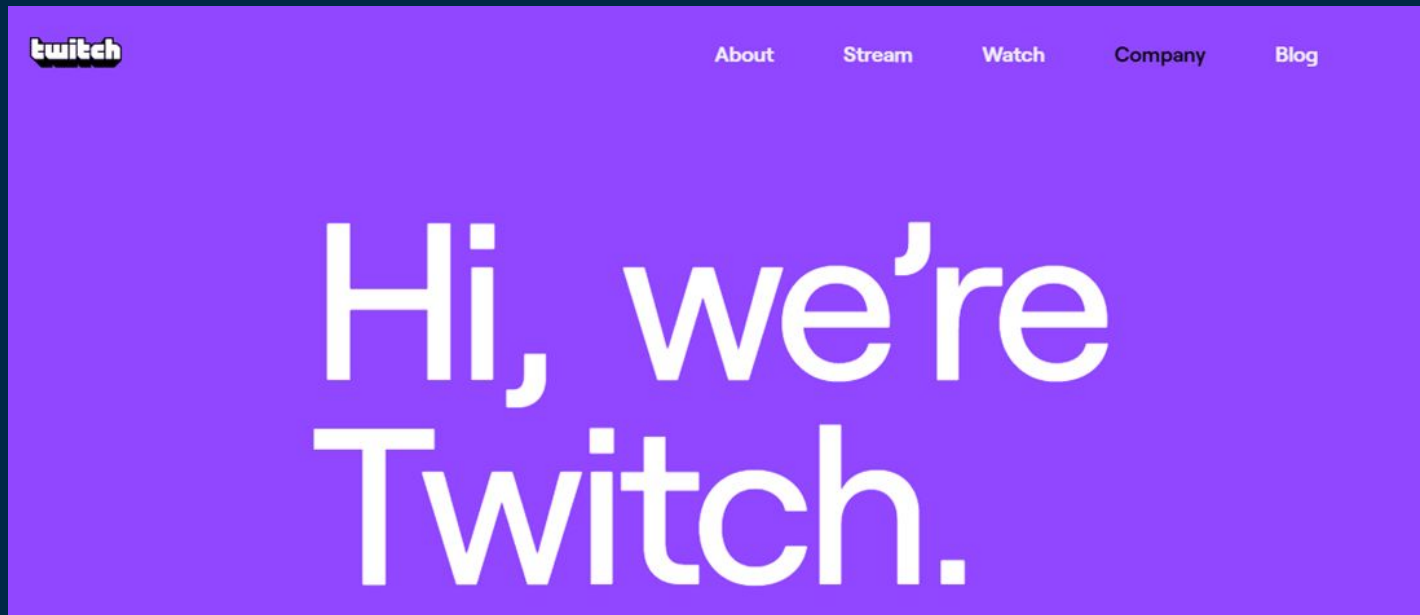


# GENERAL ASSEMBLY DS PROJECT 3

Thread Classification

27 May 2022

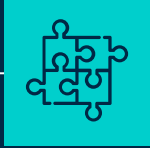
# WHO ARE WE



Data Science Team

– Calvin, Joseph, Nelson, Zhi Hong, Priscilla

# TABLE OF CONTENTS



01

PROBLEM



02

DATA  
OVERVIEW,  
CLEANING &  
VISUALIZATION



03

MODELING

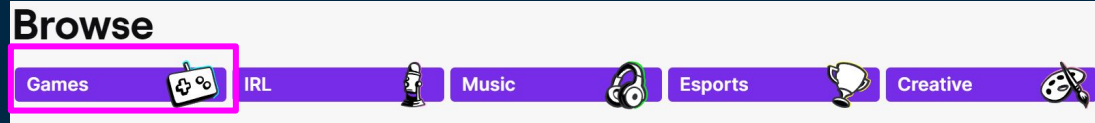


04

LIMITATIONS &  
FUTURE PLANS

RECOMMENDATIONS  
& CONCLUSIONS

# Problem



- New single **beta** forum for gamers to post discussions & comments
- High traffic of posts & comments updates daily
- Challenging for:
  - Business development / marketing team to design sales & marketing campaigns that best meet the user's needs
  - Users sieve through multiple posts to find topics that interest or are relevant to them → Poor user experience

# Top 2 Popular Games



Dota 2



League of Legends

# Problem Statement

- Use Reddit posts from Dota 2 and League of Legends to build a text classifier with an accuracy of at least 85%
  - Categorize the games in new forum
- Identify top 5 predictors for each game

[r/DotA2, r/leagueoflegends]



# DATA Scraping

Data Collected from:

- r/DotA2
- r/leagueoflegends

Using pushshift API

Dota2 : 4000 rows, 88 columns

Lol : 4000 rows, 82 columns

```
#Data collection using PushAPI
def get_reddit(subreddit, pages):
    url = 'https://api.pushshift.io/reddit/search/submission'
    params = {
        'subreddit': subreddit,
        'size' : 100
    }

    df_post = pd.DataFrame()
    for i in range(0,pages):
        if i>0:
            params['before'] = df_post['created_utc'][len(df_post)-1]

        response = requests.get(url, params)
        data=response.json()
        df_post = df_post.append(data['data'],
                                ignore_index=True)

    print (df_post.shape)
    return df_post
```

