Project 3

insertnames

Problem statement

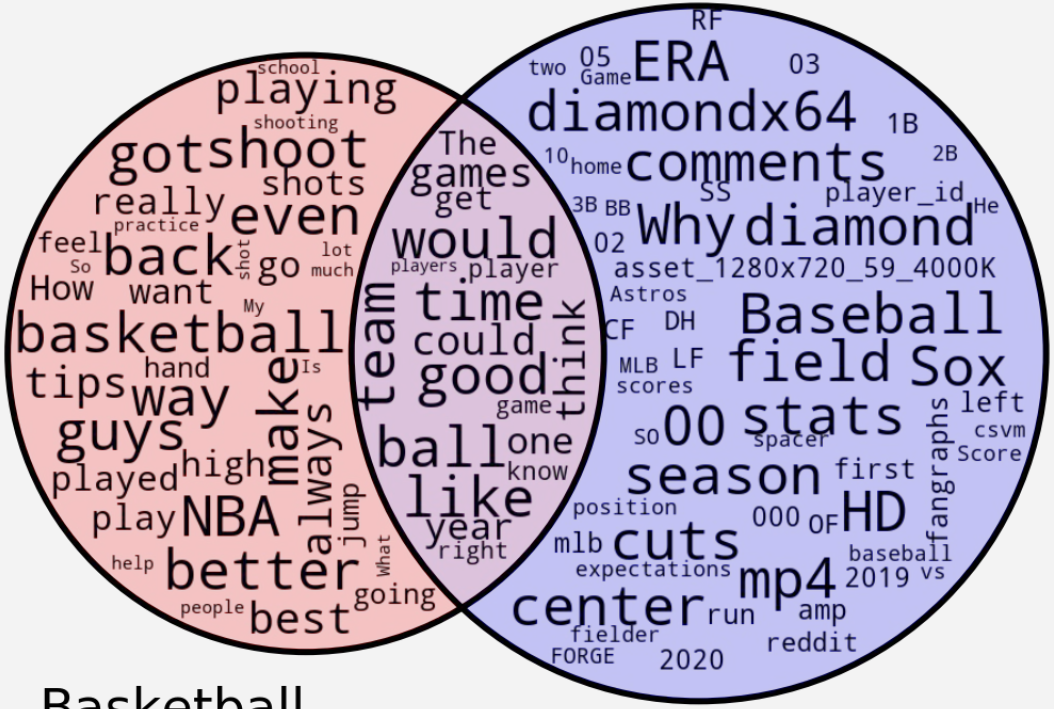
Using NLP, we created a classification model that can accurately identify social media posts into basketball and baseball interest groups to assist a VR gaming company in:

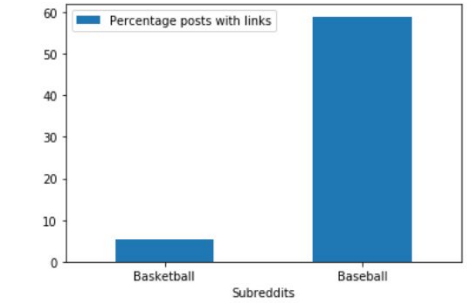
1. **cost savings on their advertising campaign**
2. **formation of B2B partnerships**

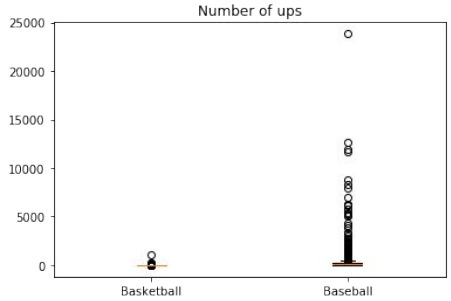
Cleaning

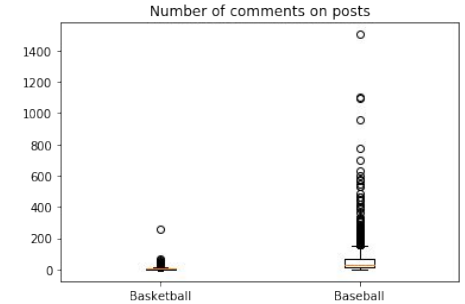
- **Dropping of duplicate posts (title)**
- **Removing admin posts**
- **Joining of selftexts and title**

