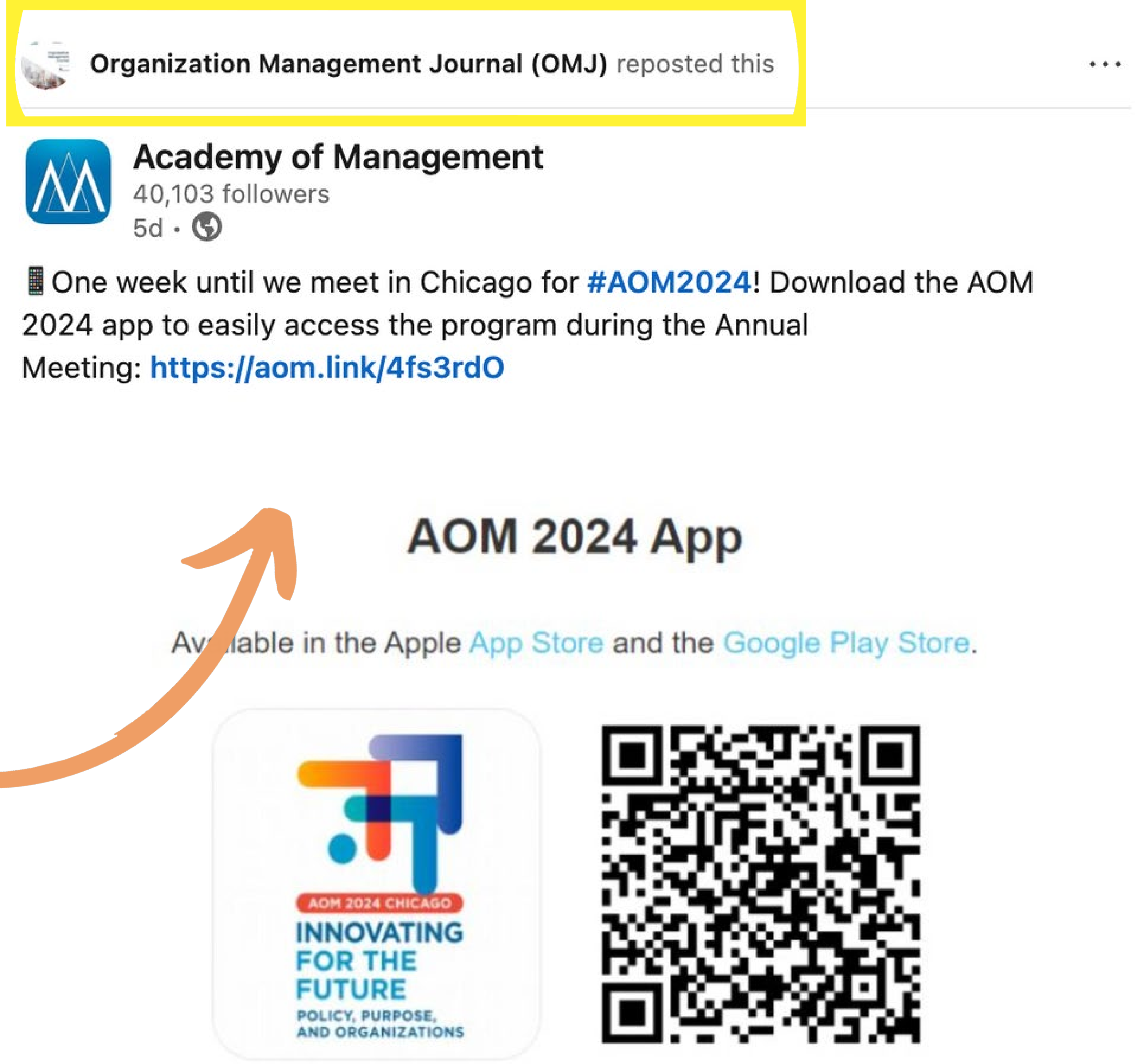
🎟️ auburntige.rs/3WUZIld

PROBLEM 3:
Cleaning Data - Reposting

Hashtags

Emojis

Reposting



Organization Management Journal (OMJ) reposted this

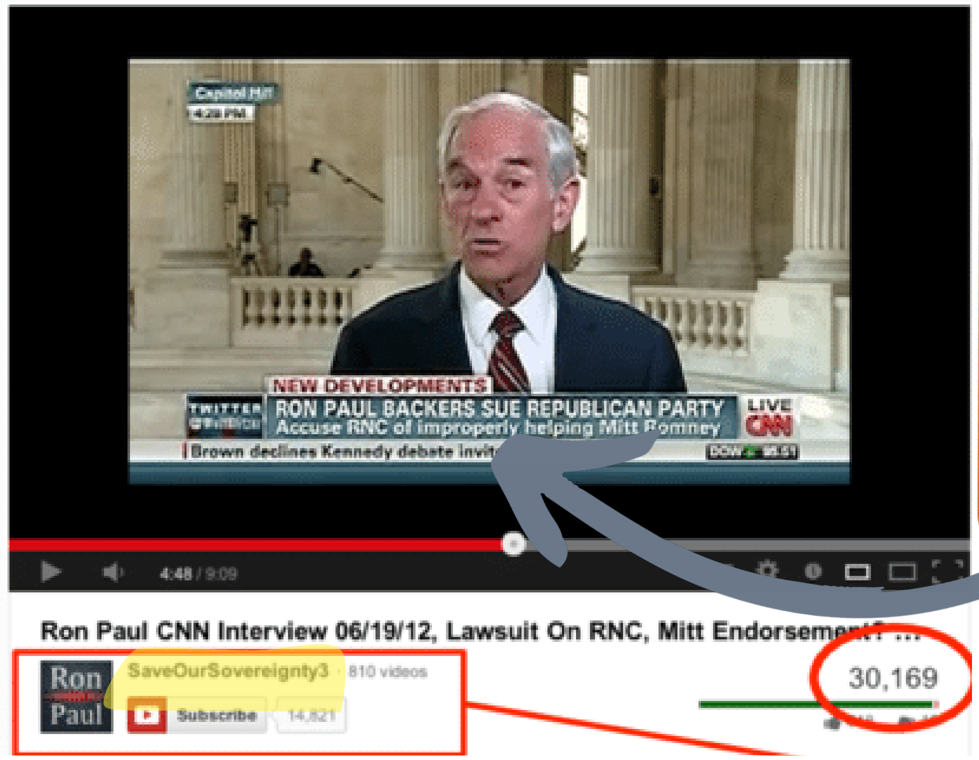
Academy of Management
40,103 followers
5d •

📱One week until we meet in Chicago for #AOM2024! Download the AOM 2024 app to easily access the program during the Annual Meeting: https://aom.link/4fs3rdO

