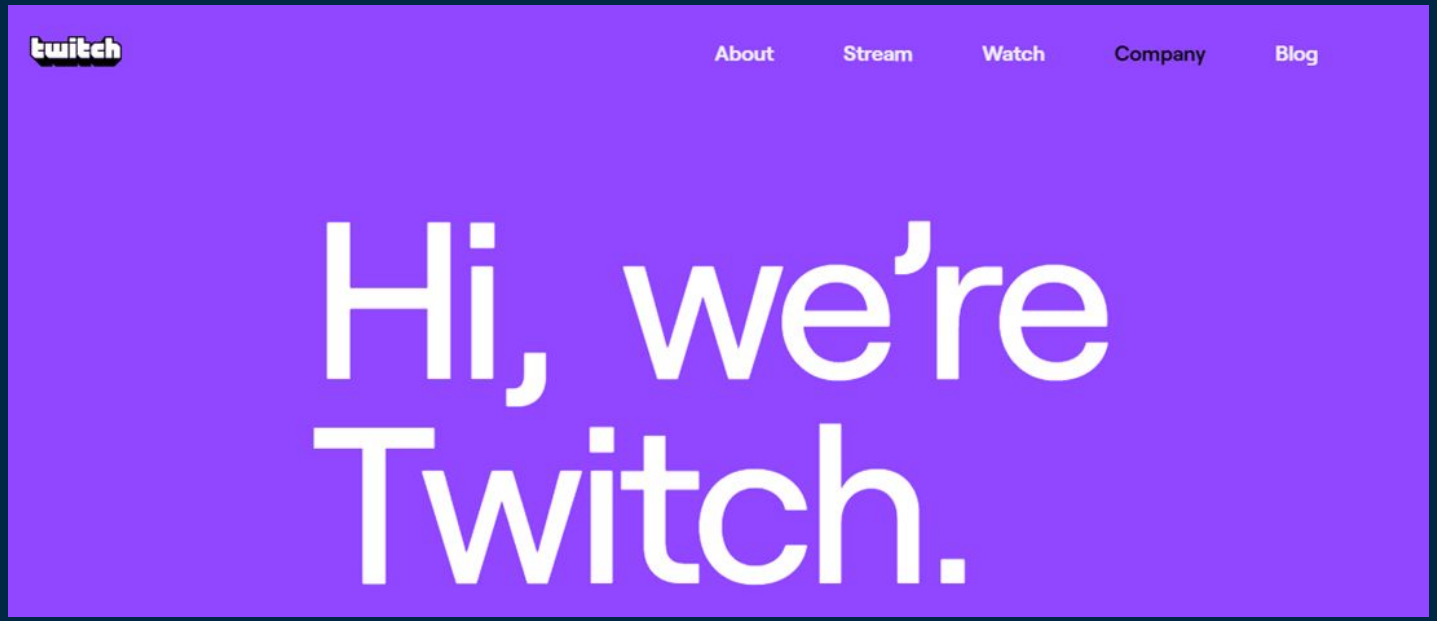
GENERAL ASSEMBLY DS PROJECT 3

Thread Classification

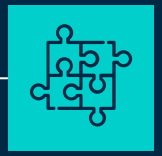
WHO ARE WE



Data Science Team

– Calvin, Joseph, Nelson, Zhi Hong, Priscilla

TABLE OF CONTENTS



01

PROBLEM

02

DATA
OVERVIEW,
CLEANING &
VISUALIZATION

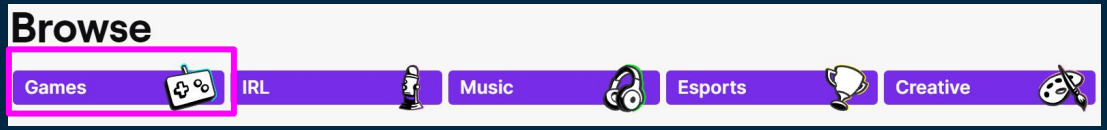
03

MODELING

04

CONCLUSIONS
LIMITATIONS
RECOMMENDATIONS
FUTURE PLANS

Problem



- New **beta** single forum for gamers to post 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 games 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	\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	wanttoseensfwcontent	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
```


```python
#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]|[$-_@.&+]|[!*\(\
```

