Subreddit Classification NLP

Auto Chess vs Teamfight Tactics

By Mingzi

Table of Contents

01 **Objective**

02 **Background**

03 **Process**

04 **EDA**

Text preprocessing

05 **Modelling**

RandomForest

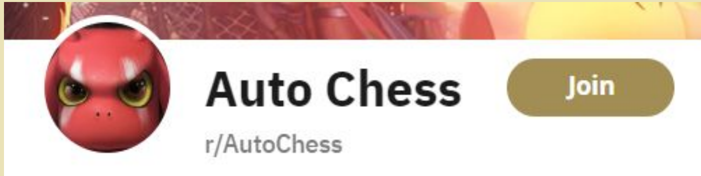
Naive Bayes

Logistic Regression

06 **Conclusion**

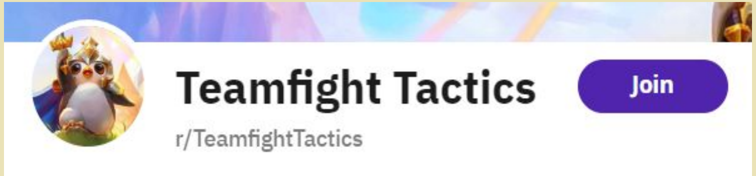
1. Objectives

Use NLP to train on classifier to identify which post belongs to which subreddit



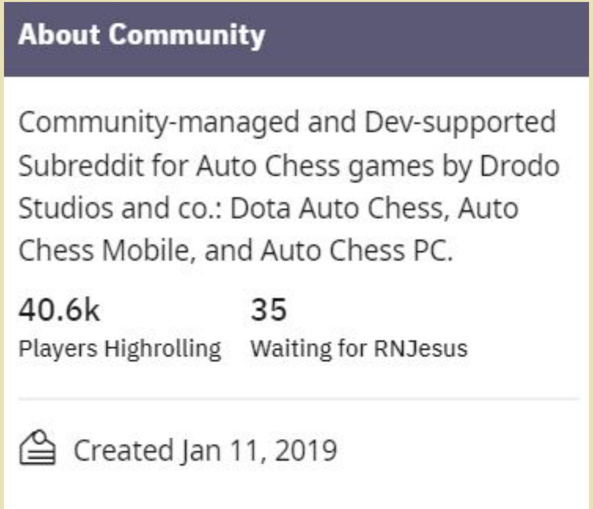
Auto Chess

r/AutoChess



Teamfight Tactics

r/TeamfightTactics



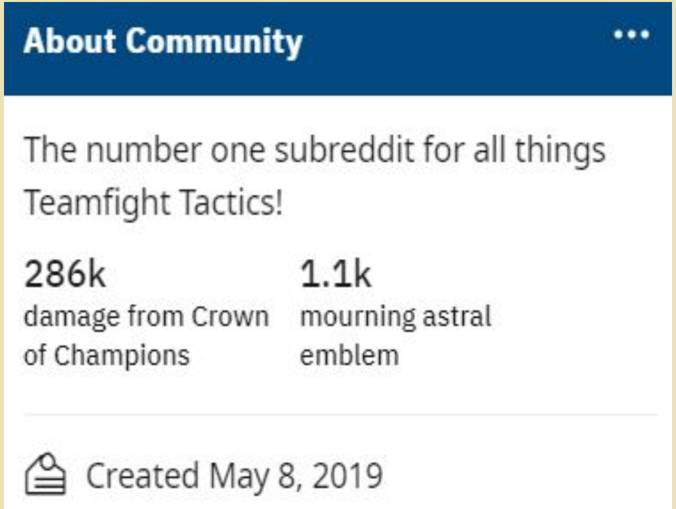
About Community

Community-managed and Dev-supported Subreddit for Auto Chess games by Drodo Studios and co.: Dota Auto Chess, Auto Chess Mobile, and Auto Chess PC.

40.6k Players Highrolling

35 Waiting for RNJesus

Created Jan 11, 2019



About Community

The number one subreddit for all things Teamfight Tactics!

