

VS



# LOL V.S. DOTA2

## SUBREDDIT CLASSIFICATION

---

DSI Project 3  
By Piyapon Pongsantisuk

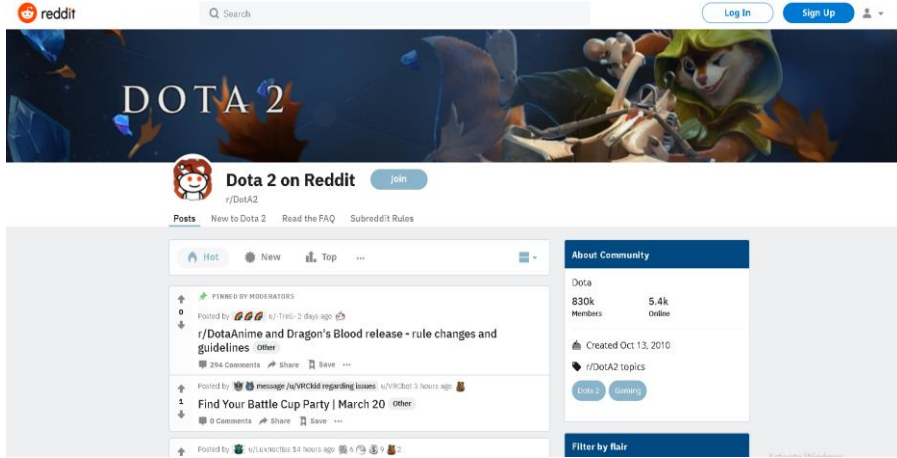
# PROBLEM STATEMENT

---

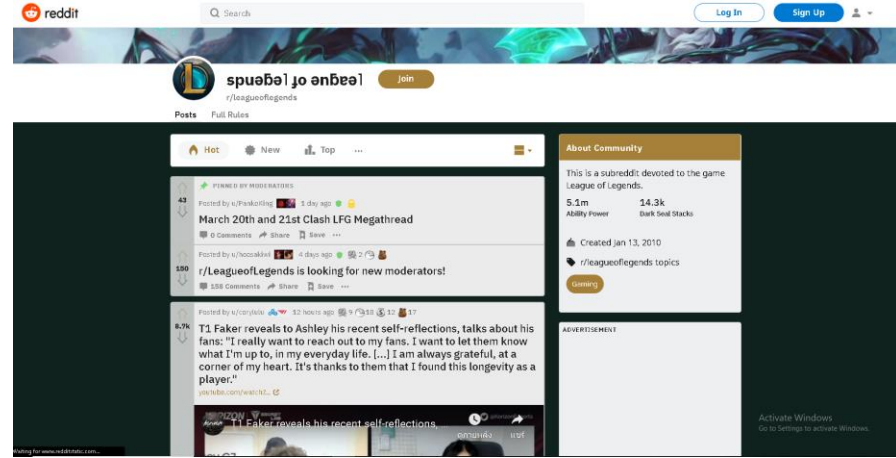
For the new game player who don't know which subreddit they should post in. I will try to classify by the text input (from post) which subreddit should the post be the reddit and auto classify to that subreddit. For instance, player on reddit no longer need to look for game subreddit themselves anymore just post and model will predict which game subreddit this post should be in!

This is the proof of concept of the project the classify subreddit game.

# SUB-REDDIT



/R/DOTA2/



/R/LEAGUEOFLEGENDS/

# DATA

---

# DATA COLLECTION

---

Collect 2,000 posts estimately from Reddit API

1. **/r/DotA2/ and /r/leagueoflegends/**
2. Gathering by 25 posts / 40 iteration is 1,000 posts in total
3. Using title and selftext for classifying subreddit



# DATA CLEANING / PREPROCESSING

---

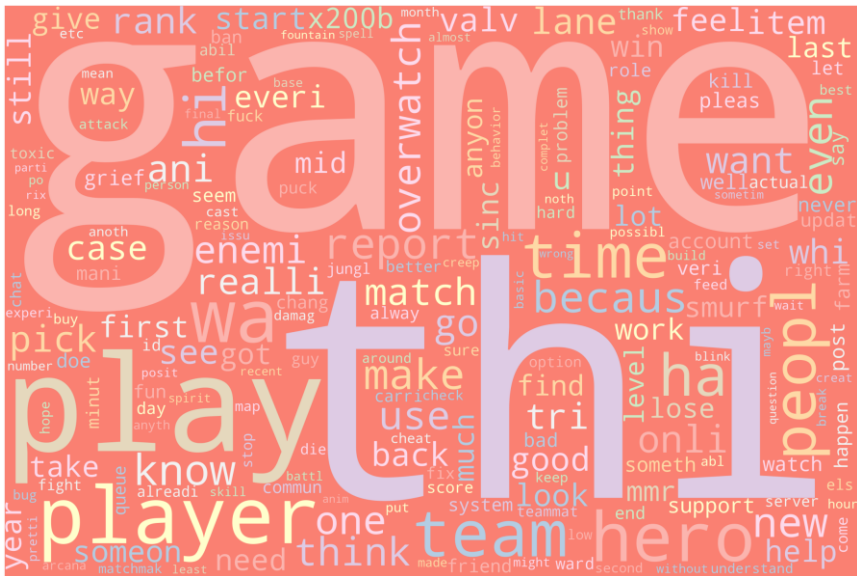
## Data cleaning

1. Dropping duplicated post by it's name (id of post) from 2,000 posts -> 1,600 posts
2. Re.sub() to remove html from the post