```
#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()
```

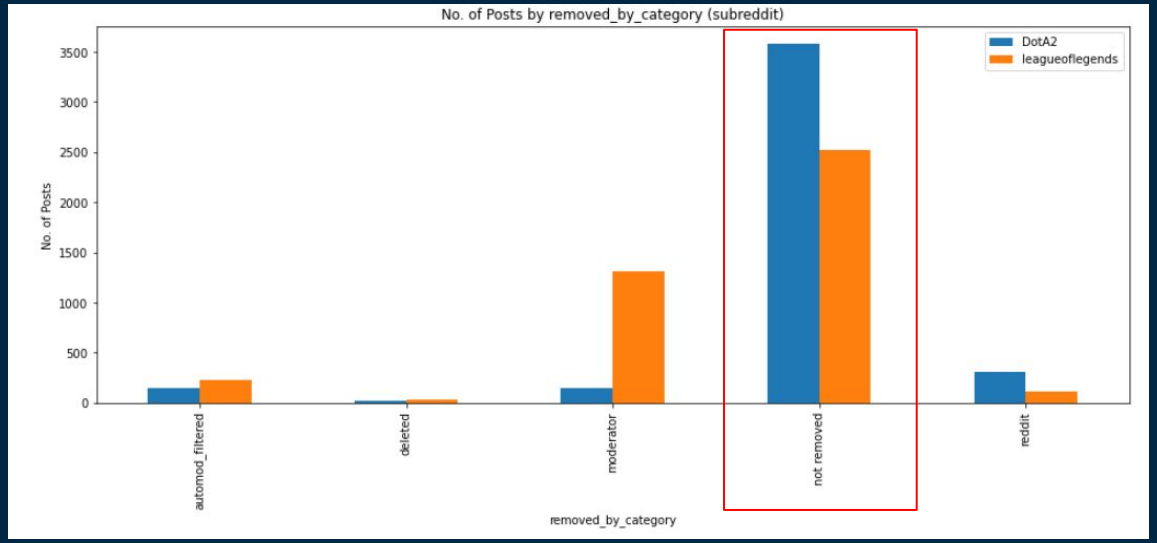
	subreddit	merged_body
0	DotA2	the c meta
1	DotA2	real talk we have enigma void spirit dark seer and faceless void as void characters idk if ta is considered a void character there is byssak or kashurra from dragon blood not sure they are the same person void characters are so unique we have like pos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

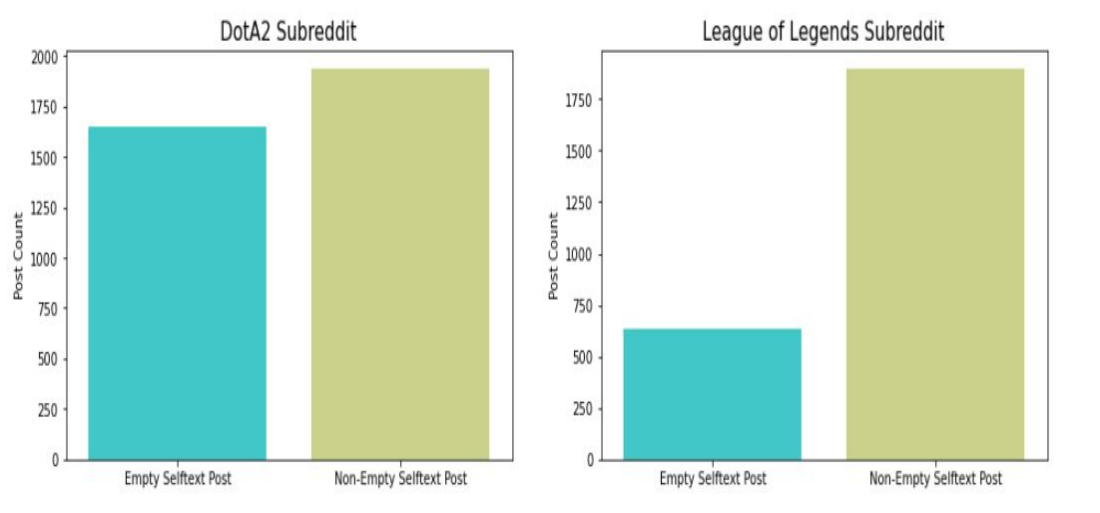
Visualisations of ‘Not Removed’ Posts



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

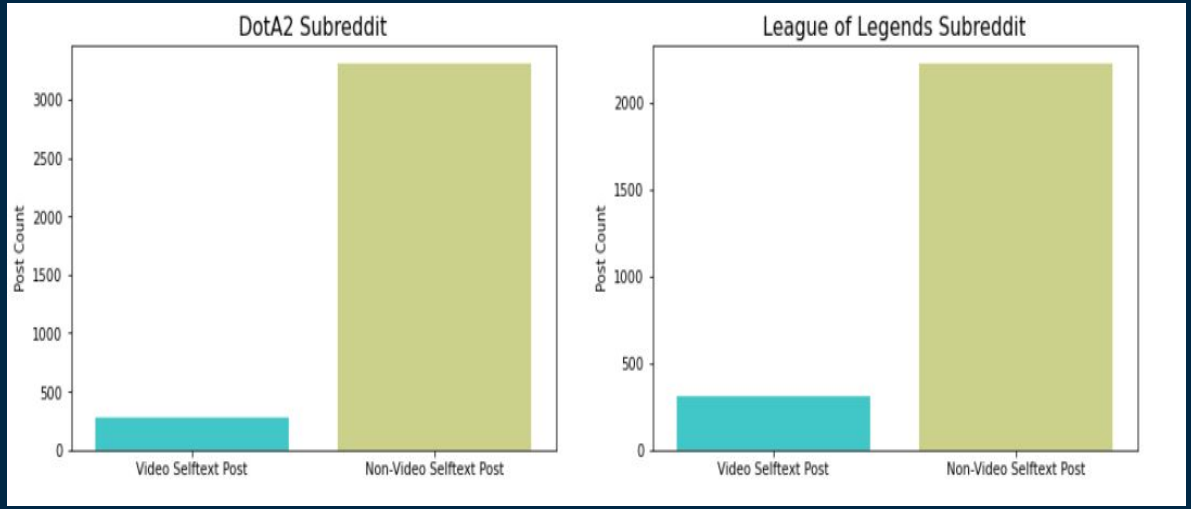
After merged, DotA2 (59%) vs League of Legends (41%) posts proportion

Visualisations of ‘Empty’ vs ‘Non Empty’ Selftext



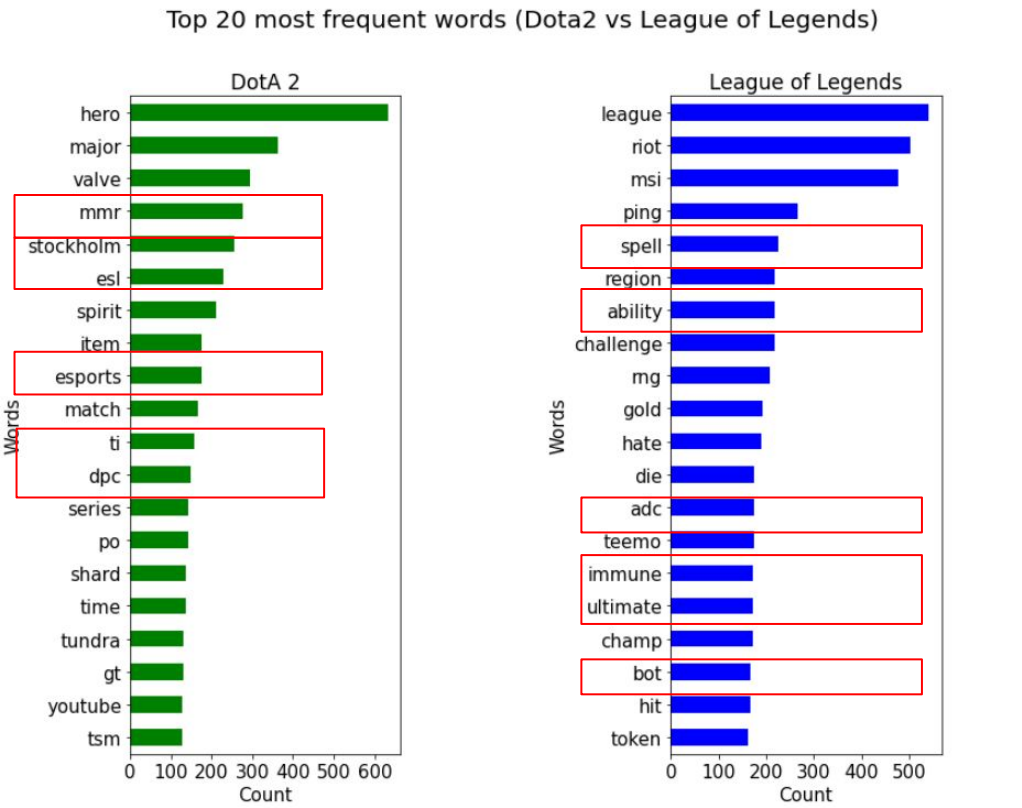
- More non-empty selftext posts
- More text to analyse -> better indicator to analyse

Visualisations of 'Video vs Non-Video' Selftext Posts



- More non-Video Selftext Posts
- Subreddits are good to go!

Data Visualisation – 1-gram Count Vectorizers



DotA2 subreddit

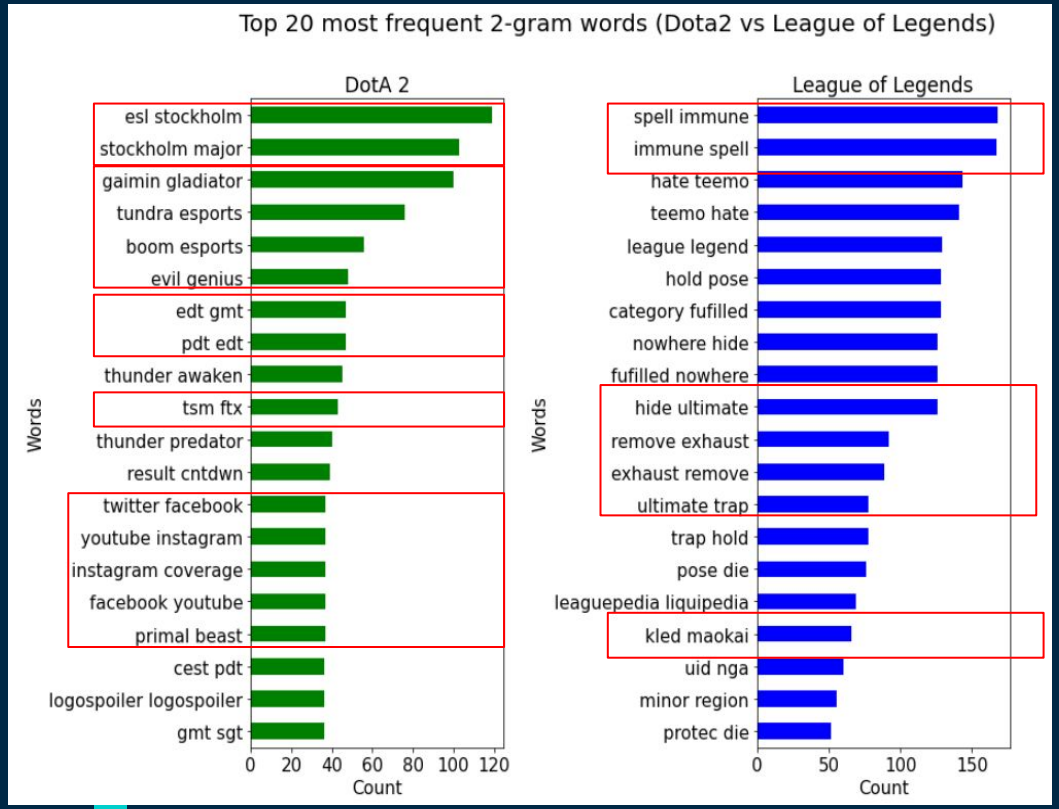
