`r/cars` vs. `r/electricvehicles` :

Rivian's Classification Problem

by Stanley Azuakola

Problem Overview:

At **Rivian**, our Comms team tracks two subreddits: **r/cars** and **r/electricvehicles**.

We have a batch of unlabeled data from those subreddits that we do not want to discard.

I have been asked to do the following:

- ❑ Collect more data
- ❑ Train a classifier on which subreddit a given post came from.
- ❑ In future, use my model to classify the unlabeled posts
- ❑ Make a presentation to my boss and the team outlining my **process** and **findings**.

Metrics.

What the boss said:

We would put this model into production if it achieves at least one of these 2 things:

❑ An accuracy of at least 0.90. We can only afford to misclassify 1 out of 10 posts.

❑ An F1 score of at least 0.90.

Why?

Process.

Data Collection

Target:

- Collect 10,000 observations (5k from each subreddit) using Pushshift API

Outcome:

- 5,000 from r/cars
- 3,660 from r/electricvehicles

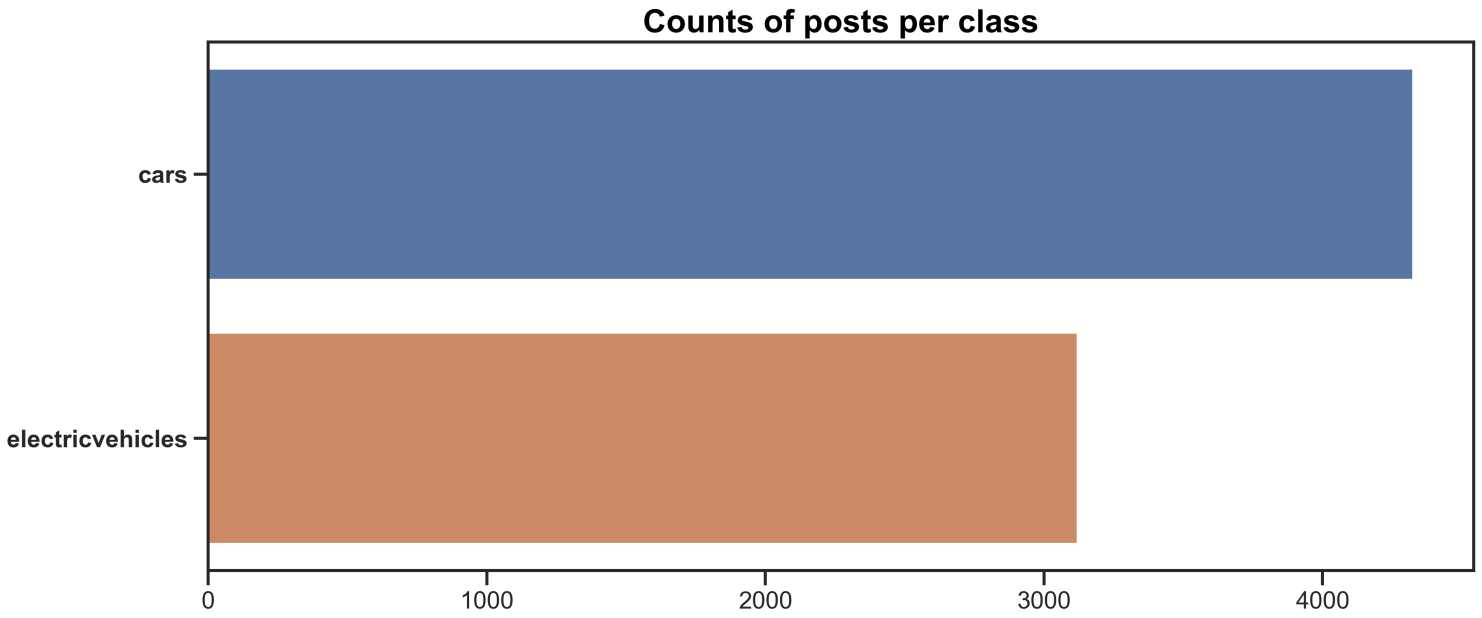
Data Cleaning

[5]:

	subreddit	title	id
481	u_daleelsayarat-cars	إيجاد اقرب محطة ديزل من موقعي بالآيفون والأندر...	108c1uk
608	u_alaimran-cars	Location de voiture Agadir AlAimran Cars	107lcjy
612	u_alaimran-cars	Al Aimran Cars : Location de voiture Agadir pa...	107l3dd
637	u_daleelsayarat-cars	مرسيدس جي كلاس 2022 من الداخل	107gnvl
638	u_daleelsayarat-cars	هذا هو الزيت الموصى به من شركة هيونداي توسان	107gf14
1457	u_daleelsayarat-cars	Is 5W30 Good For High Mileage Cars? The Comple...	1035540
1569	u_daleelsayarat-cars	إليك 6 عيوب سيارات الديزل التي ستحسم قرار شرائ...	102ibf8
3114	u_daleelsayarat-cars	أفضل منظف تابلوه السيارة 2022 كيف تختاره وكم سعره	zt9gb3