AOM 2024 App

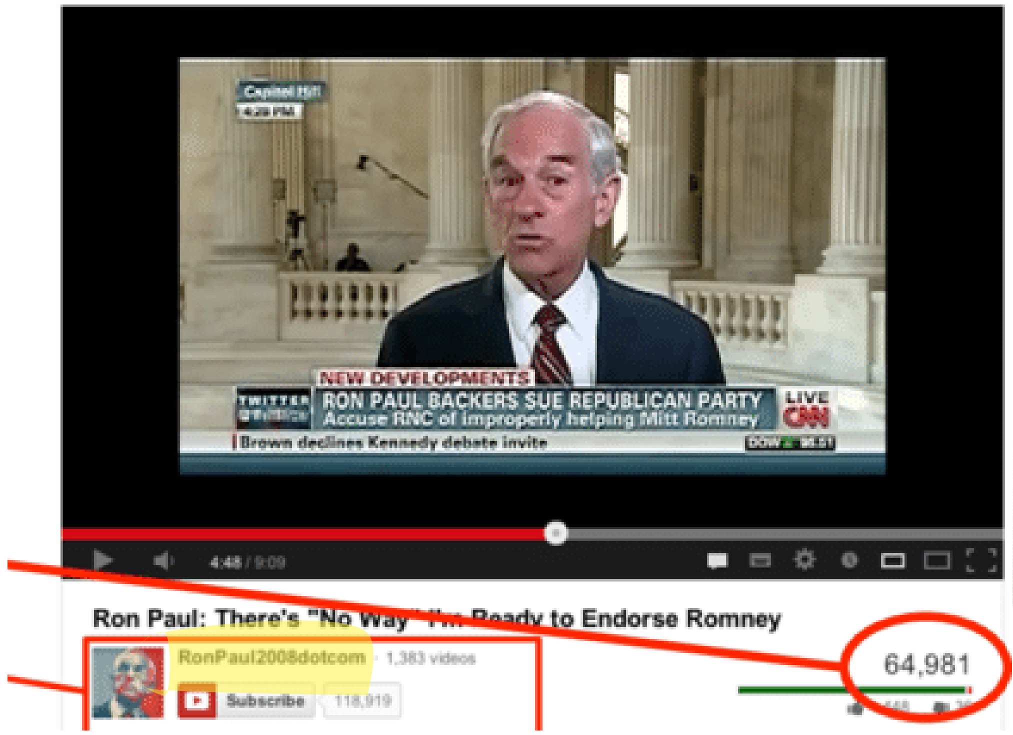
Available in the Apple App Store and the Google Play Store.

PROBLEM 4:
Cleaning Data – Duplicate Data


Capitol Hill
4:29 PM
NEW DEVELOPMENTS
RON PAUL BACKERS SUE REPUBLICAN PARTY
Accuse RNC of improperly helping Mitt Romney
LIVE
CNN
Brown declines Kennedy debate invite
4:48 / 9:09

Ron Paul CNN Interview 06/19/12, Lawsuit On RNC, Mitt Endorsement? ...

SaveOurSovereignty3 · 810 videos
Subscribe 14,821
30,169


Capitol Hill
4:29 PM
NEW DEVELOPMENTS
RON PAUL BACKERS SUE REPUBLICAN PARTY
Accuse RNC of improperly helping Mitt Romney
LIVE
CNN
Brown declines Kennedy debate invite
4:48 / 9:09

Ron Paul: There's "No Way" I'm Ready to Endorse Romney

RonPaul2008dotcom · 1,383 videos
Subscribe 118,919
64,981

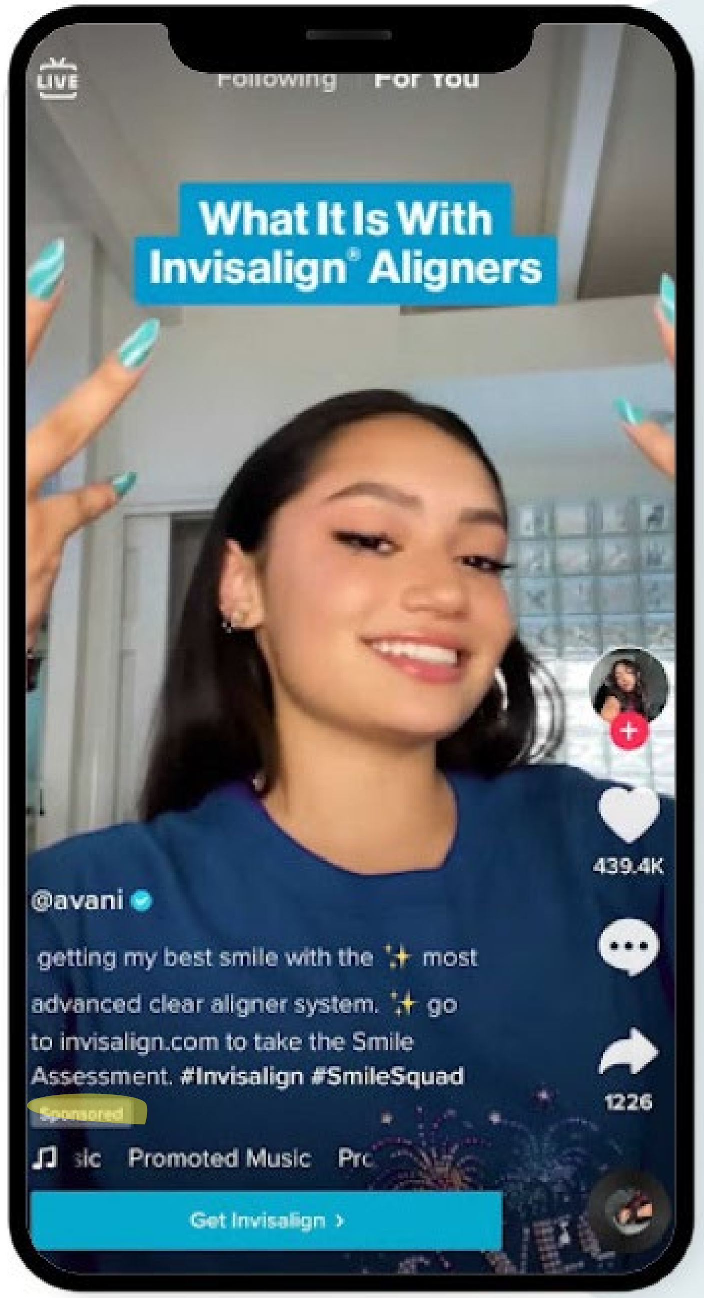
Duplicated Data

Sponsor Posts

Trolling & Sarcasm

PROBLEM 5:

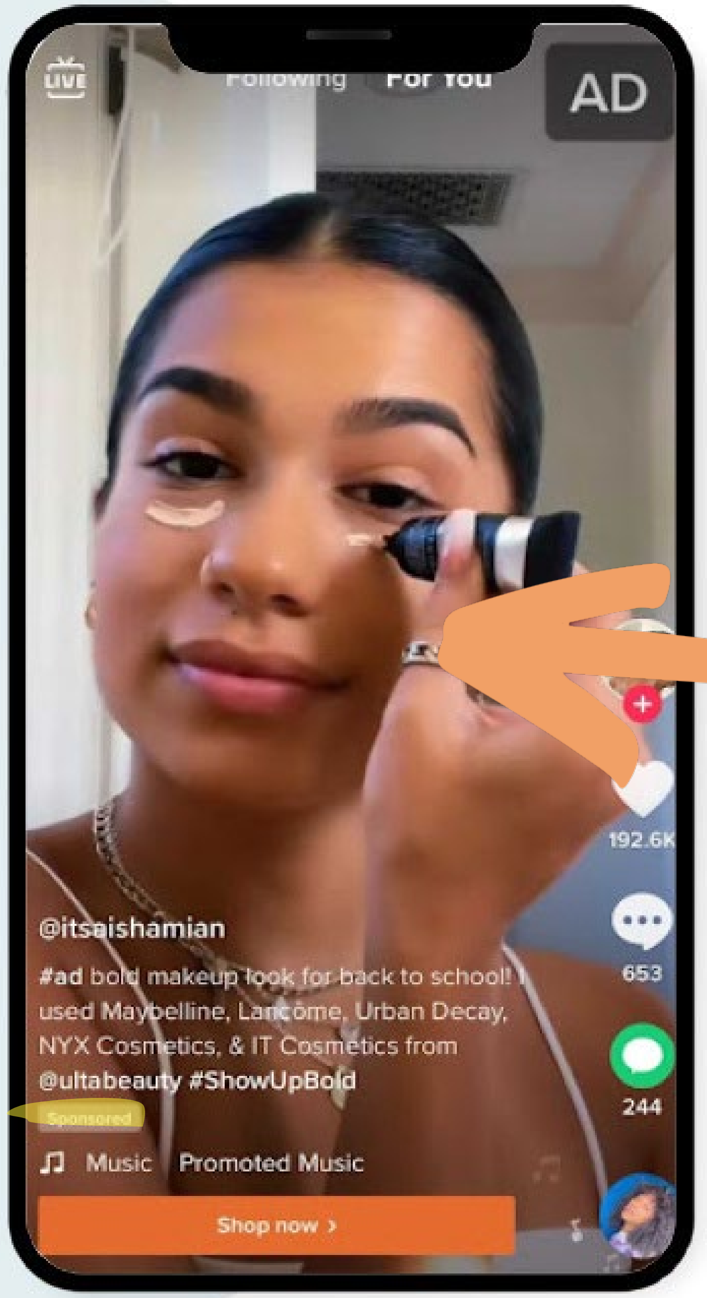
Cleaning Data – Sponsored Posts



@avani

getting my best smile with the ✨ most advanced clear aligner system. ✨ go to invisalign.com to take the Smile Assessment. #Invisalign #SmileSquad

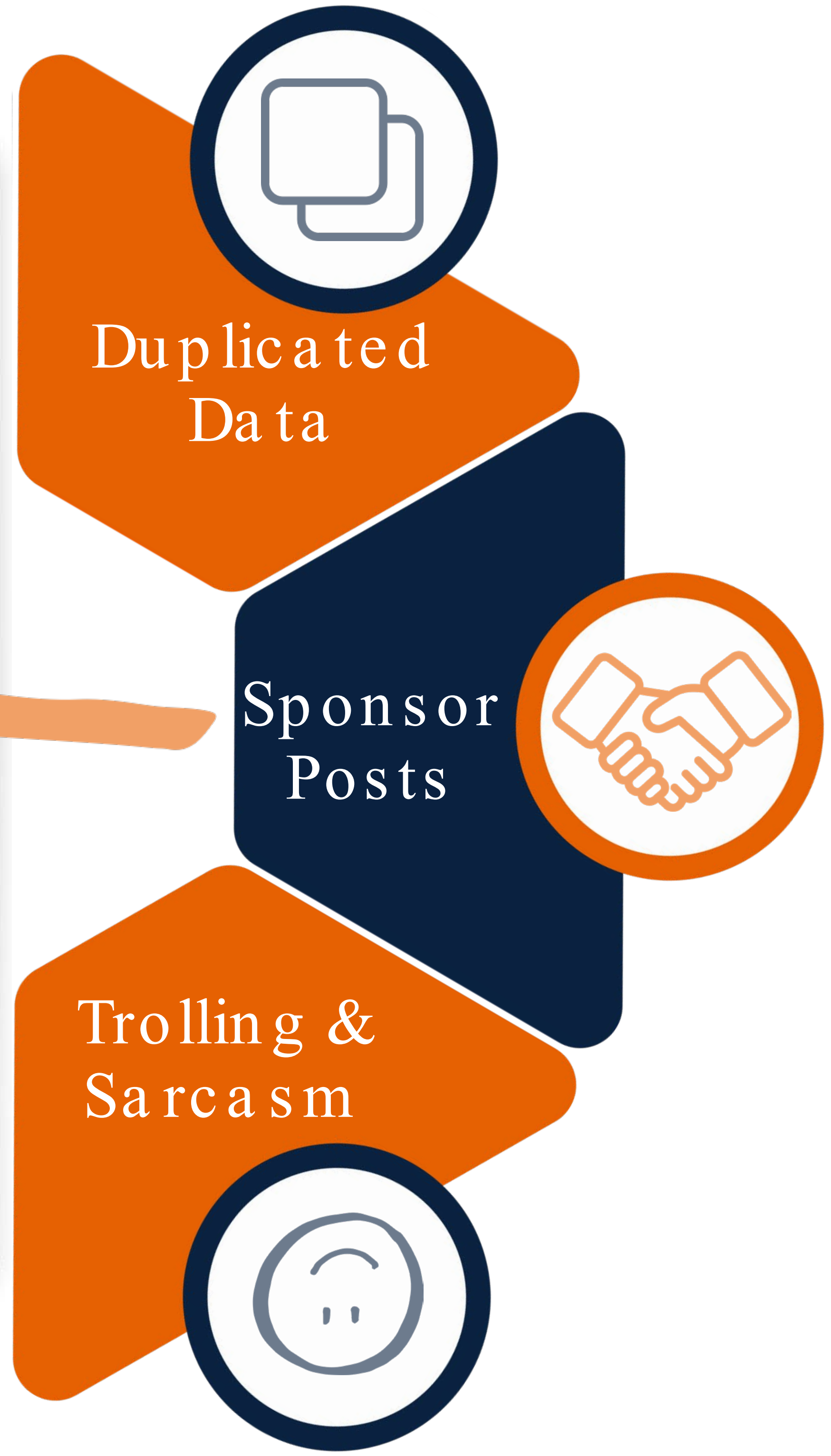
Sponsored



@itsaishamian

#ad bold makeup look for back to school! I used Maybelline, Lancôme, Urban Decay, NYX Cosmetics, & IT Cosmetics from @ultabeauty #ShowUpBold