- Centers around tournaments, e.g. ESL Stockholm, happened 12 - 22 May 2022
- TI (The International) & DPC (Dota Pro Circuit) refers to the top tournaments and placements for top teams
- MMRs - ranked match discussions

League of Legends subreddit

- 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.

Data Visualisation – 2-grams Count Vectorizers



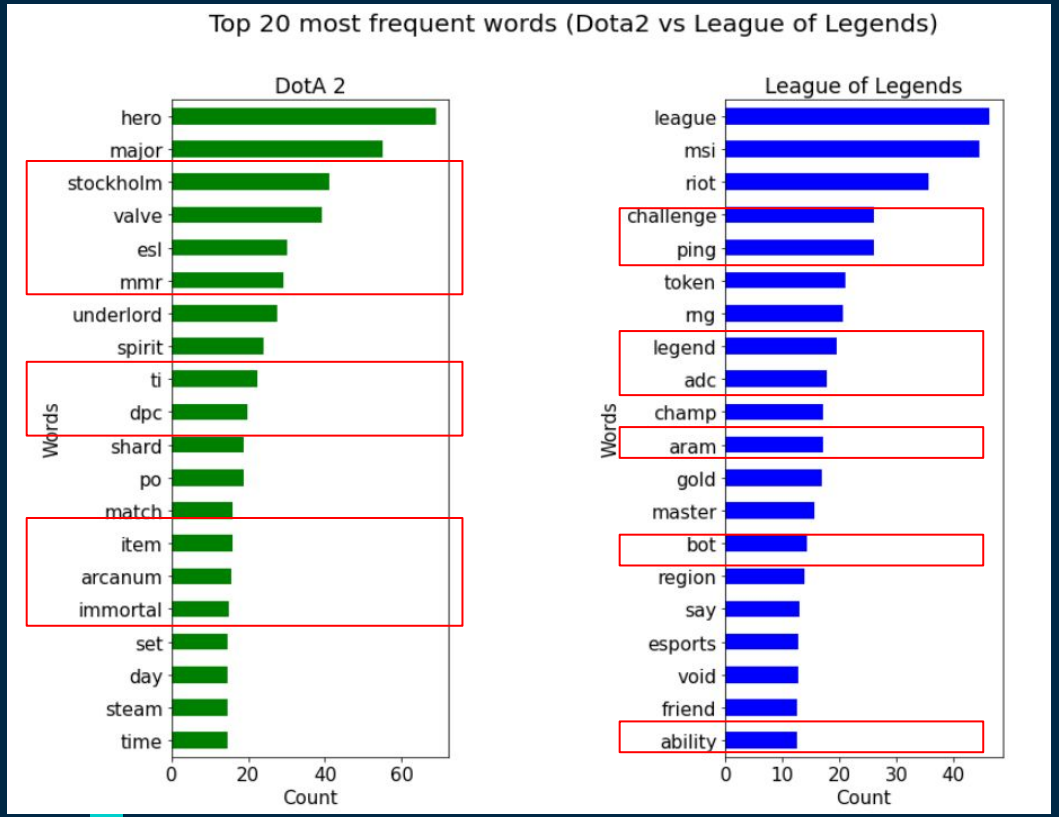
<u>DotA2 subreddit</u>

- 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)

<u>League of Legends subreddit</u>

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

Data Visualisation - 1-gram TF-IDF Vectorizers



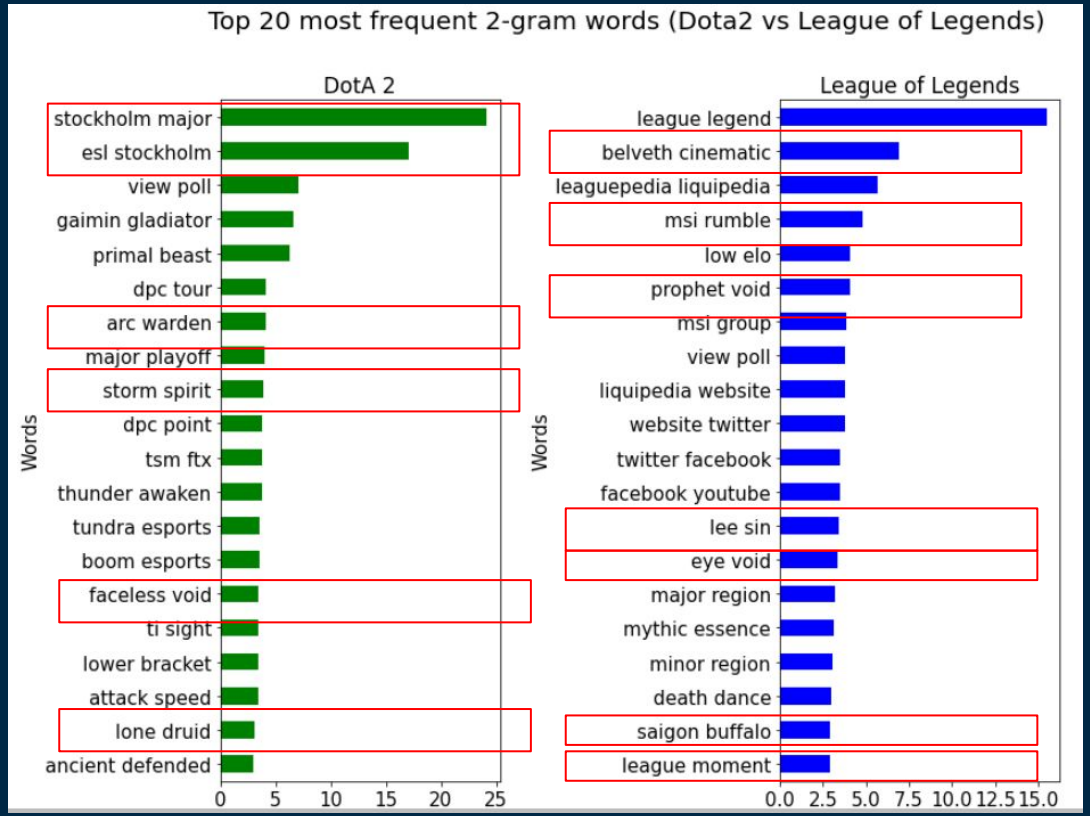
DotA2 subreddit

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

League of Legends subreddit

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

Data Visualisation – 2-grams TF-IDF Vectorizers



<u>DotA2 subreddit</u>

- 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

<u>League of Legends subreddit</u>

- 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