286k damage from Crown of Champions

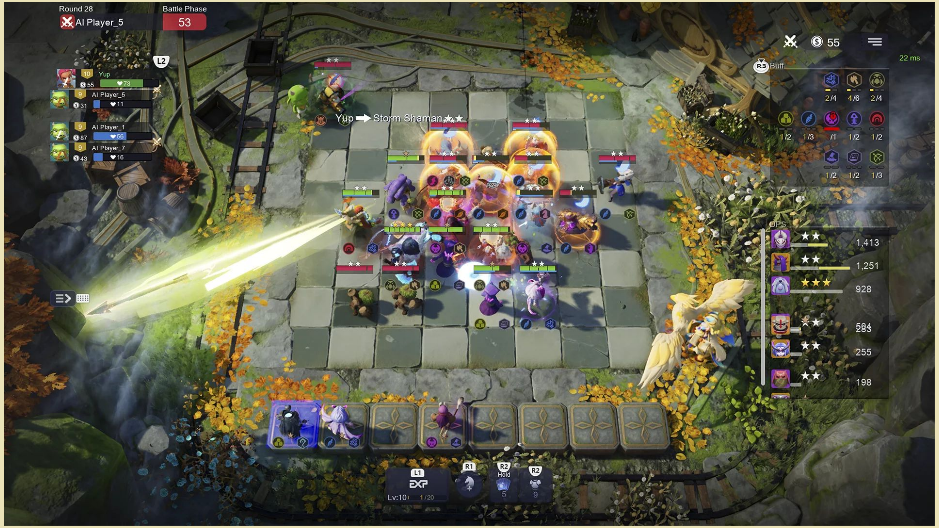
1.1k mourning astral emblem

Created May 8, 2019

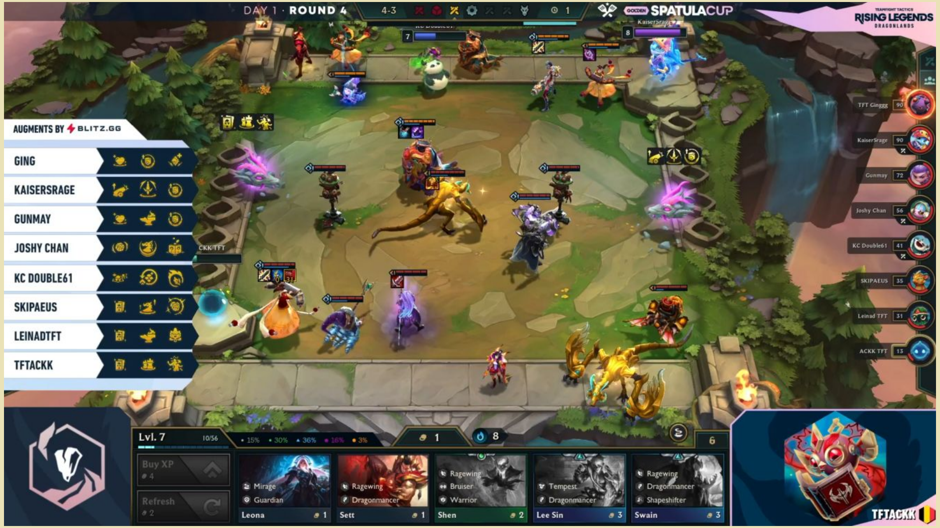
2. Background

Auto Chess & Teamfight Tactics

Auto Chess



Teamfight Tactics



- Both are auto battler online games
- Require players buy units with in-game gold, level up, upgrade