# Data Outcome

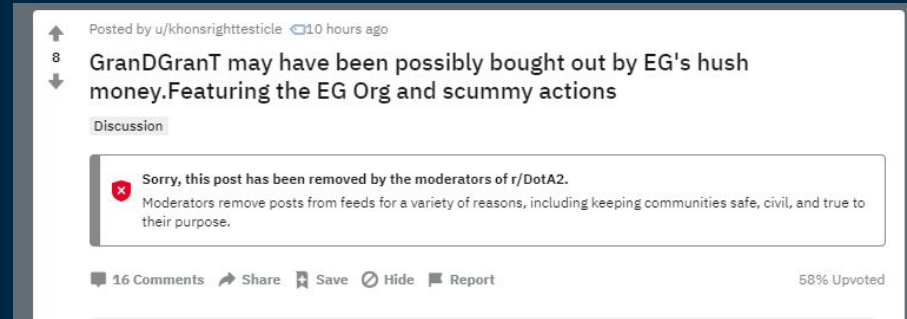
	author	created_utc	subreddit	selftext	title
0	CorinGetorix	1653494498	DotA2	NaN	The 7.31c meta
1	JohnMorgan98	1653494468	DotA2	Real talk, we have Enigma, Void Spirit, Dark Seer and Faceless Void as void characters. Idk if TA is considered a void character. There is Byssak or Kashurra from dragon blood. Not sure they are the same person. \n\nVoid characters are so unique, we have like pos 1, pos 2, pos 3 and pos 5 of void characters\n\nWhat do u guys think? This suddenly came to my mind when I saw Riot new void character	I want more void characters than demon or keen characters in dota 2
2	Filin1234	1653494382	DotA2	NaN	Love this hero, back when he was a unit in Warcraft 3
3	Sixty6Seventy7	1653494075	DotA2	Pls for once release it without making us beg (づ・...・)づ GIFF BATLLEPASS (づ・...・)づ	(づ・...・)づ GIFF BATLLEPASS (づ・...・)づ
4	DoctorHeckle	1653493973	DotA2	NaN	Busy couple weeks and missed the Stockholm Major? No worries, I made summary video that highlights the main storylines.
5	TrainTrackBallSack	1653493898	DotA2	win lane take towers choke enemy get nw advantage get rosh win game\n\nThe above is the standard formula for how dota has been played for years, where like 80% of the importance of the game lies in the first 10-15.\n\nIf you win early game the only win condition for the enemy is that you throw, otherwise the game is yours.\n\nPlease bring back the midgame. How to do so you know better than I I...	Games over at 10-15 mins is the most boring meta
6	graveyard_boy_	1653493774	DotA2	&amp;#x200B;\n\n*Processing video vabaijy38n191...*	Sunstrike!
7	DangerousLefty	1653493410	DotA2	I have a ticket for the final day of the Arlington Major, and now i can't go. Can someone please guide me.	Where can i sell my Arlington Major ticket?
8	wanttosensfwcontent	1653493375	DotA2	I remember a few years ago people would invite each other to 1vs1 games to see who is the better gamer. Why did people stop doing that?	What happened to 1vs1 mid?
10	_Drink_Bleach_	1653493219	DotA2	u/siractionslacks- please do a segment trying spicy foods in Singapore. My recommendations are the McSpicy and Mala. Also try durian which is not spicy	Request for ti10 content



# Data Cleaning

- Check for duplicated post using the ID columns
- Dropped posted that were removed by moderators
- Filter out subreddit, selftext and title columns
- Merged dataframe

```
reddit      293
moderator    155
automod_filtered  120
deleted      11
Name: removed_by_category, dtype: int64
```



```
#Removing post that were flagged out by moderators
def remove_droppedpost(df_post):
    df_post = df_post[df_post['removed_by_category'].isna()]
    return df_post
lol = remove_droppedpost(lol)
dota = remove_droppedpost(dota)
```

# Data Cleaning

## Other Data Cleaning

- Replaced null values with white spaces
- Merged self-text and title columns
- Removed Website Link
- Removed non-english text
- Removed punctuations
- 

```
#Removing Website Links
dota_lol_data['merged_body'] = dota_lol_data['merged_body'].apply(lambda x: re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+])|!*\(\n', '' , x))

#Removing Non-english text
dota_lol_data['merged_body'] = dota_lol_data['merged_body'].apply(lambda x: re.sub('[^a-zA-Z]', " ", x).lower())
dota_lol_data.head()
```

	subreddit	merged_body
0	DotA2	the c meta
1	DotA2	real talk we have enigma void spirit dark seer and faceless void as void characters ldk if ta is considered a void character there is byssak or kashurra from dragon blood not sure they are the same person void characters are so unique we have like pos pos pos and pos of void characters what do u guys think this suddenly came to my mind when i saw riot new void characteri ...
2	DotA2	love this hero back when he was a unit in warcraft
3	DotA2	pls for once release it without making us beg giff battlepass giff battlepass
4	DotA2	busy couple weeks and missed the stockholm major no worries i made summary video that highlights the main storylines