Sponsored

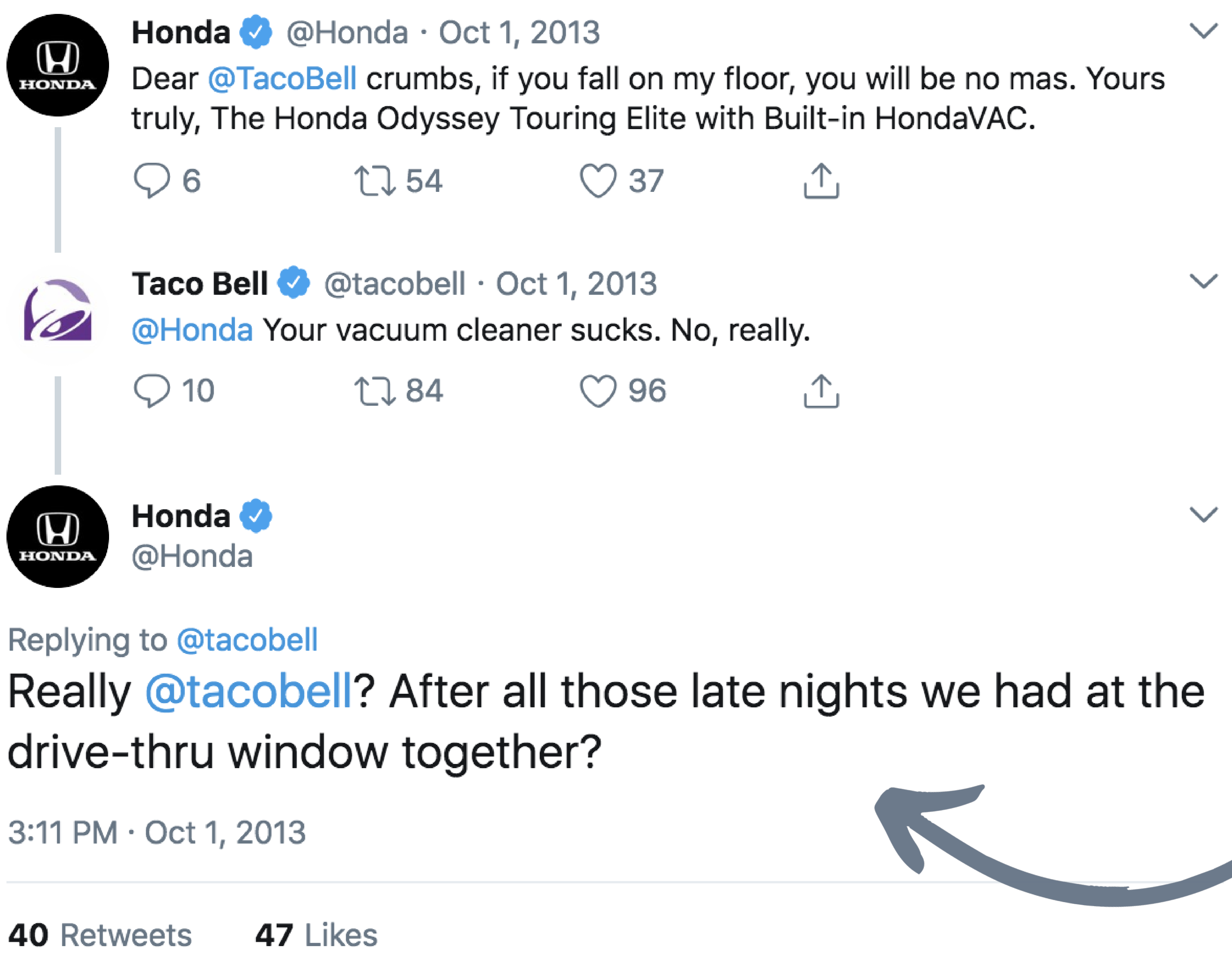


Duplicated Data

Sponsor Posts

Trolling & Sarcasm

PROBLEM 6:
Cleaning Data – Trolling & Sarcasm



Honda @Honda · Oct 1, 2013

Dear @TacoBell crumbs, if you fall on my floor, you will be no mas. Yours truly, The Honda Odyssey Touring Elite with Built-in HondaVAC.

6 54 37

Taco Bell @tacobell · Oct 1, 2013

@Honda Your vacuum cleaner sucks. No, really.

10 84 96

Honda
@Honda

Replying to @tacobell

Really @tacobell? After all those late nights we had at the drive-thru window together?

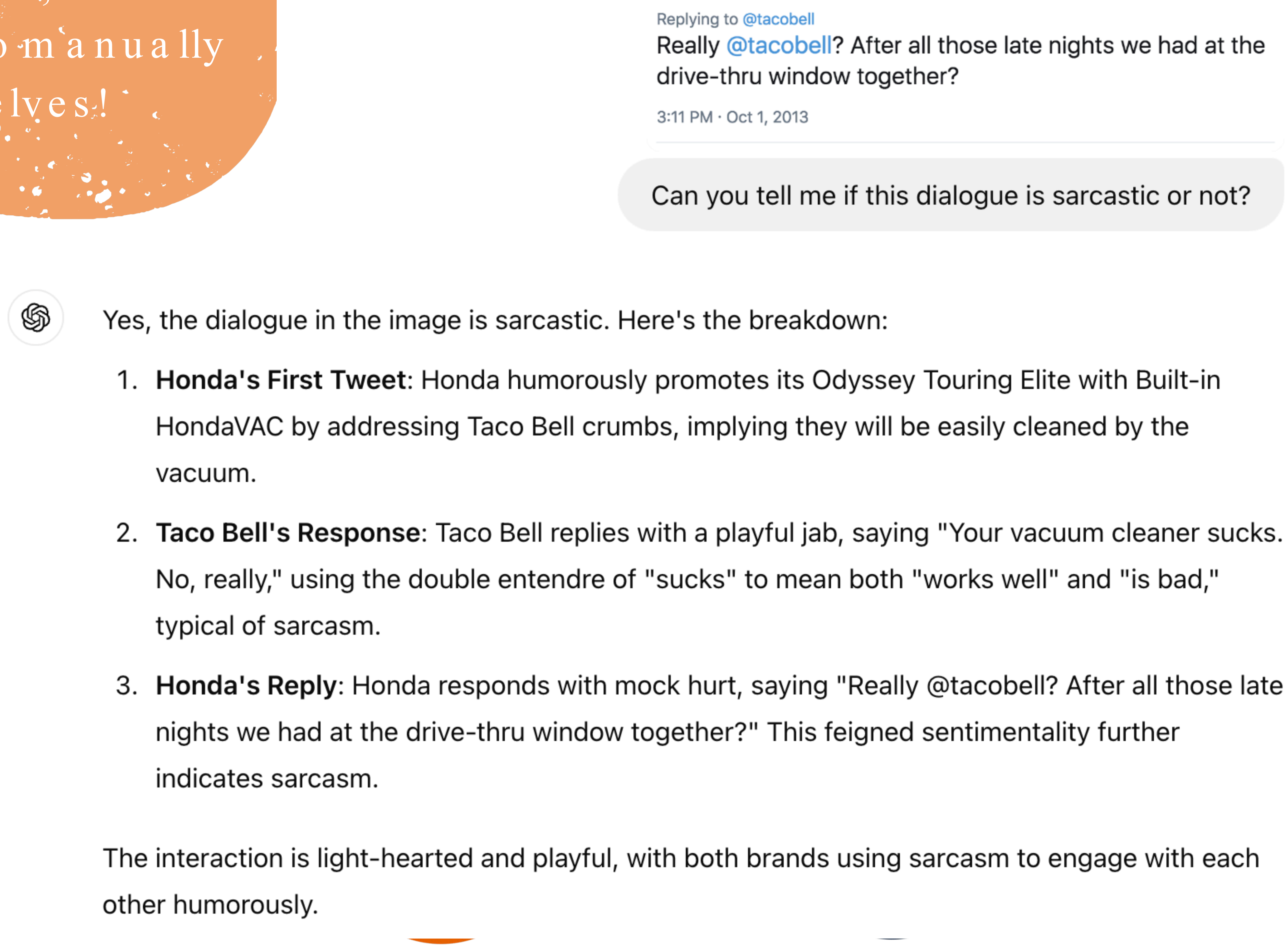
3:11 PM · Oct 1, 2013

40 Retweets **47** Likes

Duplicated Data
Sponsor Posts
Trolling & Sarcasm

NEW SOLUTIONS FOR CLEANING DATA

With new Large Language Models (LLM) many of these things can be addressed prior to using CATA, without the user having to manually do it themselves!



Replying to @tacobell

Really @tacobell? After all those late nights we had at the drive-thru window together?

3:11 PM · Oct 1, 2013

Can you tell me if this dialogue is sarcastic or not?

Yes, the dialogue in the image is sarcastic. Here's the breakdown:

1. **Honda's First Tweet**: Honda humorously promotes its Odyssey Touring Elite with Built-in HondaVAC by addressing Taco Bell crumbs, implying they will be easily cleaned by the vacuum.
2. **Taco Bell's Response**: Taco Bell replies with a playful jab, saying "Your vacuum cleaner sucks. No, really," using the double entendre of "sucks" to mean both "works well" and "is bad," typical of sarcasm.
3. **Honda's Reply**: Honda responds with mock hurt, saying "Really @tacobell? After all those late nights we had at the drive-thru window together?" This feigned sentimentality further indicates sarcasm.

The interaction is light-hearted and playful, with both brands using sarcasm to engage with each other humorously.

Using CATA for Social Media Data

1. Frequent and emotional communication
2. Speed or type of information dissemination
3. Network/ stakeholder dynamics and effects

Abrahamson & Park (1994)
Henry (2008)
Wheeler & Laham (2016)

Brigham et al. (2014)
Hubbard et al. (2018)
McKenny et al. (2018)

Banks et al. (2019)
Moss et al. (2016)
Olsen et al. (2016)

+ Linguistic Inquiry & Word Count (LIWC)

1. Etter, Ravasi, & Colleoni, 2019; Dobele et al., 2007; 2. Veil, Sellnow, & Petrun, 2012; 3. Veil et al., 2012

SUMMARY

Social Media
Data Types

How to Get Social
Media Data

Social Media
Data Problems

New Solutions to
Data Problems

Using CATA for
Social Media Data

Solutions to these problems can lead to more robust content and textual analysis using unique data!

