Influencer Scouting

Big Data Computing

Professor Gabriele Tolomei

SAPIENZA
UNIVERSITÀ DI ROMA

Faculty of Information Engineering,
Informatics and Statistics

Group Members:
Federico Barreca
Shuya Dong

Summary

1 Introduction

2 Dataset

3 Feature Extraction

4 Models

5 App Demo

6 Future Works

1 Introduction

Introduction

- **Influencer marketing:** a key marketing method for brands.
- But many brands encounter difficulties in **influencer scouting**.

- Categorizing the influencers' interests is crucial in maximizing the marketing effect.
- Aim: To determine the influencers' category based on their social media posts.

Instagram

- Instagram has become a popular platform for influencer marketing:

1) Large user base: over one billion active users.
2) Targeted audiences: specific audiences.
3) Authenticity: a personal connection with their followers.
4) Ease of use: it’s simple to create and share content.

2 Dataset

Dataset : Big Data

- Big-Data-ness: its large size and complexity.

- 33,935 Instagram influencers (labeled with 8 categories)
- 10,180,500 Instagram posts : 300 posts per influencer
- Post metadata (JSON files): ~37 GB
- Image (JPEG files): ~189 GB

https://sites.google.com/site/sbkimcv/dataset/instagram-influencer-dataset

Dataset : Big Data

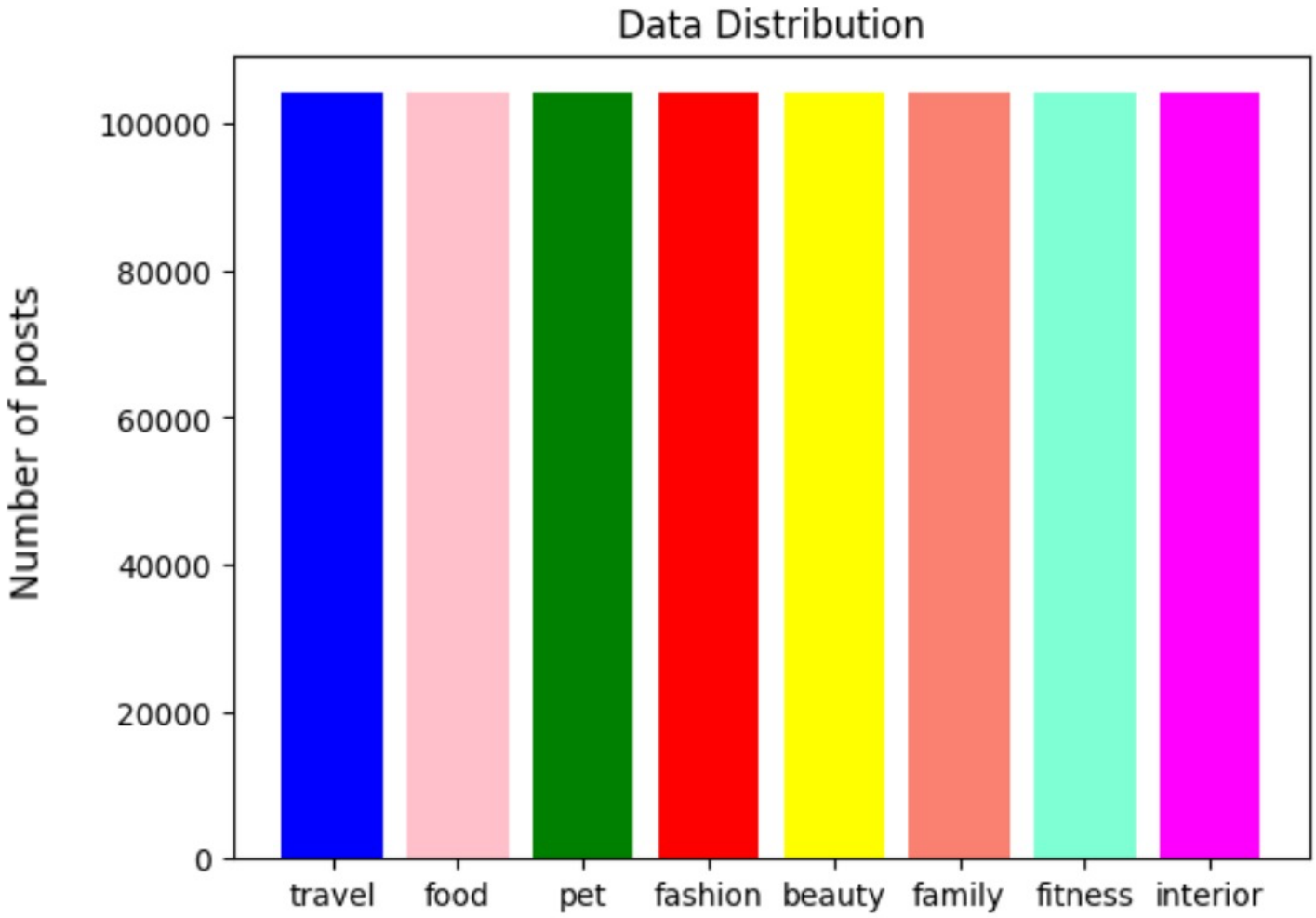
- For our project:

- 347 influencers per category ~ 832,800 Instagram posts

1) Training and validation: 200 influencers per category (8:2)
~ 480,000 posts

2) Testing: 147 influencers per category randomly selected
~ 352,800 posts

Dataset


Data Distribution
Number of posts
100000
80000
60000
40000
20000
0
travel
food
pet
fashion
beauty
family
fitness
interior

Dataset

- Each entry of the dataset includes:
  - Influencer_name
  - Label
  - Post metadata
  - Image

influencer_name	label	post_metadata	post_images
annszi	travel	vscohun magyar vs...	E:/Documenti/Big_...
annszi	travel	vscohun magyar vs...	E:/Documenti/Big_...
annszi	travel	realmadrid sdeiba...	E:/Documenti/Big_...
annszi	travel	no photo descript...	E:/Documenti/Big_...
blissfulbasil	food	alessifoods aless...	E:/Documenti/Big_...
annacichocka_offi...	travel	djkhaled dj khale...	E:/Documenti/Big_...
annacichocka_offi...	travel	sunday glam nnnnn...	E:/Documenti/Big_...
annacichocka_offi...	travel	academiededanseus...	E:/Documenti/Big_...
annacichocka_offi...	travel	as an acadmie de ...	E:/Documenti/Big_...
annacichocka_offi...	travel	please dont judge...	E:/Documenti/Big_...
annacichocka_offi...	travel	saturdays boat pa...	E:/Documenti/Big_...
annacichocka_offi...	travel	saturdays boat pa...	E:/Documenti/Big_...
annacichocka_offi...	travel	saturdays boat pa...	E:/Documenti/Big_...
annacichocka_offi...	travel	zara zara officia...	E:/Documenti/Big_...
annacichocka_offi...	travel	stylebyaniac nnnn...	E:/Documenti/Big_...
blissfulbasil	food	vegan banana brea...	E:/Documenti/Big_...
annacichocka_offi...	travel	out of blue nnnnn...	E:/Documenti/Big_...
annacichocka_offi...	travel	pkrolartist pete ...	E:/Documenti/Big_...
annacichocka_offi...	travel	pkrolartist pete ...	E:/Documenti/Big_...
annacichocka_offi...	travel	travelphotobyania...	E:/Documenti/Big_...

Dataset

- Post metadata files contain the following information:
  - caption, usertags, hashtags, timestamp, sponsorship, likes, comments, etc.


```
"taken_at_timestamp": 1545954154,
"edge_media_to_caption": {
  "edges": [
    {
      "node": {
        "text": "The fluffier the sweater the better!!\ud83e\udd19\ud83c\udffc\ud83d\udc81\ud83c\udffc\u200d\u2640\ufe0f\ud83d\udd25 Tag an extra friend\ud83d\ude02\ud83d\udc81\ud83c\udffc\u200d\u2640\ufe0f\ud83d\udc53"
      }
    }
  ]
},
"tracking_token": "eyJ2ZXJzaW9uIjo1LCJwYXlsb2FkIjp7ImlzX2FuYWx5dGljc190cmFja2VkIjp0cnVlLCJ1dWlkIjoiYTY3NjM4NTg2YzZhNDNiNWE2NzNiMDc2ZDg4Y2IzYzYxOTQzOTI2Nzk5NzM0MzQxMDA0In0sInNpZ25hdHVyZSI6IiJ9",
"has_ranked_comments": false,
"display_url": "https://scontent-lax3-1.cdninstagram.com/vp/a2884560d2f8684aa0a09f41cf0c2b29/5DA92CD8/t51.2885-15/e35/47693940_135411280792230_3949029997371547140_n.jpg?_nc_ht=scontent-lax3-1.cdninstagram.com",
"edge_web_media_to_related_media": {
  "edges": [
  ]
},
"edge_media_preview_comment": {
  "count": 588,
  "edges": [
    {
      "node": {
        "text": "\ud83d\ude0d",
        "created_at": 1558684675,
        "did_report_as_spam": false,
        "edge_liked_by": {
          "count": 0
        },
        "owner": {
          "username": "joelhackett7gmail.com2",
          "profile_pic_url": "https://scontent-lax3-1.cdninstagram.com/vp/67b5b1810480f5b7732150d200cd1155/5DA55F74/t51.2885-19/s150x150/47691490_2259235387695811_4235471854237646848_n.jpg?_nc_ht=scontent-lax3-1.cdninstagram.com",
          "is_verified": false,
          "id": "8619460382"
        },
        "viewer_has_liked": false,
        "id": "18040776964184509"
      }
    },
    {
      "node": {
        "text": "\ud83d\ude04\ud83d\ude0d",
        "created_at": 1558684692,
        "did_report_as_spam": false,
        "edge_liked_by": {
          "count": 0
        },
        "owner": {
          "username": "joelhackett7gmail.com2",
          "profile_pic_url": "https://scontent-lax3-1.cdninstagram.com/vp/67b5b1810480f5b7732150d200cd1155/5DA55F74/t51.2885-19/s150x150/47691490_2259235387695811_4235471854237646848_n.jpg?_nc_ht=scontent-lax3-1.cdninstagram.com",
          "is_verified": false,
          "id": "8619460382"
        },
        "viewer_has_liked": false,
        "id": "17887347304340999"
      }
    }
  ]
},
"comments_disabled": false,
"edge_media_to_sponsor_user": {
  "edges": [
  ]
},
"is_video": false
```