# Data Cleaning

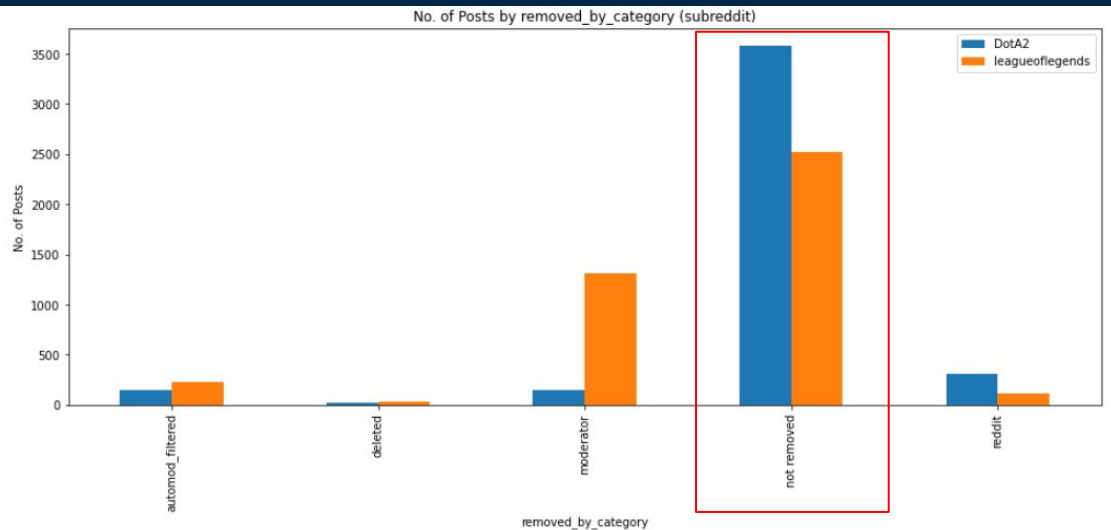
Tokenized  
the text

Stopword  
removal

Lemmatization

	subreddit	merged_body	body_text_clean	body_text_tokenized	body_text_nostop	body_text_stemmed	body_text_lemmatized	stem_sentence	lem_sentence
0	DotA2	the c meta	the c meta	[the, c, meta]	[meta]	[meta]	[meta]	meta	meta
1	DotA2	<p>real talk we have enigma void spirit dark seer and faceless void as void characters idk if ta is considered a void character there is byssak or kashurra from dragon blood not sure they are the same person void characters are so unique we have like pos pos and pos of void characters what do u guys think this suddenly came to my mind when i saw riot new void characteri ...</p>	<p>real talk we have enigma void spirit dark seer and faceless void as void characters idk if ta is considered a void character there is byssak or kashurra from dragon blood not sure they are the same person void characters are so unique we have like pos pos and pos of void characters what do u guys think this suddenly came to my mind when i saw riot new void characteri ...</p>	<p>[real, talk, we, have, enigma, void, spirit, dark, seer, and, faceless, void, as, void, characters, idk, if, ta, is, considered, a, void, character, there, is, byssak, or, kashurra, from, dragon, blood, not, sure, they, are, the, same, person, void, characters, are, so, unique, we, have, like, pos, pos, pos, and, pos, of, void, characters, what, do, u, guys, think, this, suddenly, came, to, my...</p>	<p>[real, talk, enigma, void, spirit, dark, seer, faceless, void, void, void, characters, idk, ta, considered, void, character, byssak, kashurra, dragon, blood, person, void, characters, unique, void, characters, suddenly, came, mind, saw, riot, void, characteri, void, characters, demon, keen, character]</p>	<p>[real, talk, enigma, void, spirit, dark, seer, faceless, void, void, spirit, dark, seer, faceless, void, void, character, idk, ta, consid, void, charact, byssak, kashurra, dragon, blood, person, void, charact, uniqu, void, charact, suddenli, came, mind, saw, riot, charact, demon, keen, charact]</p>	<p>[real, talk, enigma, void, spirit, dark, seer, faceless, void, void, character, idk, ta, consid, void, charact, byssak, kashurra, dragon, blood, person, void, charact, uniqu, void, charact, suddenli, came, mind, saw, riot, void, characteri, void, character, demon, keen, character]</p>	<p>meta</p>	<p>real talk enigma void spirit dark seer faceless void void character idk ta considered void character byssak kashurra dragon blood person void character uniqu void character suddenli came mind saw riot void characteri void character demon keen character</p>

# Visualisations of 'Not Removed' Posts



- DotA2 subreddit contains 3580 posts
- League of Legends subreddit contains 2526 posts