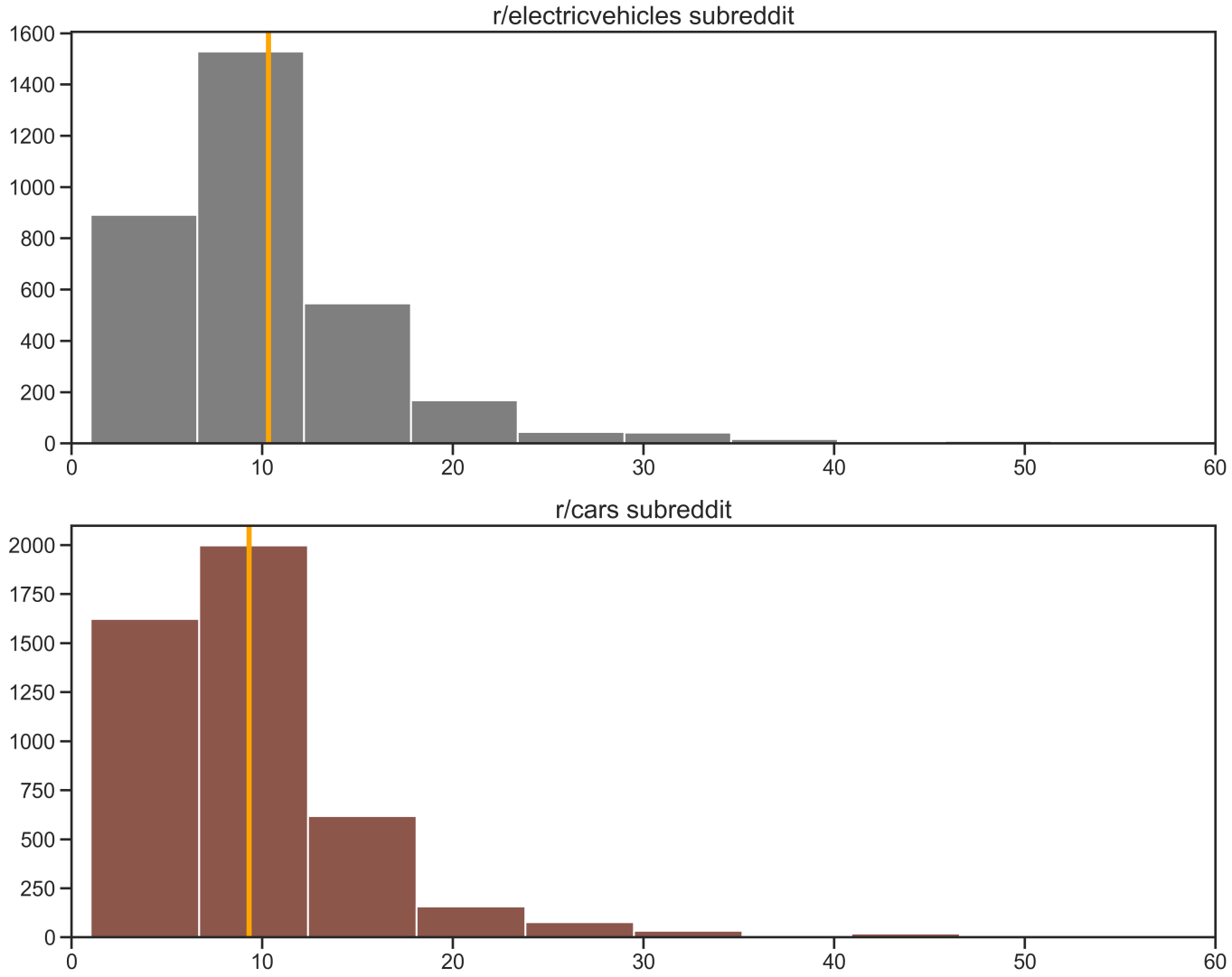
At least 3000 posts in each class

Baseline Accuracy: 58%

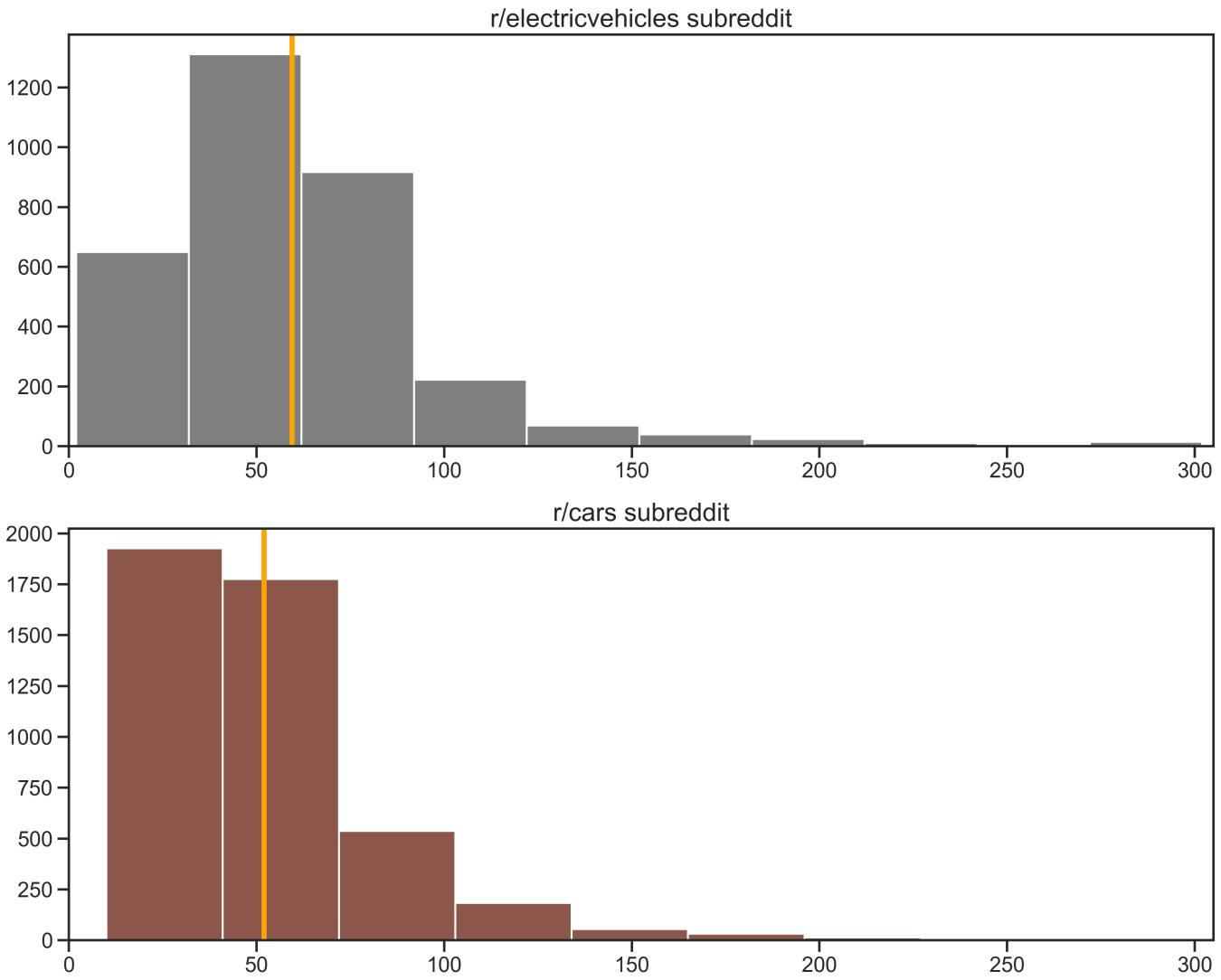


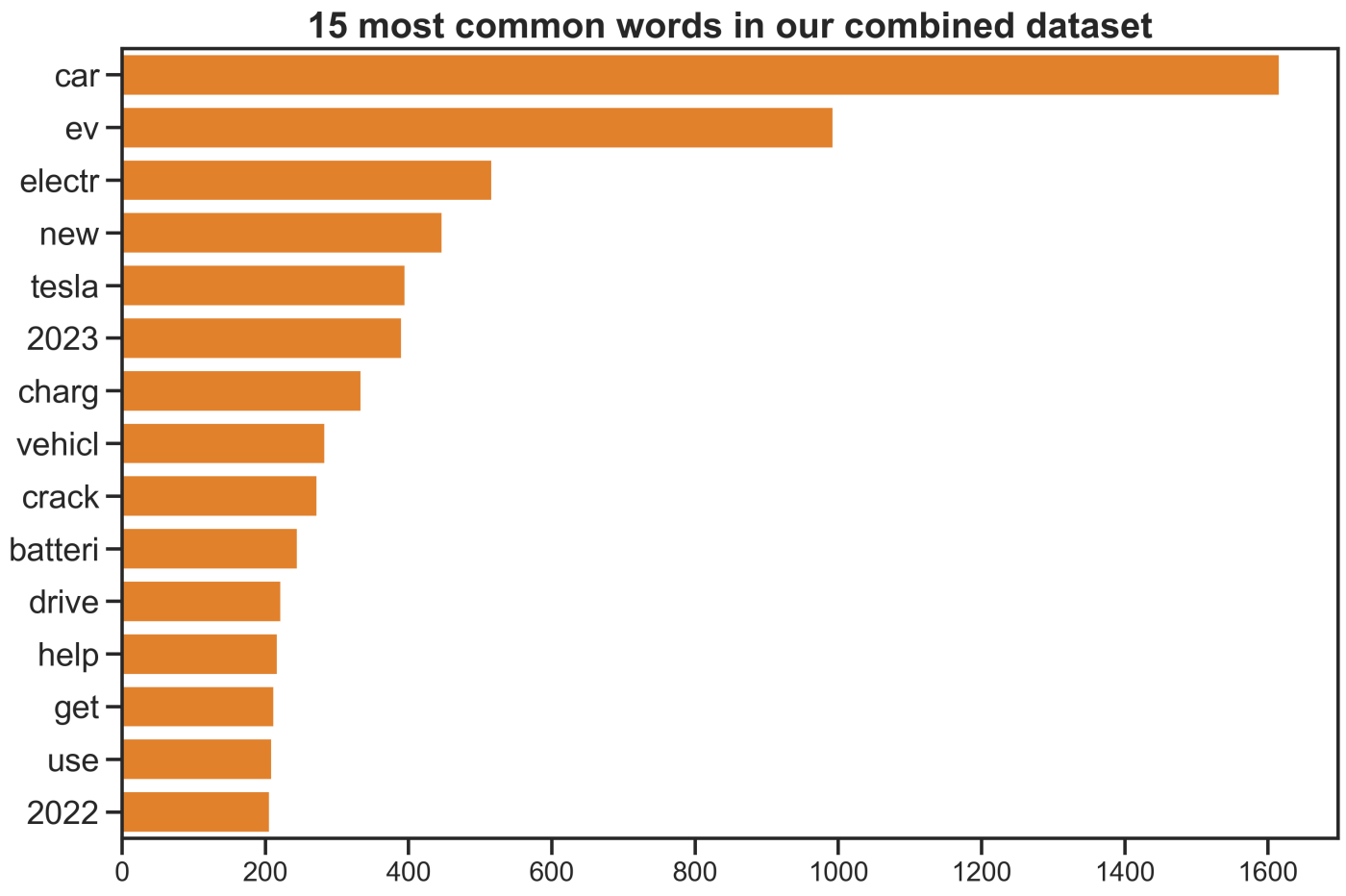
Exploratory Data Analysis

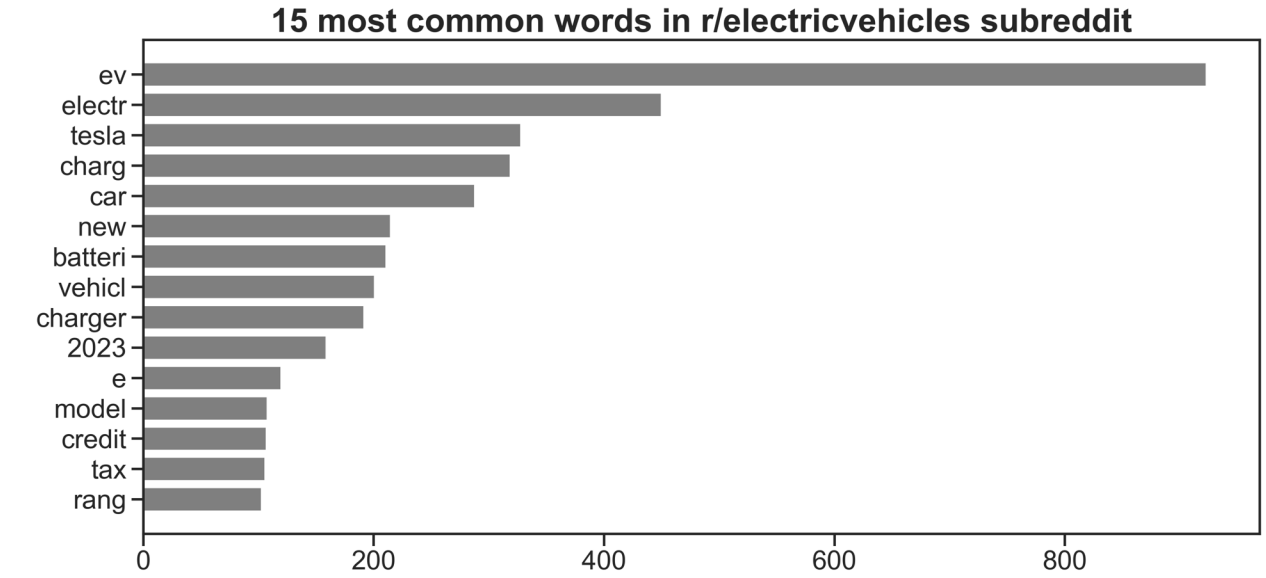
