Influencer Marketing

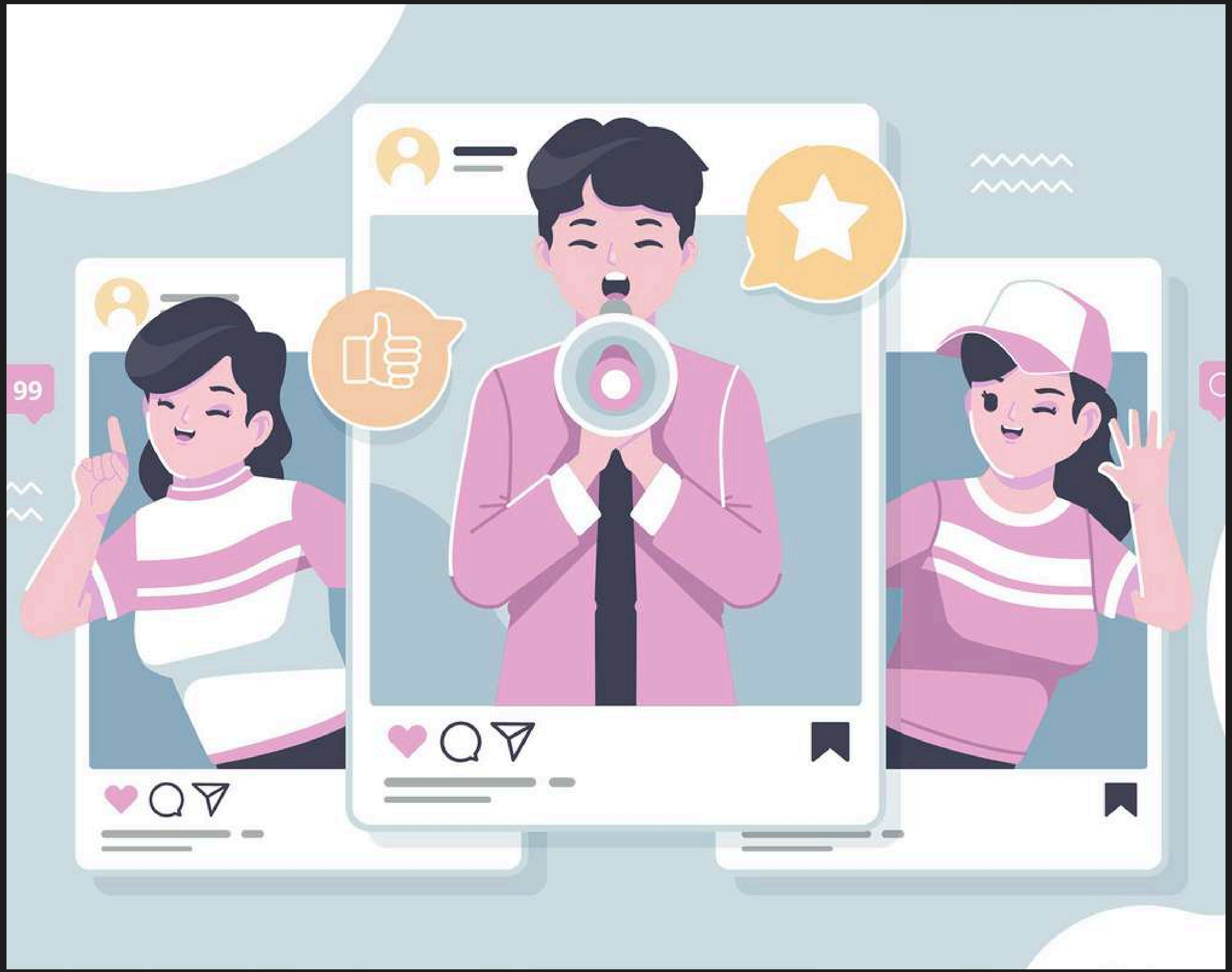
ML-BASED INFLUENCER RECOMMENDER SYSTEM

GROUP 14 - ADITYA, RUHI, DAKSH

“SECTION 1”

What is our AIM?

We aim to build a machine learning-powered recommender system to help brands find the most relevant influencers based on bio similarity, engagement, and past sponsorship data.



Problem Statement

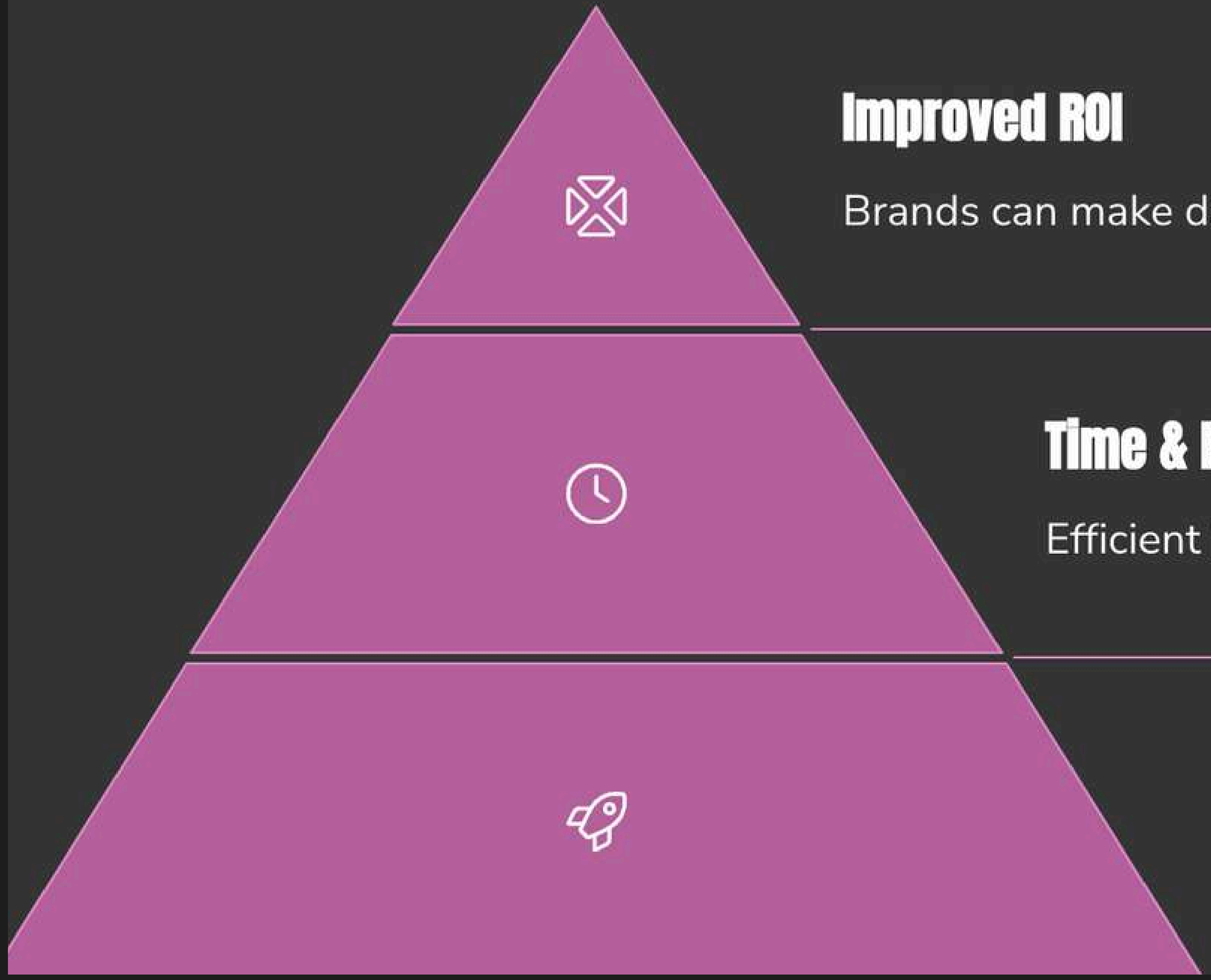
WHY?

Over 70% of brands in India are currently investing in influencer marketing .

Influencer marketing is powerful but inefficient.
Brands struggle to find the right influencer from thousands of options.

Simple metrics (like follower count) aren't enough to guarantee campaign success.
Manual selection is time-consuming and often ineffective.

This project boosts cost-effective marketing by identifying high-impact influencers and detecting fake engagement. It empowers small businesses and refines brand messaging using NLP-driven social media analysis.



Improved ROI

Brands can make data-driven decisions

Time & Bias Reduction

Efficient influencer discovery

Startup Support

Helps **startups** and **new brands** connect with relevant micro-influencers quickly

- **71% of Users more likely to purchase a given product if recommended by the right influencer**
- **This solution is scalable and can be deployed across various industries using influencer marketing.**

Metric	Insight	Source
ROI Boost	Up to $5.78 return for every $1 spent with the right influencer.	Influencer Marketing Hub (2023)
Cost Savings	Relevant micro/nano influencers can cut campaign costs by 30–40%.	Later x Fohr (2023)
Engagement	Well-matched influencers deliver 2x higher engagement rates.	SocialPubli (2022)
Conversion Rate	Proper alignment drives 43% more conversions than untargeted outreach.	Kantar x CreatorIQ (2022)

“SECTION 2”

Literature Review

How prior work tackles influencer selection?

Prior research has applied statistical regression, clustering, deep NLP, multi-task learning, and ML to identify and rank influencers. We build on these by combining lightweight transformers, CatBoost being our gradient-boosted trees, and audience sentiment into a production-ready pipeline.

Modeling Influencer Marketing Campaigns in Social Networks

Developed Solution:

- Campaign Simulation Model
- Create a fake social network
- Set parameters (category, budget, engagement)
- Allow hiring to happen
- Simulate a campaign
- Measure the outcome

Shortcomings:

- No semantic alignment between brand and influencer
- Lack of audience sentiment analysis.
- No predictive performance metrics on real-world data, instead of training on past data, they built a "what-if" model that works on assumptions (category, engagement,etc)

Our solution predicts which influencers will most likely drive sponsorship success - grounded in semantic, sentiment, and engagement signals.