Dataset

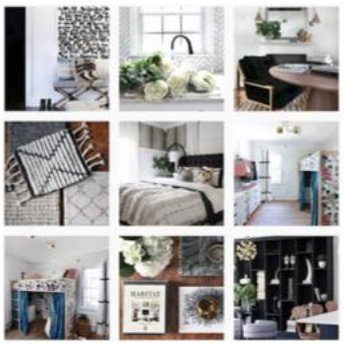
(a) Beauty

(b) Family

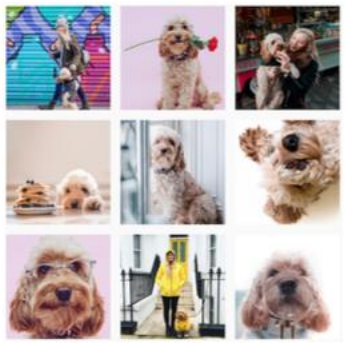
(c) Fashion

(d) Fitness

(e) Food



(f) Interior



(g) Pet

(h) Travel

How can we classify influencers?

How can we classify influencers?

How can we classify influencers?

- To obtain post representation: **Feature Extraction**
  - Text Features
  - Image Features
  - Text & Image

How can we classify influencers?

- To obtain Influencer representation:
  - Aggregate all post features by averaging values
  - Use attention mechanism to weigh higher scores on more important posts

How can we classify influencers?

- To do classification:
  - PySpark models
    1) Logistic Regression
    2) Random Forest
    3) Gaussian Naïve Bayes

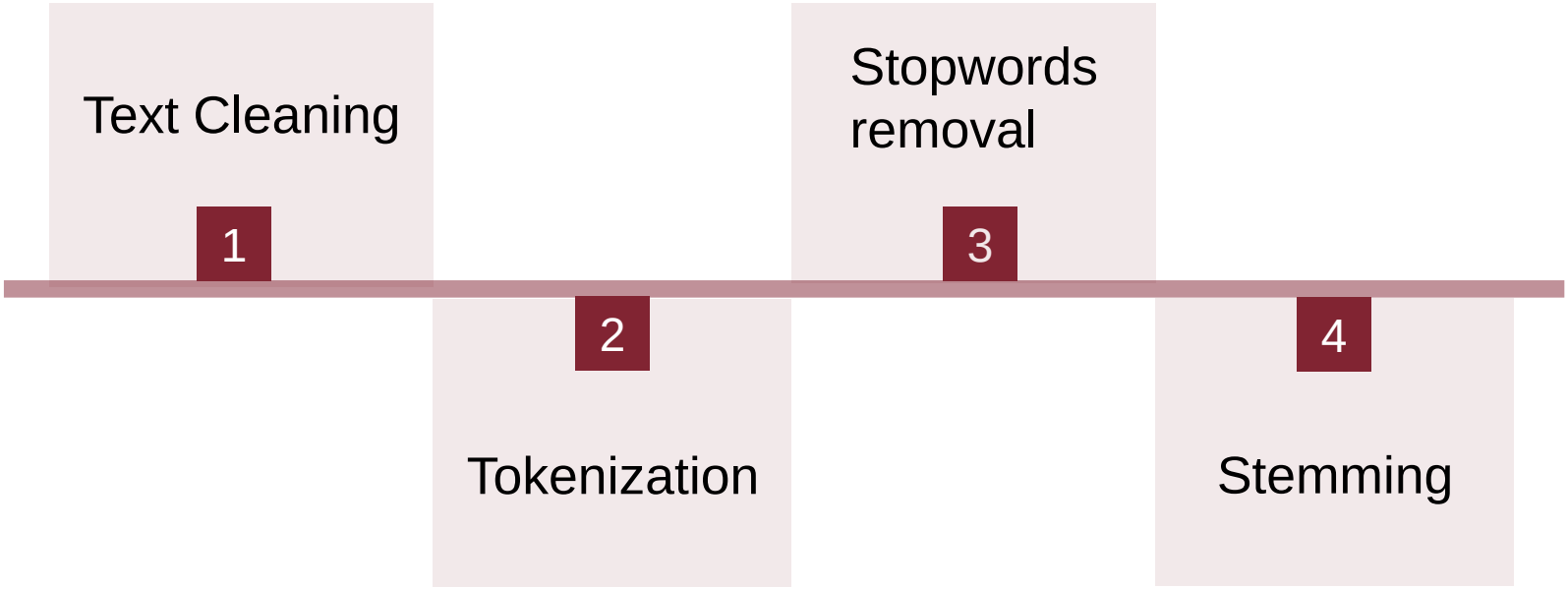
  - Influencer Profiler

    "Multimodal Post Attentive Profiling for Influencer Marketing," Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han and Wei Wang. In Proceedings of The Web Conference (WWW '20), ACM, 2020.

3 Feature Extraction

Feature Extraction: Text Features

Text preprocessing pipeline



Feature Extraction: Text Features

Text preprocessing pipeline

1 **Text Cleaning**

- **Case normalization** to lower
- **Trimming** of the outer white spaces
- **Filter out punctuation symbols**
- **Filtering out** any internal extra **whitespaces**

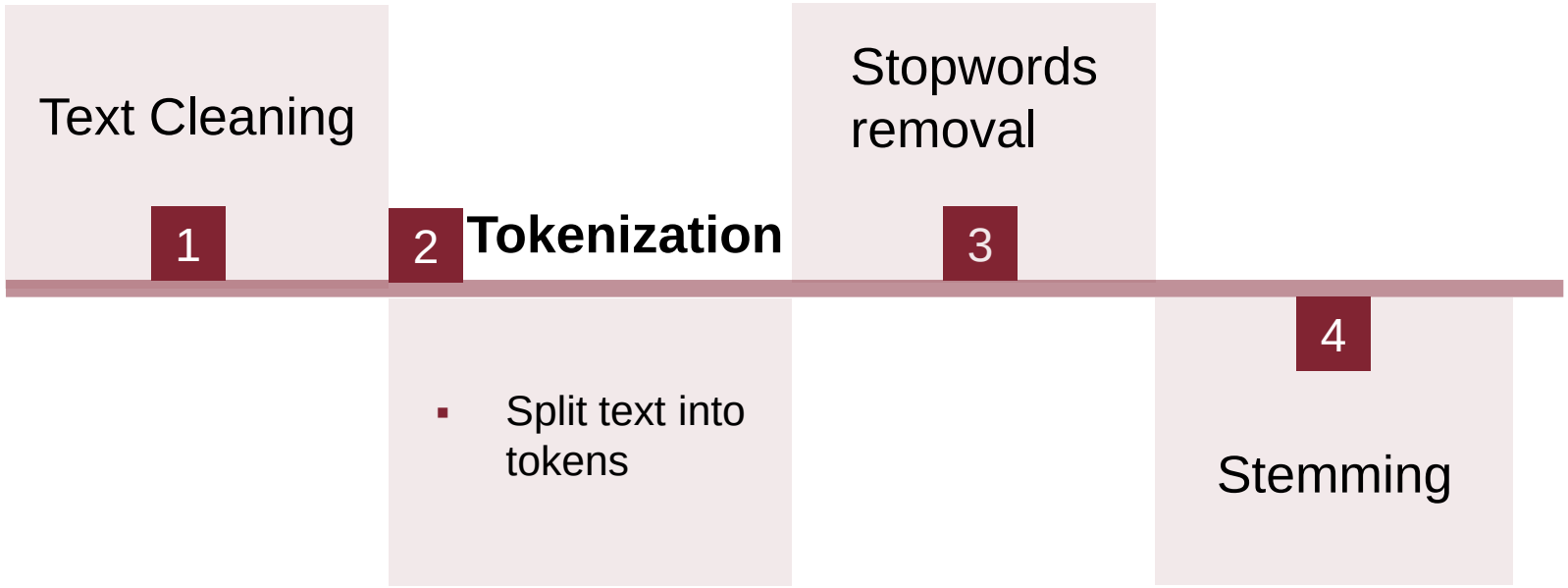
2 Tokenization

3 Stopwords removal

4 Stemming

Feature Extraction: Text Features

Text preprocessing pipeline



Feature Extraction: Text Features

Text preprocessing pipeline

1 Text Cleaning

2 Tokenization

3 **Stopwords removal**

- Remove common words that are unlikely to carry much meaning, such as "the," "and," "a," and "in.