Distribution of word counts for the two subreddits

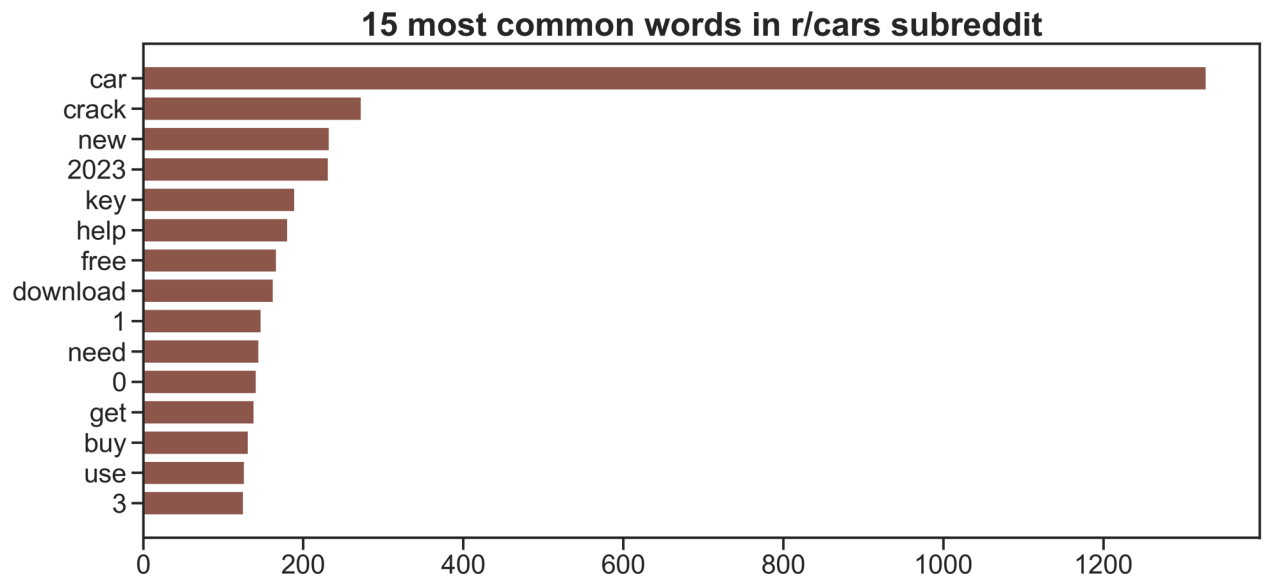


Distribution of character lengths for the two subreddits




15 most common words in our combined dataset
car
ev
electr
new
tesla
2023
charg
vehicl
crack
batteri
drive
help
get
use
2022
0
200
400
600
800
1000
1200
1400
1600


15 most common words in r/electricvehicles subreddit
ev
electr
tesla
charg
car
new
batteri
vehicl
charger
2023
e
model
credit
tax
rang
0
200
400
600
800


15 most common words in r/cars subreddit
car
crack
new
2023
key
help
free
download
1
need
0
get
buy
use
3
0
200
400
600
800
1000
1200

So should I drop those words that show up very often in my dataset?