Modeling Influencer Marketing Campaigns in Social Networks

Ronak Doshi*, Ajay Ramesh*, *Student Member, IEEE,* and Shrisha Rao, *Senior Member, IEEE*

***Abstract*—Social media are extensively used in today's world, and facilitate quick and easy sharing of information, which makes them a good way to advertise products. Influencers of a social media network, owing to their massive popularity, provide a huge potential customer base. However, it is not straightforward to decide which influencers should be selected for an advertizing campaign that can generate high returns with low investment. In this work, we present an agent-based model (ABM) that can simulate the dynamics of influencer advertizing campaigns in a variety of scenarios and can help to discover the best influencer marketing strategy. Our system is a probabilistic graph-based model that provides the additional advantage to incorporate real-world factors such as customers'**

Doshi, R., Ramesh, A., & Rao, S. (2022). Modeling influencer marketing campaigns in social networks (arXiv:2106.01750v3). arXiv. https://doi.org/10.48550/arXiv.2106.01750

Enhancing Influencer Marketing Strategies

Developed Solution:

- Takes influencer and brand features along with sponsorship history
- Data is cleaned and encoded
- Predicts sponsorship success based on sponsorships that already took place.

Shortcomings:

- Only caters to large influencers
- Lack of sentiment analysis
- Does not account for new influencers or brands

These gaps motivate our pipeline's use of explicit semantic and sentiment features, a transparent CatBoost model for clear feature importances, and a design that scales efficiently to larger, more diverse influencer populations.

Rivera, O. (2022). Enhancing influencer marketing strategies through machine learning: Predictive analysis of influencer-generated interactions (Master's thesis, KTH Royal Institute of Technology). Retrieved from https://kth.diva-portal.org/smash/get/diva2%3A1783645/FULLTEXT01.pdf

Enhancing Influencer Marketing Strategies through Machine Learning

Predictive Analysis of Influencer-Generated Interactions

OLIMPIA RIVERA

Date: July 1, 2023

Supervisor: Haibo Li
Examiner: Anders Hedman
School of Electrical Engineering and Computer Science

Swedish title: Förbättra Marknadsföringsstrategier Genom Maskininlärning

The field of influencer marketing has experienced rapid growth in recent years. However, uncovering the true effectiveness of this marketing approach remains a significant challenge. This thesis addresses the challenge of predicting the effectiveness of influencer marketing campaigns by employing advanced machine learning techniques, specifically the Auto Machine Learning framework Autogluon. With the aim of democratizing machine learning and empowering businesses in the influencer marketing domain, this work leverages Autogluon to predict the interactions generated by influencers when posting affiliate links. By evaluating various settings of AutoGluon and assessing the performance using metrics such as R-squared score, we observed promising results with good predictive accuracy. The findings from our study contribute to critical discussions in the field. This research offers a streamlined

Using Machine Learning to Connect Brands with Influencers

Developed Solution:

- Looks at all previous “successful” brand–influencer pairings
- Analyzes patterns in those past collaborations
- Clusters influencers from similar collaborations together
- A brand fills out a three-step form (budget, category, reach)
- Model picks an influencer that falls in the matching cluster and recommends them

Shortcomings:

- No test set or quantitative metrics to validate clustering quality or recommendation precision.
- Unsupervised clustering can lead to groupings without clear labels so we dont really know that new recommendations truly match brand needs or not.
- It doesn’t address each brand’s unique category or needs
- Lack of audience sentiment analysis.

We frame it as a supervised CatBoost classifier with clear evaluation metrics, expose feature importances for explainabiliy.

Master of Science and Engineering in Interaction Technology and Design

Using Machine Learning to Connect Brands with Influencers

by Jonathan Hedlund

Master Thesis

February 9, 2022

Fall 2021
Master Thesis, 30 ECTS
Master of Science in Interaction Technology and Design, 300 ECTS

Examiner – Thomas Mejtoft
Supervisor at UmU – Kalle Prorok

Hedlund, J. (2021). Using machine learning to connect brands with influencers (Master’s thesis, Umeå University). Retrieved from https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1636469

Ranking Micro-Influencers

Developed Solution:

- Given a brand's profile, which micro-influencers should I partner with?
- Learns what makes a good match by looking at both their post captions and images.
- Gives a ranked list of - most likely to fit the brand.

Shortcomings:

- It only counts likes/comments, but never looks at whether those reactions were positive or negative.
- Does not cosider an influener or brand's bio
- Only trained for small influencers.

Performance Metrics:

- Recall@10 ≥ 0.19
- Recall@50 = 0.55

Our model adds sentiment scores and deep text understanding, and uses CatBoost classifier to give clear influencer matches.

Ranking Micro-Influencers: a Novel Multi-Task Learning and Interpretable Framework

Adam Elwood
lastminute.com
Chiasso, Switzerland
adam.elwood@lastminute.com

Alberto Gasparin
lastminute.com
Chiasso, Switzerland
alberto.gasparin@lastminute.com

Alessandro Rozza
lastminute.com
Chiasso, Switzerland
alessandro.rozza@lastminute.com

***Abstract*—With the rise in use of social media to promote branded products, the demand for effective influencer marketing has increased. Brands are looking for improved ways to identify valuable influencers among a vast catalogue; this is even more challenging with "micro-influencers", which are more affordable than mainstream ones but difficult to discover. In this paper, we propose a novel multi-task learning framework to improve the state of the art in micro-influencer ranking based on multimedia content. Moreover, since the visual congruence between a brand and influencer has been shown to be good measure of compatibility, we provide an effective visual method for interpreting our models' decisions, which can also be used to inform brands' media strategies. We compare with the current state-of-the-art on a recently constructed public dataset and we show significant improvement both in terms of accuracy and model complexity. The techniques for ranking and interpretation presented in this work can be generalised to arbitrary multimedia ranking tasks**

Elwood, A., Gasparin, A., & Rozza, A. (2021). Ranking micro-influencers: A novel multi-task learning and interpretable framework (arXiv:2107.13943v1). arXiv. https://doi.org/10.48550/arXiv.2107.13943

Our Approach model pipeline

Data Collection

Publicly available data of 25,282 brands and 38,113 influencers and 16,01,074 posts.

Pre-processing

Handle missing data, Standardize features, Merge datasets, Compute metrics

Feature Engineering

Engagement metrics, Influencer Sentiment, Brand & influencer categories, Bio similarity, Comment Sentiments.

Model Prediction

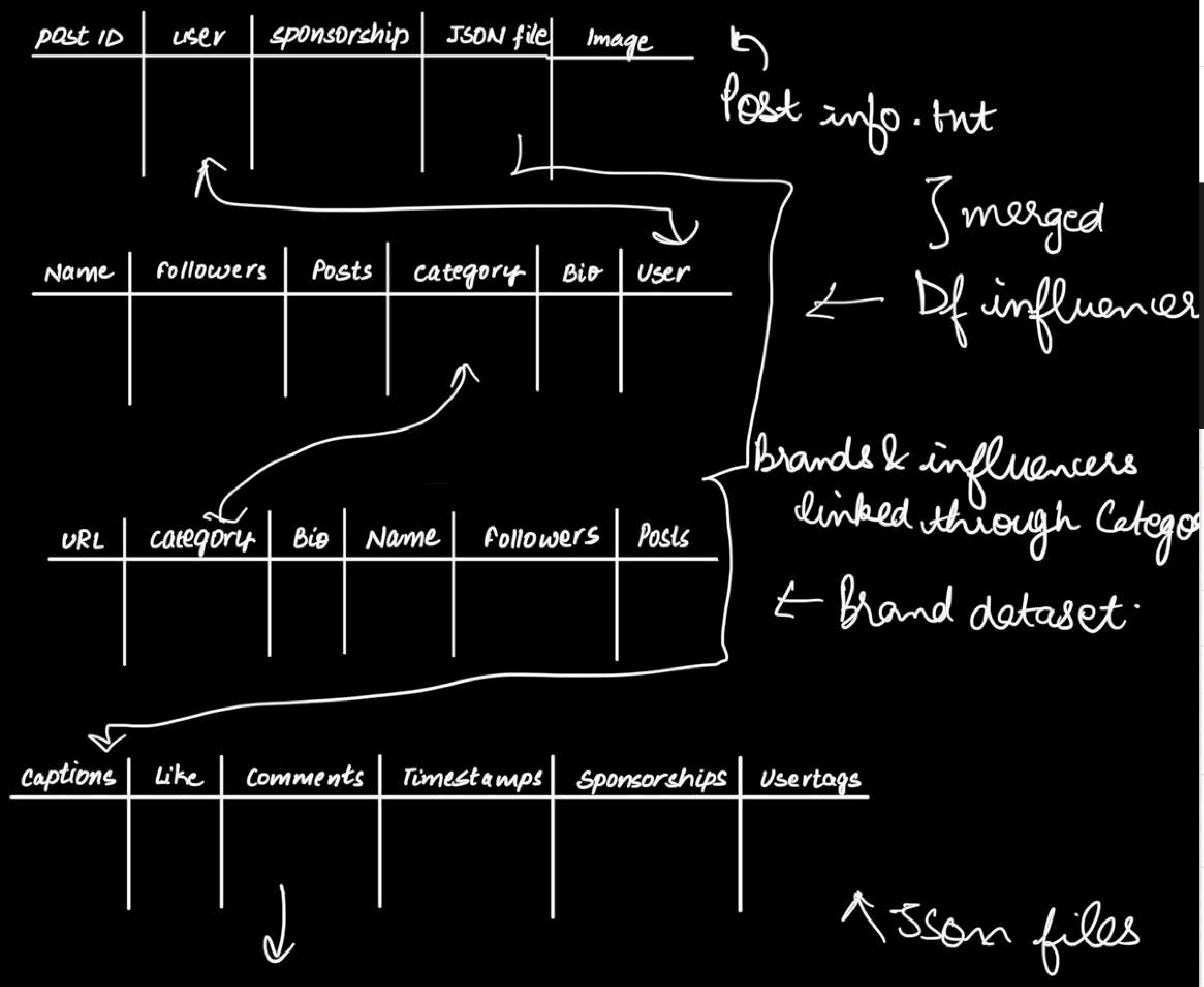
The CatBoostClassifier model (trained on historical sponsorship data) predicts the probability of sponsorship for each influencer.