4 Stemming

Feature Extraction: Text Features

Text preprocessing pipeline

1 Text Cleaning

2 Tokenization

3 Stopwords removal

4 **Stemming**

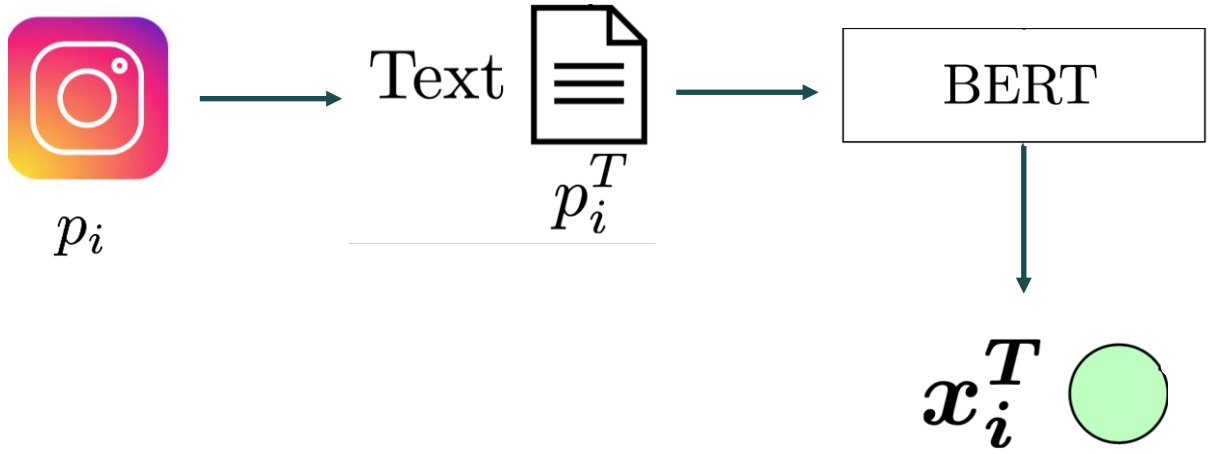
- Reduce words to their base or root form, by removing suffixes and prefixes.

Feature Extraction: Text Features

Text feature embedding

- Use SparkNLP library:

  We exploit the pre-trained text model to derive text features.



Feature Extraction: Image Features

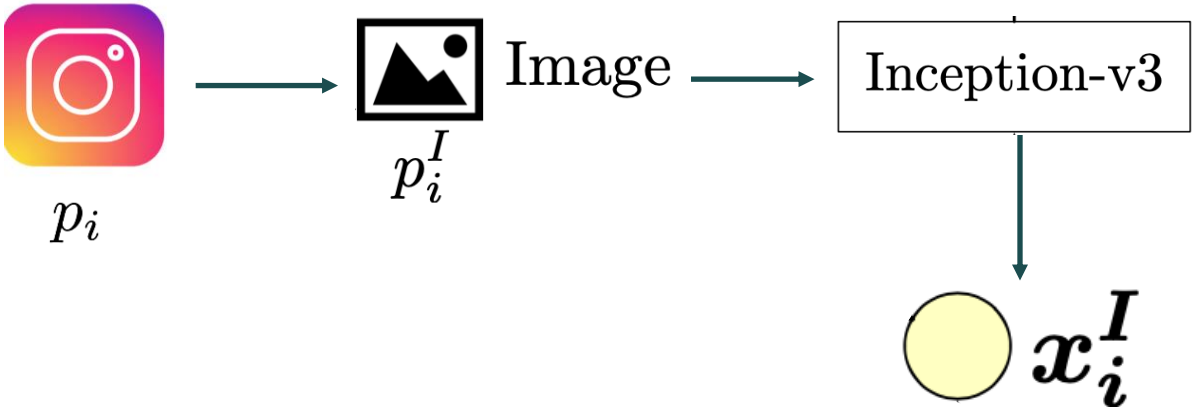
Image preprocessing pipeline

- **Resizing**
  - To ensure that all input images have the same size
- **Data augmentation** using random transformations:
  - To increase the size of the training dataset and improve the generalization performance of model
  - RandomRotation
  - RandomHorizontalFlip
  - RandomResizedCrop
  - Random color jittering
- **Normalization**

Feature Extraction: Image Features

Image feature embedding

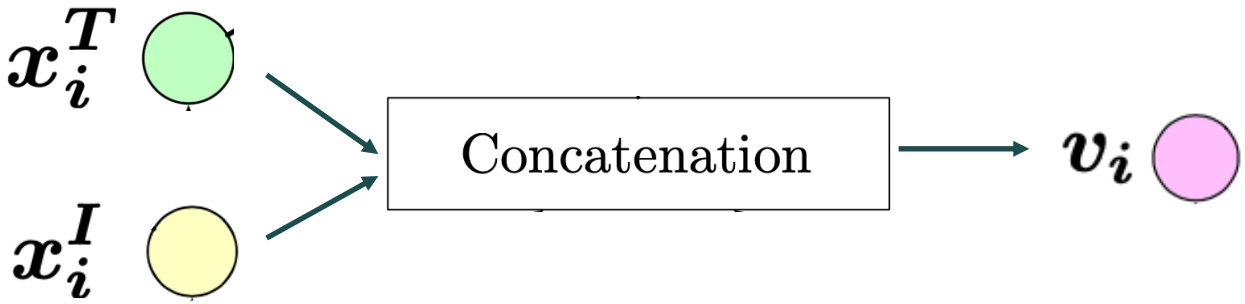
- Apply the transfer learning technique using the pre-trained model
  - Fine Tuning



Feature Extraction: Text & Image

- The feature vector is derived by concatenating the text features and image features.

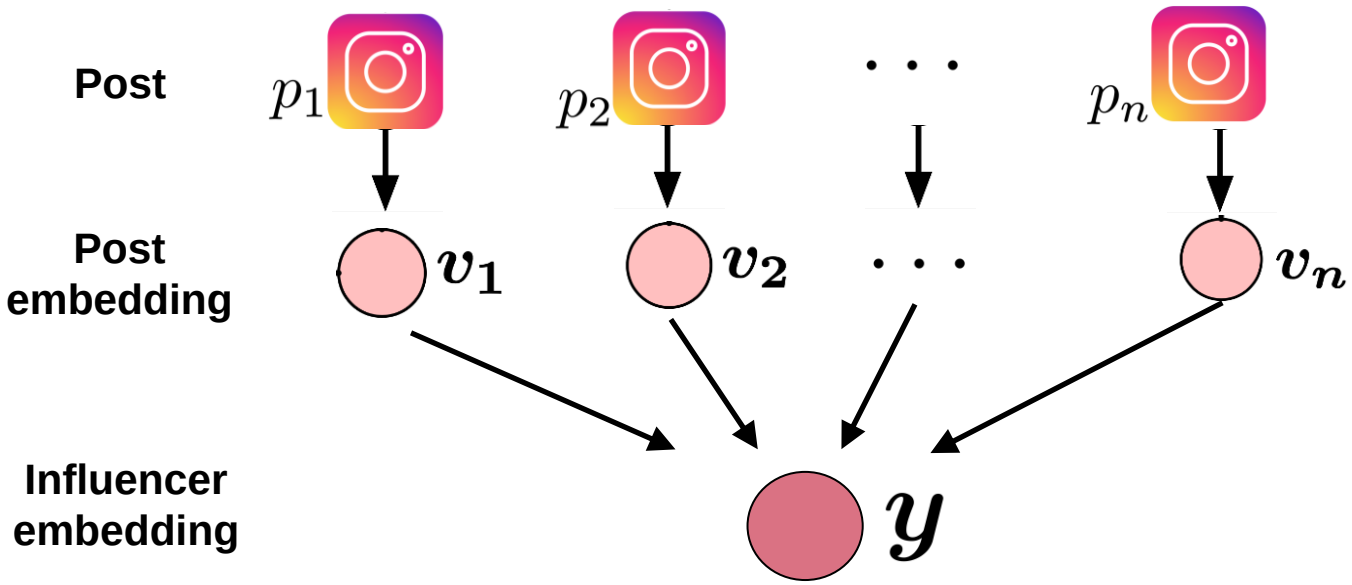
$$x_i = [x_i^I ; x_i^T].$$



4 Models

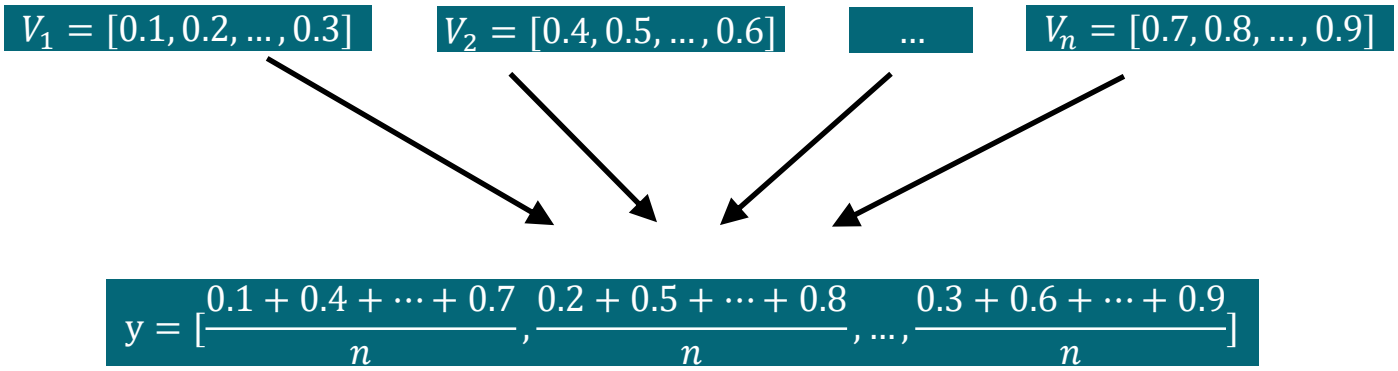
PySpark Models

- Input features: influencer embeddings using average values.