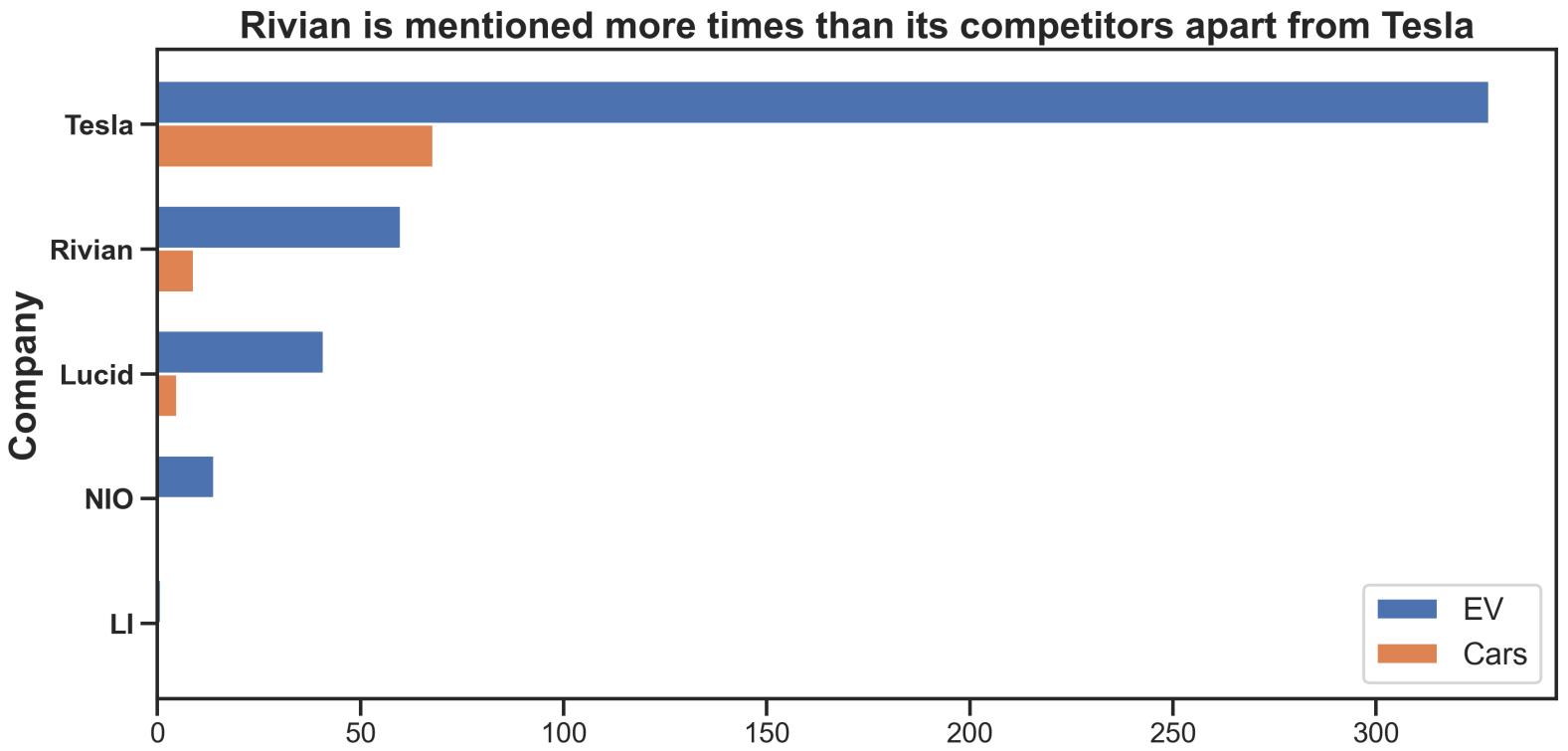
Words in Top 10 of both classes:

	r/cars	r/electricvehicles
car	1329	288
new	233	215
2023	232	159

The competition:

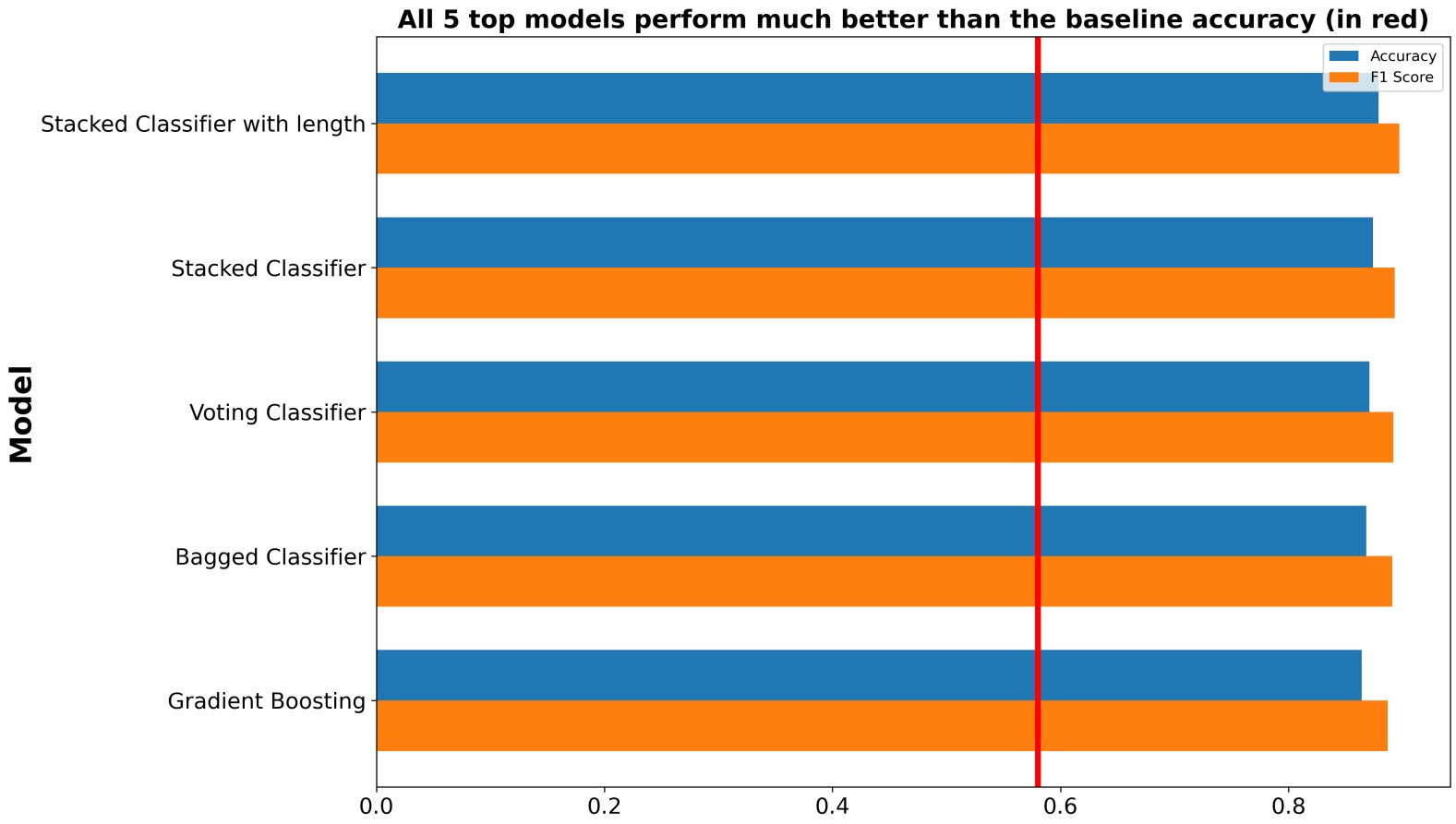
Rank	Company	MarketCap	Country
1	Tesla	$355.90bn	US
2	Li Auto	$19.52bn	China
3	NIO	$18.34bn	China
4	Rivian	$15.67bn	US
5	Lucid Motors	$10.72bn	US