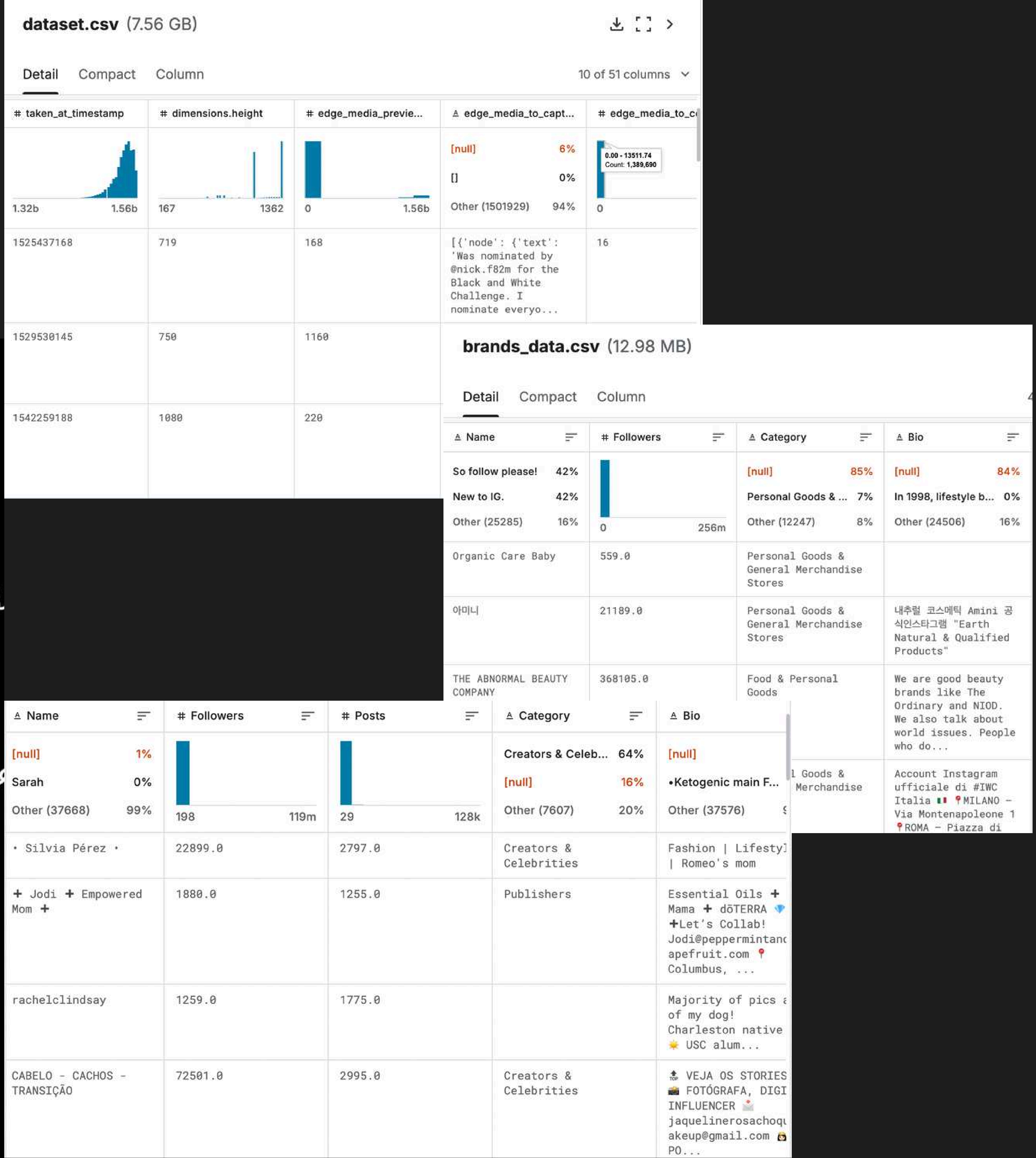
Recommendation System

Predict sponsorship probability for each influencer, Rank influencers by predicted probability and recommend top relevant influencers (hopefully) .

“SECTION 3”

Dataset

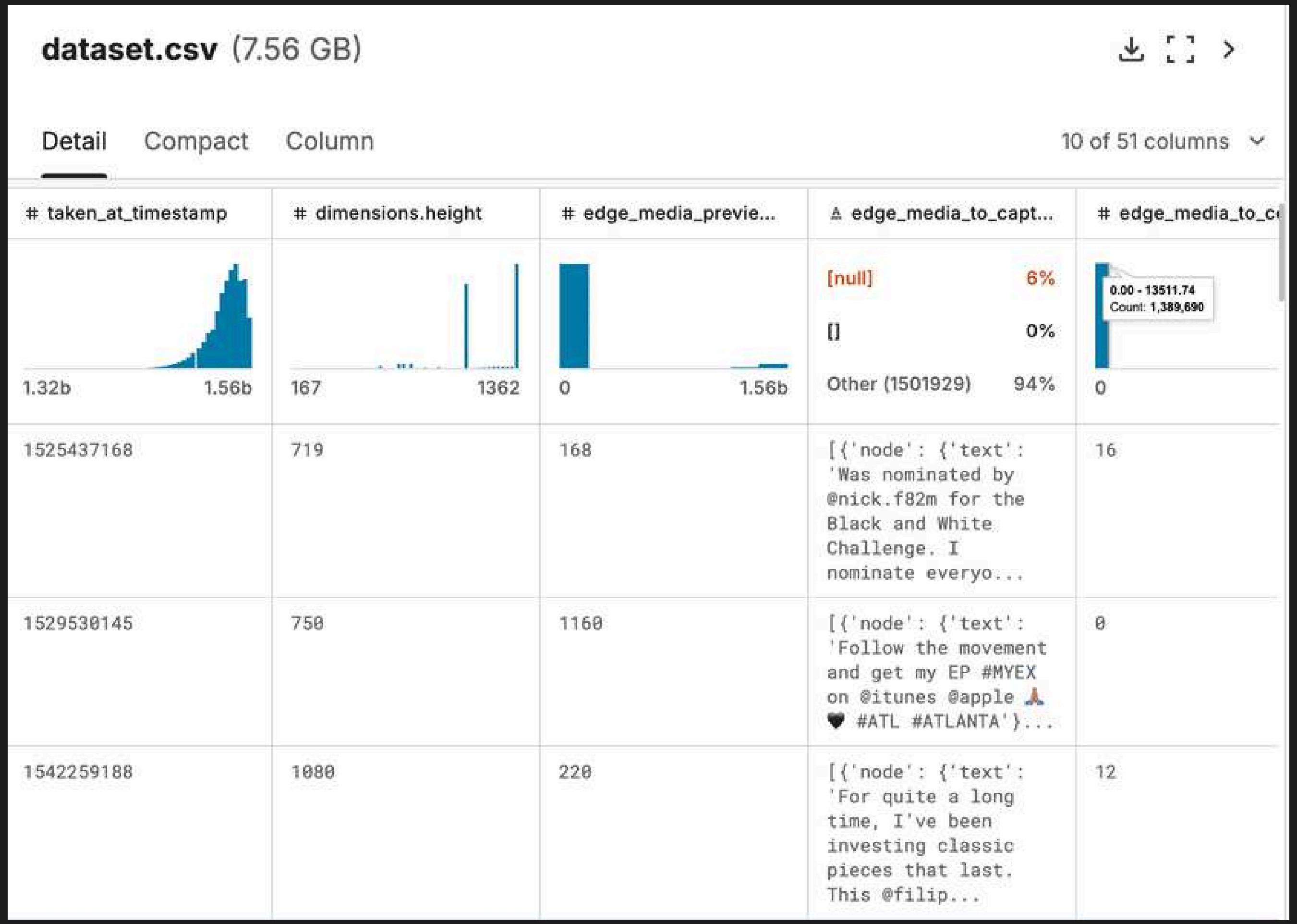

post ID | user | sponsorship | JSON file | Image
Post info.txt
merged
Name | followers | Posts | category | Bio | User
Df influencer
Brands & influencers linked through Catego
URL | category | Bio | Name | followers | Posts
brand dataset.
Captions | Like | Comments | Timestamps | Sponsorships | Usertags
Json files


dataset.csv (7.56 GB)
Detail Compact Column
10 of 51 columns
taken_at_timestamp | dimensions.height | edge_media_previe... | edge_media_to_capt... | edge_media_to_c
1.32b 1.56b | 167 1362 | 0 1.56b | [null] 6% [] 0% Other (1501929) 94% | 0.00 - 13511.74 Count: 1,389,690 0
1525437168 | 719 | 168 | [{'node': {'text': 'Was nominated by @nick.f82m for the Black and White Challenge. I nominate everyo... | 16
1529530145 | 750 | 1160
1542259188 | 1080 | 220
brands_data.csv (12.98 MB)
Detail Compact Column
Name | Followers | Category | Bio
So follow please! 42% New to IG. 42% Other (25285) 16% | 0 256m | [null] 85% Personal Goods & ... 7% Other (12247) 8% | [null] 84% In 1998, lifestyle b... 0% Other (24506) 16%
Organic Care Baby | 559.0 | Personal Goods & General Merchandise Stores
아미니 | 21189.0 | Personal Goods & General Merchandise Stores | 내추럴 코스메틱 Amini 공식인스타그램 "Earth Natural & Qualified Products"
THE ABNORMAL BEAUTY COMPANY | 368105.0 | Food & Personal Goods | We are good beauty brands like The Ordinary and NIOD. We also talk about world issues. People who do...
l Goods & Merchandise | Account Instagram ufficiale di #IWC Italia 🇮🇹 📍MILANO – Via Montenapoleone 1 📍ROMA – Piazza di
Name | Followers | Posts | Category | Bio
[null] 1% Sarah 0% Other (37668) 99% | 198 119m | 29 128k | Creators & Celeb... 64% [null] 16% Other (7607) 20% | [null] •Ketogenic main F... Other (37576)
• Silvia Pérez • | 22899.0 | 2797.0 | Creators & Celebrities | Fashion | Lifestyl | Romeo's mom
✚ Jodi ✚ Empowered Mom ✚ | 1880.0 | 1255.0 | Publishers | Essential Oils ✚ Mama ✚ dōTERRA 💎 ✚Let's Collab! Jodi@peppermintanc apefruit.com 📍 Columbus, ...
rachelclindsay | 1259.0 | 1775.0 | | Majority of pics a of my dog! Charleston native 🌟 USC alum...
CABELO - CACHOS - TRANSIÇÃO | 72501.0 | 2995.0 | Creators & Celebrities | VEJA OS STORIES 📸 FOTÓGRAFA, DIGI INFLUENCER 📩 jaquelinerosachoqu akeup@gmail.com PO...