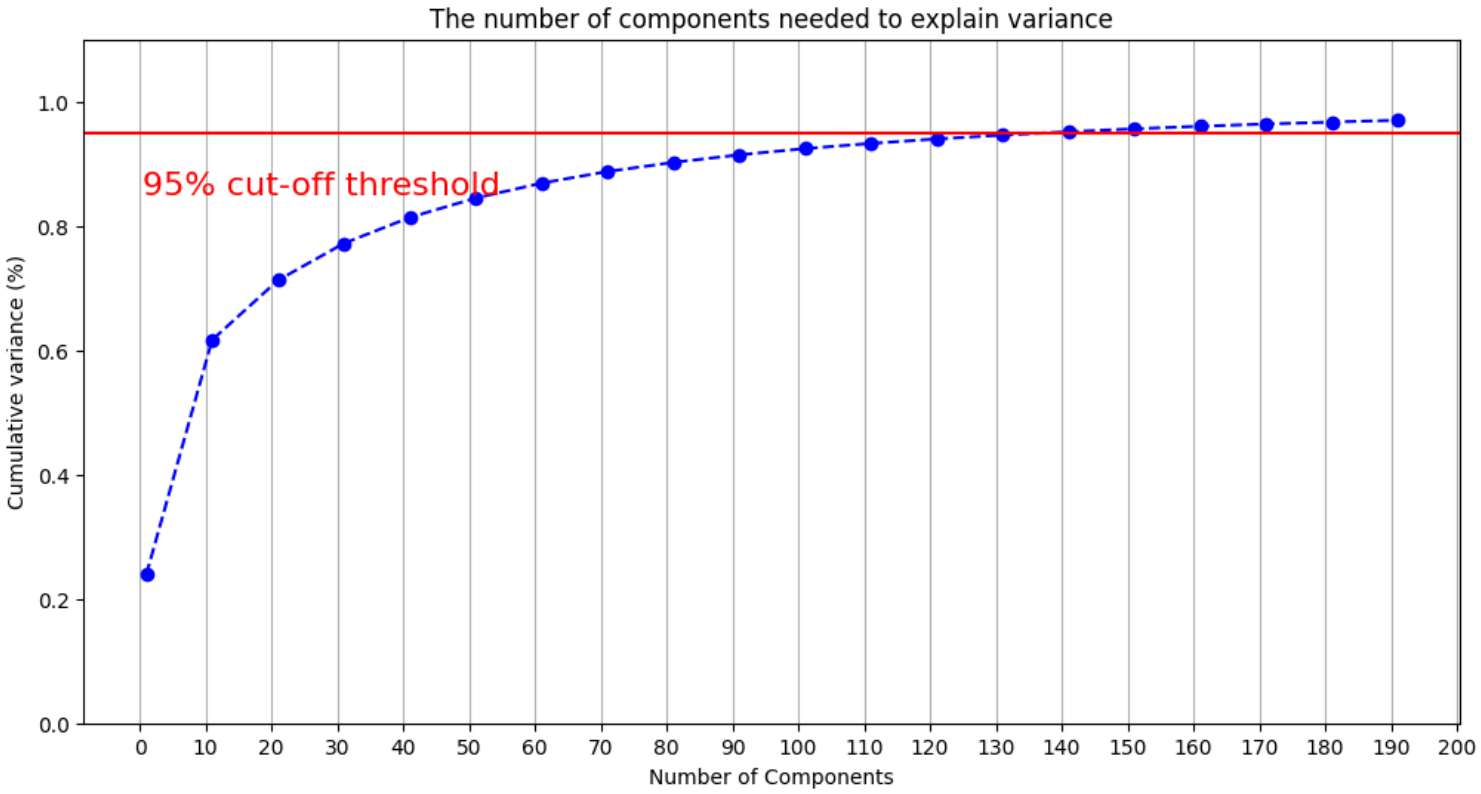


PySpark Models

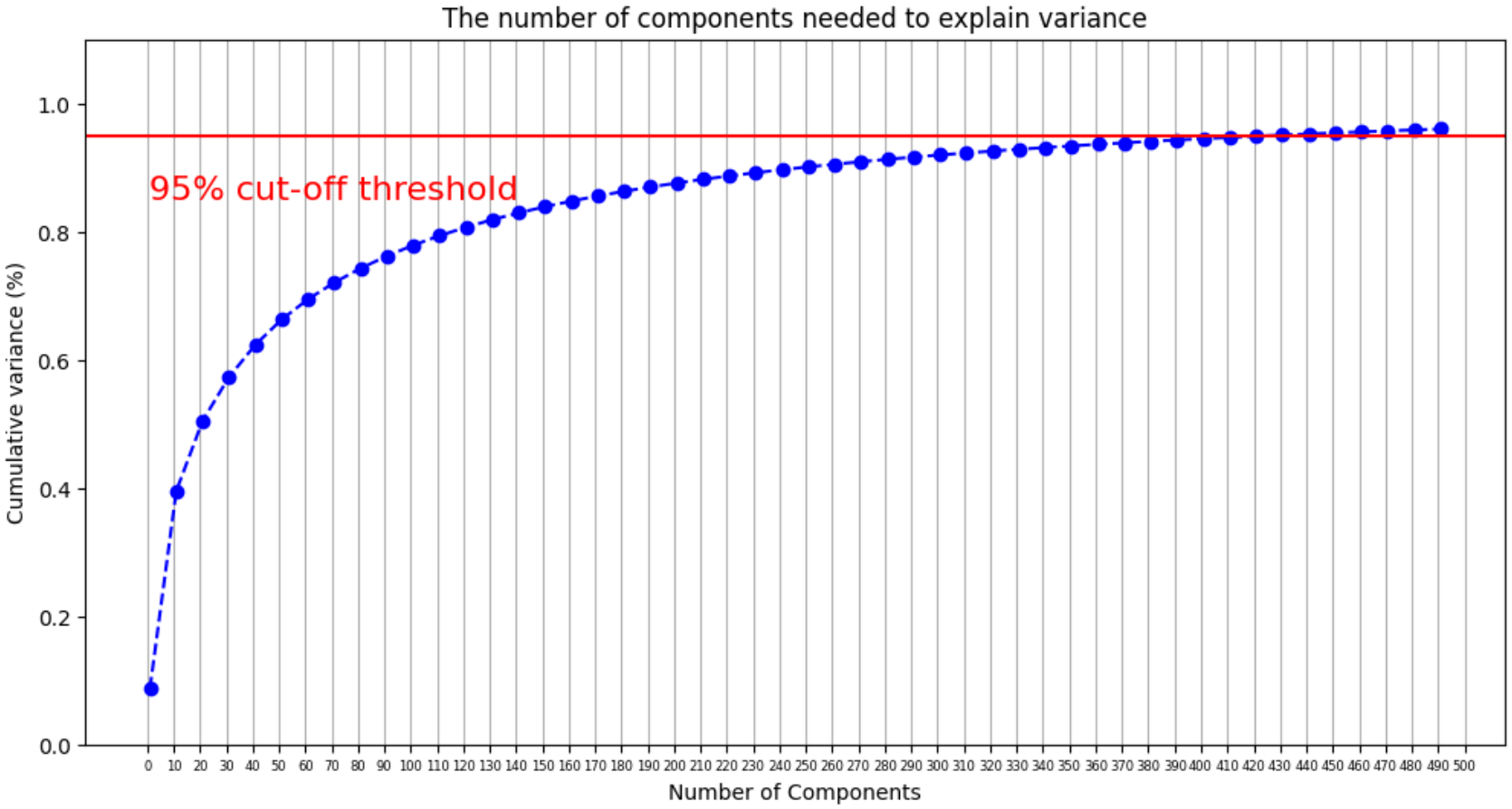
- Input features: influencer embeddings using average values.