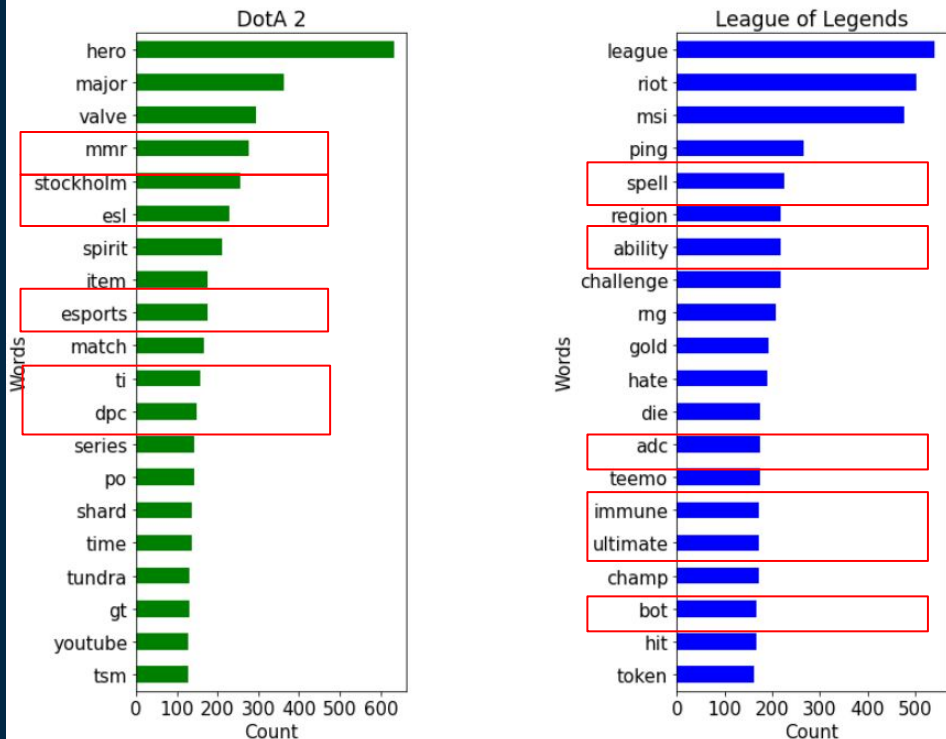
After merged, DotA2 (59%) vs League of Legends (41%) posts proportion

# Vectorizers to generate word counts

Vectorizers	n-gram
Count Vectorizer	1-gram, 2-grams
TF-IDF Vectorizer	1-gram, 2-grams

# Data Visualisation - 1-gram Count Vectorizers

Top 20 most frequent words (Dota2 vs League of Legends)



## DotA2

- Centers around tournaments, e.g. ESL Stockholm, happened 12 - 22 May 2022
- TI (The International) & DPC (Dota Pro Circuit) refers to the top tournaments and placements for top teams
- MMRs - ranked match discussions

## League of Legends

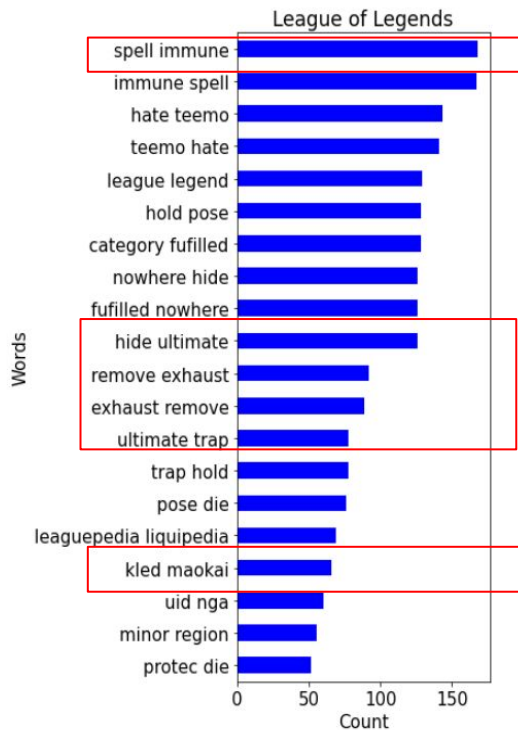
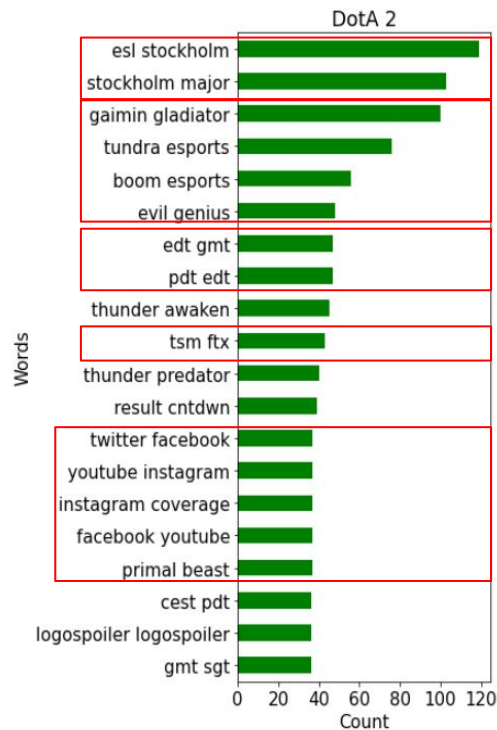
- Centers around gameplay & game highlights as there are more discussions on ability, challenge, spell, ultimate, immune
- Strategy talks as shown in the word adc (Attack Damage Carry) and bot (Bottom lane)

Both game publishers (Valve for Dota2 and Riot for League of Legends) are in the top words list which is expected.

Surprisingly, only 1 hero is in the list.

