Automatic Prediction of Owned Car Brand Based on Tweets Data

Wei Xu, Renhan Zhang, Weizi Liu, Bohan Zhao

CONTENT

- Goal
- Description
- Dataset
- Methods & Results
- Analysis
- Future Work
- Application

Goal

Goal

Car Brand Prediction

Predict user's car brand ownership based on his/her Twitter tweets

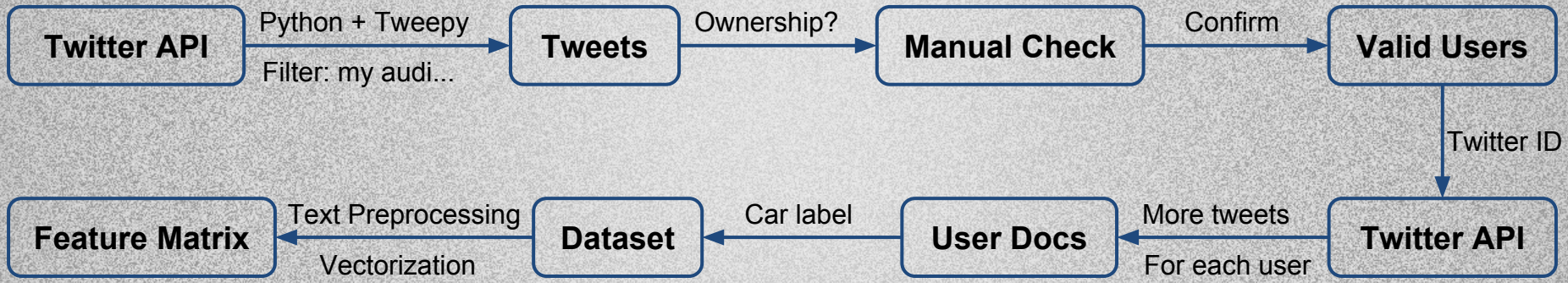
Understand differences among people driving different cars

Want to know what car someone is driving? Give us his/her Twitter ID !