Principal Component Analysis : Text



Principal Component Analysis: Text & Image



PySpark Models

- Logistic Regression
- Random Forest
- Gaussian Naïve Bayes

PySpark Models

Logistic Regression

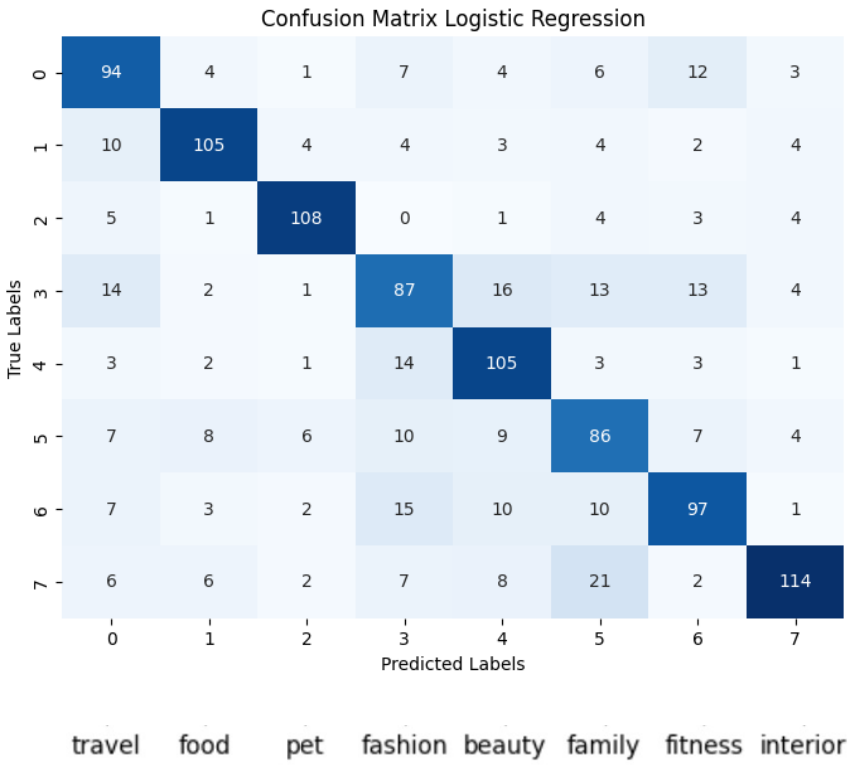
- Random Forest
- Gaussian Naïve Bayes

PySpark Models : Logistic Regression

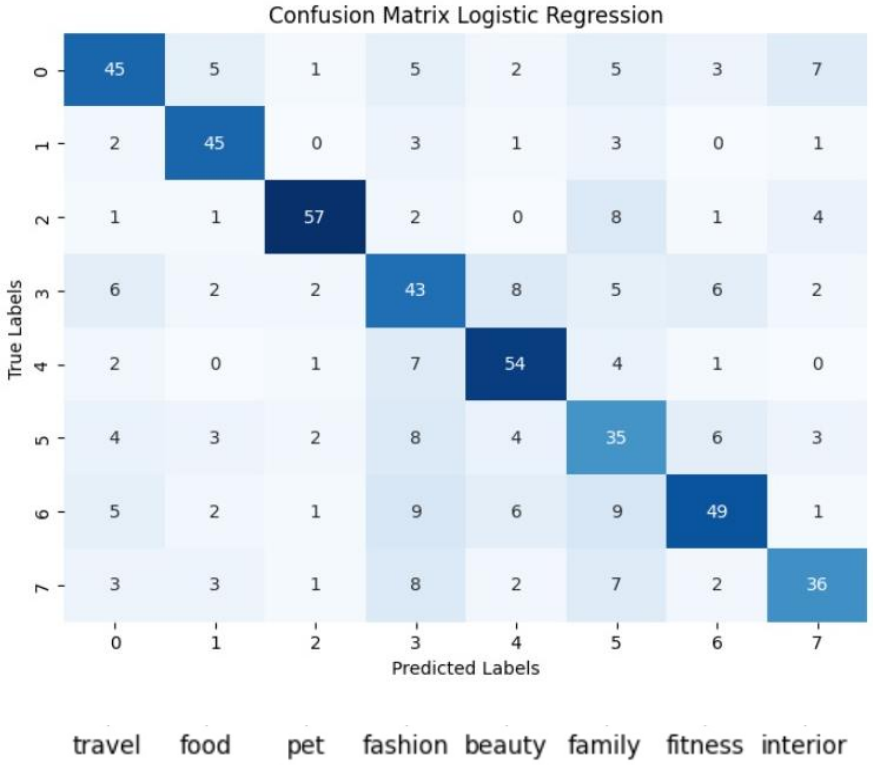
	Precision		Recall		F1 Score		Accuracy	
	Text	Text & Image	Text	Text & Image	Text	Text & Image	Text	Text & Image
Travel	0.64	0.66	0.72	0.62	0.68	0.64	71%	66%
Food	0.80	0.74	0.77	0.82	0.79	0.78		
Pet	0.86	0.88	0.86	0.77	0.86	0.82		
Fashion	0.60	0.51	0.58	0.58	0.59	0.54		
Beauty	0.67	0.70	0.80	0.78	0.73	0.74		
Family	0.59	0.46	0.63	0.54	0.61	0.50		
Fitness	0.70	0.72	0.67	0.60	0.68	0.65		
Interior	0.84	0.67	0.69	0.58	0.76	0.62		

PySpark Models : Logistic Regression

Text



Text & Image



PySpark Models :

- Logistic Regression

Random Forest

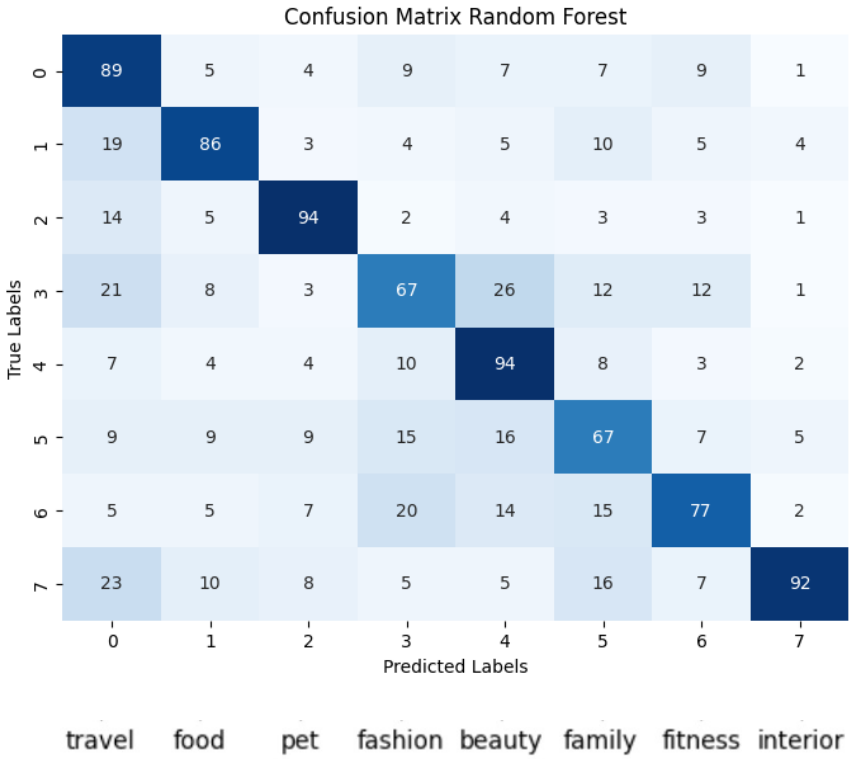
- Gaussian Naïve Bayes

PySpark Models : Random Forest

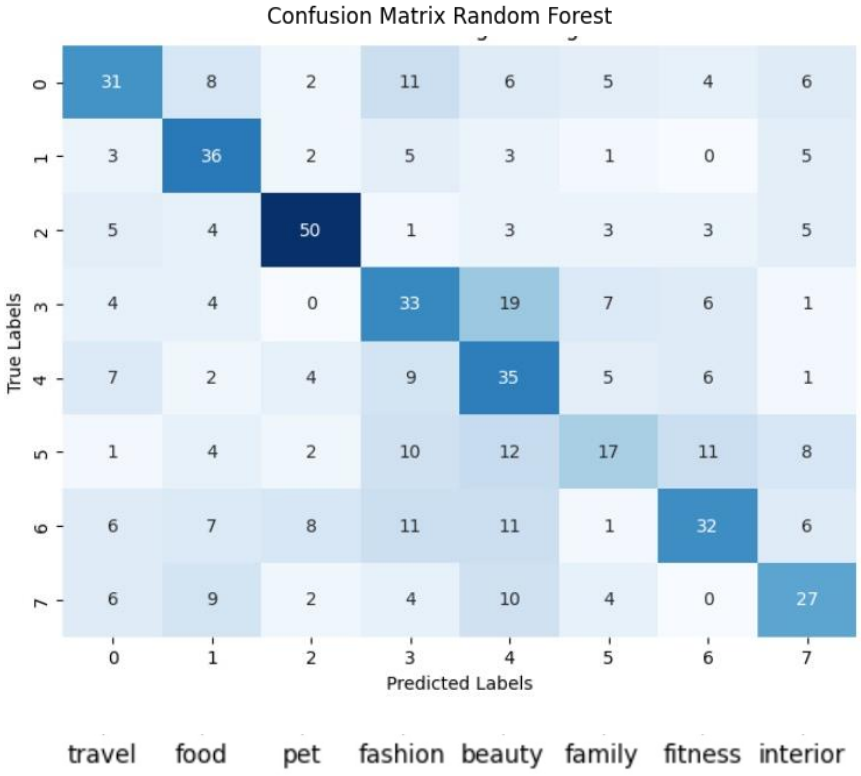
	Precision		Recall		F1 Score		Accuracy	
	Text	Text & Image	Text	Text & Image	Text	Text & Image	Text	Text & Image
Travel	0.48	0.49	0.68	0.42	0.56	0.46	59%	47%
Food	0.65	0.49	0.63	0.65	0.64	0.56		
Pet	0.71	0.71	0.75	0.68	0.73	0.69		
Fashion	0.51	0.39	0.45	0.45	0.48	0.42		
Beauty	0.55	0.35	0.71	0.51	0.62	0.42		
Family	0.49	0.40	0.49	0.26	0.49	0.31		
Fitness	0.63	0.52	0.53	0.39	0.57	0.44		
Interior	0.85	0.46	0.55	0.44	0.67	0.45		

PySpark Models : Random Forest

Text



Text & Image



PySpark Models :