# Data Visualisation - 2-grams Count Vectorizers

Top 20 most frequent 2-gram words (Dota2 vs League of Legends)



## DotA2

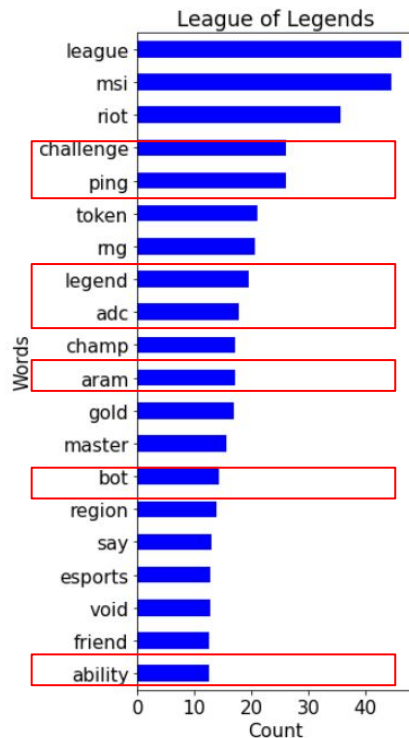
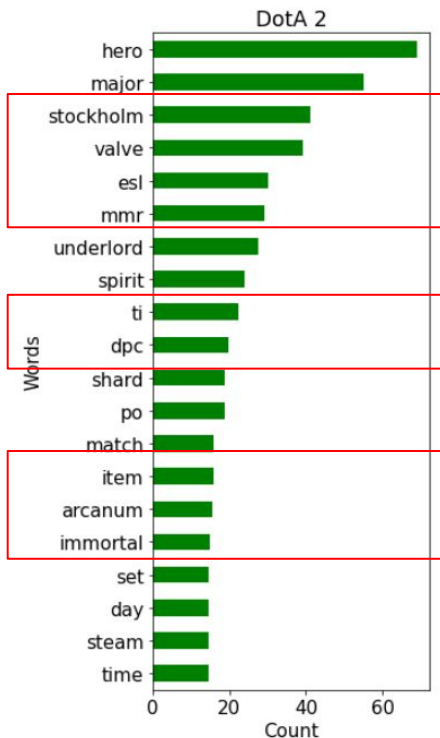
- Similar to 1-gram, subreddit focuses on tournaments discussions (ESL Stockholm and Stockholm majors).
- More esports teams are mentioned:
  - Gaimin Gladiator
  - Tundra Esports
  - Boom Esports
  - Evil Genius
  - TSM FTX
- Several time zones are mentioned
- Social media coverage
- Primal Beast (newest born hero)

## League of Legends

- Similar to 1-gram, subreddit focus on gameplay through spell immune, hide ultimate, remove exhaust
- Head to head top lane heroes (Kled and Maokai)

# Data Visualisation - 1-gram TF-IDF Vectorizers

Top 20 most frequent words (Dota2 vs League of Legends)



## DotA2

- DotA2 centers around tournament talks
- More emphasis/weightage on items, e.g. Arcanum and immortal as item rarity level

## League of Legends

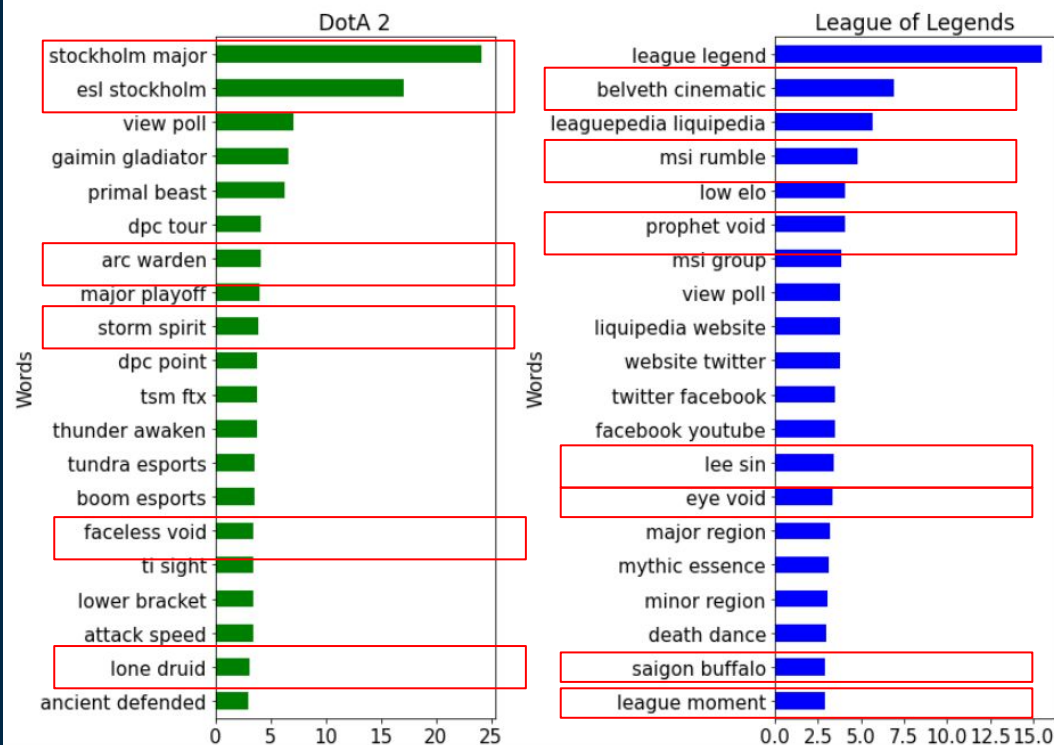
- Similar to Count Vectorizers, League of legends community discusses about gameplay, ability and strategies