Thank You!

Questions?

CONNECT:

AUBURN
Harbert College of Business

APPENDIX

PROBLEM 2:
ANALYZING DATA

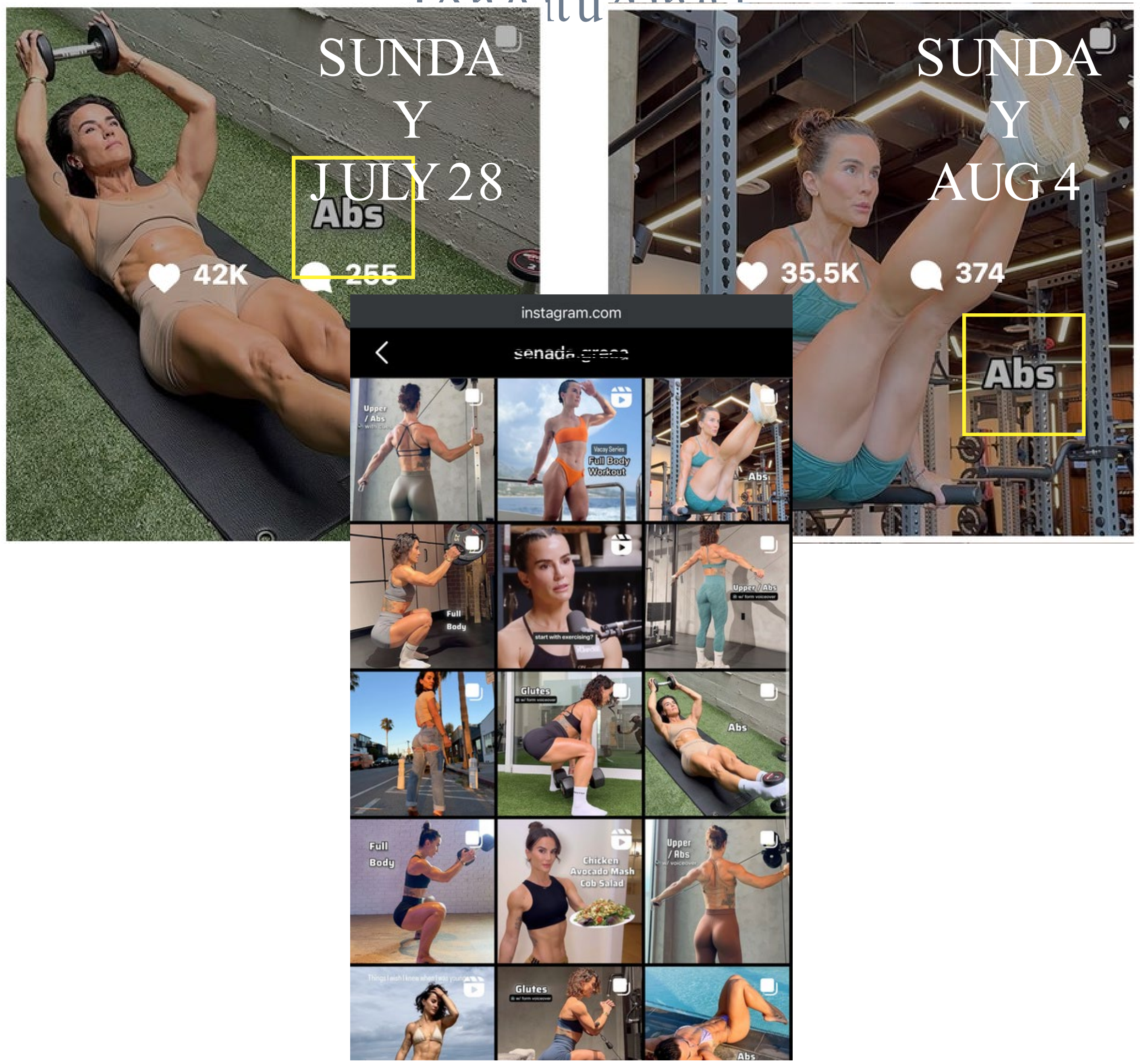
Data Types

Algorithms

Data Problems

PROBLEM 2:
ANALYZING DATA

Cross-Sectional/ Longitudinal

