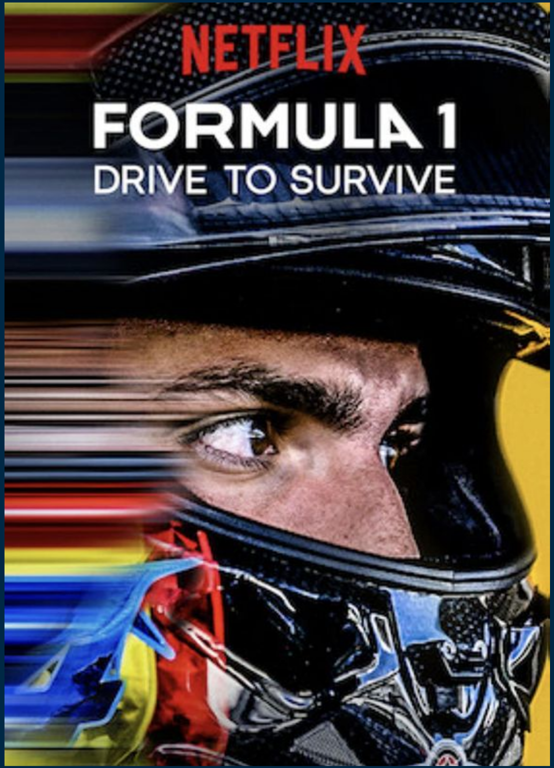
Building a Tweet Classifier (Binary)