Dataset

- Brand Username
- Brand Bio
- Brand Category
- Influencer Username
- Influencer Name
- Influencer Category
- Influencer Bio
- Influencer Followers
- Engagement Rate
- Sponsorship
- Bio Similarity
- Brand Name
- Followers
- Sentiment Score

	Brand Username	Brand Bio	Brand Category	Influencer Username	Influencer Name	Influencer Category	Influencer Bio	Influencer Followers	Influencer Engagement Rate	sponsorship_final	bio_similarity	Brand Name	Followers	sentiment_score
0	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	wellness_ed	Amy Hopkinson	Publishers	I can't dance but I can burpee 💻 Digital Ed @W...	46736.0	1.076717	1	0.268495	Pukka Herbs	61353.0	0.9783
1	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	niomismart	Niomi Smart	Unknown	🇬🇧 Lifestyle Vlogger 📕 Author of Eat Smart 🍓 C...	1685570.0	0.421798	1	0.334331	Pukka Herbs	61353.0	0.9993
2	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	bbaldz	brooke baldwin	Creators & Celebrities	LA. 💌: brookebaldwin99@gmail.com	26254.0	31.429285	0	0.067289	Pukka Herbs	61353.0	0.0000
3	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	lauraschra	Lau	Home Goods Stores	-🌿 - our house in pictures '1904' - Drenthe -...	31572.0	31.429285	0	0.054145	Pukka Herbs	61353.0	0.0000
4	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	beautyandthebambino	A L I A E L L I S O N	Unknown	Living ➕ Loving ➕ Life aliajacksonc@gmail.com	9219.0	31.429285	0	0.205278	Pukka Herbs	61353.0	0.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38230	europeanwax	Revealing You. Revealing Beautiful Skin.	Transportation & Accomodation Services	_davidecorsini_	Davide Corsini	Creators & Celebrities	#lifestyle #sportaddicted #mediainfluencer 🏋...	62339.0	3.185205	0	0.217633	European Wax Center	35206.0	0.9775
38231	europeanwax	Revealing You. Revealing Beautiful Skin.	Transportation & Accomodation Services	rheamendezona	Maria Regina Ramas Mendezona	Creators & Celebrities	@thesocialtumbleweed Wife to @marcomendezona ...	6014.0	29.794501	0	0.175660	European Wax Center	35206.0	0.9972

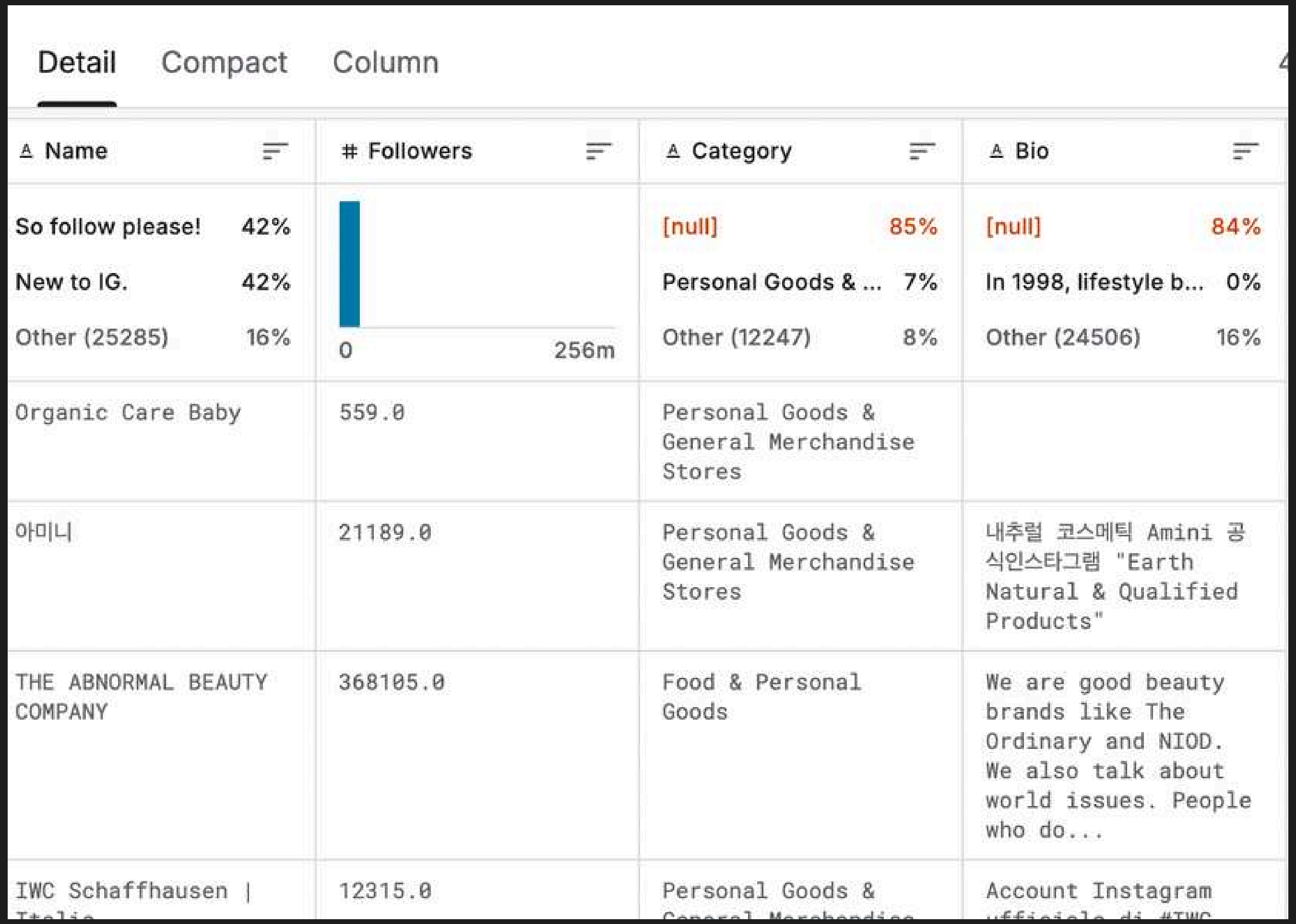


dataset.csv (7.56 GB)

Detail Compact Column — 10 of 51 columns

# taken_at_timestamp	# dimensions.height	# edge_media_previe...	≜ edge_media_to_capt...	# edge_media_to_c...
1.32b – 1.56b	167 – 1362	0 – 1.56b	[null] 6%; [] 0%; Other (1501929) 94%	0 (0.00 - 13511.74 Count: 1,389,690)
1525437168	719	168	[{'node': {'text': 'Was nominated by @nick.f82m for the Black and White Challenge. I nominate everyo...	16
1529530145	750	1160	[{'node': {'text': 'Follow the movement and get my EP #MYEX on @itunes @apple 🙏 🖤 #ATL #ATLANTA'}...	0
1542259188	1080	220	[{'node': {'text': 'For quite a long time, I've been investing classic pieces that last. This @filip...	12

- Contains JSON files of the 16,01,074 posts.
- JSON files have various information such as captions, likes, comments, timestamps, sponsorship, usertags, etc.
- Narrowed down to relevant attributes (Name, Followers, Category, Bio)
- Data preprocessing done on Kaggle.



Detail Compact Column

≜ Name	# Followers	≜ Category	≜ Bio
So follow please! 42%; New to IG. 42%; Other (25285) 16%	0 – 256m	[null] 85%; Personal Goods & ... 7%; Other (12247) 8%	[null] 84%; In 1998, lifestyle b... 0%; Other (24506) 16%
Organic Care Baby	559.0	Personal Goods & General Merchandise Stores	
아미니	21189.0	Personal Goods & General Merchandise Stores	내추럴 코스메틱 Amini 공식인스타그램 "Earth Natural & Qualified Products"
THE ABNORMAL BEAUTY COMPANY	368105.0	Food & Personal Goods	We are good beauty brands like The Ordinary and NIOD. We also talk about world issues. People who do...
IWC Schaffhausen |	12315.0	Personal Goods &	Account Instagram

- Contains Instagram profiles of the 25,282 brands scraped by authors.
- A dataset of 10 brand attributes (eg. Name, Followers,Following, No. of posts, Category, URLs, Bio, Email, Phone no., Profile picture.)
- Narrowed down to 4 relevant attributes (Name, Followers, Category, Bio)

Name	# Followers	# Posts	Category	Bio
[null] 1% Sarah 0% Other (37668) 99%	198 119m	29 128k	Creators & Celeb... 64% [null] 16% Other (7607) 20%	[null] •Ketogenic main F... Other (37576)
• Silvia Pérez •	22899.0	2797.0	Creators & Celebrities	Fashion | Lifestyl | Romeo's mom
✚ Jodi ✚ Empowered Mom ✚	1880.0	1255.0	Publishers	Essential Oils ✚ Mama ✚ dōTERRA 💎 ✚Let's Collab! Jodi@peppermintanc apefruit.com 📍 Columbus, ...
rachelclindsay	1259.0	1775.0		Majority of pics a of my dog! Charleston native ☀️ USC alum...
CABELO - CACHOS - TRANSIÇÃO	72501.0	2995.0	Creators & Celebrities	🔝 VEJA OS STORIES 📸 FOTÓGRAFA, DIGI INFLUENCER 📩 jaquelinerosachoqu akeup@gmail.com 👸 PO...