Description

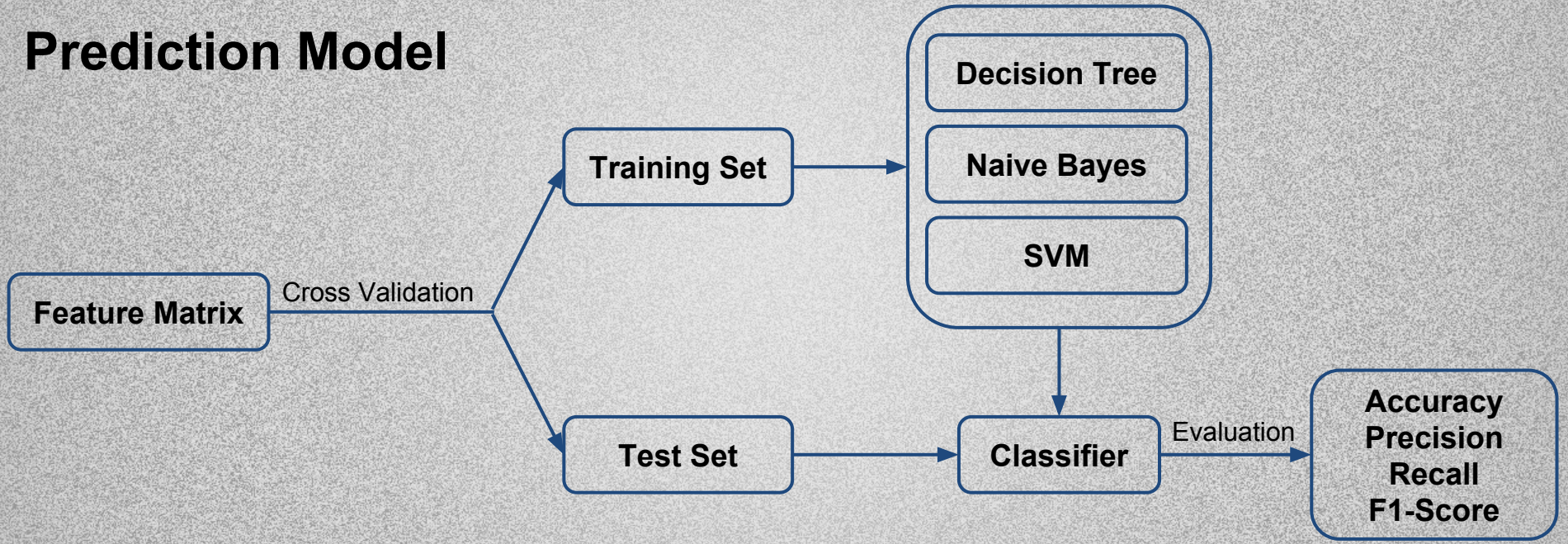
Description

Data Collection



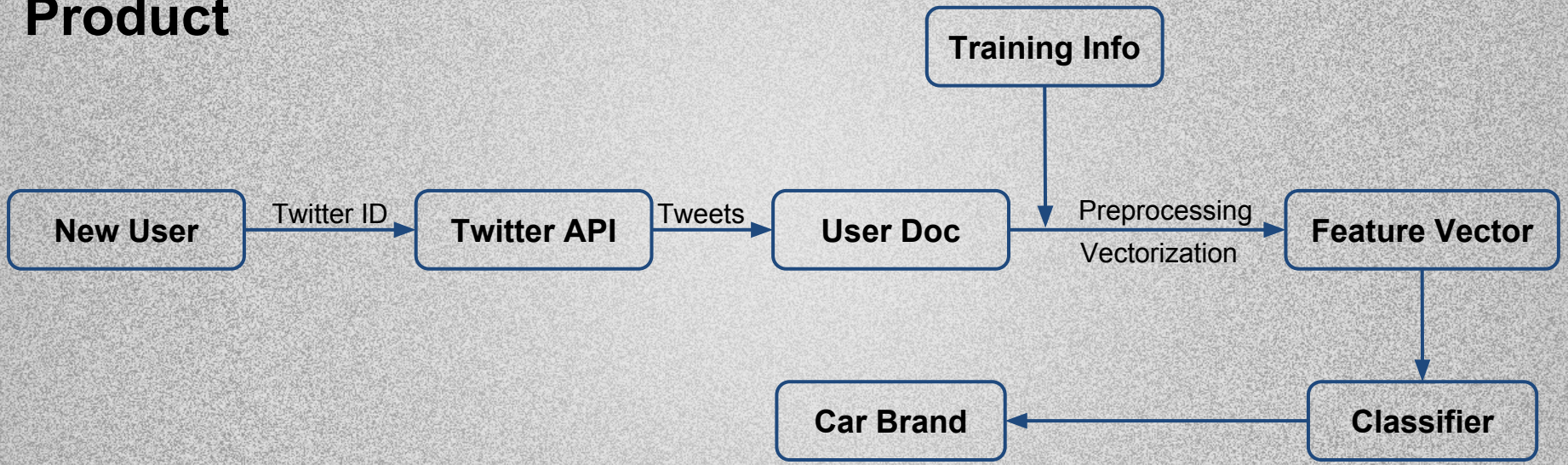
Description

Prediction Model



Description

Product



Dataset

Dataset

Twitter API

Crawling tweets that contain the keywords (Filter: my + *car brand*...)
Data form: User ID | car brand | tweet

Manual Check

Manually check the tweet which indicates user's car ownership

Twitter API

Crawling 200 more tweets for each qualified user
Combine the tweets as a document(data sample) for each user

Dataset

Data Statistics(Total 1813 samples, 12 car labels)

Jeep	Nissan	Honda	Toyota	Hyundai	Ford	Mazda	BMW	Lexus	KIA	Audi	Ferrari
28%	9%	7%	8%	2%	6%	5%	5%	6%	4%	13%	2%

Data Format

- ➔ label: car brand
- ➔ data: combined user tweets

Baseline Accuracy: 0.28

i'm at hilton garden inn atlanta north point in alpharetta, ga
having lunch with a very hip lady. sal's italian restaurant in burlington, nc)
getting my haircut because someone at home said i look like an old man with a comb over. supercuts)
getting an oil change for my rogue carolina nissan in burlington, nc)
can you all guess how many at tonight's dance? only has the answer!#winthesejellybeans
friends forever #fallonticketfriday
my favorite strategy game is the professor layton series of games. #hopetowintheamiibo
how do you enter to win this?
played some #codenamesteam and loving it so far. #nintendo #nintendo3dsxl
why do i always have to ask you to go get things from your back room on release day ? #areyouremployeeslazy
mozart is visiting me
...