- Logistic Regression
- Random Forest

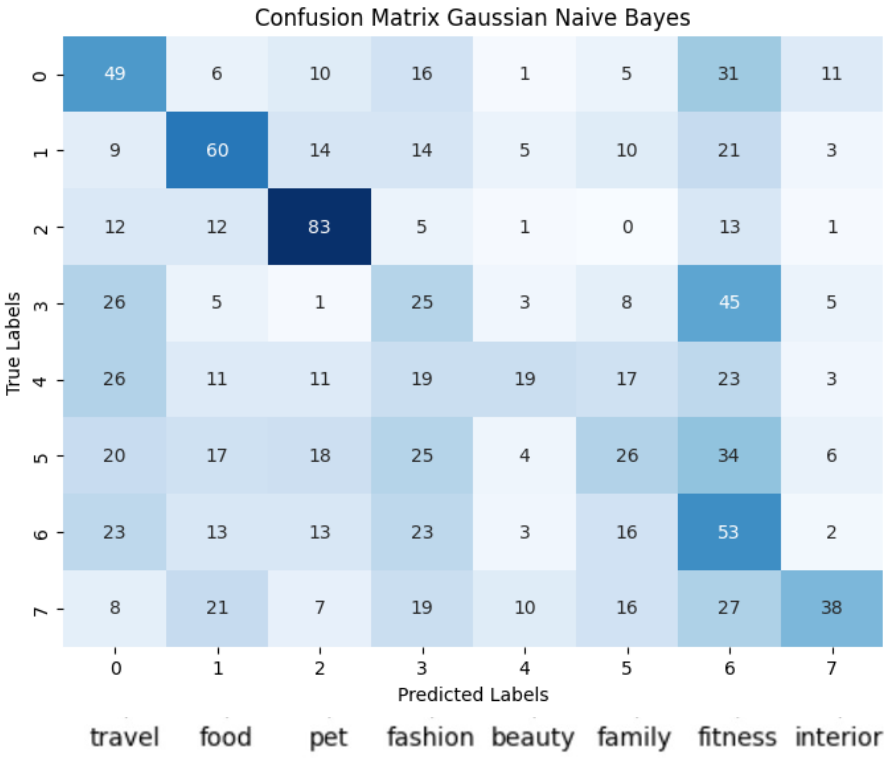
Gaussian Naïve Bayes

PySpark Models : Gaussian Naïve Bayes

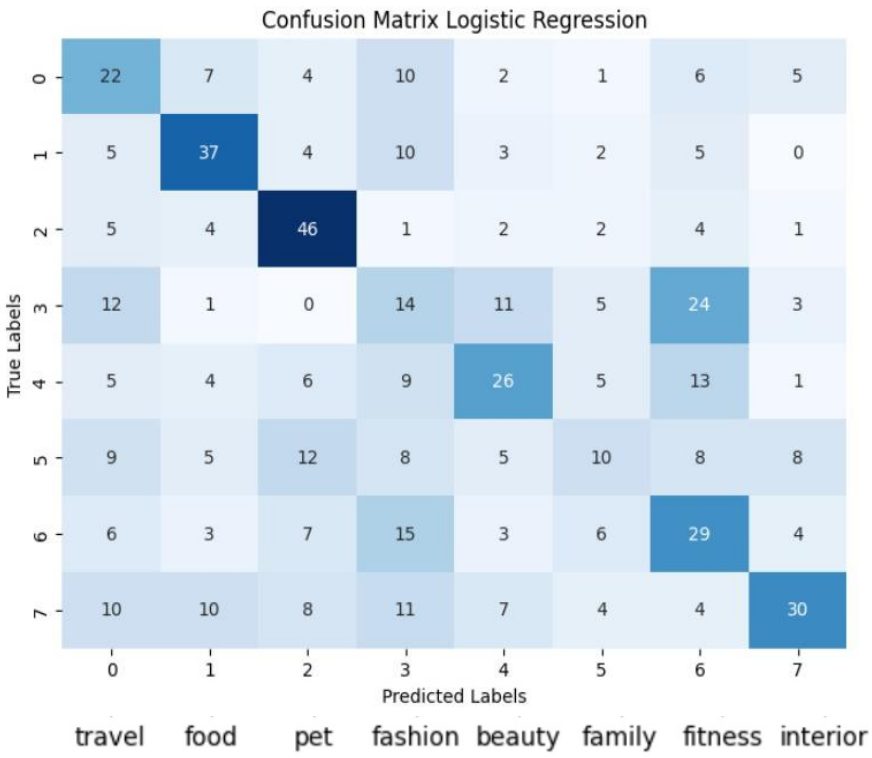
	Precision		Recall		F1 Score		Accuracy	
	Text	Text & Image	Text	Text & Image	Text	Text & Image	Text	Text & Image
Travel	0.28	0.30	0.38	0.39	0.32	0.34	33%	39%
Food	0.41	0.52	0.44	0.56	0.43	0.54		
Pet	0.53	0.53	0.65	0.71	0.58	0.61		
Fashion	0.17	0.18	0.21	0.20	0.19	0.19		
Beauty	0.41	0.44	0.15	0.38	0.22	0.41		
Family	0.27	0.29	0.17	0.15	0.21	0.20		
Fitness	0.21	0.31	0.36	0.40	0.27	0.35		
Interior	0.55	0.58	0.26	0.36	0.35	0.44		

PySpark Models : Gaussian Naïve Bayes

Text



Text & Image



Influencer Profiler

- All posts are not equally important to represent the category of the given influencer.

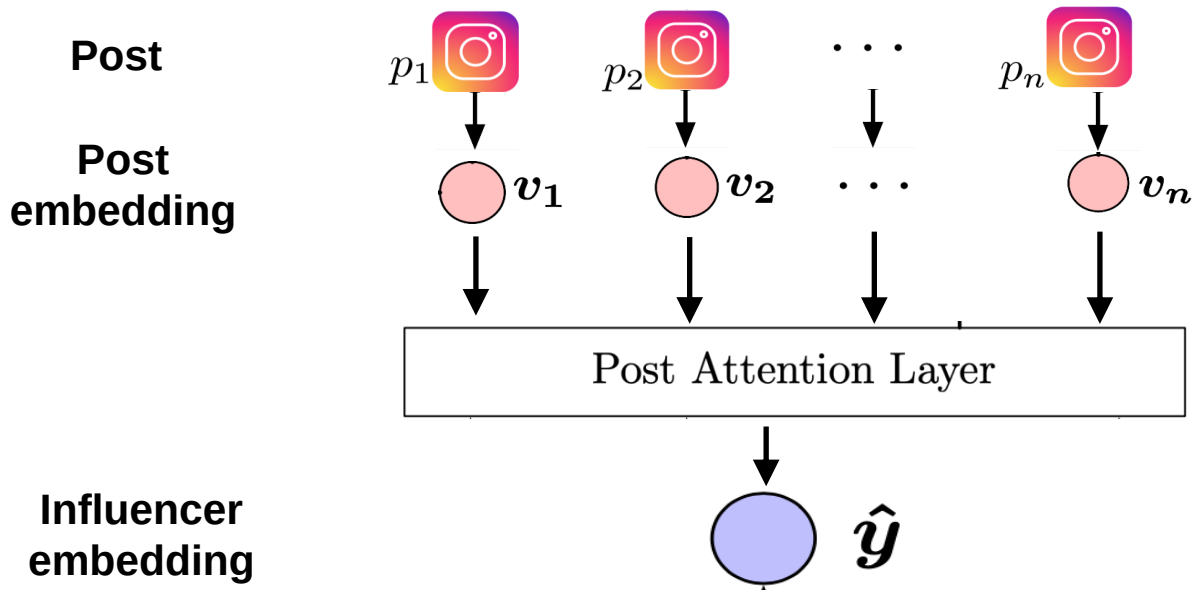
- An attention mechanism is applied to the feature maps to emphasize the most important parts of the plots.

- To estimate the importance α_i of each post with softmax function:
  - The more important the post, the higher the weight.

Influencer Profiler

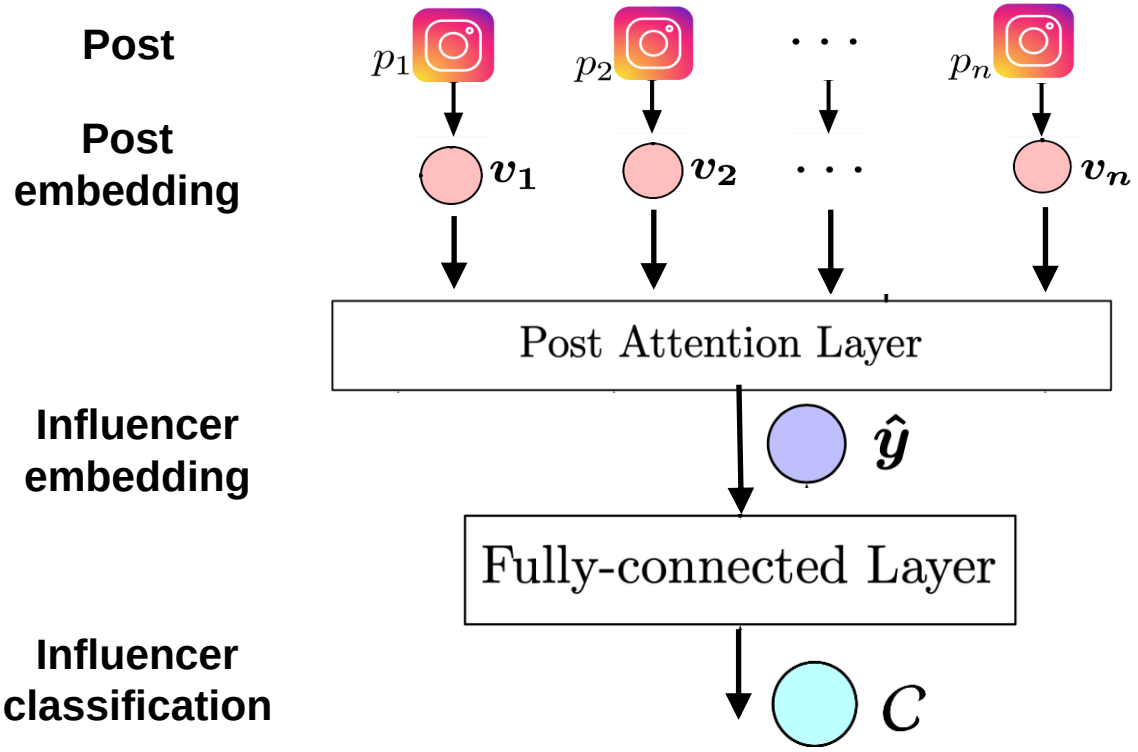
- **Influencer embedding:** a weighted combination of post features.