- Contains Instagram profiles of the 38,113 influencers scraped by author.
- A dataset of 11 influencer attributes (Name, followers, No. of posts, URLs, category, bio, email, phone no., profile picture, username).
- Narrowed down to 5 relevant attributes (Name, Followers, No. of posts, Category, Bio)

	Brand Username	Brand Bio	Brand Category	Influencer Username	Influencer Name	Influencer Category	Influencer Bio	Influencer Followers	Influencer Engagement Rate	sponsorship_final	bio_similarity	Brand Name	Followers	sentiment_score
0	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	wellness_ed	Amy Hopkinson	Publishers	I can't dance but I can burpee 🖥️ Digital Ed @W...	46736.0	1.076717	1	0.268495	Pukka Herbs	61353.0	0.9783
1	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	niomismart	Niomi Smart	Unknown	🇬🇧 Lifestyle Vlogger 📕 Author of Eat Smart 🍓 C...	1685570.0	0.421798	1	0.334331	Pukka Herbs	61353.0	0.9993
2	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	bbaldz	brooke baldwin	Creators & Celebrities	LA. 💌: brookebaldwin99@gmail.com	26254.0	31.429285	0	0.067289	Pukka Herbs	61353.0	0.0000
3	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	lauraschra	Lau	Home Goods Stores	-🌿 - our house in pictures '1904' - Drenthe -...	31572.0	31.429285	0	0.054145	Pukka Herbs	61353.0	0.0000
4	pukkaherbs	Celebrating how delicious, organic herbal teas...	Grocery & Convenience Stores	beautyandthebambino	A L I A E L L I S O N	Unknown	Living ➕ Loving ➕ Life 📧 aliajacksonc@gmail.com	9219.0	31.429285	0	0.205278	Pukka Herbs	61353.0	0.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38230	europeanwax	Revealing You. Revealing Beautiful Skin.	Transportation & Accomodation Services	_davidecorsini_	Davide Corsini	Creators & Celebrities	#lifestyle #sportaddicted #mediainfluencer 🏋️...	62339.0	3.185205	0	0.217633	European Wax Center	35206.0	0.9775
38231	europeanwax	Revealing You. Revealing Beautiful Skin.	Transportation & Accomodation Services	rheamendezona	Maria Regina Ramas Mendezona	Creators & Celebrities	@thesocialtumbleweed Wife to @marcomendezona ...	6014.0	29.794501	0	0.175660	European Wax Center	35206.0	0.9972

- Obtained after merging and preprocessing the initial datasets.
- For every brand, there are all sponsored influencers + 10 unsponsored influencers

Ethical considerations

The dataset used in our project was originally collected by the authors through web scraping from Instagram

- Scraping enabled collection of attributes like followers, posts, bios, and engagement metrics that are not easily accessible via APIs.
- Violates Instagram's (Meta's) Terms of Service, which strictly prohibit unauthorised automated access.
- Unstable as Instagram can change its structure or block IPs at any time

- To avoid these issues, the author used the official Instagram Graph API instead of scraping, which requires approval and access tokens, for legitimate and scalable data collection.
- Access only allowed to Instagram Business or Creator accounts

Pre-Processing

```
     sponsor_user_username                      edge_media_to_tagged_user.edges  \
122       drogariasaopaulo  [{'node': {'user': {'full_name': 'Vichy Brasil...
140       uniquemonbijoux  [{'node': {'user': {'full_name': 'Maysa Leão ✨...
580       homesensecanada  [{'node': {'user': {'full_name': 'Style at Hom...
600               zalando  [{'node': {'user': {'full_name': 'ALLSAINTS', ...
710                 stage  [{'node': {'user': {'id': '1298729246', 'usern...
806         outletdelcorso  [{'node': {'user': {'full_name': 'OUTLET DEL C...
3250                iorane                                                   []
3658              shopiixc  [{'node': {'user': {'full_name': 'Erika Munro ...
4428          veuveclicquot  [{'node': {'user': {'full_name': 'Veuve Clicqu...
4698      constance_calcados  [{'node': {'user': {'full_name': 'Constance Ca...

              sponsor_user_full_name
122                     Vichy Brasil
140    Maysa Leão ✨ Marketing Digital
580                    Style at Home
600                        ALLSAINTS
710                             None
806                  OUTLET DEL CORSO
3250                            None
3658       Erika Munro Kennerly, Esq.
4428                Veuve Clicquot USA
4698                Constance Calçados
```

- Cleaned the tagged user text to remove unwanted characters.
- Isolated the relevant part of the tagged text.
- Extracted sponsor full names from the tagged users column.
- Matched the extracted names with the sponsor usernames already provided.

Pre-Processing

```
                              Category  \
0  Personal Goods & General Merchandise Stores
1  Personal Goods & General Merchandise Stores
2                      Food & Personal Goods
3  Personal Goods & General Merchandise Stores
4       Transportation & Accomodation Services

                                                 Bio
0                                                NaN
1  내추럴 코스메틱 Amini 공식인스타그램 "Earth Natural & Qualif...
2  We are good beauty brands like The Ordinary an...
3  Account Instagram ufficiale di #IWC Italia 🇮🇹 ...
4  Welcome to the Instagram account for Four Seas...

Final Dataset Sample:
       User               JSON_File       Name  Followers  Posts  \
0  alisasia  1309041812857818435.json  Alisa Sia    41099.0  691.0
1  alisasia  1311846669234786866.json  Alisa Sia    41099.0  691.0
2  alisasia  1315560311952470229.json  Alisa Sia    41099.0  691.0
3  alisasia  1318531733175899446.json  Alisa Sia    41099.0  691.0
4  alisasia  1343280729400051114.json  Alisa Sia    41099.0  691.0

               Category                                        Bio
0  Creators & Celebrities  Los Angeles Snapchat & Twitter: AlisaSia alisa...
1  Creators & Celebrities  Los Angeles Snapchat & Twitter: AlisaSia alisa...
2  Creators & Celebrities  Los Angeles Snapchat & Twitter: AlisaSia alisa...
3  Creators & Celebrities  Los Angeles Snapchat & Twitter: AlisaSia alisa...
4  Creators & Celebrities  Los Angeles Snapchat & Twitter: AlisaSia alisa...
```

Cleaning missing data

```
   owner.username  Engagement Rate  Engagement Rate (Standardized)  \
0             NaN              NaN                             NaN
1             NaN              NaN                             NaN
2     essiesophie        11.460780                       -0.045592
3     essiesophie        66.744951                        0.041928
4     essiesophie         7.092532                       -0.052507
5     essiesophie         9.112259                       -0.049310
6     essiesophie         0.798497                       -0.062471
7     essiesophie       164.208549                        0.196221
8     essiesophie         1.033349                       -0.062100
9     ellaxarnold         3.579401                       -0.058069
10    ellaxarnold         2.563313                       -0.059677
11  madison_silotti       5.489163                       -0.055046
12  madison_silotti      10.735264                       -0.046741
13  madison_silotti      67.956249                        0.043845
14  madison_silotti      66.214300                        0.041088

    Average Engagement Rate  Average Engagement Rate (Standardized)
0                       NaN                                     NaN
1                       NaN                                     NaN
2                 37.207274                                0.157577
3                 37.207274                                0.157577
4                 37.207274                                0.157577
5                 37.207274                                0.157577
6                 37.207274                                0.157577
7                 37.207274                                0.157577
8                 37.207274                                0.157577
9                 54.962384                                0.344262
10                54.962384                                0.344262
11                84.151615                                0.651171
12                84.151615                                0.651171
13                84.151615                                0.651171
14                84.151615                                0.651171
```

- Computed engagement rate by (Likes + 2*comments)/followers
- Any missing engagement values were filled by calculating the mean

Pre-Processing

```python
# Display 10-15 samples from both datasets
sample_influencers = influencers_df.head(7)  # First 7 influencer bios
sample_brands = brands_df.head(8)  # First 8 brand bios

# Print processed samples
print("\nInfluencer Bios (Processed):\n", sample_influencers[['Bio', 'Processed_Bio']])
print("\nBrand Bios (Processed):\n", sample_brands[['Bio', 'Processed_Bio']])
```

```
[nltk_data] Error loading punkt: <urlopen error [Errno -3] Temporary
[nltk_data]     failure in name resolution>
<ipython-input-18-4f6f14a264ba>:10: DtypeWarning: Columns (6) have mixed types. Specify dtype option on import or set low_me
  brands_df = pd.read_csv('/kaggle/input/influencer-dataset/brands_data.csv', usecols=['Bio'])

Influencer Bios (Processed):
                                                  Bio  \
0                     Fashion | Lifestyle | Romeo's mom
1  Essential Oils ➕ Boy Mama ➕ dōTERRA 💎 ➕Let’s C...
2  Majority of pics are of my dog!                ...
3  ⚓ VEJA OS STORIES⚓ 📸 FOTÓGRAFA, DIGITAL INFLUE...
4  📍Auckland, New Zealand 📓Honest reviews / Routi...
5  I can't dance but I can burpee 💻 Digital Ed @W...
6  Editing a novel• previously oldfashionedsus Bo...

                                       Processed_Bio
0                       fashion lifestyle romeos mom
1  essential oils boy mama dterra lets collab jod...
2  majority of pics are of my dog charleston nati...
3  veja os stories fotgrafa digital influencer ja...
4  auckland new zealand honest reviews routines f...
5  i cant dance but i can burpee digital ed @ wom...
6  editing a novel previously oldfashionedsus boo...
```

```python
    # Apply function to column
    df['edge_media_to_comment.edges'] = df['edge_media_to_comment.edges'].fillna("[]")  # Handle NaN values
    df['edge_media_to_comment.edges'] = df['edge_media_to_comment.edges'].apply(extract_clean_text)

    # Keep only the processed text column
    df = df[['edge_media_to_comment.edges']]

    # Rename for clarity
    df.rename(columns={'edge_media_to_comment.edges': 'Comments_Text'}, inplace=True)

    # Display sample output
    print(df.head(10))
```

```
                                       Comments_Text
0                                                 []
1                                                 []
2  [nice photo there is no need for more words i ...
3                                                 []
4                                                 []
5                                                 []
6                                                 []
7                                                 []
8                                                 []
9  [travspringer, oh wow love the watch plus you ...
```