Most Frequent Words from Basketball & Baseball Corpus







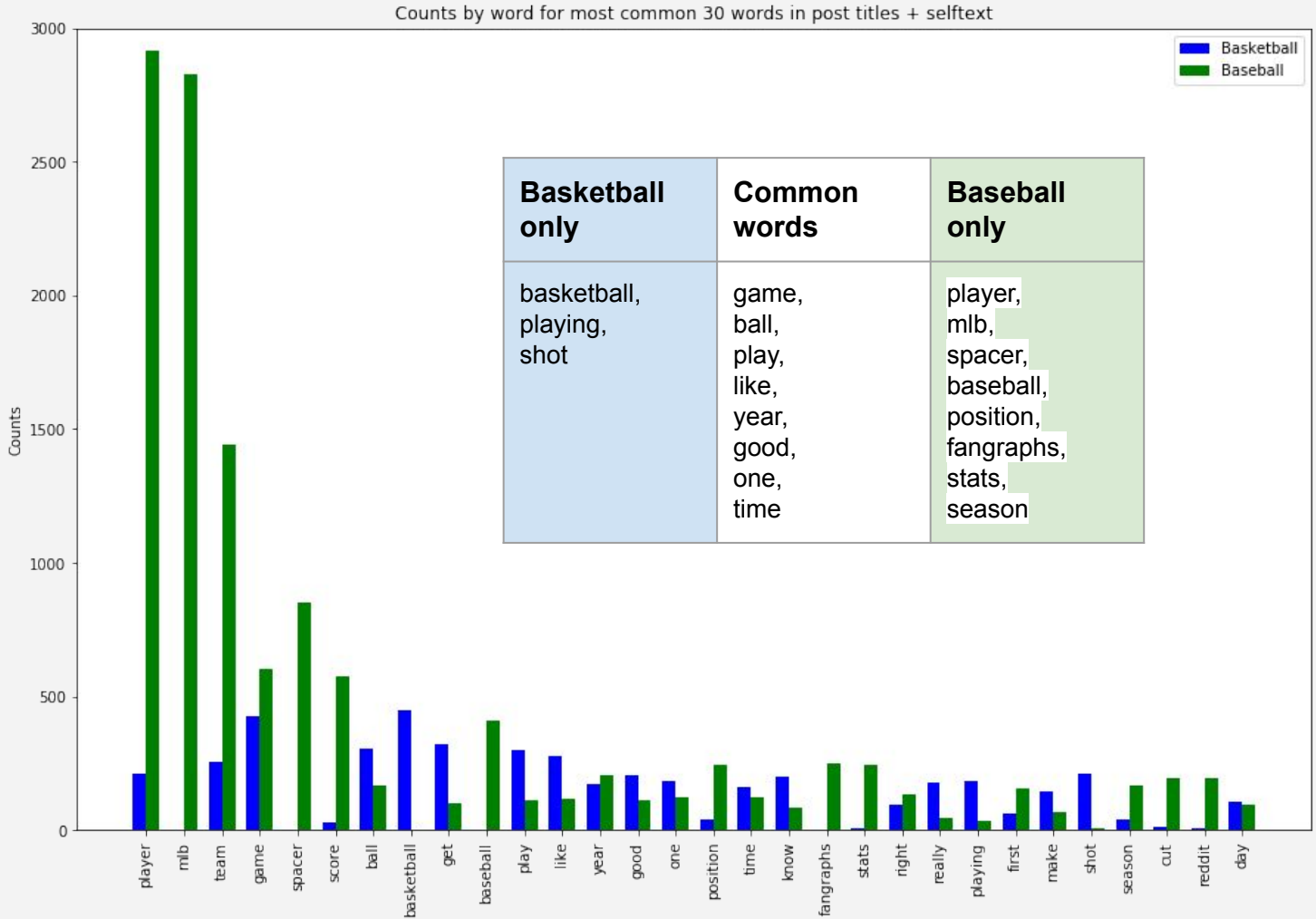


EDA - General observations

- **Baseball posts are more popular than basketball (1.1m vs 51k members)**

- **Baseball posts contain more links, while basketball posts contain more original texts**