- Tournaments (ESL, TI, DPC) - MMRs - ranked match discussions	- Tournaments (ESL Stockholm and Stockholm majors) - Esports teams (Gaimin Gladiator, Evil Genius, etc)	- Tournaments (ESL, TI, DPC) - Items and items rarity (Arcanum and Immortal)	- Tournaments (ESL, TI, DPC) - Hero names (Arc Warden, Storm Spirit, Faceless Void, Lone Druid)

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
- Gameplay (ability, challenge, spell, ultimate, immune) - Strategy talks (Attack Damage Carry) and bot (Bottom lane)	- Gameplay (spell immune, remove exhaust, ultimate trap) - Head to head top lane heroes discussion (Kled & Maokai)	- Gameplay (ability, challenge, spell, ultimate, immune) - Strategy talks (Attack Damage Carry) ,bot (Bottom lane)	- Hero names (Prophet Void, Eye Void, Lee Sin, Belveth Cinematics) - Tournament jargons (Saigon Buffalo, MSI rumble)

Model Building & Testing

Base Model

Logistic Regression

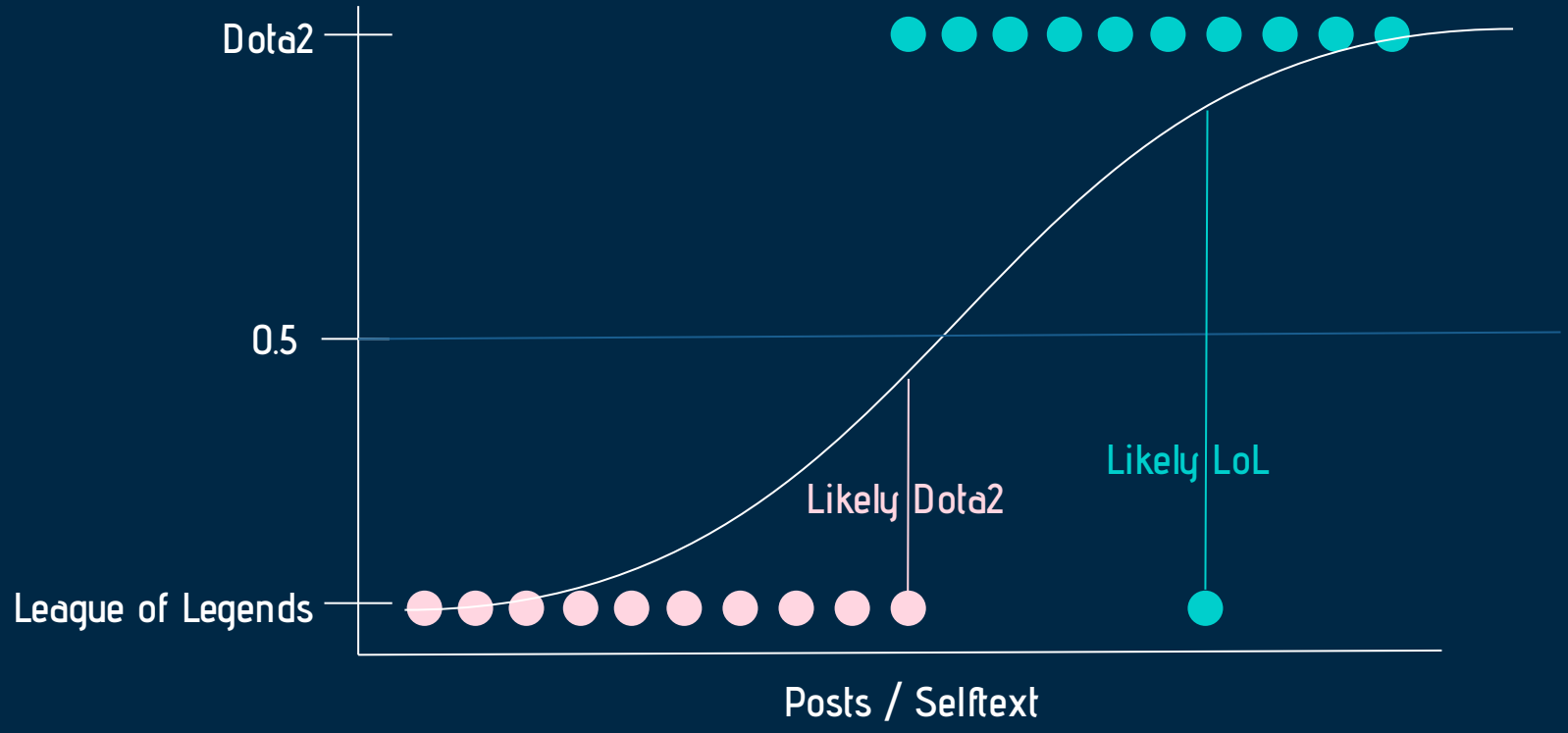
Random Forest

Multinomial Bay Naives

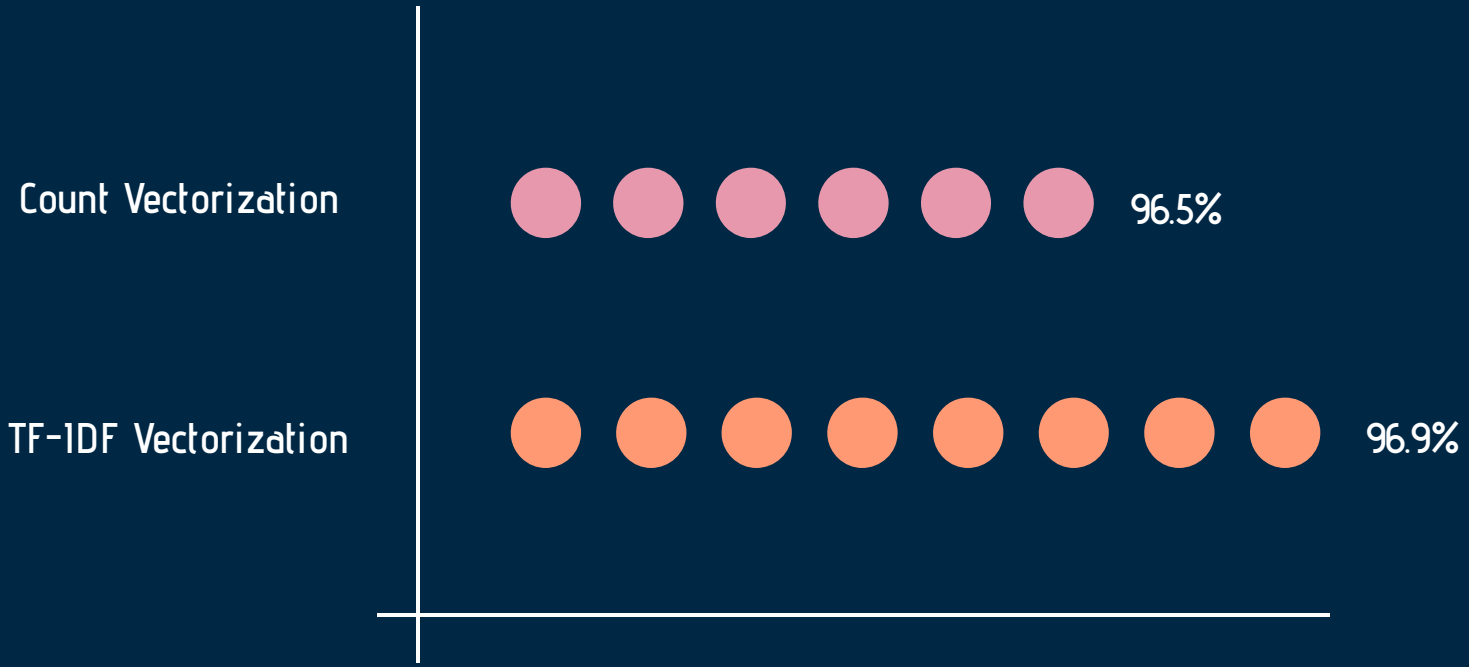
Count Vectorization

TFIDF Vectorization

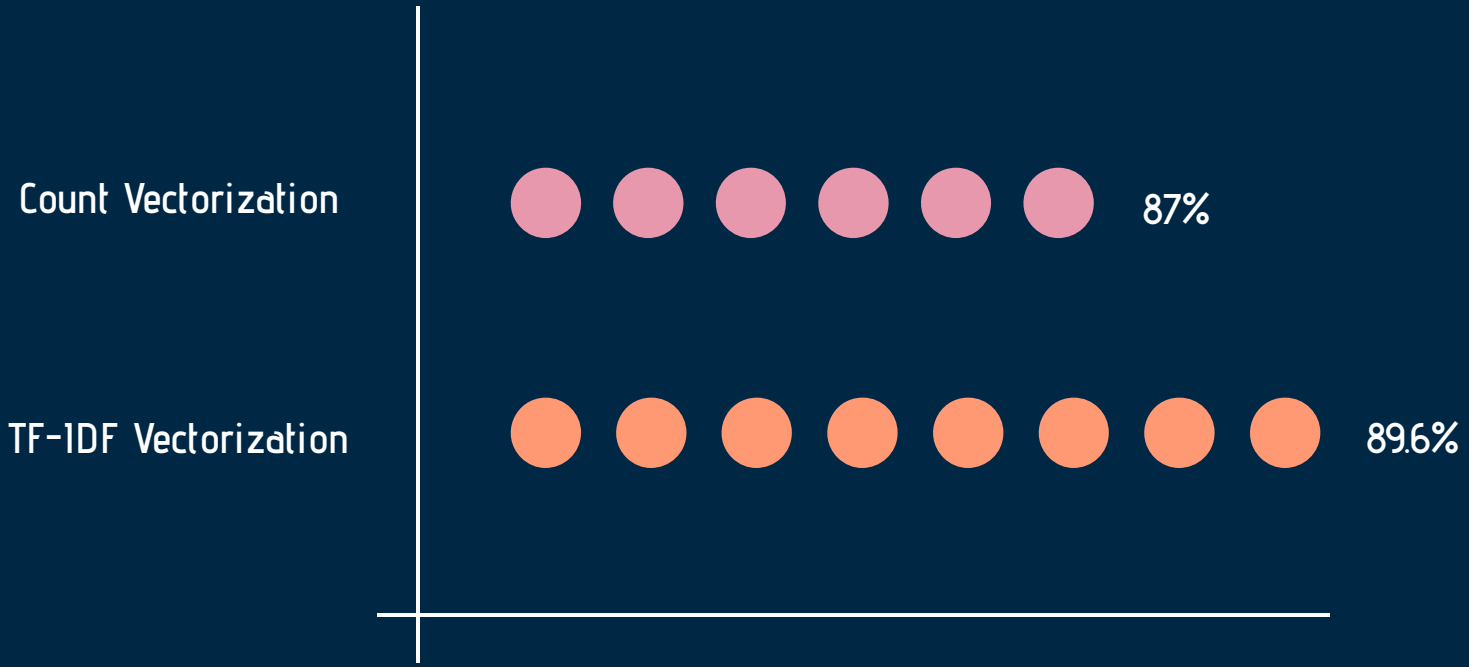
Logistic Regression



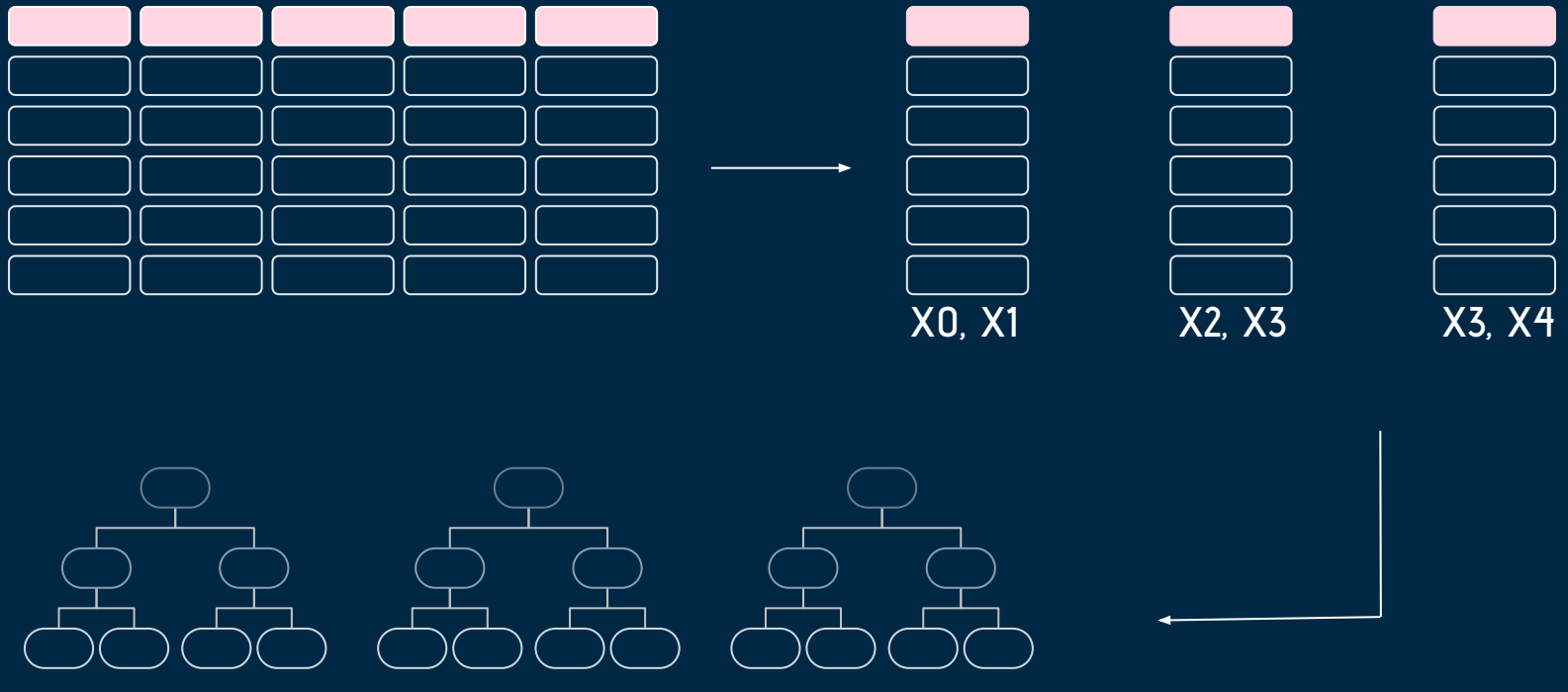
Logistic Regression (Train)