Methods & Results

Preprocessing

Scikit Learn

Built in function: TfidfVectorizer

CMU Tokenizer

Using tokenizer from Tweet NLP of CMU, detail refers to

http://www.ark.cs.cmu.edu/TweetNLP/#pos

Model & Result

Decision Tree

5-fold Cross Validation (StratifiedKFold)

Evaluation: accuracy, precision, recall, f1-score

Times	precision	recall	f1-score	accuracy
1	0.18	0.19	0.19	0.19
2	0.23	0.22	0.22	0.22
3	0.17	0.16	0.16	0.16
4	0.23	0.23	0.23	0.23
5	0.26	0.26	0.26	0.26
Average	0.214	0.212	0.212	0.212

Baseline Accuracy: 0.28

Car brand	Precision	Recall	F1-socre	Support
BMW	0.08	0.10	0.09	21
Honda	0.28	0.17	0.21	29
Jeep	0.41	0.40	0.41	102
Audi	0.23	0.24	0.24	50
Ford	0.17	0.18	0.18	22
Hyundai	0.09	0.10	0.10	10
KIA	0.00	0.00	0.00	18
Lexus	0.16	0.17	0.17	23
Mazda	0.05	0.05	0.05	20
Nissan	0.13	0.15	0.14	34
Toyota	0.20	0.21	0.20	29
Ferrari	0.00	0.00	0.00	9
avg/total	0.23	0.22	0.22	367

Model & Result

Naive Bayes

5-fold Cross Validation (StratifiedKFold)

Evaluation: accuracy, precision, recall, f1-score

Times	precision	recall	f1-score	accuracy
1	0.35	0.33	0.23	0.33
2	0.24	0.32	0.21	0.32
3	0.30	0.36	0.26	0.36
4	0.31	0.34	0.21	0.33
5	0.27	0.34	0.22	0.33
Average	0.294	0.338	0.226	0.334

Baseline Accuracy: 0.28

Car brand	Precision	Recall	F1-socre	Support
BMW	0.67	0.10	0.17	21
Honda	0.50	0.03	0.06	29
Jeep	0.33	0.96	0.49	102
Audi	0.35	0.24	0.26	50
Ford	0.00	0.00	0.00	22
Hyundai	0.00	0.00	0.00	10
KIA	0.00	0.00	0.00	18
Lexus	1.00	0.04	0.08	23
Mazda	0.00	0.00	0.00	20
Nissan	0.50	0.09	0.15	34
Toyota	0.31	0.17	0.22	29
Ferrari	0.00	0.00	0.00	9
avg/total	0.35	0.33	0.23	367

Model & Result

Support Vector Machines (SVM)

5-fold Cross Validation (StratifiedKFold)

Evaluation: accuracy, precision, recall, f1-score

Times	precision	recall	f1-score	accuracy
1	0.41	0.40	0.31	0.40
2	0.40	0.41	0.32	0.41
3	0.42	0.44	0.35	0.44
4	0.42	0.40	0.30	0.40
5	0.31	0.39	0.29	0.39
Average	0.392	0.408	0.314	0.408