3. Process

General Flow of Notebook

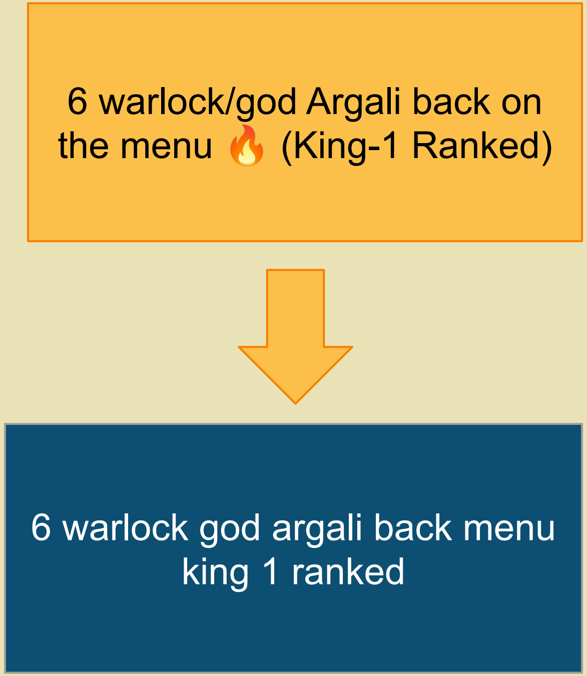
1. Obtain data Pushshift's API (49.9% - 50.0%)
   a. 2000 posts from Auto Chess
   b. 2000 posts from TFT
2. Exploratory Data Analysis
   a. Data Cleaning
   b. Detailed Text Preprocessing
   c. Lemmatization
3. Modelling
   a. Random Forest
   b. Naive Bayes
   c. Logistic Regression
4. Conclusion

4. EDA

Data Cleaning Text Processing

General Steps for text preprocessing

1. Convert words to lowercase
2. Remove newlines and tabs
3. Strip HTML tags
4. Remove links
5. Dealing with expand contractions (didn’t -> did not)
6. Remove stopwords
7. Remove special Characters (#@)
8. Remove whitespace
9. Lemmatization



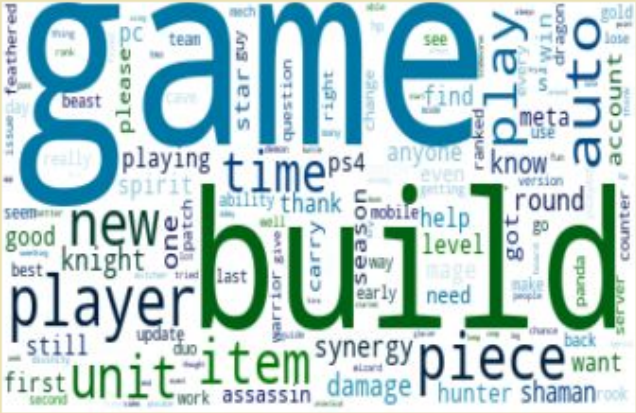
6 warlock/god Argali back on the menu 🔥 (King-1 Ranked)