EDA - High word counts



Basketball only	Common words	Baseball only
basketball, playing, shot	game, ball, play, like, year, good, one, time	player, mlb, spacer, baseball, position, fangraphs, stats, season

Preprocessing

1. **Converted HTML links to text**
2. **Removed non-letters**
3. **Converted to lowercase, split into individual words**
4. **Removed stop words (included extra words: ‘http’, ‘www’, ‘com’, ‘id’)**
5. **Lemmatized words**

Modeling

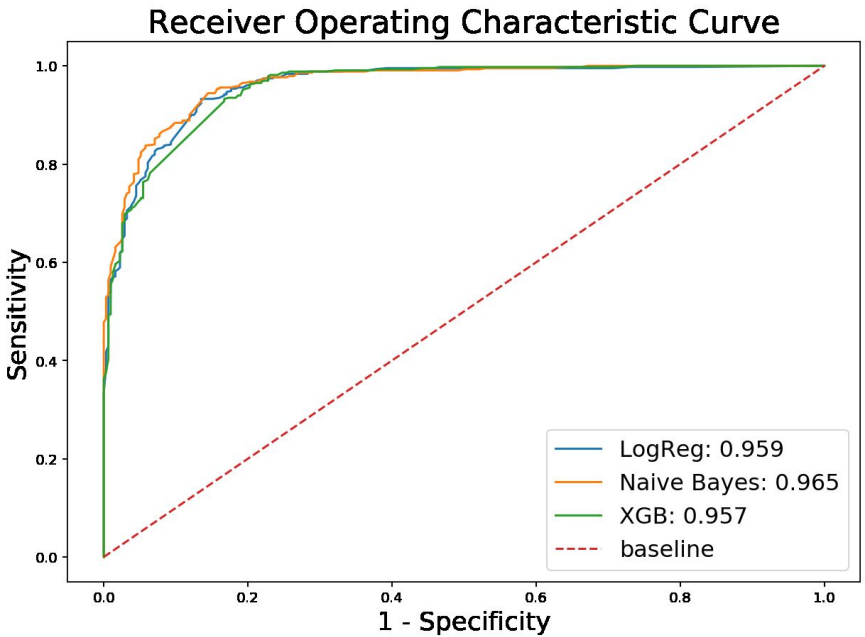
- **Logistic regression:** Classification via a linear equation like algorithm that produces a binary output
- **Naive Bayes:** Classification via a probabilistic classifier like algorithm
- **XGBoost:** Classification via a decision tree like algorithm

Which model to choose?