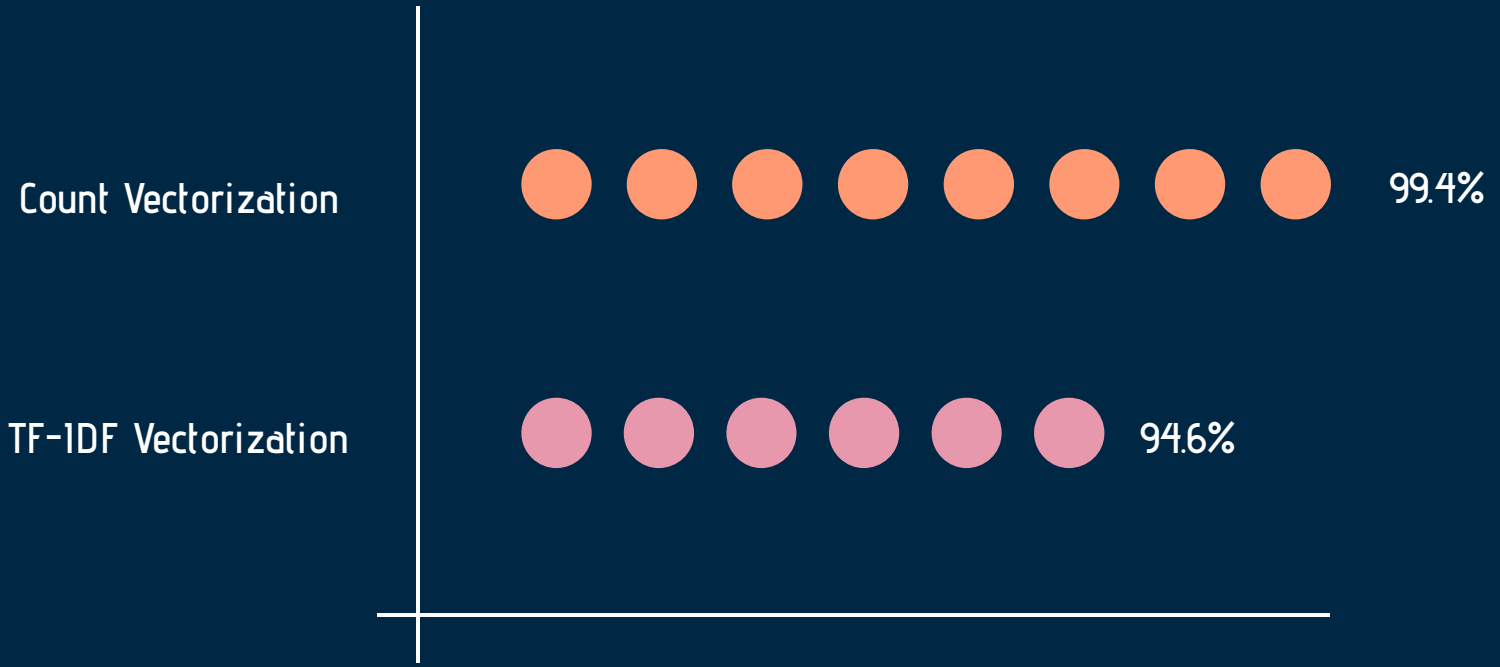
Logistic Regression (Test)



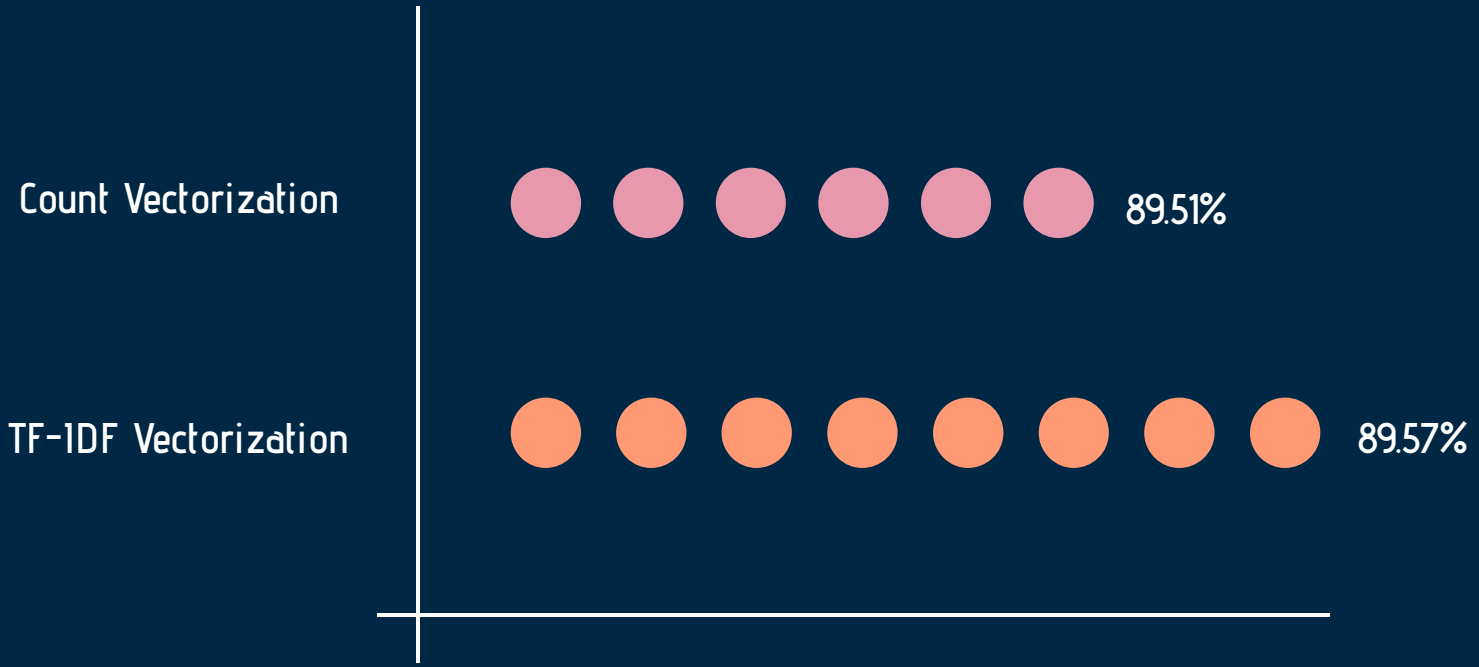
Random Forest Algorithm



Random Forest (Train)



Random Forest (Test)



Multinomial Naive Bayes

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

Step 1

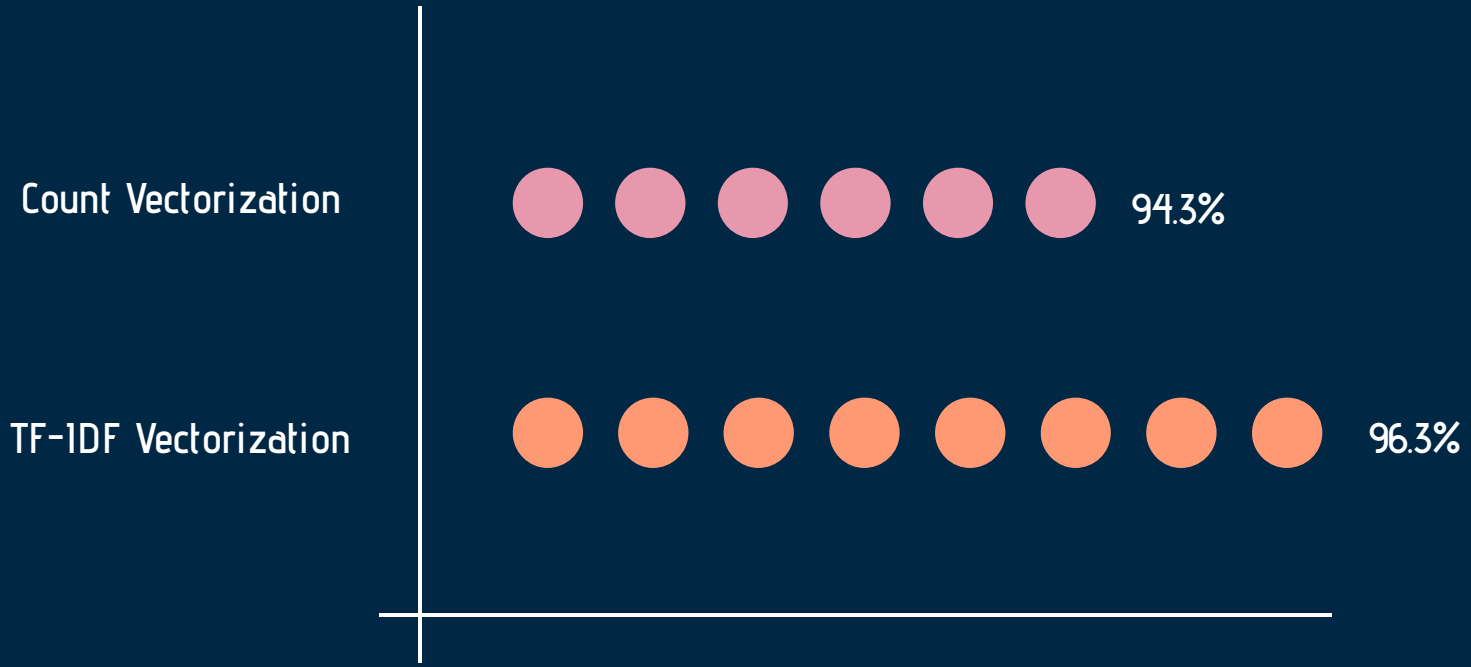
Measure prior probability of the respective category, and the probability of each discrete words given the respective categories

Step 2

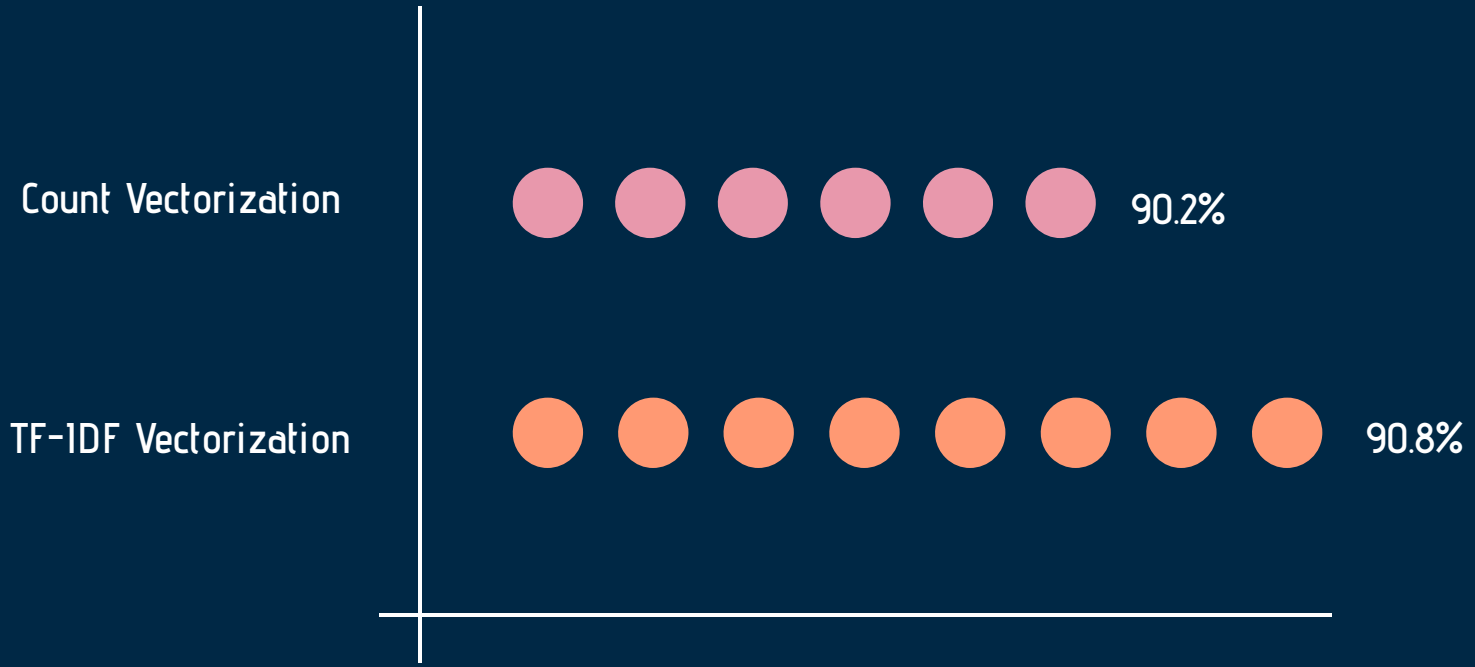
For a given sentence, calculate and compare the probability of the sentence being classified as each of the respective category, given each of the sentence's discrete words

Step 3

Multinomial Naive Bayes (Train)



Multinomial Naive Bayes (Test)



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 + TF1DF

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

####Day 3 (May 14th) - Group Stage

ID | Team | vs | Team | Result | Cntdwn (CEST) | PDT | EDT | GMT | SGT | AEST | Frmt | Stream
:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:|:--:
A5 | [Evil Geniuses](/logo-eg "Evil Geniuses") | vs | BetBoom Team | >!0:0!< | [**12:00**](https://www.timeanddate.com/countdown/generic?iso=20220514T12&p0=239&font=cursive&csz=1) | 3:00 | 6:00 | 10:00 | 18:00 | 20:00 | Bo2 | [/esl_dota2](https://www.twitch.tv/esl_dota2)
A5 | [Boom Esports](/logo-boomid "Boom Esports") | vs | [T1](/logo-t1 "T1") | >!0:0!< | [**12:00**](https://www.timeanddate.com/countdown/generic?iso=20220514T12&p0=239&font=cursive&csz=1) | 3:00 | 6:00 | 10:00 | 18:00 | 20:00 | Bo2 | [/esl_dota2b](https://www.twitch.tv/esl_dota2b)
A5 | [beastcoast](/logo-beastcoast "beastcoast") | vs | [Tundra Esports](/logo-tundra "Tundra Esports") | >!0:0!< |
[**12:00**](https://www.timeanddate.com/countdown/generic?iso=20220514T12&p0=239&font=cursive&csz=1) | 3:00 | 6:00 | 10:00 | 18:00 | 20:00 | Bo2 | [/esl](https://www.twitch.tv/esl)
||||||||||||
A6 | [Evil Geniuses](/logo-eg "Evil Geniuses") | vs | [OG](/logo-og "OG") | >!0:0!< | [**14:30**](https://www.timeanddate.com/countdown/generic?iso=20220514T1430&p0=239&font=cursive&csz=1) | 5:30 | 8:30 | 12:30 | 20:30 | 22:30 | Bo2 | [/esl_dota2](https://www.twitch.tv/esl_dota2)