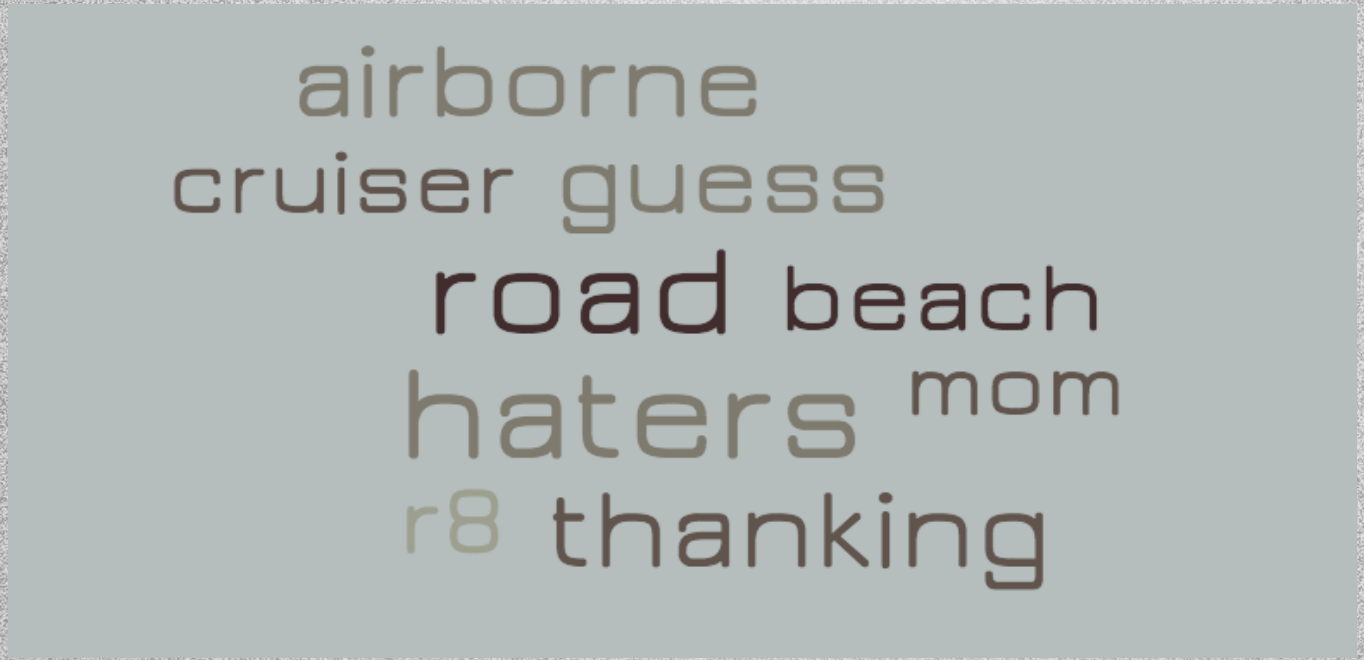
Baseline Accuracy: 0.28

Car brand	Precision	Recall	F1-socre	Support
BMW	0.50	0.10	0.16	21
Honda	0.60	0.11	0.18	28
Jeep	0.44	0.95	0.60	102
Audi	0.49	0.63	0.55	49
Ford	0.29	0.09	0.14	22
Hyundai	0.00	0.00	0.00	10
KIA	0.00	0.00	0.00	17
Lexus	0.00	0.00	0.00	22
Mazda	1.00	0.05	0.10	19
Nissan	0.33	0.41	0.37	34
Toyota	0.62	0.34	0.44	29
Ferrari	0.00	0.00	0.00	8
avg/total	0.42	0.44	0.35	361

Analysis

Analysis

Distinguishing Words



Analysis

Most Frequent Words

Audi: audi, like, good

BMW: BMW, time, change

Ferrari: race, play, love

Ford: new, can, ford, people, work

Honda: best, work, thank

Hyundai: reality, work, life

Jeep: don’t, girl, lol, want

Kia: happy, think, best

Future Work

Linguistic Analysis

Finding dominant word classes of each car brand

Foreground: one chosen car label

Background: the other car labels

$C = \{W_1, W_2, \ldots, W_n\}$

$$Coverage_F = \frac{\sum_{W_i \in C} Frequency(W_i)}{Size(F)}$$

$$Coverage_B = \frac{\sum_{W_i \in C} Frequency(W_i)}{Size(B)}$$

$$Dominance_F = \frac{Coverage_F(C)}{Coverage_B(C)}$$

Application

Commercial Applications

Identify people's opinions on certain car brand, providing consumer feedback based on social media for car companies

Identify the Twitter users driving expensive cars, and send them related luxury ads or car insurance ads accordingly.

Thanks