Pang Hong Xiang

18th Mar 2023



…ed stock (NASDAQ: FWONA) has also
…argin, returning +62% since *Drive To*

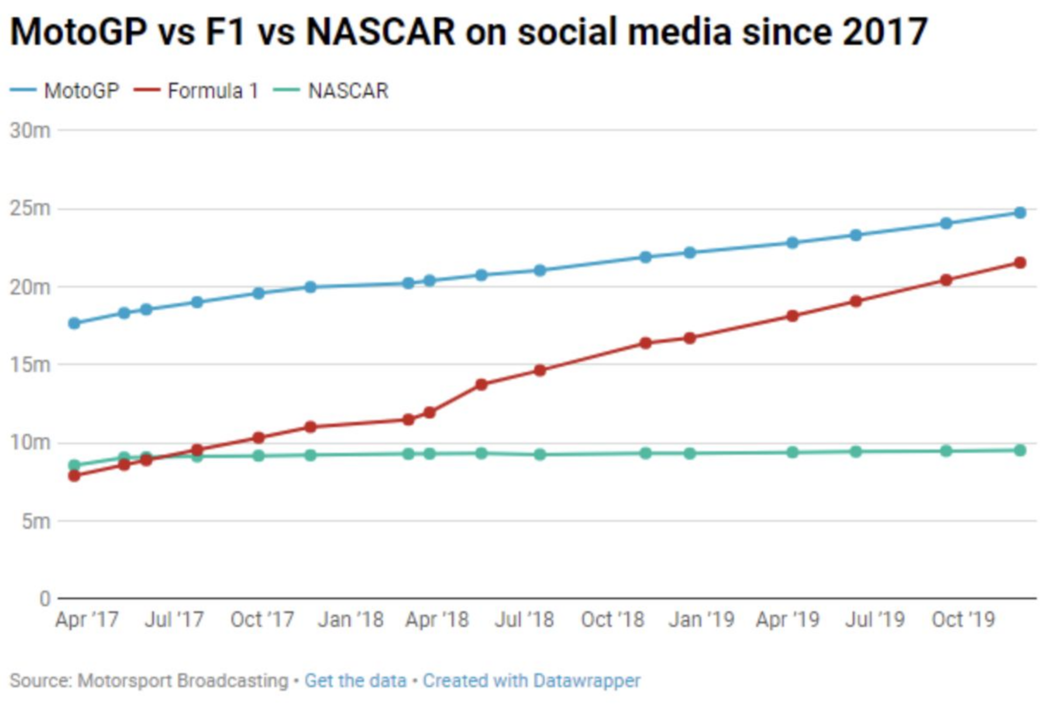
US Grand Prix attendance numb…

Formula 1's Libe…
…sed in price by …
…premiered…

attendance numbers …

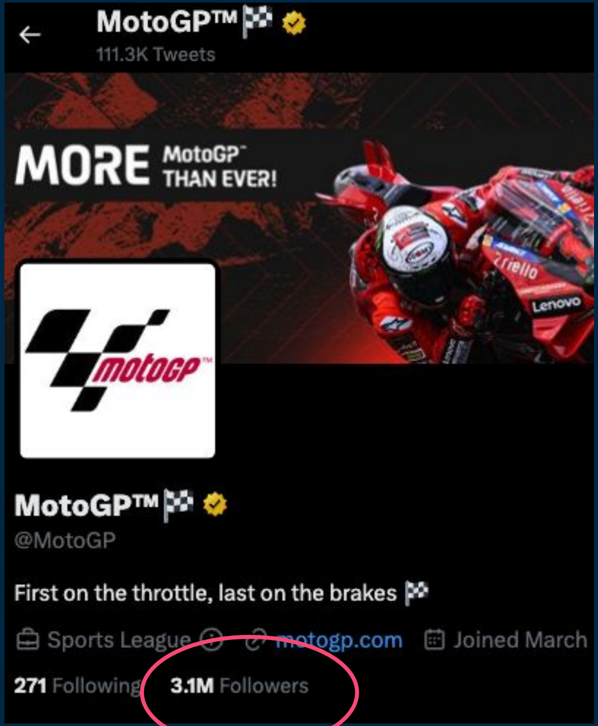
The cherry on top? Social media growth: 49 million followers with 40% growth annually and 1.5 billion aggregate social media engagements – in other words, all F1 content is practically viral content

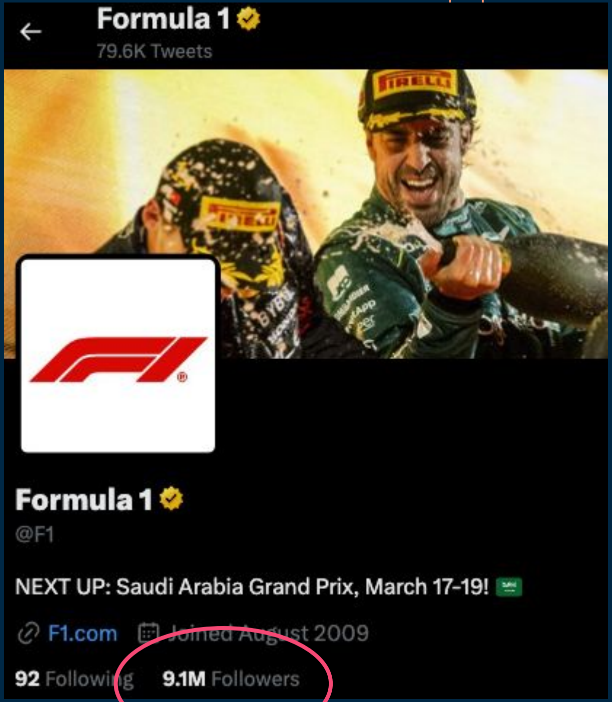
Races now average approximat…
2021 season (Abu Dhabi Grand Prix) beating Supe…
million simultaneous viewers vs 101 million simultaneous viewers
…Prix


MotoGP vs F1 vs NASCAR on social media since 2017
MotoGP
Formula 1
NASCAR
30m
25m
20m
15m
10m
5m
0
Apr '17
Jul '17
Oct '17
Jan '18
Apr '18
Jul '18
Oct '18
Jan '19
Apr '19
Jul '19
Oct '19
Source: Motorsport Broadcasting • Get the data • Created with Datawrapper


MORE
MotoGP
THAN EVER!