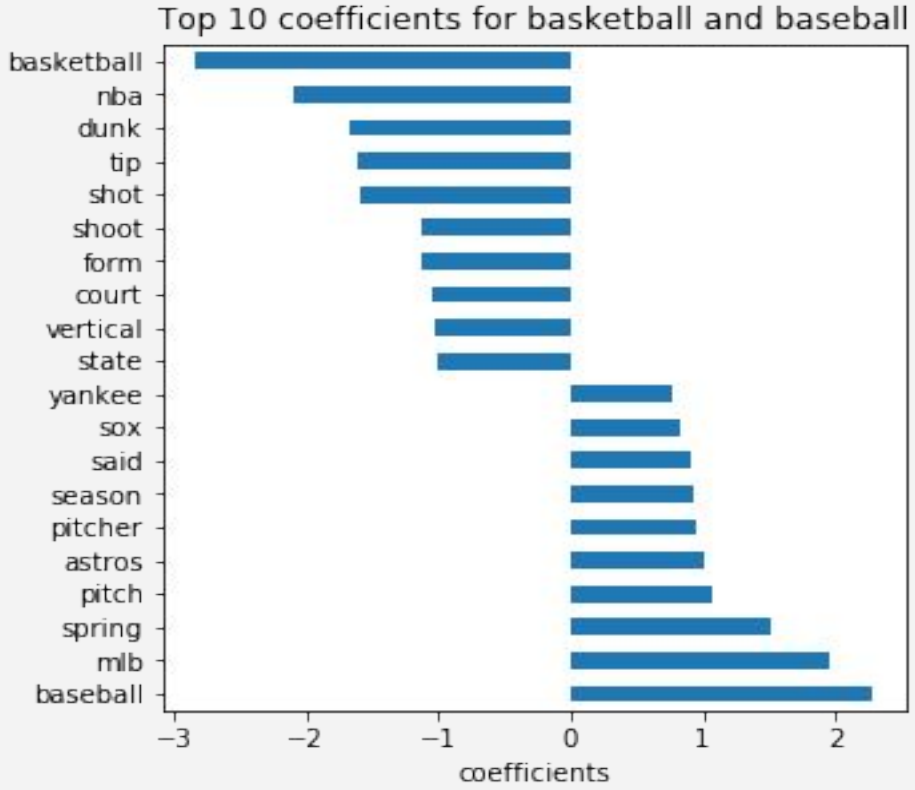
	Logistic regression		Multinomial Naïve Bayes		XGboost	
	Count vectorizer	**TF*IDF vectorizer**	**Count vectorizer**	**TF*IDF vectorizer**	**Count vectorizer**	**TF*IDF vectorizer**
Accuracy Scores	RanSearch CV scores: 0.899 train score: 0.953 test score: 0.896	RanSearch CV scores: 0.764 train score: 0.819 test score: 0.815	RanSearch CV scores: 0.871 train score: 0.878 test score: 0.872	RanSearch CV scores: 0.899 train score: 0.926 test score: 0.900	RanSearch CV scores: 0.876 train score: 0.930 test score: 0.894	RanSearch CV scores: 0.875 train score: 0.943 test score: 0.881
True Negatives: **False Positives:** **False Negatives:** **True Positives:**	True Negatives: 259 False Positives: 52 False Negatives: 25 True Positives: 406	True Negatives: 180 False Positives: 131 False Negatives: 6 True Positives: 425	True Negatives: 229 False Positives: 82 False Negatives: 13 True Positives: 418	True Negatives: 253 False Positives: 58 False Negatives: 16 True Positives: 415	True Negatives: 248 False Positives: 63 False Negatives: 16 True Positives: 415	True Negatives: 245 False Positives: 66 False Negatives: 22 True Positives: 409
Sensitivity: **Specificity:** **Precision:** **F1:**	Sensitivity: 0.942 Specificity: 0.833 Precision: 0.886 F1: 0.913	Sensitivity: 0.986 Specificity: 0.579 Precision: 0.764 F1: 0.861	Sensitivity: 0.97 Specificity: 0.736 Precision: 0.836 F1: 0.898	Sensitivity: 0.963 Specificity: 0.814 Precision: 0.877 F1: 0.918	Sensitivity: 0.963 Specificity: 0.797 Precision: 0.868 F1: 0.913	Sensitivity: 0.949 Specificity: 0.788 Precision: 0.861 F1: 0.903
AUC:	AUC: 0.959	AUC: 0.953	AUC: 0.926	AUC: 0.965	AUC: 0.957	AUC: 0.950

Baseline: 0.58

Which model to choose?



What words were the best discriminators?



Recommendations

Insights

1. We created an accurate Naive Bayes classifier to identify social media posts into basketball and baseball interest groups.
2. Basketball posts - techniques in improving the skills as evidenced by popular words such as dunk, shoot & jumpshot.
3. Baseball posts - baseball teams and games eg fangraphs, 'Mets', 'Astros', 'SOX', 'Red', 'Yankee', 'Angels' etc.

Recommendations

Applications

VR content/ add-on modules : basketball techniques, baseball teams and players

Partnerships with :

1. Trainers/coaches for techniques
2. Sports merchandisers to target baseball team-centered merchandise
3. Teams to feature player dialogues, snippets, tips and techniques

Conclusion

1. Our model predicts very well with text

2. Our insights provide :

   - Scope for business opportunities to offer specific products
   - Targeted cost effective emplacement of advertisements

3. Future improvements - detect images and videos in posts to further sharpen advertising edge.