# Data Visualisation - 2-grams

## TF-IDF Vectorizers

Top 20 most frequent 2-gram words (Dota2 vs League of Legends)



### DotA2

- DotA2 centers around tournament talks and teams
- Notably observed more topics on hero names e.g. Arc Warden, Storm Spirit, Faceless Void, Lone Druid, which is more unique to the game

### League of Legends

- There are new topics with more weightage on hero names, e.g. Prophet Void, Eye Void, Lee Sin, Belveth cinematic
- Tournament jargons (Saigon Buffalo, MSI rumble)

# Count Vectorizer vs TF-IDF Vectorizer (DotA2)

DotA2			
Count Vectorizer		TF-IDF Vectorizer	
1-gram	2-grams	1-gram	2-grams
<ul style="list-style-type: none"><li>- Tournaments (ESL, TI, DPC)</li><li>- MMRs - ranked match discussions</li></ul>	<ul style="list-style-type: none"><li>- Tournaments (ESL Stockholm and Stockholm majors)</li><li>- Esports teams (Gaimin Gladiator, Evil Genius, etc)</li></ul>	<ul style="list-style-type: none"><li>- Tournaments (ESL, TI, DPC)</li><li>- Items and items rarity level (Arcanum and Immortal)</li></ul>	<ul style="list-style-type: none"><li>- Tournaments (ESL, TI, DPC)</li><li>- Hero names (Arc Warden, Storm Spirit, Faceless Void, Lone Druid)</li></ul>

# Count Vectorizer vs TF-IDF Vectorizer (League of Legends)

League of Legends			
Count Vectorizer		TF-IDF Vectorizer	
1-gram	2-grams	1-gram	2-grams
<ul style="list-style-type: none"><li>- Gameplay (ability, challenge, spell, ultimate, immune)</li><li>- Strategy talks (Attack Damage Carry) and bot (Bottom lane)</li></ul>	<ul style="list-style-type: none"><li>- Gameplay (spell immune, remove exhaust, ultimate trap)</li><li>- Head to head top lane heroes discussion (Kled &amp; Maokai)</li></ul>	<ul style="list-style-type: none"><li>- Gameplay (ability, challenge, spell, ultimate, immune)</li><li>- Strategy talks (Attack Damage Carry) ,bot (Bottom lane)</li></ul>	<ul style="list-style-type: none"><li>- Hero names (Prophet Void, Eye Void, Lee Sin, Belveth Cinematics)</li><li>- Tournament jargons ( Saigon Buffalo, MSI rumble)</li></ul>

