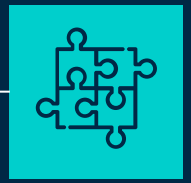
API and Reddit Classification

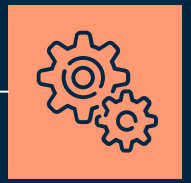
By Marcus Tan

TABLE OF CONTENTS



01

Problem Statement
& EDA



02

Model and
features

03

Conclusion
&
Recommendations

Problem Statement & EDA

01

Background

- 2 of the most popular Multiplayer online battle arena (MOBA)
- Dota 2 and League of Legends (LoL)
- Prize pool of over USD 40 million and 6 million respectively
- 5 players against 5 players
- Winning condition - Destroy the opponent’s 'Headquarter building'
- Requires mechanical skills and good teamwork.

Problem Statement

- Razor - Singapore grown company specializing in gaming hardware
- Market their gaming products
- Target areas that best differentiates between Dota 2 and LoL
- Budget to target only 1 gaming community
- Provide recommendations on areas to target
- Model with high accuracy

Data Cleaning and EDA

- Dataset is from subreddit posts of Dota 2 and LoL
- Drop posts with duplicated titles
- Combined all words from title and post
- Removed:
  - HTML tags
  - URL
  - Non-letters
  - Stopwords

Model and features

02

Model selection (baseline score – 56%)

Model	GridSearch_CV	Train with optimal model	Test with optimal model
Logistic Regression (Count)	0.8242320500451278	0.9992119779353822	0.839622641509434
Logistic Regression (TF-IDF)	0.8439575487846629	0.9676910953506698	0.8553459119496856
Naive Bayes (Count)	0.8400018673555133	0.9369582348305753	**0.8742138364779874**
Naive Bayes (TF-IDF)	**0.8644237652111668**	0.9818754925137904	0.8679245283018868
Random Forest (Count)	0.8108711213469857	0.9992119779353822	0.7955974842767296
Random Forest (TF-IDF)	0.8045501229342378	0.9897557131599685	0.8113207547169812
Ada Boost (Count)	0.8321496374218045	0.9637509850275807	0.8270440251572327
Ada Boost (TF-IDF)	0.8029939933397653	0.9747832939322301	0.8238993710691824
Support vector machine (Count)	0.8006068905418443	0.9921197793538219	0.8144654088050315
Support vector machine (TF-IDF)	0.8463166412498833	0.9984239558707644	0.8616352201257862

- Naive Bayes model highest accuracy
- Best CV score of 86%
- Able to interpret the features with log probability
- SVM log probability created using cross validation
- Results might be different from those obtained by predict

Naive Bayes – Top 15 features for Dota 2 (Log probability)

Feature	LoL	Dota2	Difference
animajor	-9.613901	-5.346606	-4.267295
ti	-9.613901	-5.464396	-4.149505
weplay	-9.613901	-5.719307	-3.894594
dpc	-9.613901	-6.260720	-3.353181
nigma	-9.613901	-6.340948	-3.272953
abed	-9.613901	-6.392057	-3.221844
valve	-9.613901	-6.488106	-3.125795
iceiceice	-9.613901	-6.549924	-3.063977
sea	-9.613901	-6.575001	-3.038900
major	-8.705811	-5.686325	-3.019486
rtz	-9.613901	-6.603803	-3.010098
finals	-9.349339	-6.366255	-2.983083
hero	-8.655439	-5.685026	-2.970412
arteezy	-9.613901	-6.653775	-2.960126
psg	-9.613901	-6.681452	-2.932449

- Top 15 features for Dota 2
- Log probability of each word classifying the post as LoL or Dota 2
- Bigger -ve Difference value, 'stronger' in classifying as Dota2
- Competitions - (Animajor, TI, weplay, DPC)
- Professional Players (iceiceice, Abed, Arteezy)
- All players from the same team "Evil Geniuses"
- Professional teams (Nigma, PSG)

Naive Bayes – Top 15 features for LoL (Log probability)

Feature	LoL	Dota2	Difference
mt	-5.583515	-9.606221	4.022706
adc	-6.411998	-9.606221	3.194223
skin	-6.461247	-9.606221	3.144974
sett	-6.501881	-9.606221	3.104340
penta	-6.580306	-9.606221	3.025915
champ	-6.590252	-9.606221	3.015969
ezreal	-6.665979	-9.606221	2.940242
akali	-6.693899	-9.606221	2.912322
leaguepedia	-6.733373	-9.606221	2.872848
mundo	-6.772850	-9.606221	2.833371
lec	-6.841683	-9.606221	2.764538
lucian	-6.889228	-9.606221	2.716993
rumble	-6.907031	-9.606221	2.699190
jungler	-6.949193	-9.606221	2.657028
viego	-6.966945	-9.606221	2.639276

- Top 15 features for LoL
- Log probability of each word classifying the post as LoL or Dota 2
- Bigger Difference value, 'stronger' in classifying as LoL
- Mostly gaming terms
- Game characters (skin, sett, ezreal, akali, mundo, lucian, rumble and viego)
- Professional player (mt)
- Competition (lec)

Conclusion & Recommendations

03

Conclusion and Recommendations

Conclusion:

- Naive Bayes model with TF-IDF Vectorizer
- Best accuracy (87%)
- Easy to interpret
- To improve model - more data, more time tuning hyperparameters
- Not able to differentiate other topics

Recommendation:

- Target Dota 2 (LoL features mostly on game characters/terms)
- Large exposure to potential customers through Competitions, players and teams

Some targets to sponsor:

- The International 10 (TI) (prize pool over 40 million USD)
- Viewership will most likely be very high.
- Professional Singaporean player “iceiceice”
- Resonate with Razor (Singapore grown company)

Do you have any questions?

THANKS