6 warlock god argali back menu king 1 ranked

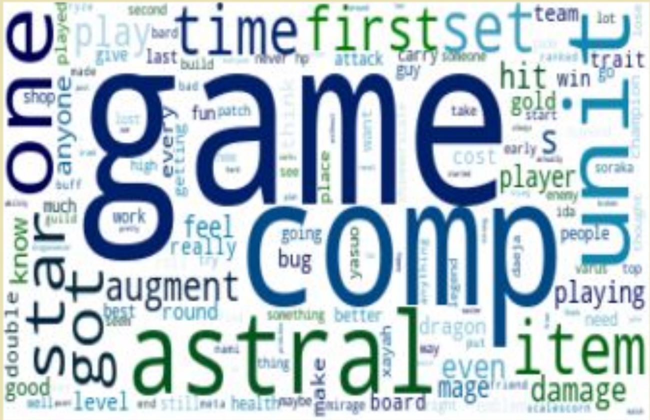
4. EDA

Visualisation

Auto Chess



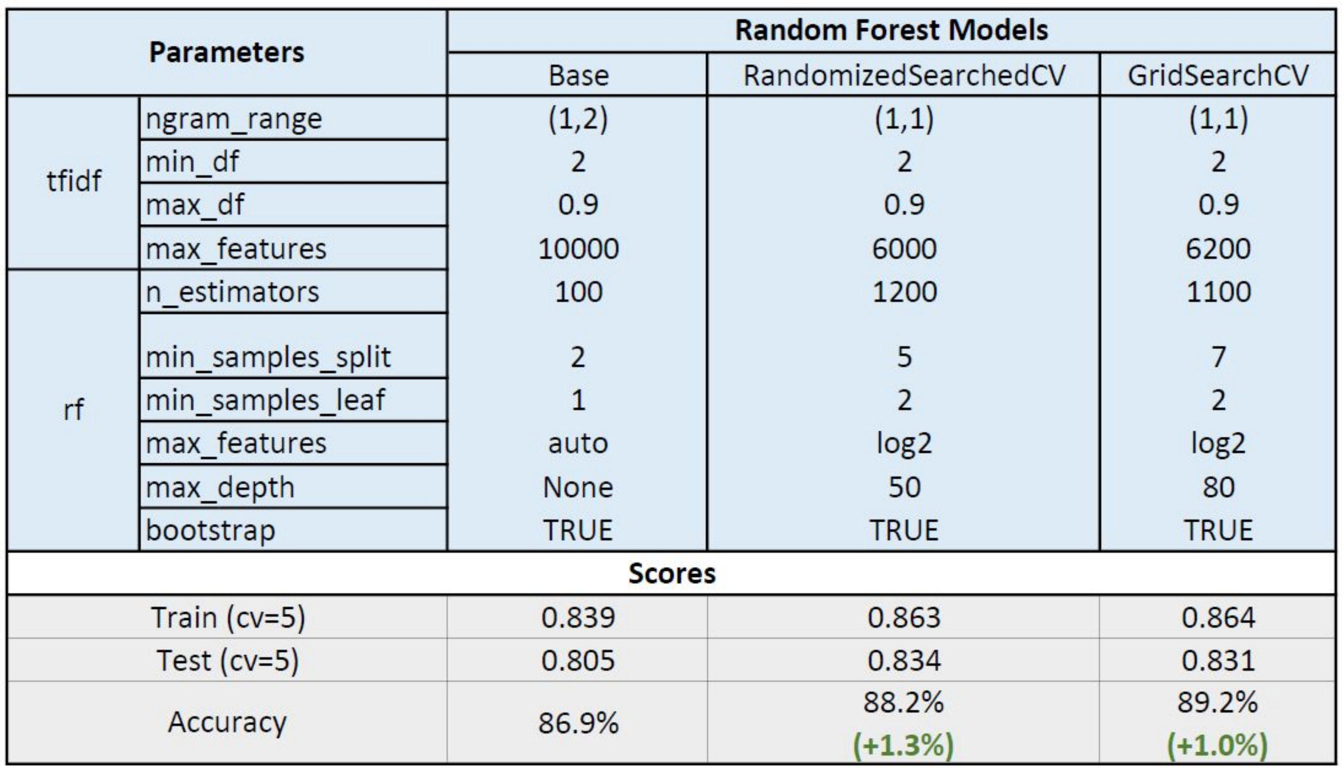
Teamfight Tactics



- There are a few words that occur quite frequently : ‘game’, ‘build’, ‘unit’, ‘time’, ‘item’.