>


MotoGP™
111.3K Tweets
MORE MotoGP THAN EVER!
motoGP
MotoGP™
@MotoGP
First on the throttle, last on the brakes
Sports League
motogp.com
Joined March
271 Following
3.1M Followers


Formula 1
79.6K Tweets
Formula 1
@F1
NEXT UP: Saudi Arabia Grand Prix, March 17-19!
F1.com
Joined August 2009
92 Following
9.1M Followers

Understanding the Problem

↑ twitter users following topic of Formula 1

↑ auto-racing related accounts and tweets

↑ complicated to differentiate tweet topics

Fans are flocking to Twitter to discuss the twists and turns of motorsport's most prestigious competition.

To maximize relevant tweets on news feeds and minimize spam,

Build a classifier that is able to differentiate Formula 1 related tweets from MotoGP related tweets.

Problem-Solving Process

01

Data Collection, Exploration & Processing

02

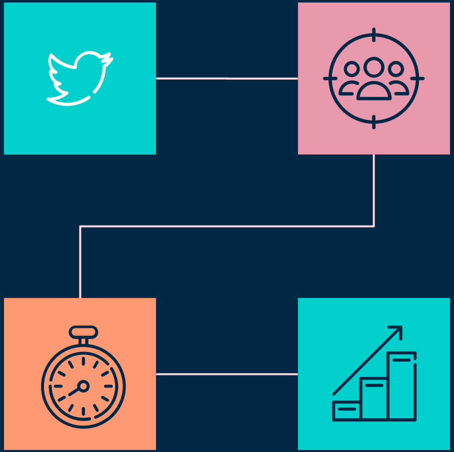
Modeling & Evaluation

03

Conclusion & Recommendations

Data Collection

19,850 tweets

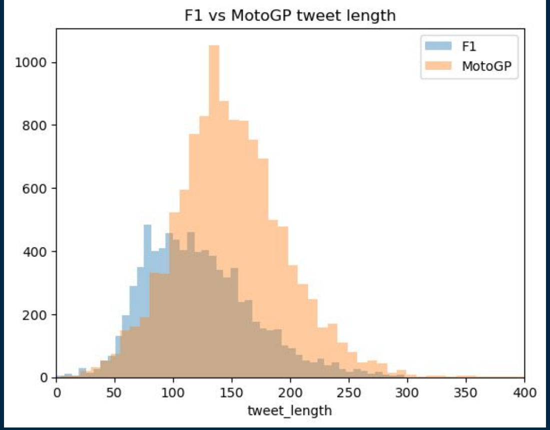


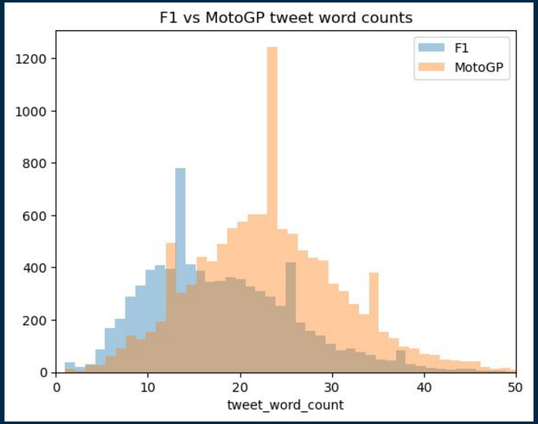
@F1 @MotoGP

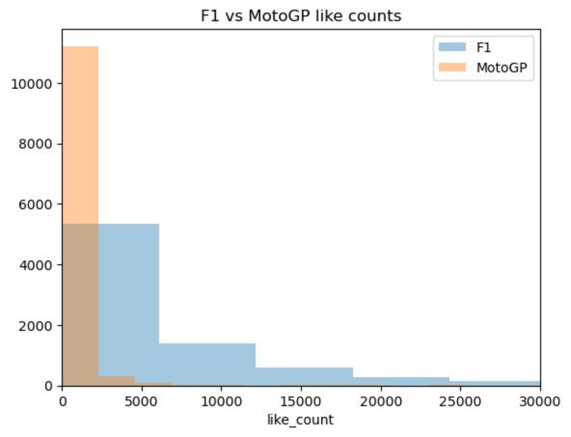
1st Jan 2022 –
28th Feb 2023

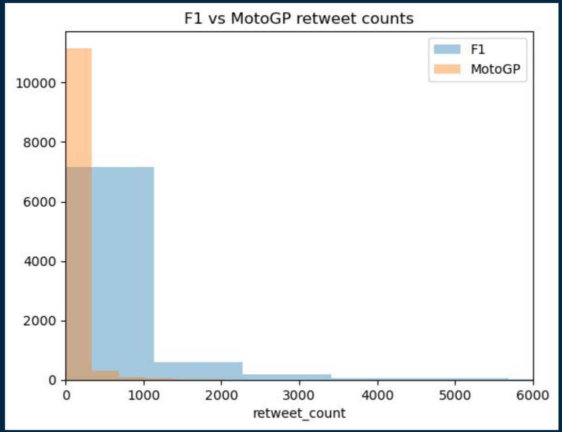
Tweets, likes,