$$\hat{\boldsymbol{y}} = \sum_i \alpha_i \cdot \boldsymbol{v_i}.$$



Influencer Profiler

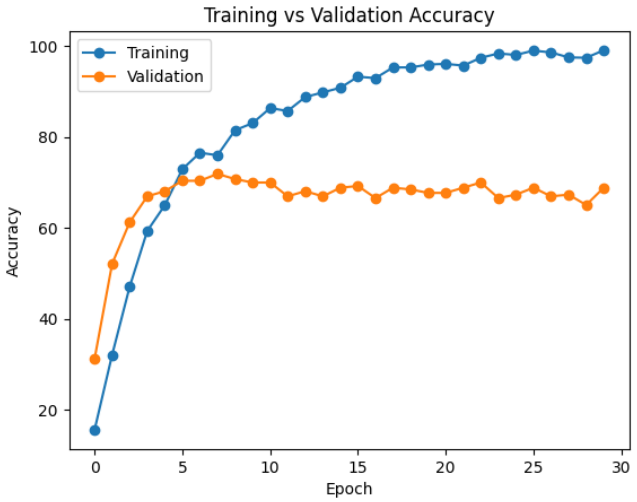
- **Influencer classification:** Influencer embeddings are passed through one or more fully connected layers to produce the final classification output.

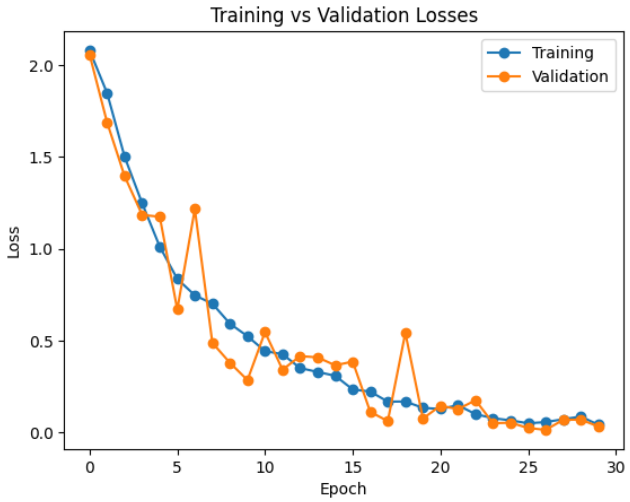


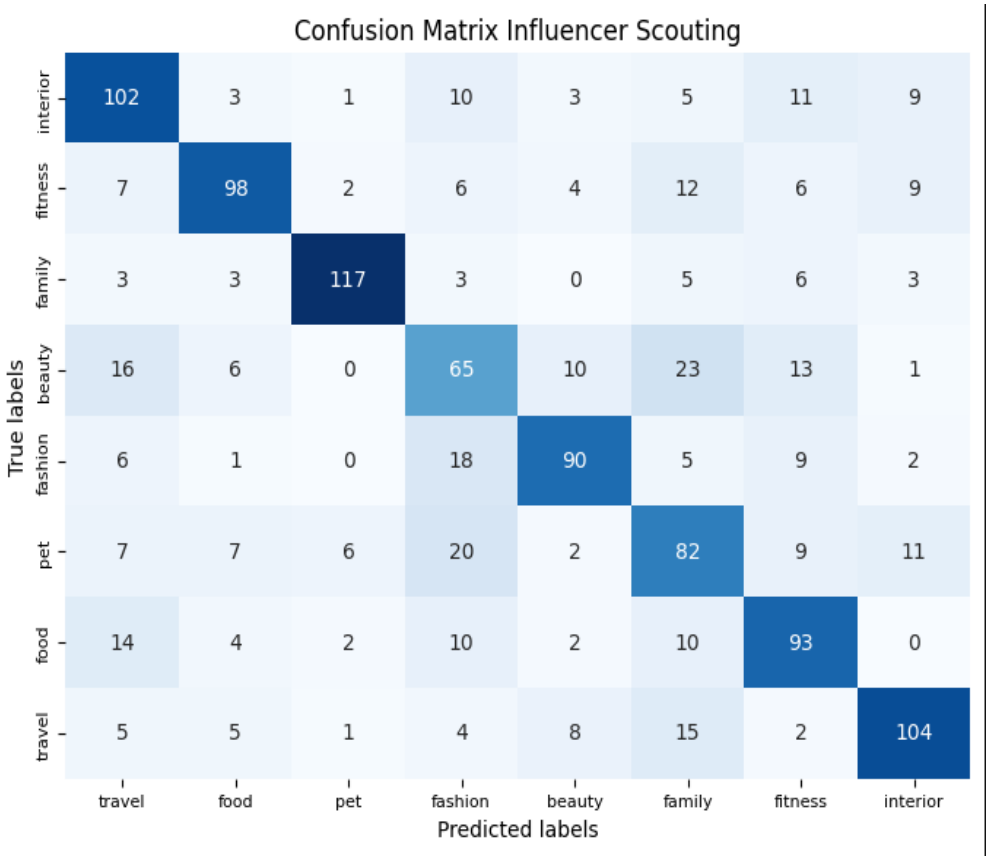
Influencer Profiler

	Precision		Recall		F1 Score		Accuracy	
	Text	Text & Image	Text	Text & Image	Text	Text & Image	Text	Text & Image
Travel	0.64	0.60	0.71	0.68	0.67	0.63	67%	60%
Food	0.77	0.71	0.68	0.64	0.72	0.67		
Pet	0.91	0.88	0.84	0.80	0.87	0.85		
Fashion	0.48	0.46	0.49	0.45	0.48	0.44		
Beauty	0.76	0.70	0.69	0.64	0.72	0.66		
Family	0.52	0.45	0.57	0.55	0.54	0.48		
Fitness	0.62	0.54	0.69	0.63	0.65	0.59		
Interior	0.75	0.67	0.72	0.67	0.73	0.68		