Cleaning comments are other text data

Pre-Processing

```python
import pandas as pd
from nltk.sentiment import SentimentIntensityAnalyzer
import nltk
from langdetect import detect, DetectorFactory

# Ensure consistent language detection
DetectorFactory.seed = 0

# Download VADER lexicon
nltk.download('vader_lexicon')

# Load data
df = pd.read_csv("/kaggle/input/comments/merged_df-2.csv")
df['extracted_comments'] = df['extracted_comments'].fillna("")

# Initialize VADER
sia = SentimentIntensityAnalyzer()

# Function: Detect language and apply sentiment
def get_vader_sentiment(text):
    try:
        if detect(text) == 'en':
            return sia.polarity_scores(text)['compound']
        else:
            return 0  # Non-English
    except:
        return 0  # Empty or undetectable text

# Apply sentiment analysis
df['sentiment_score'] = df['extracted_comments'].apply(get_vader_sentiment)

# Average sentiment per influencer
influencer_scores = df.groupby('User').agg({
    'sentiment_score': 'mean',
    'extracted_comments': lambda x: list(x)[:3]  # first 3 comments for display
}).reset_index()

# Sort by sentiment (ascending)
influencer_scores = influencer_scores.sort_values(by='sentiment_score', ascending=True)
```

- VADER LEXICON for sentiment analysis.
- Widely used for sentiment analysis on social media text

```python
import pandas as pd
import numpy as np
from catboost import CatBoostClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.metrics.pairwise import cosine_similarity
from sentence_transformers import SentenceTransformer
```

- Standardisation done using StandardScaler
- SentenceTransformer and Cosine similarity were used for bio similarty

“SECTION 4”

ML
Methodology

Semantic Similarity using Sentence Transformers and Cosine Similarity

Why?

- Bio Similarity Score: to quantify alignment
- Captures Meaning, Not Just Keywords: Unlike TextBlob (which is rule-based and keyword-driven)

How?

- Sentence Embeddings: Convert brand and influencer bios into dense vectors.
- Cosine Similarity: Measure semantic closeness between each brand–influencer bio pair.

ML
Methodology

CatBoostClassifier

Why?

- Trains a CatBoost classifier to learn patterns that distinguish sponsored vs. non-sponsored influencer-brand pairs.

How?

CatBoost uses gradient boosting — that builds a sequence of decision trees, where each new tree tries to fix the errors of the previous trees.

- Start with a basic prediction (like the average label).
- Compute residuals (the mistakes).
- Train new trees to correct those mistakes.
- Repeat for many rounds — each tree gradually improves the model.

ML Methodology

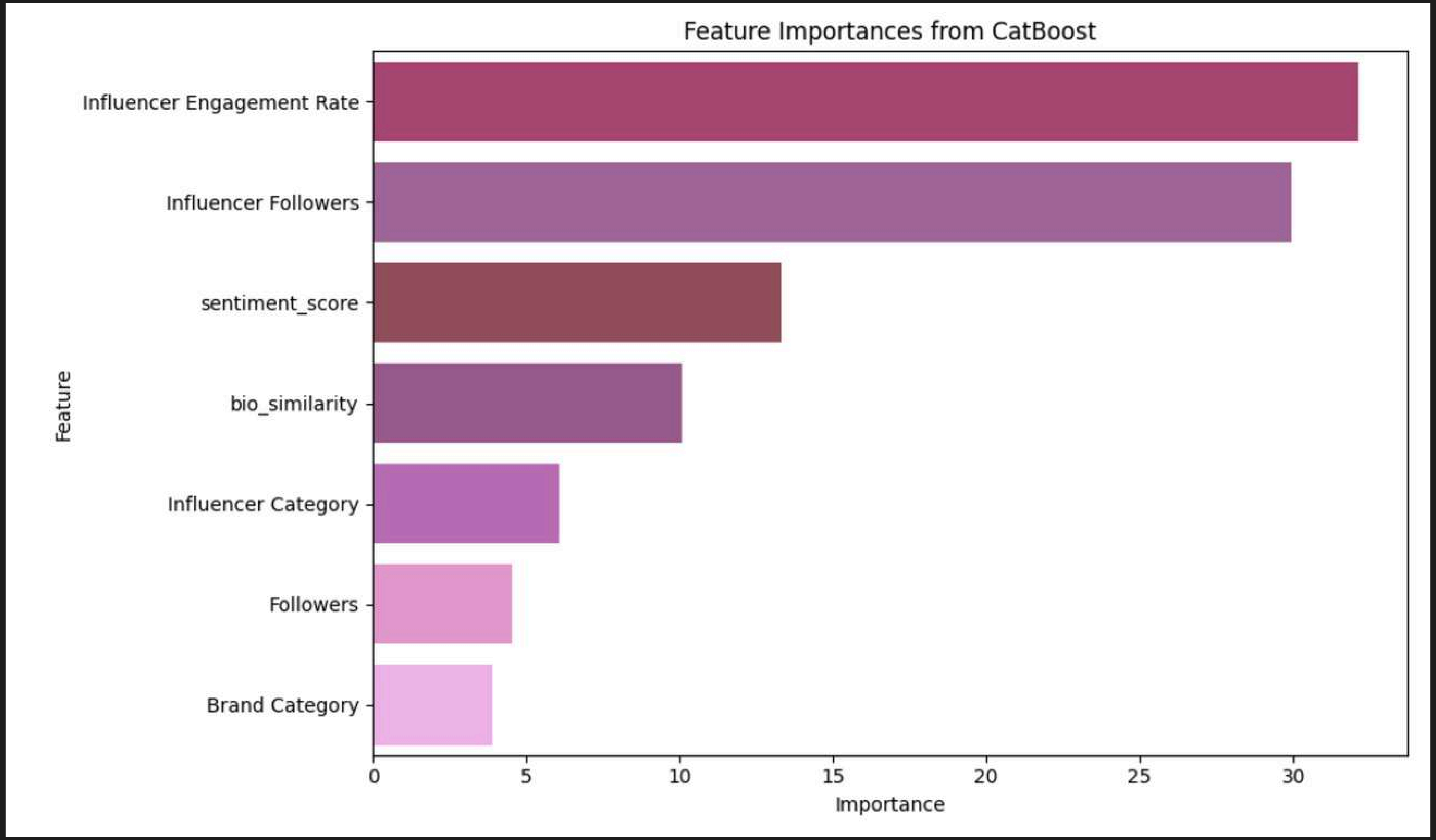
CatBoostClassifier

Feature	Benefit
Native categorical handling	No need for encoding; better performance
High accuracy on tabular data	Usually outperforms random forests
Low overfitting	Thanks to ordered boosting
Easy to use	Minimal preprocessing

- Traditional Boosting uses the full dataset to train at each stage — leading to overfitting whereas CatBoost's Ordered Boosting trains each sample using only the samples before it, preventing data leakage and improving generalization.

ML Methodology

CatBoostClassifier



Challenges Faced & How We Tackled Them

Dataset Challenges

- Brand usernames missing & Sponsorships not directly mentioned → Extracted from custom preprocessing.
- Massive dataset size (~7 crore records) → Couldn’t load in Kaggle due to RAM limits.

Solution:

- Used smart pairing: for each brand, included all sponsored influencers + 10 random unsponsored ones → reduced data size without loss of logic.