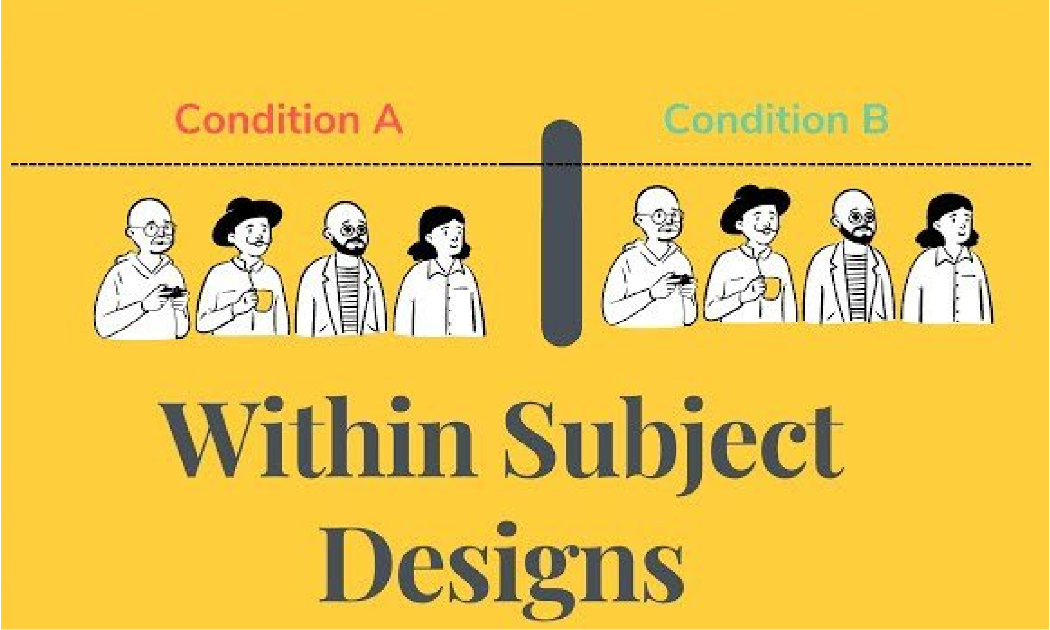

SUNDAY JULY 28

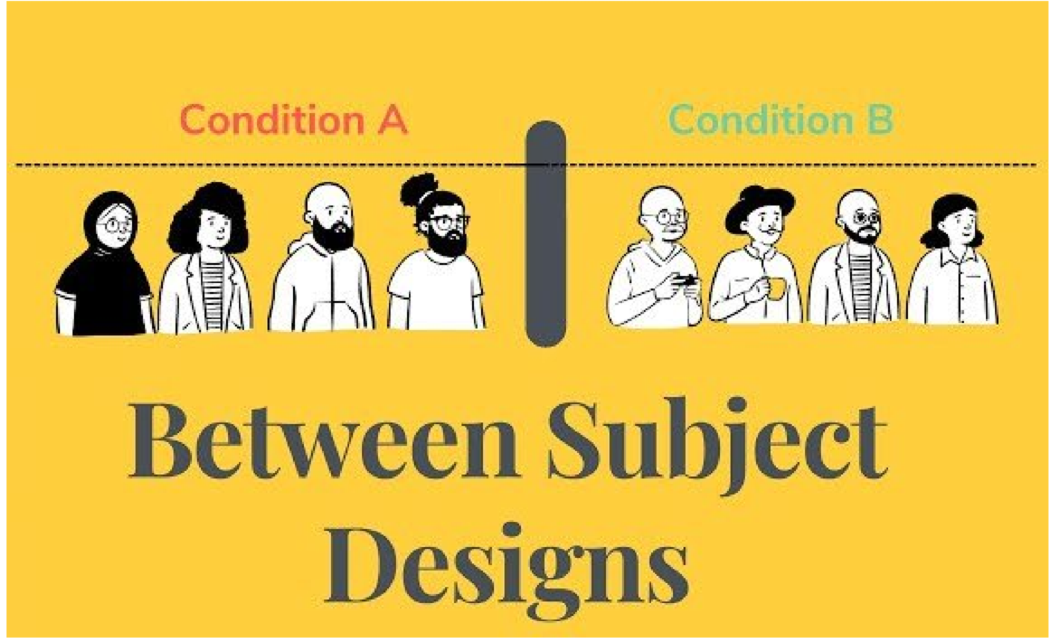
SUNDAY AUG 4

Data Types

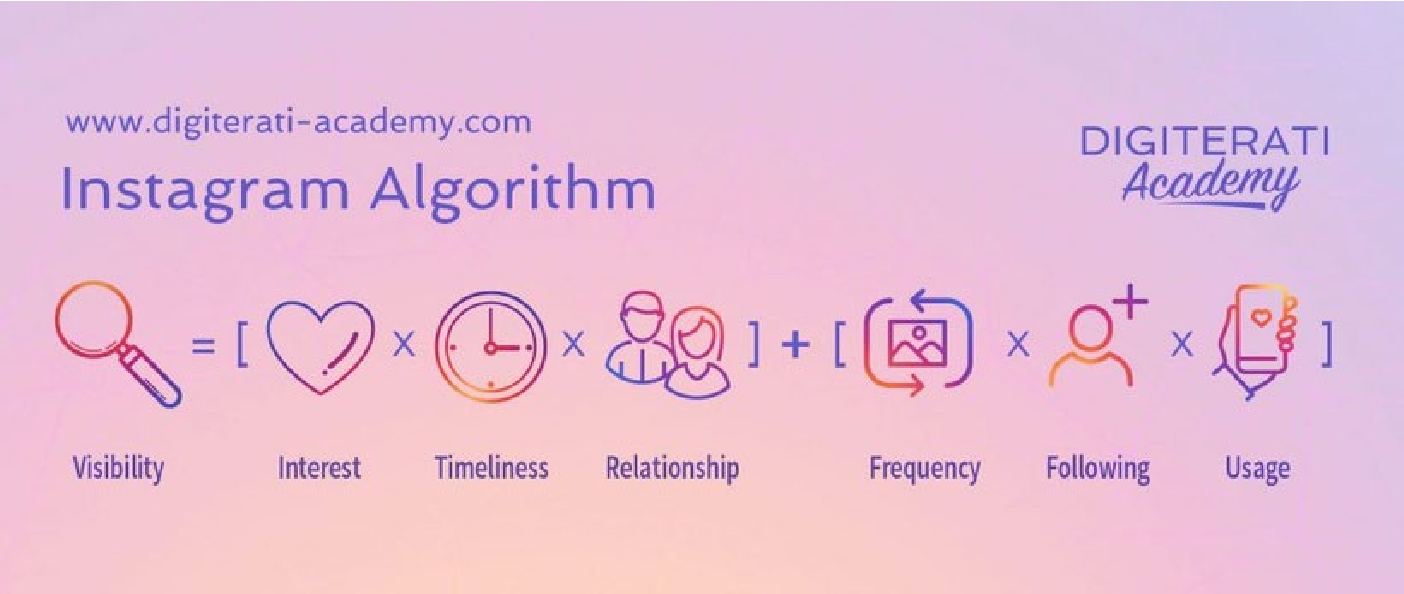
Differences depend on the research question being addressed and the specific industry or context being explored.