5. Modelling

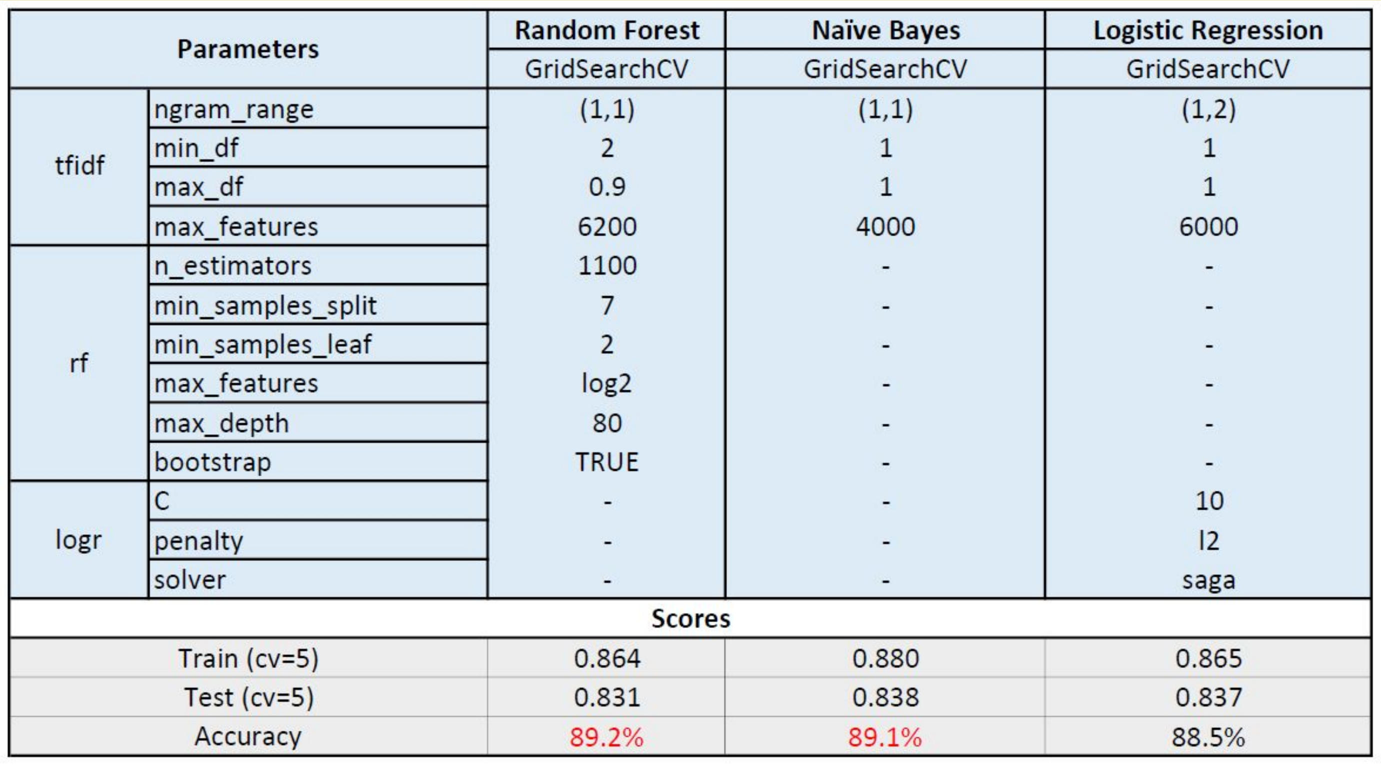
RandomForest



Parameters		Random Forest Models		
		Base	RandomizedSearchedCV	GridSearchCV
tfidf	ngram_range	(1,2)	(1,1)	(1,1)
	min_df	2	2	2
	max_df	0.9	0.9	0.9
	max_features	10000	6000	6200
rf	n_estimators	100	1200	1100
	min_samples_split	2	5	7
	min_samples_leaf	1	2	2
	max_features	auto	log2	log2
	max_depth	None	50	80
	bootstrap	TRUE	TRUE	TRUE
Scores				
Train (cv=5)		0.839	0.863	0.864
Test (cv=5)		0.805	0.834	0.831
Accuracy		86.9%	88.2% (+1.3%)	89.2% (+1.0%)

5. Modelling

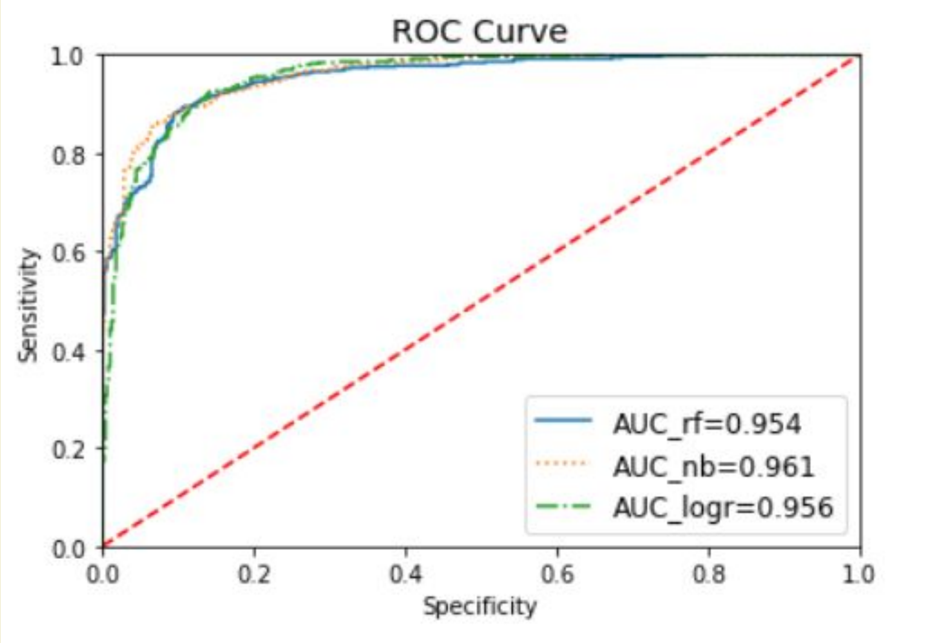
Comparing RF with Naive Bayes and Logistic Regression