# Model Building & Testing

Base Model

Logistic Regression



Random Forest



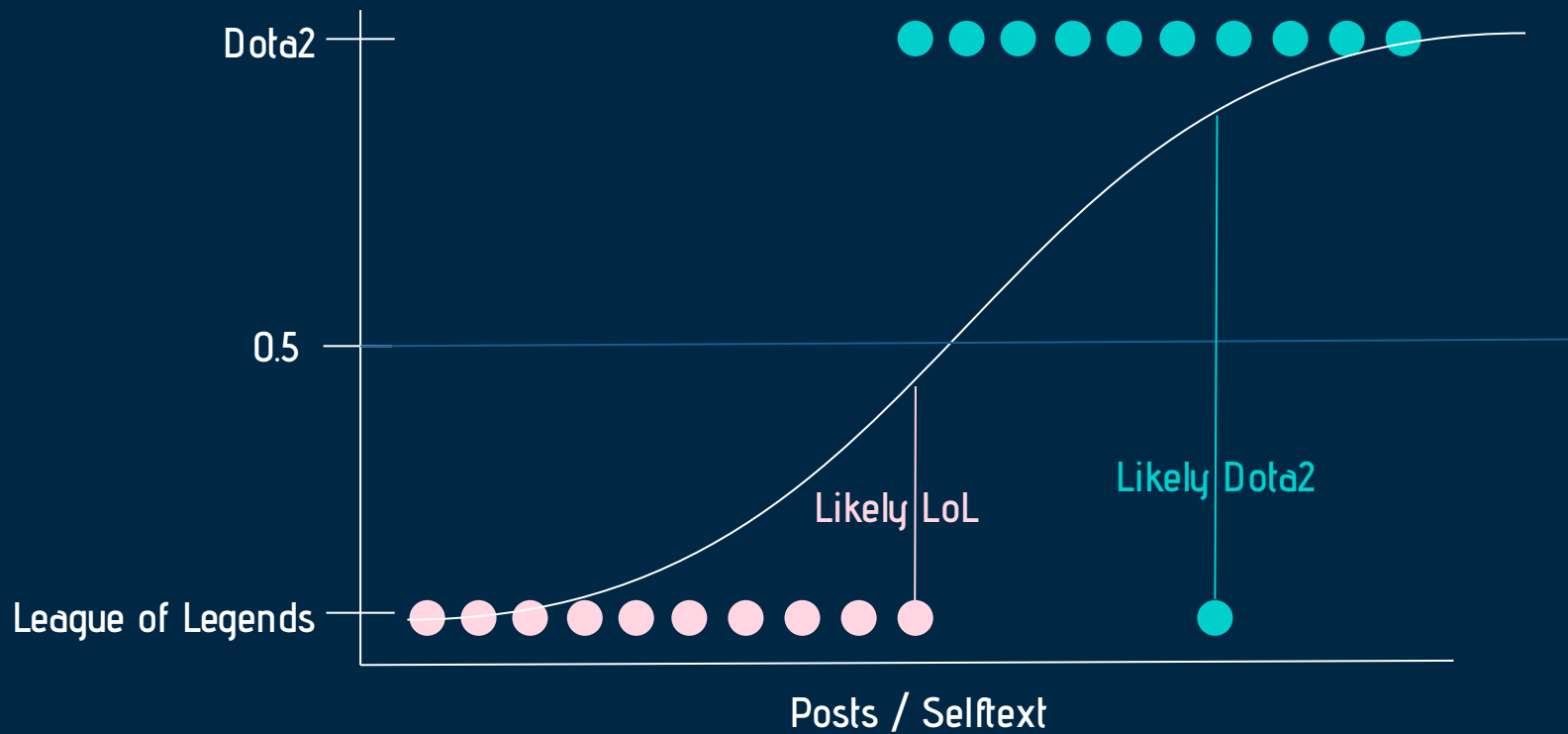
Multinomial Bay Naives



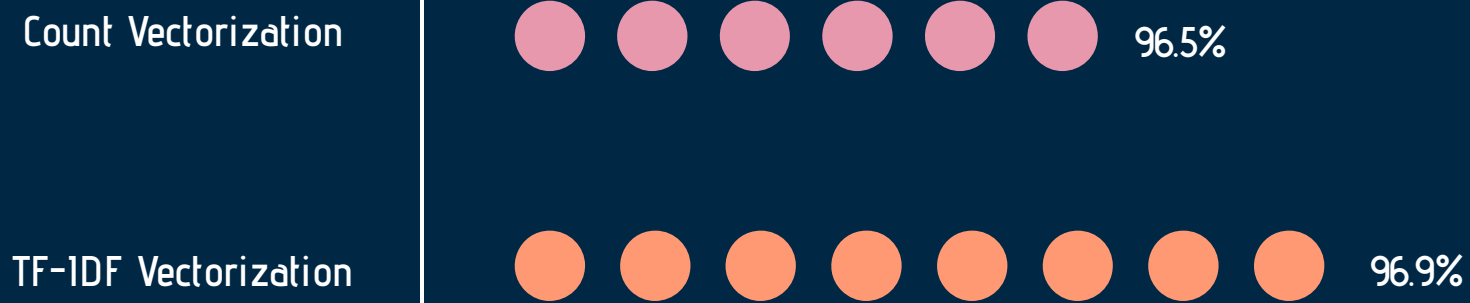
Count Vectorization

TFIDF Vectorization

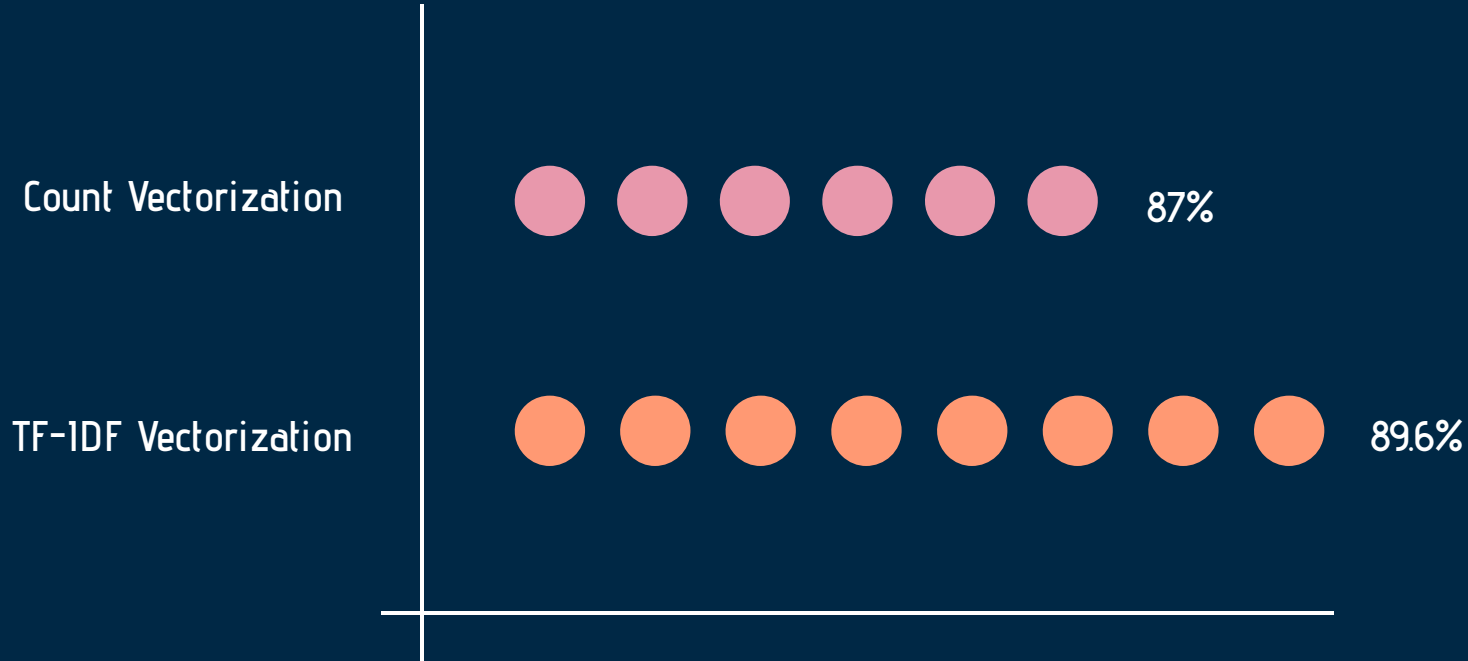
# Logistic Regression



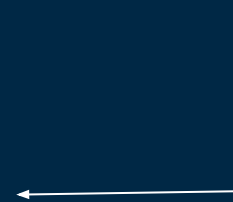
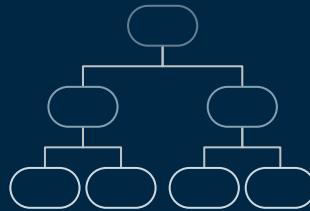
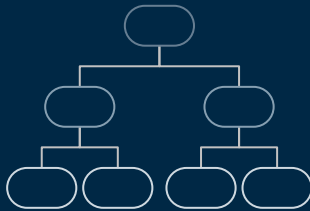
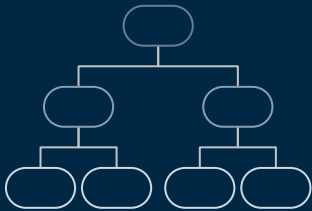
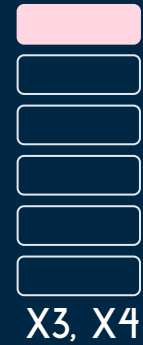
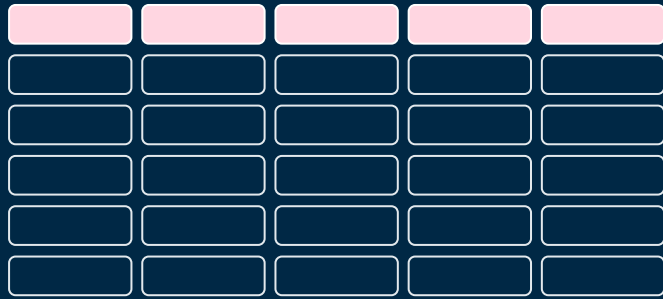
# Logistic Regression (Train)



# Logistic Regression (Test)



# Random Forest Algorithm





# Random Forest (Train)

Count Vectorization



TF-IDF Vectorization



# Random Forest (Test)

Count Vectorization



TF-IDF Vectorization



# Multinomial Naive Bayes



Calculate the frequency of each discrete individual words in their respective category

**Step 1**



Measure prior probability of the respective category, and the probability of each discrete words given the respective categories

**Step 2**



For a given sentence, calculate and compare the probability of the sentence being classified as each of the respective category, given each of the sentence's discrete words

**Step 3**

# Multinomial Naive Bayes (Train)

Count Vectorization



TF-IDF Vectorization



# Multinomial Naive Bayes (Test)

Count Vectorization



TF-IDF Vectorization



# Test Results Overlook

	model
0	TF-IDFVec + Multinomial Naive Bayes
1	CountVec + Multinomial Naive Bayes
2	TF-IDF Vec + random forest
3	TF-IDFVec + Logistic Regression
4	CountVec + Random Forests
5	CountVec + Logistic Regression

train_accuracy_score	test_accuracy_score
0.963163	0.908362
0.943271	0.902062
0.993861	0.899198