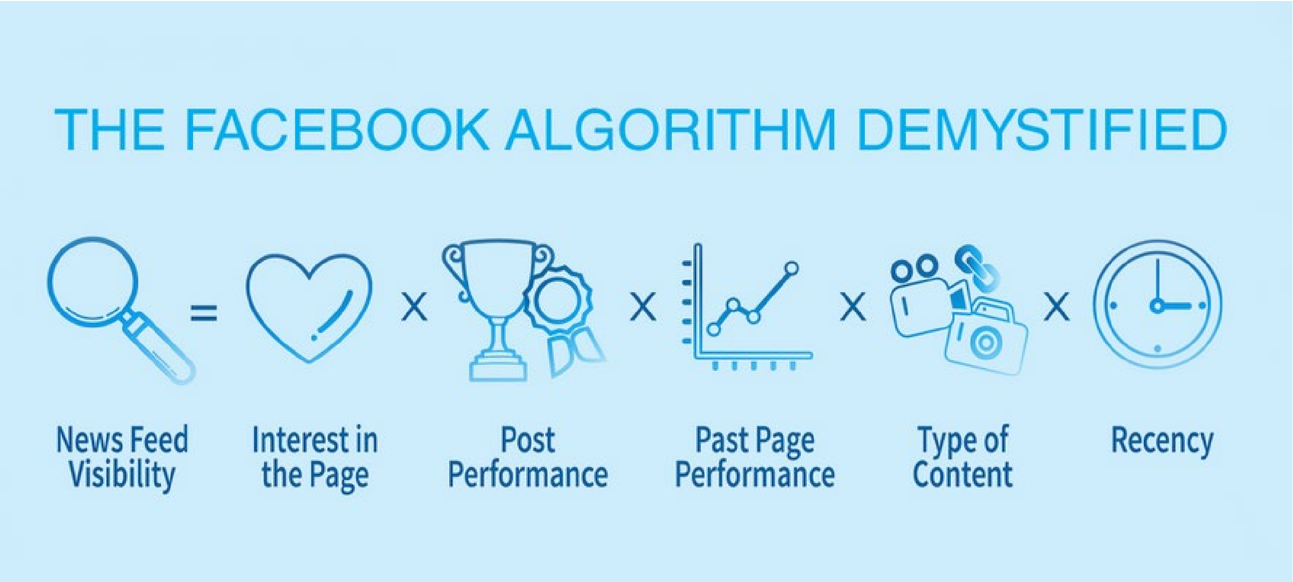
Fixed or Random Effects



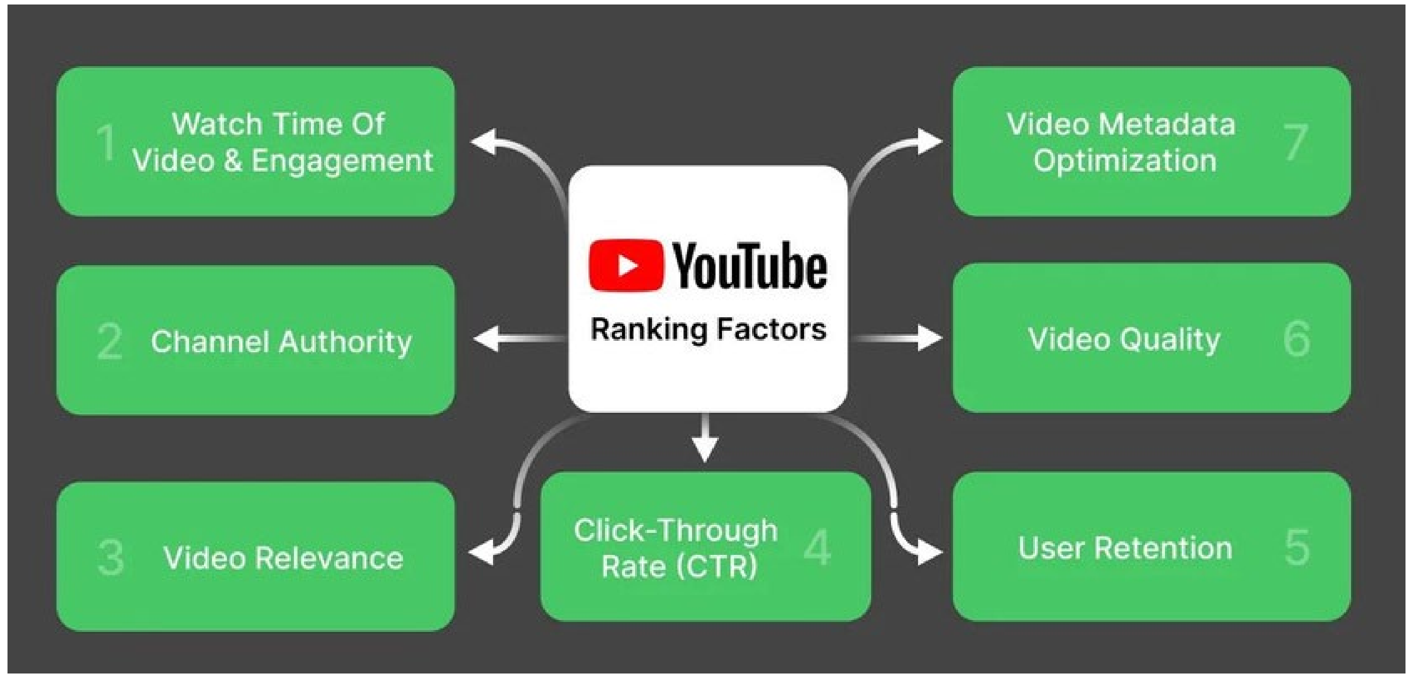


PROBLEM 2:
ANALYZING DATA



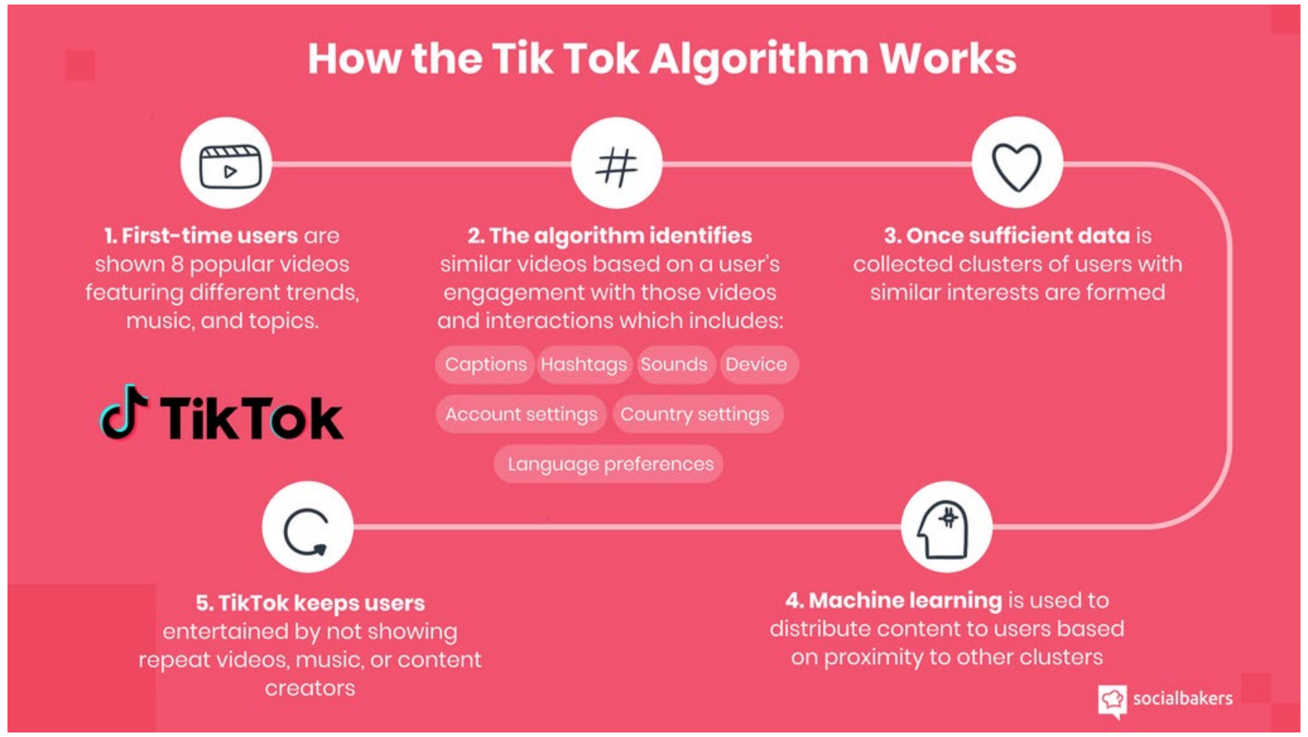


Algorithms


YouTube
Ranking Factors
1 Watch Time Of Video & Engagement
2 Channel Authority
3 Video Relevance
Click-Through Rate (CTR) 4
User Retention 5
Video Quality 6
Video Metadata Optimization 7

Different analysis methods can account for unobserved variables like the algorithm.

Some examples:
Hausman Taylor
2SLS
Latent Variables



1. First-time users are shown 8 popular videos featuring different trends, music, and topics.

2. The algorithm identifies similar videos based on a user's engagement with those videos and interactions which includes:

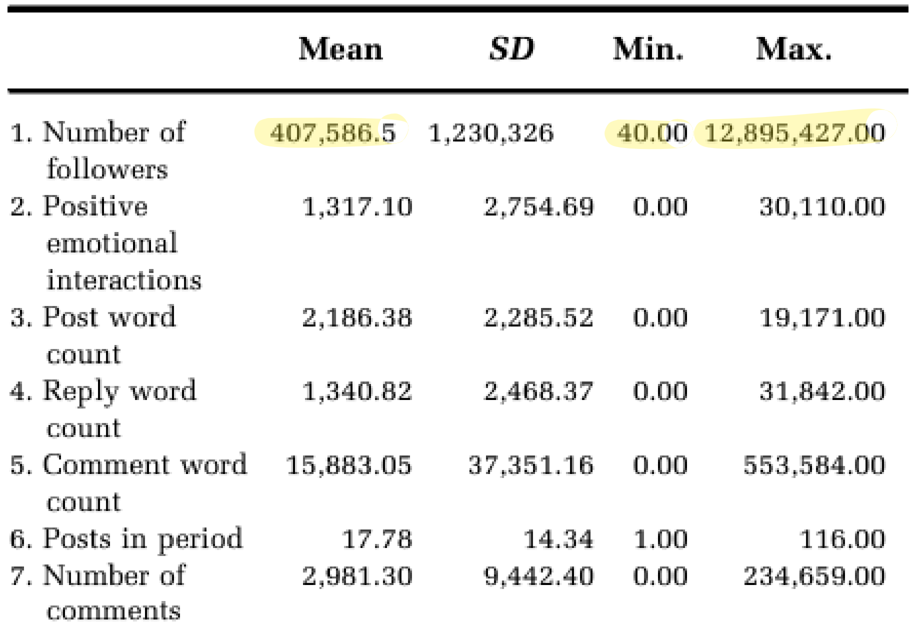
3. Once sufficient data is collected clusters of users with similar interests are formed