A6 | [Boom Esports](/logo-boomid "Boom Esports") | vs | [Tundra Esports](/logo-tundra "Tundra Esports") | >!0:0!< |
[**14:30**](https://www.timeanddate.com/countdown/generic?iso=20220514T1430&p0=239&font=cursive&csz=1) | 5:30 | 8:30 | 12:30 | 20:30 | 22:30 | Bo2 |
[/esl_dota2b](https://www.twitch.tv/esl_dota2b)
A6 | [T1](/logo-t1 "T1") | vs | BetBoom Team | >!0:0!< | [**14:30**](https://www.timeanddate.com/countdown/generic?iso=20220514T1430&p0=239&font=cursive&csz=1) | 5:30 | 8:30 | 12:30 | 20:30 | 22:30 | Bo2 | [/esl](https://www.twitch.tv/esl)
||||||||||||
B5 | [Team Liquid](/logo-liquid "Team Liquid") | vs | [TSM](/logo-tsm "TSM") | >!0:0!< |
[**17:20**](https://www.timeanddate.com/countdown/generic?iso=20220514T1720&p0=239&font=cursive&csz=1) | 8:20 | 11:20 | 14:20 | 23:20 | 1:20 | Bo2 |
[/esl_dota2b](https://www.twitch.tv/esl_dota2b)
B5 | [Fnatic](/logo-fnatic "Fnatic") | vs | [Thunder Awaken](/logo-thunder "Thunder Awaken") | >!0:0!< |
[**17:20**](https://www.timeanddate.com/countdown/generic?iso=20220514T1720&p0=239&font=cursive&csz=1) | 8:20 | 11:20 | 14:20 | 23:20 | 1:20 | Bo2 | [/esl_dota2](https://www.twitch.tv/esl_dota2)