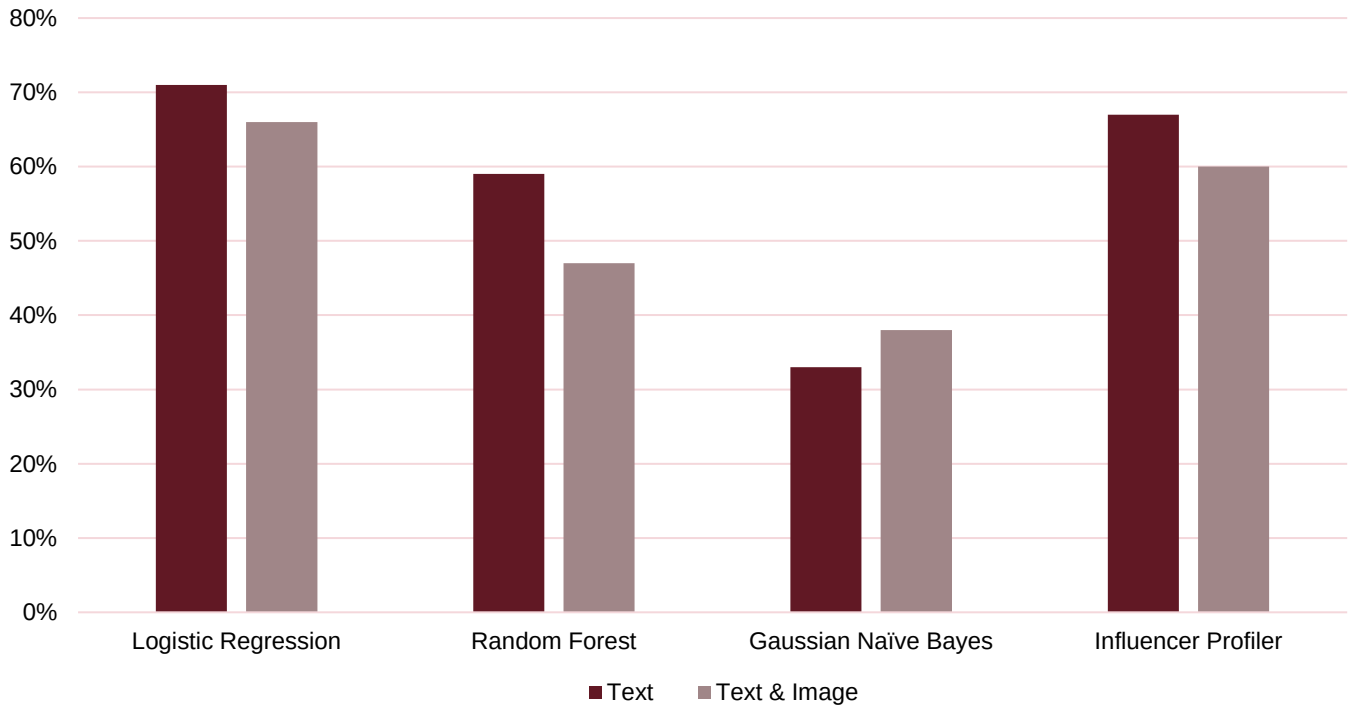
Influencer Profiler: Text





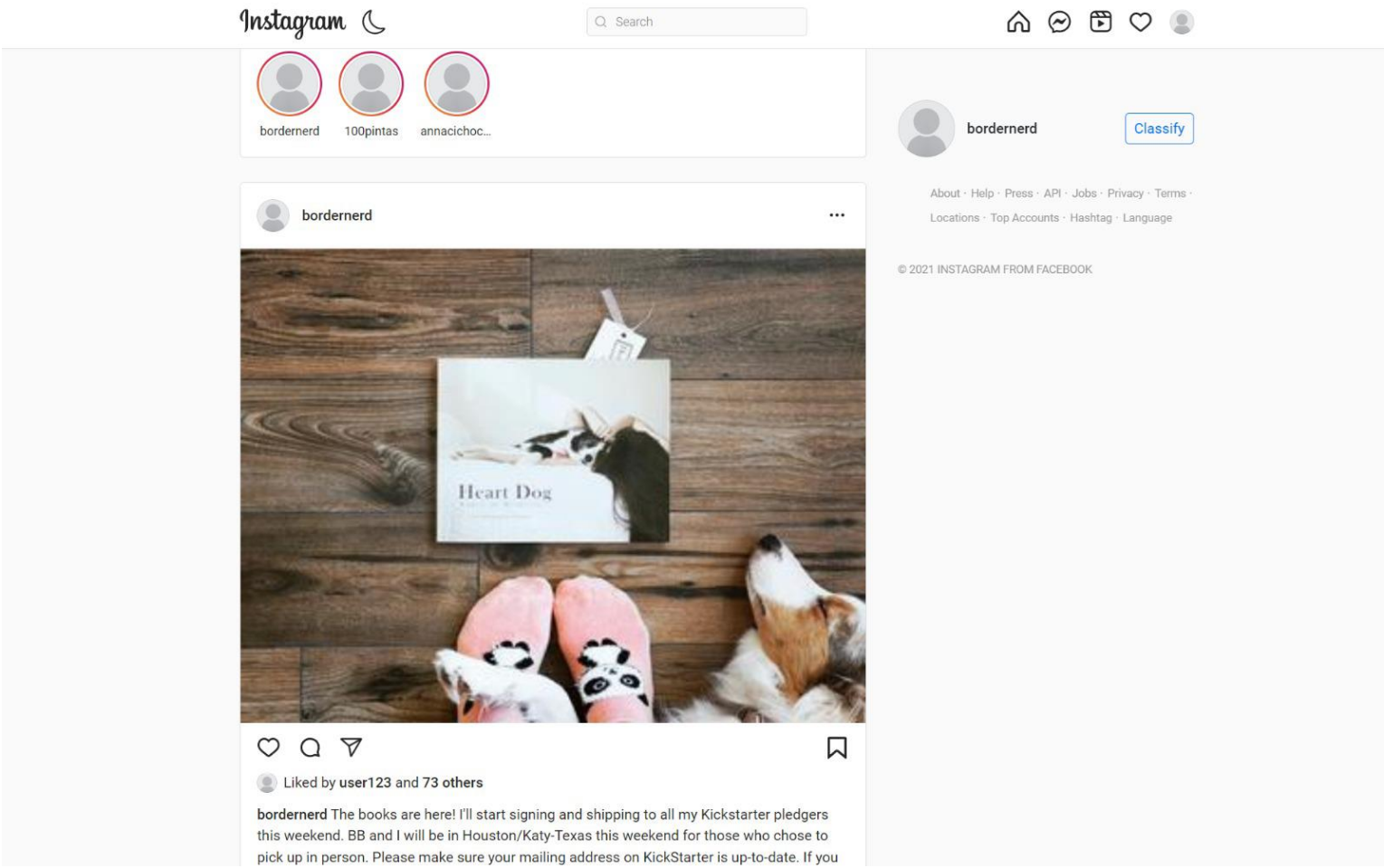


Comparison: Accuracy

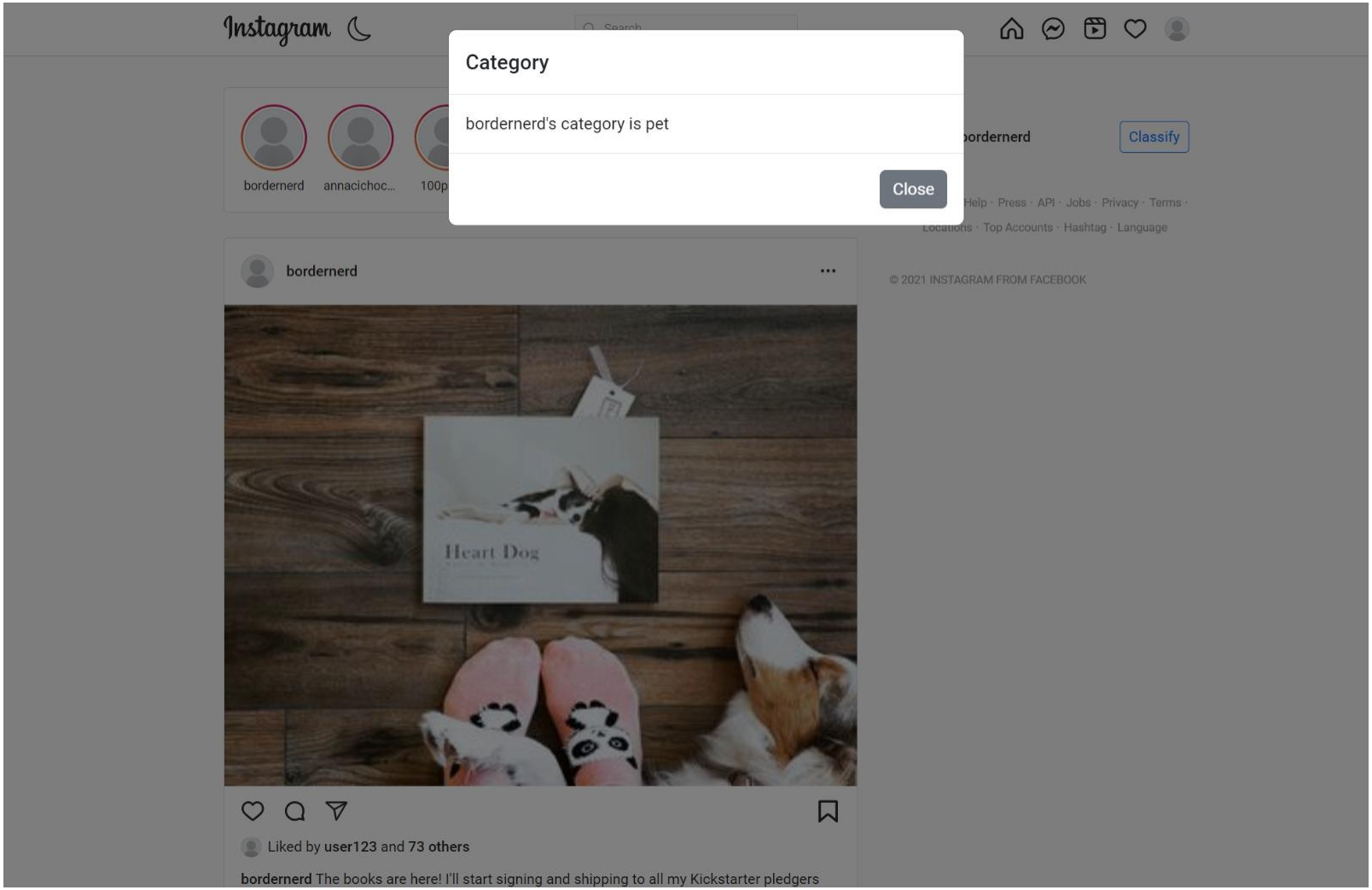


5 App Demo

App Demo


Instagram
Search
bordernerd
100pintas
annacichoc...
bordernerd
Heart Dog
Liked by user123 and 73 others
bordernerd The books are here! I'll start signing and shipping to all my Kickstarter pledgers this weekend. BB and I will be in Houston/Katy-Texas this weekend for those who chose to pick up in person. Please make sure your mailing address on KickStarter is up-to-date. If you
bordernerd
Classify
About · Help · Press · API · Jobs · Privacy · Terms · Locations · Top Accounts · Hashtag · Language
© 2021 INSTAGRAM FROM FACEBOOK

App Demo


Instagram
Category
bordernerd's category is pet
Close
bordernerd
annacichoc...
bordernerd
Classify
Help · Press · API · Jobs · Privacy · Terms ·
Locations · Top Accounts · Hashtag · Language
© 2021 INSTAGRAM FROM FACEBOOK
bordernerd
Heart Dog
Liked by user123 and 73 others
bordernerd The books are here! I'll start signing and shipping to all my Kickstarter pledgers

6 Future Works

Future Works

- Develop a user-friendly platform for brands to easily access and analyze influencer data.

- Refine the models to improve accuracy.

Thank you for your attention

SAPIENZA
UNIVERSITÀ DI ROMA

Federico Barreca
Shuya Dong