Algorithmic & Modeling Decisions

- Initially used TF-IDF for bio similarity → Sentence Transformers for deeper semantic understanding after testing.
- Initially considered XGBoost → CatBoost after deeper domain research

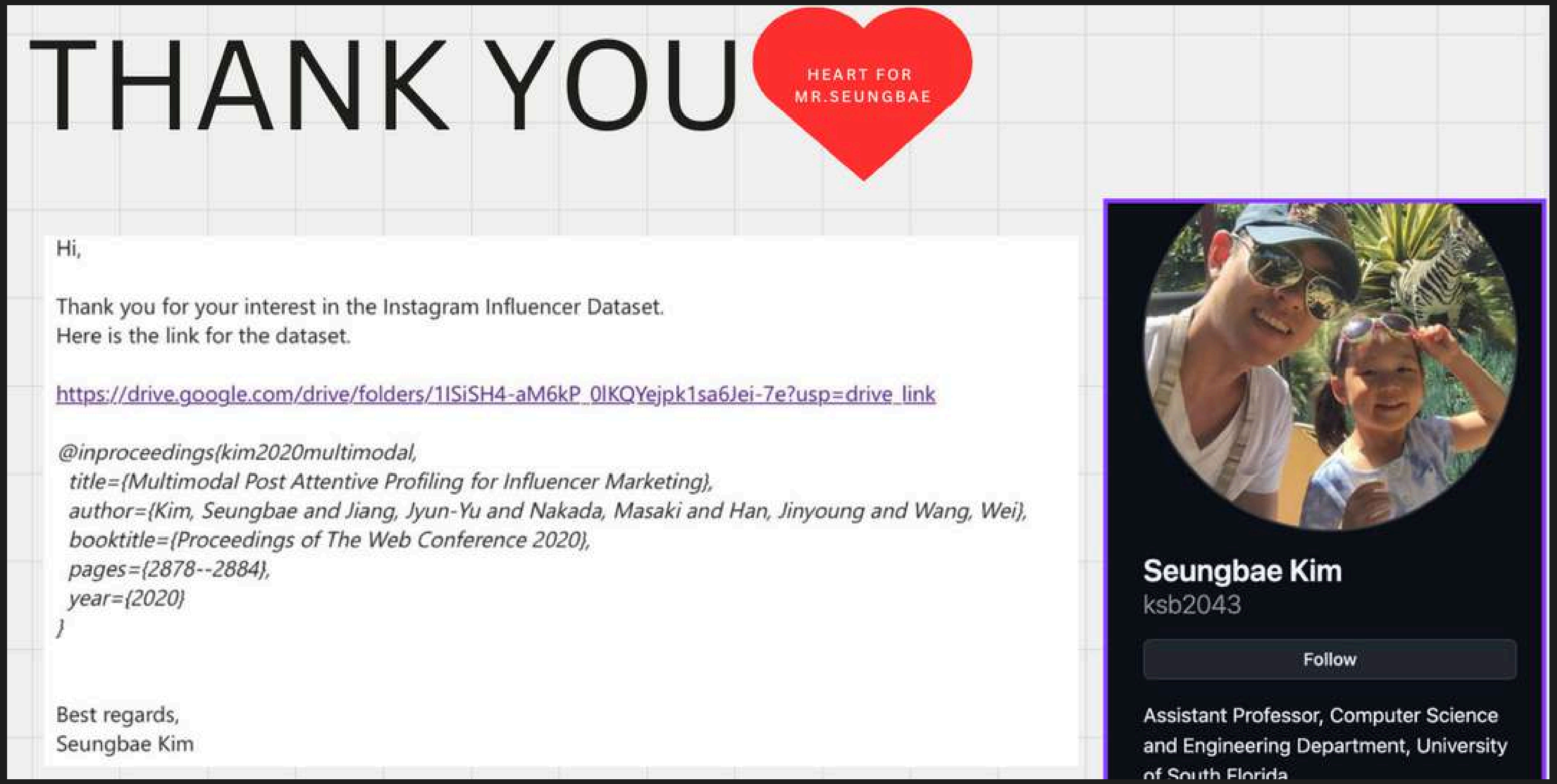
Challenges Faced & How We Tackled Them

Data Availability & Quality

- Dataset acquired after extensive research and cold emailing.
- Limited sponsorship history → Augmented slightly but avoided overdoing to maintain consistency.
- Some influencer usernames and metrics changed over time (e.g., follower count) → Accepted as part of dataset limitation.

Solution:

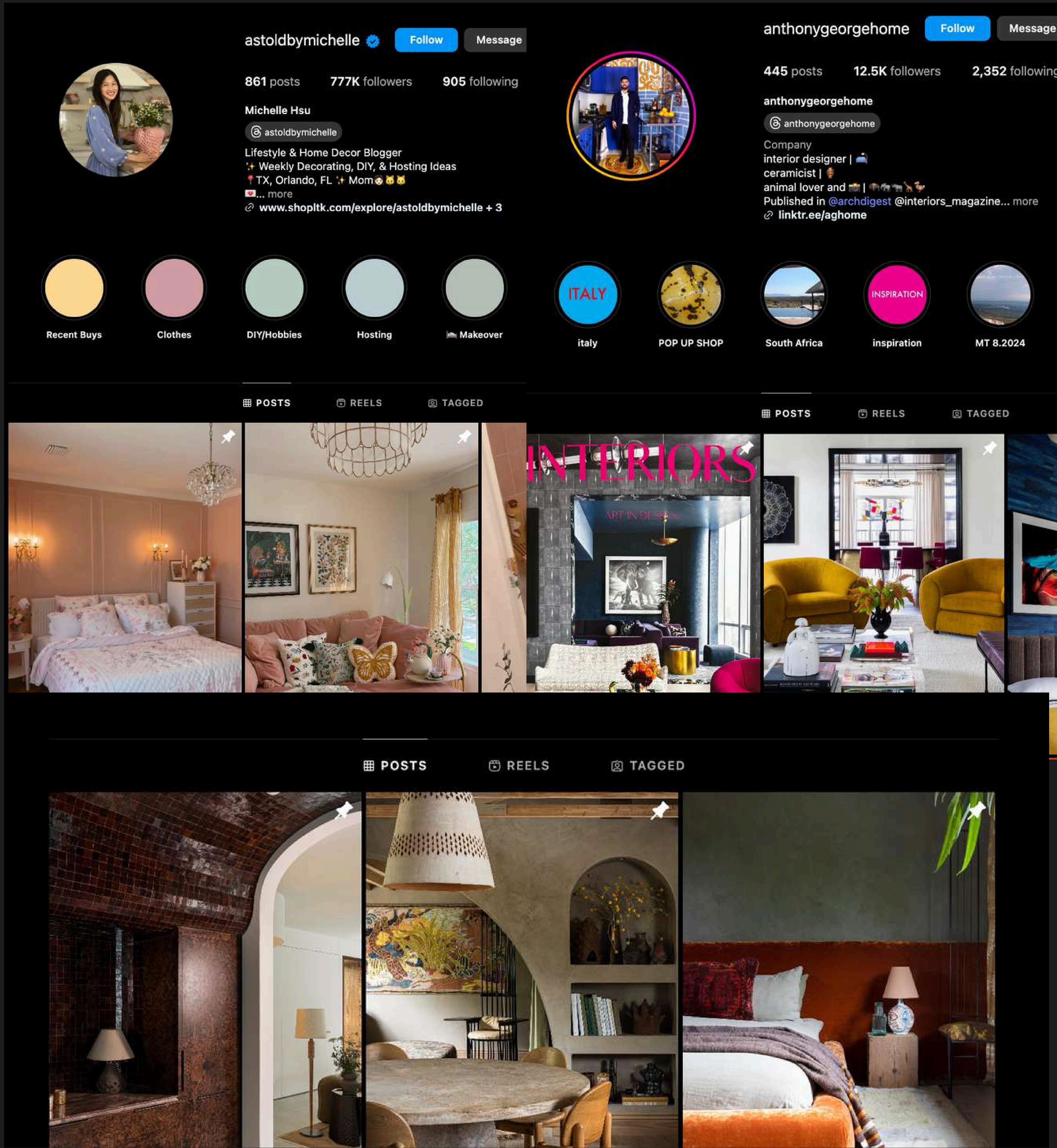
- Treated historical sponsorships as ground truth.
- Acknowledge that with an up-to-date dataset, these issues can be fully resolved.



The Fun Part

```
# === 2. Input new brand data ===
brand_bio = "we have the best furniture and interior design in the industry"
brand_category = "furniture"
brand_followers = 5000
```

```
Batches: 100%|██████████| 1/1 [00:00<00:00, 15
       Influencer Username  Influencer Followers  sponsorship_prob
10308            hommeboys             -0.079969          0.999412
8256      anthonygeorgehome            -0.099726          0.999356
35281      astoldbymichelle            -0.085095          0.999113
10854    centered_by_design            -0.088690          0.998724
4910            coco.and.jack          -0.075875          0.998537
```



The Fun Part

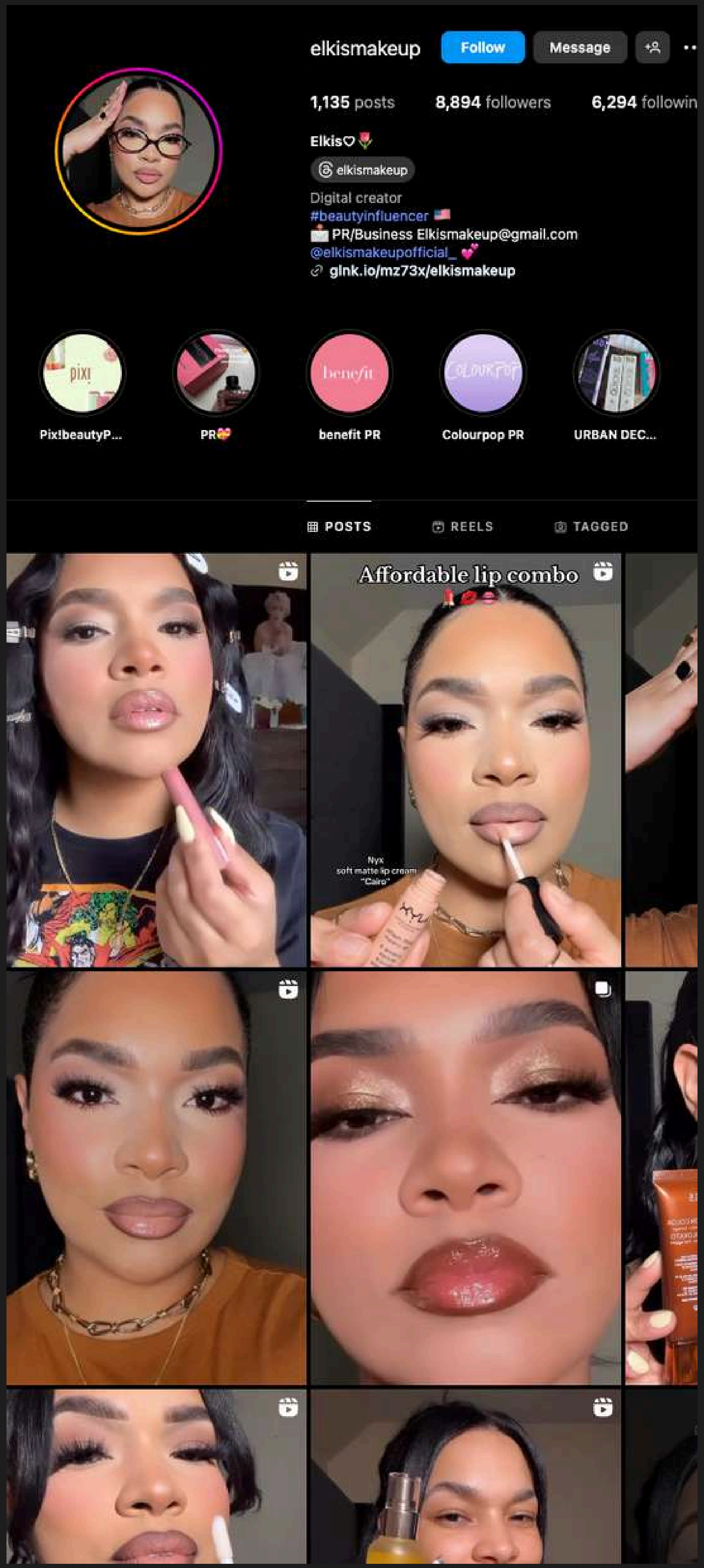
```python
# === 2. Input new brand data ===
brand_bio = "beauty brand focuses on makeup and specializes in lipsticks "
brand_category = "makeup"
brand_followers = 5000
```