retweets

Data Exploration

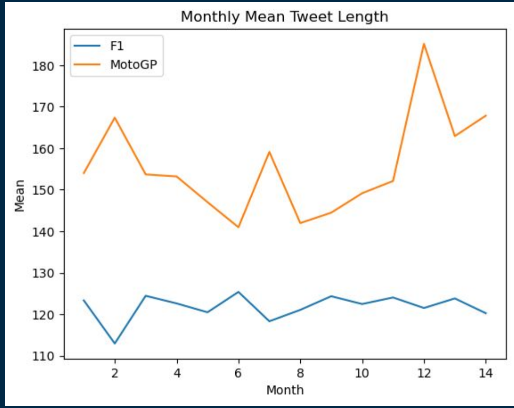


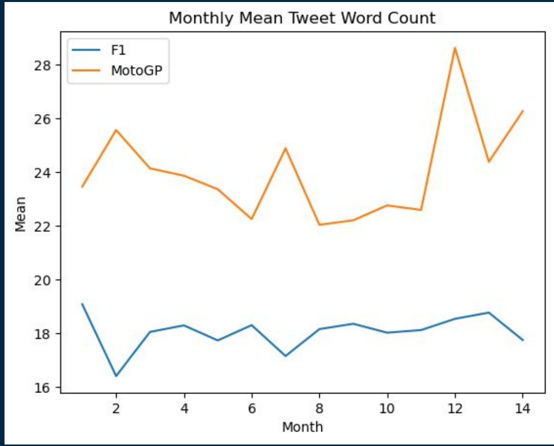


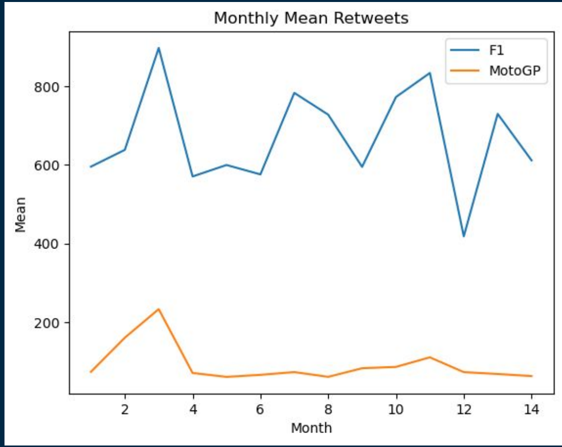


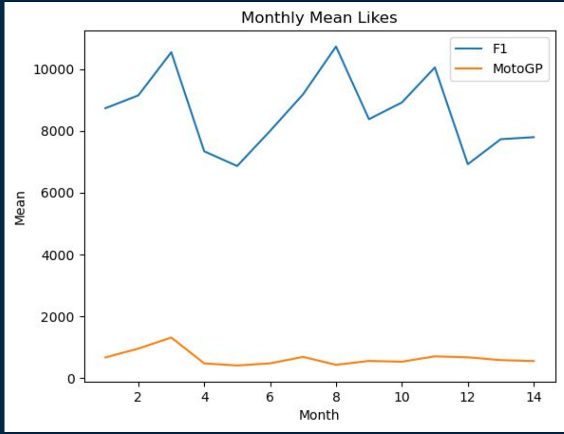


Data Exploration





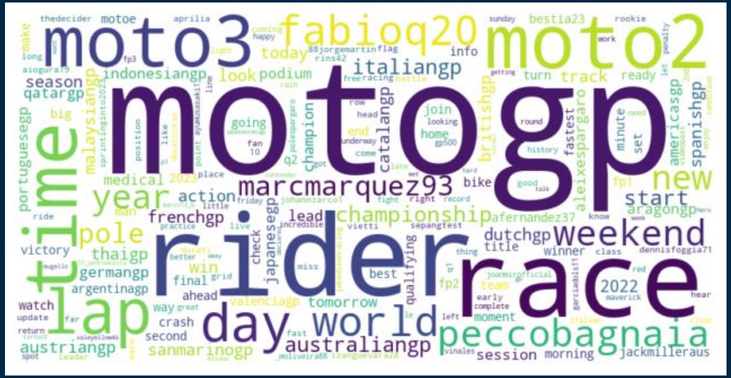




Data Exploration

Formula 1

motoGP



Data Processing

STEP 01

Create tweet length and tweet word count

STEP 02

Remove newlines, stopwords, emoticons, quotations, urls

STEP 03

Split tweets into text-only, hashtag, mentions

STEP 04

Revisit EDA using processed data to see if anything changed

Modeling Techniques

MultiNomial Naive-Bayes

Probability of a class, given the occurrence rate of features

Logistic Regression

Models the relationship between the class and the features

Modeling Approach

Build a model (M1) using tweet as feature and evaluate the model.

Build a model (M2) using hashtag as feature and evaluate the model.

If hashtag is overly dominant as a predictor, re-evaluate performance of M1 on text-only.

Improve on M1 (M3) by incorporating other features such as tweet length, tweet word count, number of likes, and number of retweets.

Determine best threshold to maximize recall and F1 scores.