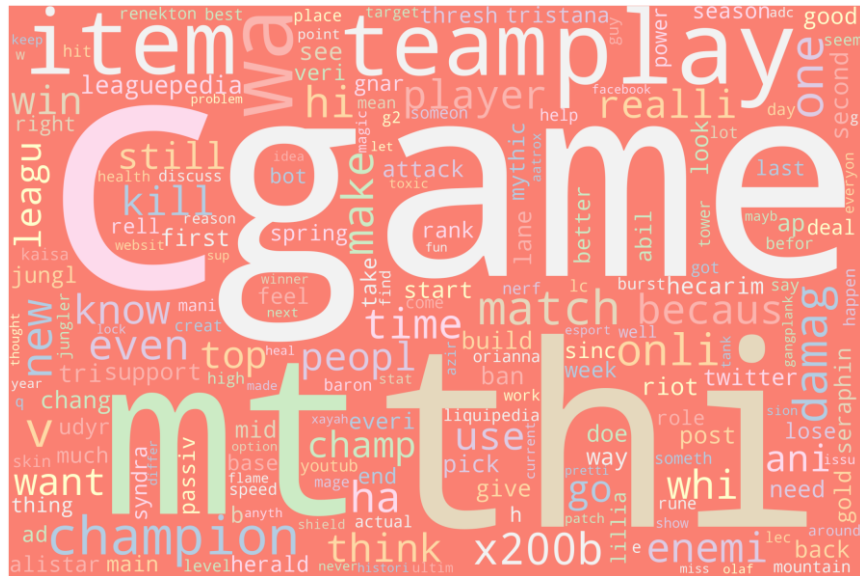
## Preprocessing

1. Remove hyper link (Detect from start of word which start with http)
2. Remove punctuation (text will remain only character a-z and number 0-9)
3. Stem the word - stem it to root of it (I have try lemmatiser too their is only a slightly different on accuracy, So, I choose stemmertise)
4. Remove stop word - stop words are a set of commonly used words in any language. For example, in English, “the”, “is” and “and” we remove it so it will not effect our model
5. Remove League of legend and DotA2 - Remove word that exactly the relate with subreddit will not bias on model

# DotA2



# League of Legend (LoL)





# MODEL & TUNING

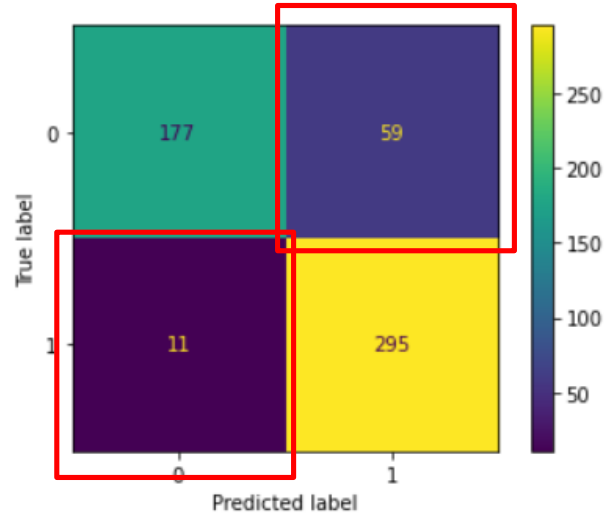
---

# MODEL & ACCURACY

---

Model	CVEC (Training)	CVEC (Testing)	Different	TFIDF (Training)	TFIDF (Testing)	Different
Baseline	56.49%	56.49%	56.49%	56.49%	56.49%	56.49%
LR	98.544%	86.162%	12.382%	96.451%	85.608%	10.843%
Naive-Bayes	90.991%	87.084%	3.907%	94.085%	84.501%	9.584%
KNN	99.090%	64.206%	34.884%	84.349%	80.996%	4.353%
Random Forest	100%	85.240%	14.76%	99.818%	83.395%	16.423%