Source: Capital.com

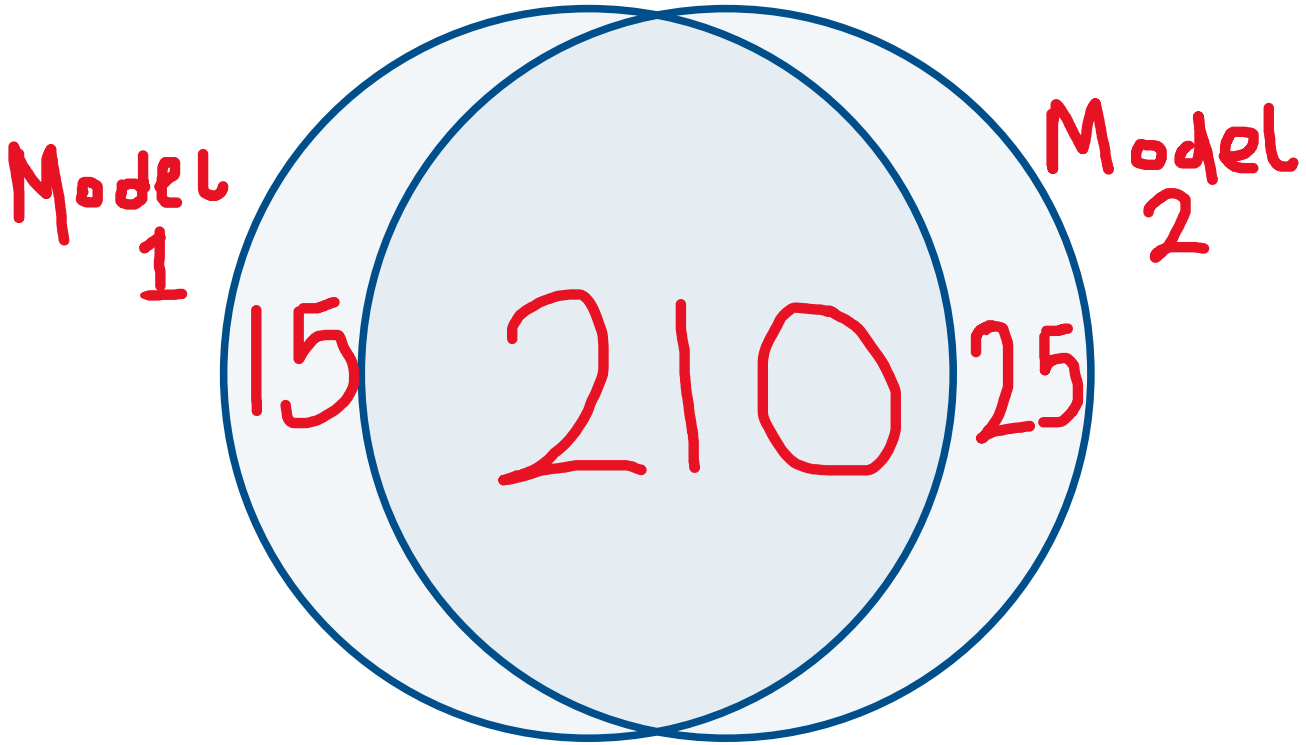

Rivian is mentioned more times than its competitors apart from Tesla
Company
Tesla
Rivian
Lucid
NIO
LI
0
50
100
150
200
250
300
EV
Cars

Findings.

Model No.	Model Type	Accuracy	Misclassification Rate	F1 Score
1	Logistic Regression	0.86	0.14	0.88
2	Random Forest	0.80	0.20	0.85
3	Multinomial Naïve Bayes	0.85	0.15	0.87
4	Bagged Classifier with LR Base Estimator	0.87	0.13	0.89
5	Bagged Classifier with RF base estimator	0.87	0.13	0.89
6	Gradient Boosting	0.86	0.14	0.89
7	Voting Classifier	0.87	0.13	0.89
8	Stacked Classifier	0.87	0.13	0.89
9	Stacked Classifier with character length feature	0.88	0.12	0.90


All 5 top models perform much better than the baseline accuracy (in red)
Accuracy
F1 Score
Model
Stacked Classifier with length
Stacked Classifier
Voting Classifier
Bagged Classifier
Gradient Boosting
0.0
0.2
0.4
0.6
0.8

What my top 2 models are misclassifying:



Misclassifications:

Model_1 = 225 misclassifications

Model_2 = 235 misclassifications

Model_1 ∩ Model_2 = 210

Let’s play a guessing game:

Between r/cars and r/electricvehicles…

Hello! I am a research student researching the opinions on electric cars.
Please help me by completing my 1-minute survey for this project.
Thank you.

For U.S. Companies, the Race for the New EV Battery Is On

Conclusions & Recommendations

Problem: Collect more data from the two subreddits, using Pushshift's API.

Outcome: I collected almost 9000 observations, giving me confidence about my models' abilities to solve the task.

Problem: Train a classifier on which subreddit a given post came from.

Outcome: I trained several high performing predictive models with accuracies far above the baseline.

Recommendations

- My Stacked Classifier which got the highest F1 score should be put into production. Meets the 0.90 F1 score criteria. Its accuracy of 88% and misclassification rate of 12 out of 100 are also solid enough to proceed with.
- If there was more time, I would love to incorporate the texts of the posts as well.

Thank you for listening