Best Naive-Bayes Model

Using tweet, tweet length, word count, number of likes, number of retweets

Model (M3, 0.5)

Tweet: 91.9%
Text-Only: 91.6%

Accuracy

Tweet: 87.9%
Text-Only: 87.5%

Recall

Tweet: 91.9%
Text-Only: 91.6%

Precision

Tweet: 89.9%
Text-Only: 89.5%

F1 Score

Tweet: 96.1%
Text-Only: 95.7%

ROC AUC

Best Log-Regression Model

Using tweet, tweet length, word count, number of likes, number of retweets

Model (M3, 0.3)

Tweet: 99.4%
Text-Only: 94.0%

Accuracy

Tweet: 99.1%
Text-Only: 87.1%

Recall

Tweet: 99.4%
Text-Only: 98.1%

Precision

Tweet: 99.2%
Text-Only: 92.3%

F1 Score

Tweet: 99.9%
Text-Only: 99.3%

ROC AUC

Recommendation

To maximize relevant tweets on news feeds and minimize spam,

Build a classifier that is able to differentiate Formula 1 related tweets from MotoGP related tweets.

Logistic Regression Model
(M3, 0.3)

Conclusion

Key Limitation

- Tweet length, word count, number of likes, number of retweets are all features which could vary greatly among individual users, hence performance may drop further.

Suggested area for improvement

- Scrape tweets from individual users instead and manually classify for training data.

THANK YOU