# REMOVE GENERAL WORD



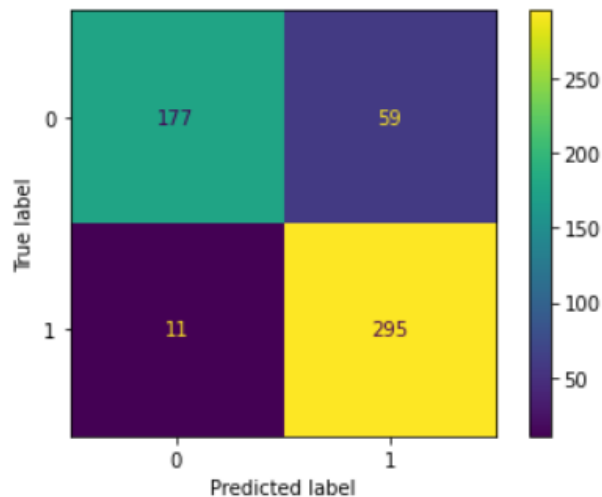
I focus on misclassification and list out the word which occur in the most classification

```
game 103
thi 73
play 40
get 35
like 32
wa 31
player 27
would 24
time 22
rank 22
peopl 21
toxic 21
leagu 20
team 20
know 19
think 17
realli 17
hi 17
back 16
```

After that I remove top 20 words from the text and use Naives-Bayes again

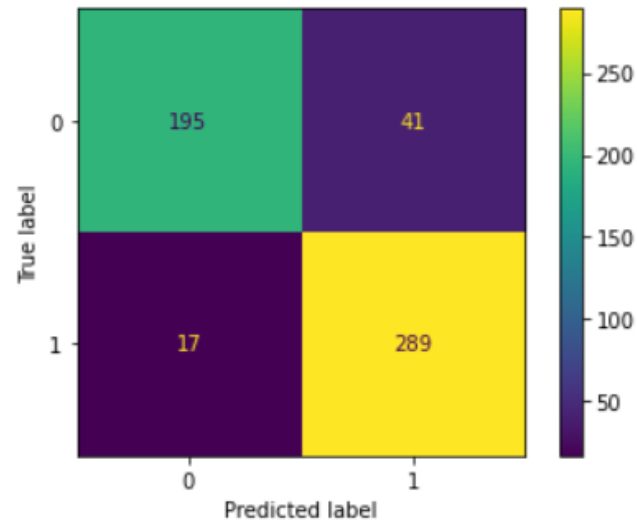
# REMOVE GENERAL WORD

---



**Naive bayes before remove general word**

Accuracy: 87.084%



**Naive bayes before remove general word**

Accuracy: 89.3%

# MOST EFFECTED WORD

---

Coef	LoL word	Coef	DotA2 word
-10.2353	100	-5.3755	fon
-10.2353	11	-5.3990	trackpad
-10.2353	120	-5.5079	tf
-10.2353	150	-5.6814	damag
-10.2353	26	-5.7355	coronian
-10.2353	48	-5.8164	kill
-10.2353	56	-5.9312	face
-10.2353	absurd	-5.9586	true
-10.2353	act	-6.0764	perspect
-10.2353	ad	-6.1081	bunch
-10.2353	address	-6.1409	yone
-10.2353	adult	-6.1748	depth
-10.2353	adventur	-6.3434	featur
-10.2353	afford	-6.3641	short
-10.2353	ago	-6.3851	within
-10.2353	almost	-6.4066	00
-10.2353	along	-6.4286	terribl
-10.2353	anywher	-6.5217	level
-10.2353	apart	-6.5464	cours
-10.2353	around	-6.5464	calm



## DotA

- Most of them is a general word which occur in MOBA game
- Some of unique word to classify such as
  - Hero (in LoL call it as a champion) or
  - Match (in LoL more frequently use a game and etc.)

## LoL

- They talk lots about the number in game like stats (lifest - lifesteal) / gold which use to buy item in game / number of AD and AP (role like carry and mage)
- And also talk to famous person in game like Ceiran – shoutcaster or Czekolad – Gamer

# MODEL & ACCURACY

Model	CVEC (Training)	CVEC (Testing)	Different	TFIDF (Training)	TFIDF (Testing)	Different
Baseline	56.49%	56.49%	-	56.49%	56.49%	-
LR	98.544%	86.162%	12.382%	96.451%	85.608%	10.843%
Naive-Bayes	90.991%	87.084%	3.907%	94.085%	84.501%	9.584%
KNN	99.090%	64.206%	34.884%	84.349%	80.996%	4.353%
Random Forest	100%	85.240%	14.76%	99.818%	83.395%	16.423%
Naive-Bayes (remove general word)	91.811%	89.299% 	2.511% 	-	-	-

# CONCLUSION

---

## DOTA2 VS. LOL

We can classify these 2 games quite accurate due to each game have different unique word for classify in NLP

## BEST SCORING MODEL

- Naives-Bayes with CountVectorizer()
- Accuracy on Train: 91.811%
- Accuracy on Test: 89.3%

## POTENTIAL FOR IMPROVEMENT

- Try other ensemble models, such as using boosting , SVM
- Ability for model to classify more than two subreddits
- Improve the false positive value (predicted LoL as DotA)



# THANK YOU

Do you have any question?