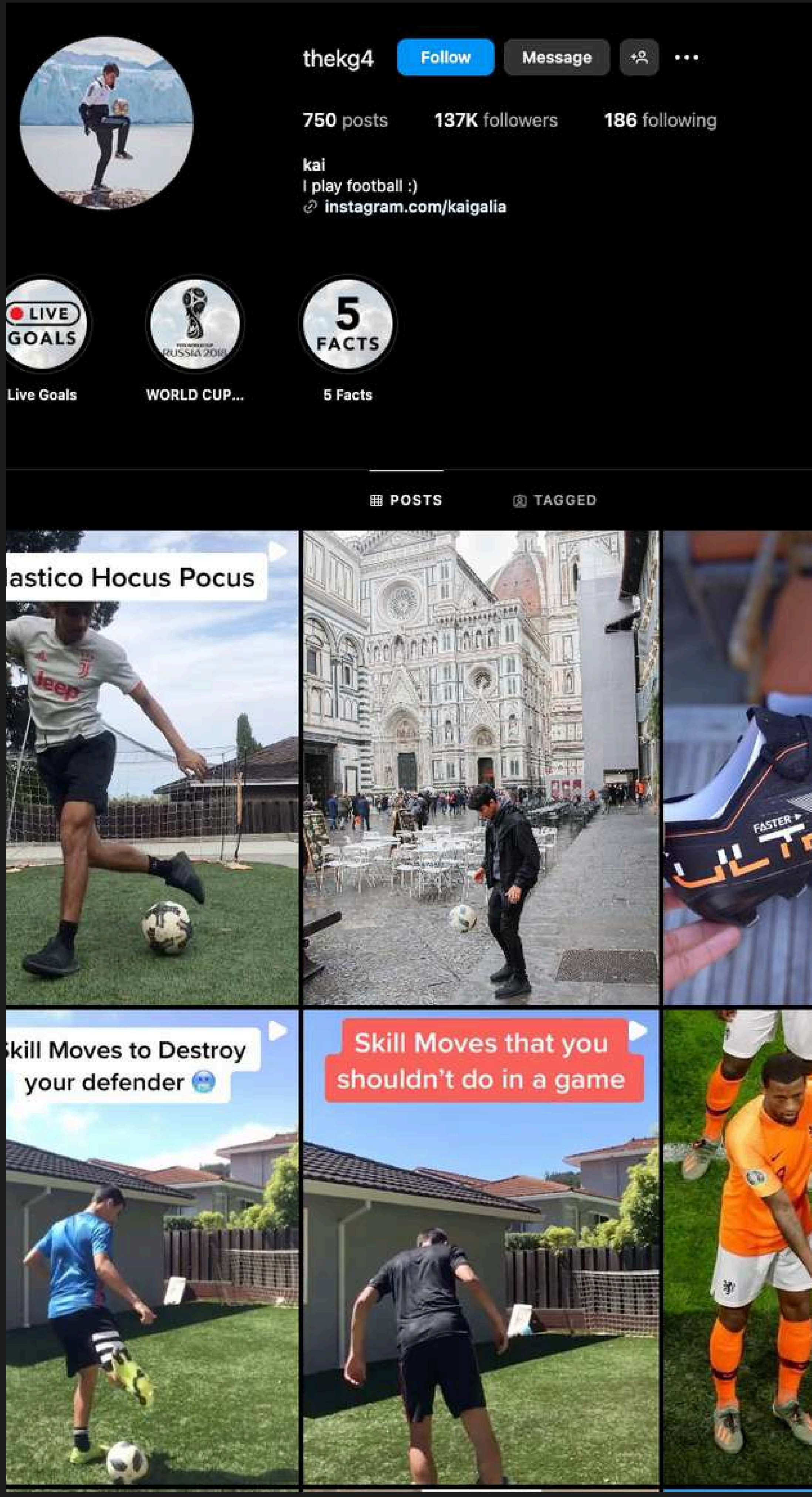
Batches: 100% 1/1 [00:00<00:00, 32.64i

	Influencer Username	Influencer Followers	sponsorship_prob \
35902	wong.jpg	-0.077716	0.999188
27841	elkismakeup	-0.100610	0.998936
12539	_make_up_looks	-0.099955	0.998865
30536	tulipheels95	-0.096729	0.998858
21034	mandiglitter	-0.095591	0.998313

```python
# === 2. Input new brand data ===
brand_bio = "passioante brand about sports football specefically"
brand_category = "football"
brand_followers = 5000
```

22914	thekg4	0.033219	0.996388
14447	_han_banan_	-0.054909	0.991702
28511	angeliquefiske	-0.100603	0.987175
5639	thefootballrepublic	-0.012542	0.986366





“SECTION 5”

Performance Metrics of our model

Mean Reciprocal Rank (MRR): 0.473

- On average, the first true sponsored influencer appears around the 2nd position in our ranked list.

```
    return sum(mrr_list) / len(mrr_list)
mrr_score = calculate_mrr(final_df)
print("MRR Score:", mrr_score)

MRR Score: 0.4729697599196747
```

Recall@3: 0.8902

- In nearly 9 out of 10 cases, at least one correct influencer is in the top three recommendations.

```
k = 3
recall_scores = final_df.groupby('Brand Username').apply(lambda x: recall_at_k(x, k))
overall_recall_at_k = recall_scores.mean()
print(f"Overall Recall@{k}: {overall_recall_at_k:.4f}")

Overall Recall@3: 0.8902
```

Classification Metrics

- Precision ≈ 0.83 Recall ≈ 0.87 (sponsored)
- Balanced Accuracy ≈ 0.9081, F1 ≈ 0.93

Evaluation Protocol

- Computed on a stratified 80/20 hold-out split to preserve sponsorship ratios and ensure robust performance estimates.

Why it matters: High MRR and Recall@3 minimize the list brands need to review & Balanced Precision/Recall shows we avoid both irrelevant picks and missed opportunities.

Deploying Our Influencer Recommendation Engine: Transforming Influencer Marketing

Vision:

- Tap into the booming $40B influencer marketing industry with a first-of-its-kind ML-powered recommendation system that helps brands find the most effective, cost-efficient influencers.

Why Now?

- No commercialized influencer recommendation model exists yet.
- Backed by our passion, rigorous development, and support from Prof. Brainerd — we are ready to scale this innovation to market.

<table>
<tr><th>Business Model</th><th>Description</th></tr>
<tr><td>B2B SaaS Platform</td><td>Subscription-based web interface for startups and agencies to find influencers.</td></tr>
<tr><td rowspan="3">Full-Service Marketing Agency</td><td>Operate as a dedicated agency recommending influencers and earning:</td></tr>
<tr><td>Commission from brand-influencer deals</td></tr>
<tr><td>Revenue share from successful collaborations</td></tr>
</table>

Scaling Considerations for our model

Global Expansion	Add multilingual support and integrate additional platforms (e.g., TikTok, YouTube) to serve new markets.
Advanced Capabilities	Incorporate visual-content matching and fraud/bot detection as load and scope grow.
Go-to-Market Scaling	Launch a freemium tier to drive early adoption and forge agency partnerships for rapid traction.
Data Freshness	Recompute embeddings and sentiment scores daily or weekly; monitor input distributions and retrain when performance degrades.

“The End”

```
# === 2. Input new brand data ===
brand_bio = "Machine Learning & Pattern Recognition"
brand_category = "AI"
brand_followers = 5000
```

```
Batches: 100%                                  1/1 [00:00<00:00, 3
       Influencer Username  Influencer Followers  sponsorship_prob
35902
27841
12539
30536
```

Influencer Username	Influencer Followers	sponsorship_prob
Mr. Siddharth	1.00	1.00
Mr. Pushpinder Singh	1.00	1.00

References :

- https://doi.org/10.48550/arXiv.2106.01750
- https://kth.diva-portal.org/smash/get/diva2%3A1783645/FULLTEXT01.pdf
- https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1636469
- https://www.statista.com/topics/2496/influence-marketing/#topicOverview
- https://doi.org/10.48550/arXiv.2107.13943
- https://www.tcs.com/what-we-do/industries/consumer-goods-distribution/white-paper/reimagining-influencer-marketing-with-machine-learning-nlp
- https://prohibitionpr.co.uk/digital-marketing/influencer-marketing/the-impact-of-working-with-the-wrong-influencers-in-marketing-campaigns/
- https://dl.acm.org/doi/abs/10.1145/3366423.3380052