Parameters		Random Forest	Naïve Bayes	Logistic Regression
		GridSearchCV	GridSearchCV	GridSearchCV
tfidf	ngram_range	(1,1)	(1,1)	(1,2)
	min_df	2	1	1
	max_df	0.9	1	1
	max_features	6200	4000	6000
rf	n_estimators	1100	-	-
	min_samples_split	7	-	-
	min_samples_leaf	2	-	-
	max_features	log2	-	-
	max_depth	80	-	-
	bootstrap	TRUE	-	-
logr	C	-	-	10
	penalty	-	-	l2
	solver	-	-	saga
Scores				
Train (cv=5)		0.864	0.880	0.865
Test (cv=5)		0.831	0.838	0.837
Accuracy		89.2%	89.1%	88.5%

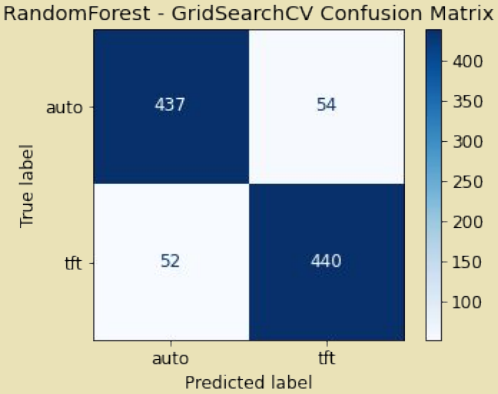
5. Modelling

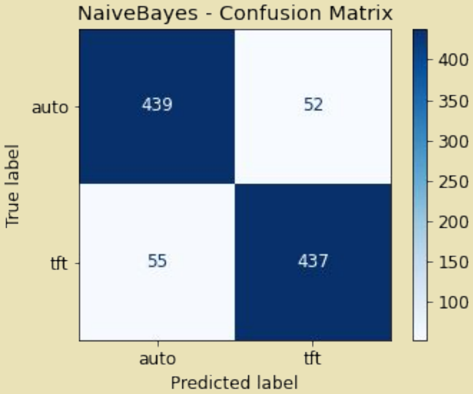
Evaluation of Models

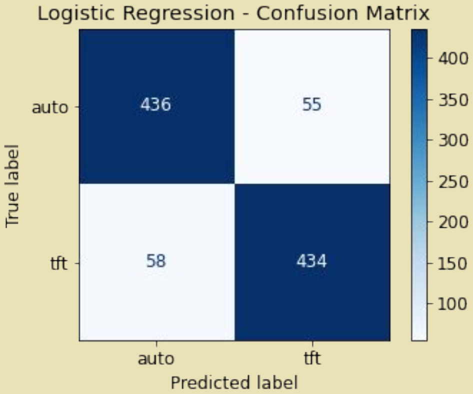


5. Modelling

Evaluation of Models







Specificity	0.890	0.894	0.887
Sensitivity	0.894	0.888	0.882
Accuracy	89.2%	89.1%	88.5%

6. Conclusion

Both random forest and Naive Bayes models perform well for this classification problem.

Recommendations

1. Text preprocessing - misspelled words, non-english languages
2. Compare with other models (SVM, Bayesian Network etc.)
3. Consider scraping TFT post by date due to frequent patch updates

Thanks

Do you have any questions?