4. Machine learning is used to distribute content to users based on proximity to other clusters

5. TikTok keeps users entertained by not showing repeat videos, music, or content creators

PROBLEM 2:
ANALYZING DATA

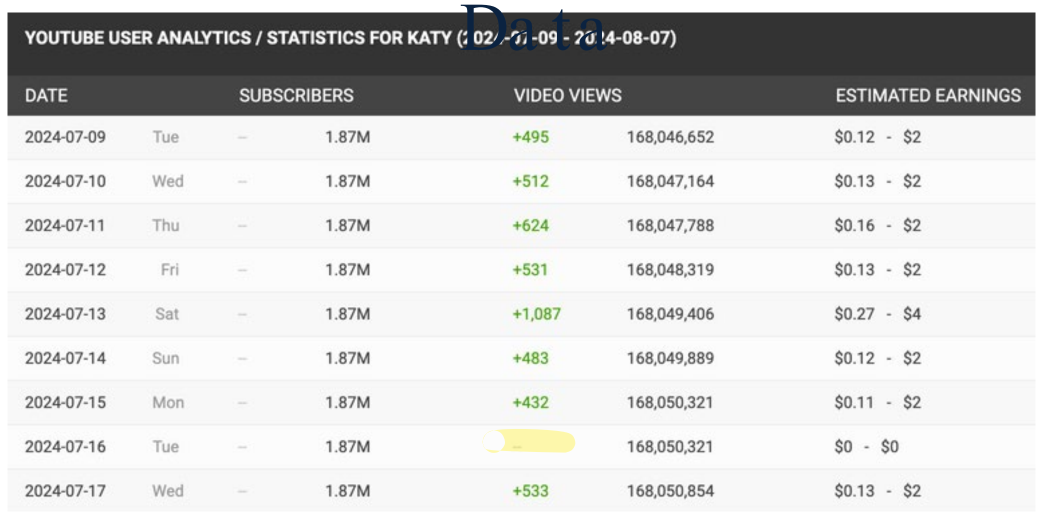
Outliers



	Mean	*SD*	Min.	Max.
1. Number of followers	407,586.5	1,230,326	40.00	12,895,427.00
2. Positive emotional interactions	1,317.10	2,754.69	0.00	30,110.00
3. Post word count	2,186.38	2,285.52	0.00	19,171.00
4. Reply word count	1,340.82	2,468.37	0.00	31,842.00
5. Comment word count	15,883.05	37,351.16	0.00	553,584.00
6. Posts in period	17.78	14.34	1.00	116.00
7. Number of comments	2,981.30	9,442.40	0.00	234,659.00

Roccapriore & Pollock, 2023

Missing Data



YOUTUBE USER ANALYTICS / STATISTICS FOR KATY (2024-07-09 - 2024-08-07)

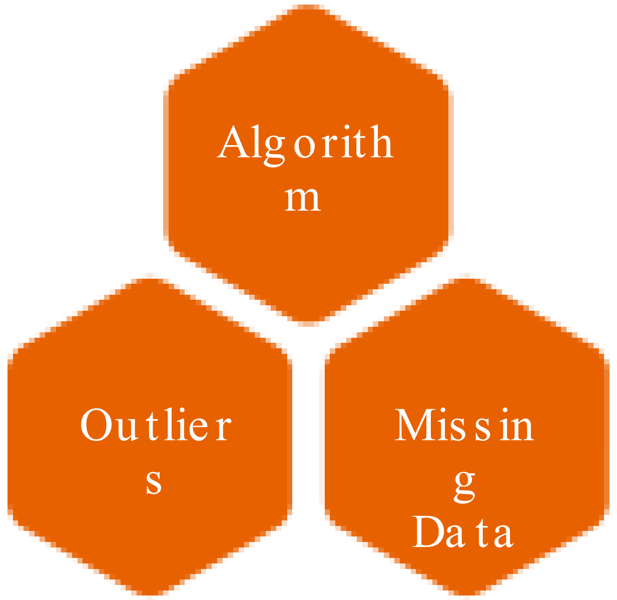
DATE		SUBSCRIBERS		VIDEO VIEWS		ESTIMATED EARNINGS
2024-07-09	Tue	–	1.87M	+495	168,046,652	$0.12 - $2
2024-07-10	Wed	–	1.87M	+512	168,047,164	$0.13 - $2
2024-07-11	Thu	–	1.87M	+624	168,047,788	$0.16 - $2
2024-07-12	Fri	–	1.87M	+531	168,048,319	$0.13 - $2
2024-07-13	Sat	–	1.87M	+1,087	168,049,406	$0.27 - $4
2024-07-14	Sun	–	1.87M	+483	168,049,889	$0.12 - $2
2024-07-15	Mon	–	1.87M	+432	168,050,321	$0.11 - $2
2024-07-16	Tue	–	1.87M	–	168,050,321	$0 - $0
2024-07-17	Wed	–	1.87M	+533	168,050,854	$0.13 - $2

Data Problems

Different robustness tests can allow that your data will not change from these data problems.

Some examples:
Winsorizing
Variable mean between periods
ITCV or RIR

Engogeneity



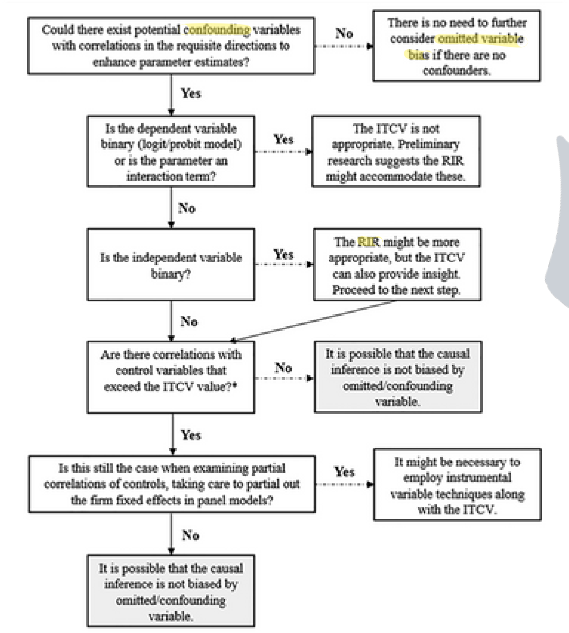


Figure 3
Decision Tree of When to Employ the ITCV

Could there exist potential confounding variables with correlations in the requisite directions to enhance parameter estimates?
No → There is no need to further consider omitted variable bias if there are no confounders.
Yes ↓
Is the dependent variable binary (logit/probit model) or is the parameter an interaction term?
Yes → The ITCV is not appropriate. Preliminary research suggests the RIR might accommodate these.
No ↓
Is the independent variable binary?
Yes → The RIR might be more appropriate, but the ITCV can also provide insight. Proceed to the next step.
No ↓
Are there correlations with control variables that exceed the ITCV value?*
No → It is possible that the causal inference is not biased by omitted/confounding variable.
Yes ↓
Is this still the case when examining partial correlations of controls, taking care to partial out the firm fixed effects in panel models?
Yes → It might be necessary to employ instrumental variable techniques along with the ITCV.
No ↓
It is possible that the causal inference is not biased by omitted/confounding variable.

Note. *It is important to reiterate that the ITCV represents the square root of the product of correlations between a potential omitted variable and both the independent and dependent variables. This is therefore the case when examining control variables as potential proxies for an omitted variable. Specifically, it is essential to compare the square root corr [control, y] times corr [control, x] against the ultimate ITCV value.

Busenbark et al., 2022