||||||||||||
B6 | [Team Liquid](/logo-liquid "Team Liquid") | vs | [Team Spirit](/logo-spirit "Team Spirit") | >!0:0!< |
[**19:40**](https://www.timeanddate.com/countdown/generic?iso=20220514T1940&p0=239&font=cursive&csz=1) | 10:40 | 13:40 | 16:40 | 1:40 | 3:40 | Bo2 | [/esl_dota2](https://www.twitch.tv/esl_dota2)
B6 | [Thunder Awaken](/logo-thunder "Thunder Awaken") | vs | [Gaimin Gladiators](/logo-gg "Gaimin Gladiators") | >!0:0!< |
[**19:40**](https://www.timeanddate.com/countdown/generic?iso=20220514T1940&p0=239&font=cursive&csz=1) | 10:40 | 13:40 | 16:40 | 1:40 | 3:40 | Bo2 |
[/esl_dota2b](https://www.twitch.tv/esl_dota2b)

* Mind Games are disqualified due to visa issues

Countdown times are in CEST. All times are subject to change based on the length of matches and delays.
Other match discussions: [/r/dota2 on Discord](https://discord.gg/ctRYVpW)

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 to

-
- Better data cleaning steps e.g. remove Character Entities

2) Include New Features

- Sentiment analysis

Do you have any questions?

youremail@freepik.com
+91 620 421 83
yourcompany.com

THANKS