0.946709	0.895762
0.994352	0.895189
0.916749	0.870561

Base Model : Logistic Regression

Best Model (Chosen): Multinomial Naive Bayes + TFIDF

# Top 5 Predictors

## Dota2

- 1) Hero
- 2) Valve
- 3) Major
- 4) MMR
- 5) Stockholm

## League of Legends

- 1) Riot
- 2) Challenger
- 3) Champion
- 4) Adc
- 5) Aram



# Top 3 Most Frequent Words in Misclassified Posts

Words	DESCRIPTION
Ping	In gaming ping refers to the delay between a players input and the servers response
Hero	Primarily refers to characters in Dota2. A lot of player base in LOL comes from Dota2 and refers to LOL champions as Heros
Discord	A popular group-chatting platform that is built for gamers to communicate in-game



# Limitations & Future Plans

## 1) Model Accuracy Improvement

- Some posts were misclassified e.g. ping, hero, discord
- Better data cleaning steps e.g. remove Character Entities

## 2) Include New Model Features

- Analyse moderated posts to implement auto regulation of potential user rogue behaviour e.g. profanities, spams
- Sentiment analysis



# Recommendations & Conclusion

## Internal

- 1) Roll out classification to split posts into 2 separate threads
- 2) Establish timeline for future model features roll-out
- 3) Tease out campaign specific insights e.g. new users acquisition

## External

- 1) **Developers**: Insights on sentiments, new games &/or features, etc.
- 2) **Esports/Events**: Insights on interests & mechanics, partnerships, etc.
- 3) **Gaming Streamers**: Insights on peak online user activities, etc.

Do you have any questions?

boonleonglim@gmail.com

# THANKS

CREDITS: This presentation template was created by [Slidesgo](#),  
including icons by [Flaticon](#), and infographics & images by [Freepik](#)