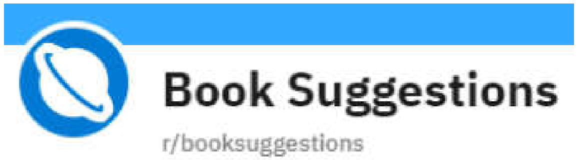
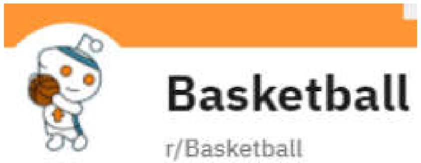
DSI Project 3
Classification Model in Reddit posts

Objectives:

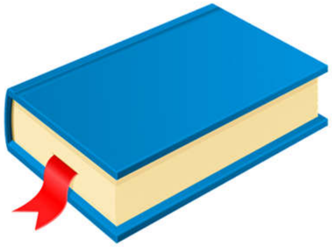
- To create the classification model to classify posts in subreddits

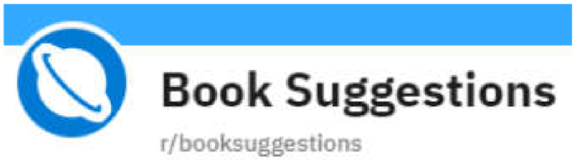
1.

Book Suggestions
r/booksuggestions

2.


Basketball
r/Basketball

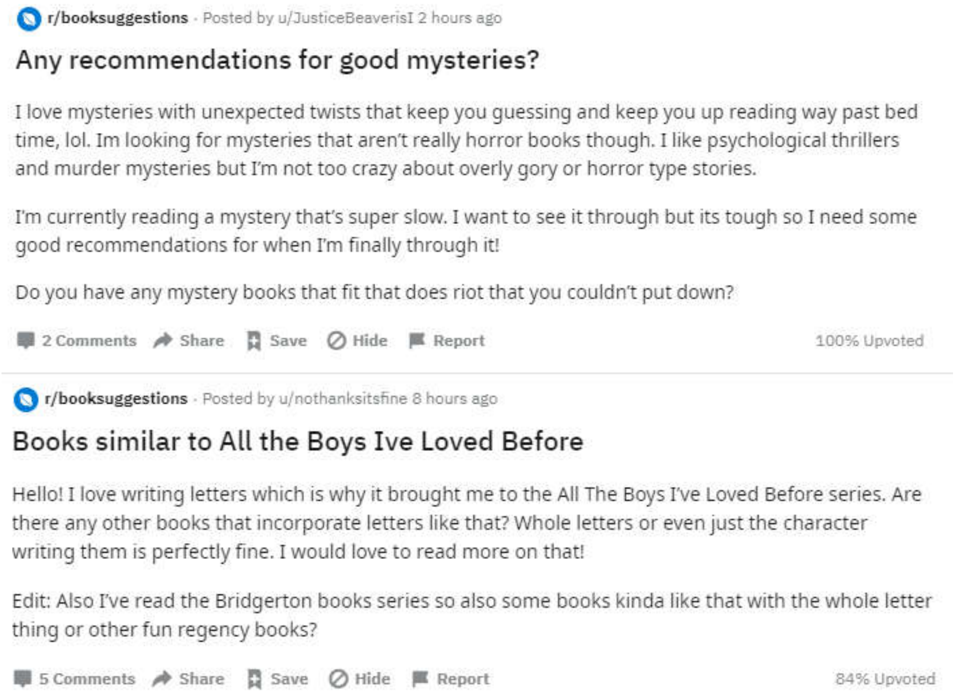
- Find a list of main key words which can identify the post is booksuggestion or basketball post





Book Suggestions

r/booksuggestions



r/booksuggestions · Posted by u/JusticeBeaverisI 2 hours ago

Any recommendations for good mysteries?

I love mysteries with unexpected twists that keep you guessing and keep you up reading way past bed time, lol. Im looking for mysteries that aren't really horror books though. I like psychological thrillers and murder mysteries but I'm not too crazy about overly gory or horror type stories.

I'm currently reading a mystery that's super slow. I want to see it through but its tough so I need some good recommendations for when I'm finally through it!

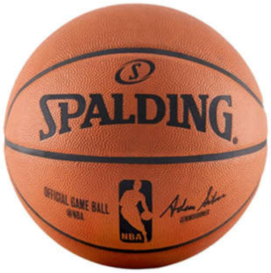
Do you have any mystery books that fit that does riot that you couldn't put down?

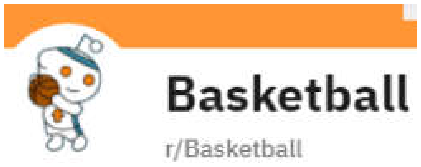
r/booksuggestions · Posted by u/nothanksitsfine 8 hours ago

Books similar to All the Boys Ive Loved Before

Hello! I love writing letters which is why it brought me to the All The Boys I've Loved Before series. Are there any other books that incorporate letters like that? Whole letters or even just the character writing them is perfectly fine. I would love to read more on that!

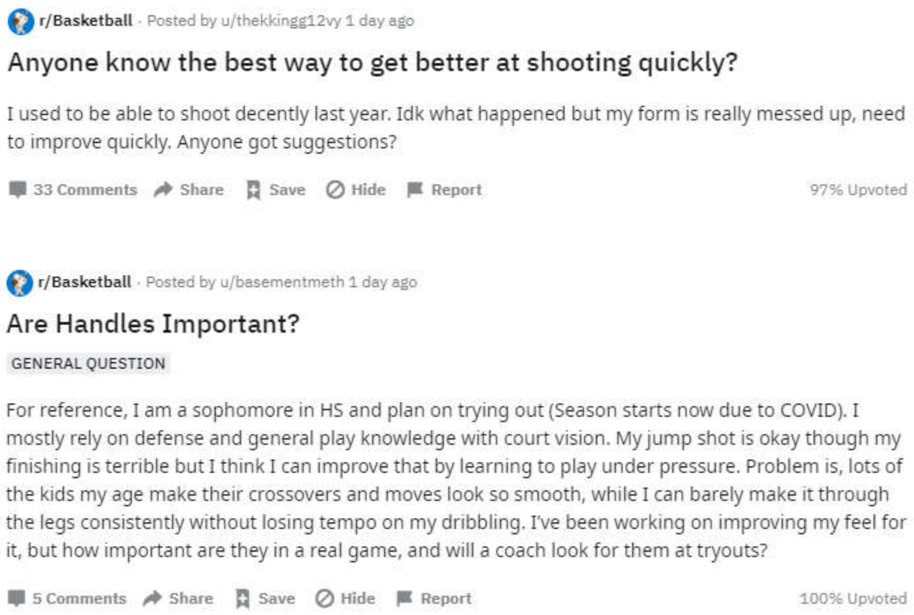
Edit: Also I've read the Bridgerton books series so also some books kinda like that with the whole letter thing or other fun regency books?





Basketball

r/Basketball



r/Basketball · Posted by u/thekkingg12vy 1 day ago

Anyone know the best way to get better at shooting quickly?

I used to be able to shoot decently last year. Idk what happened but my form is really messed up, need to improve quickly. Anyone got suggestions?

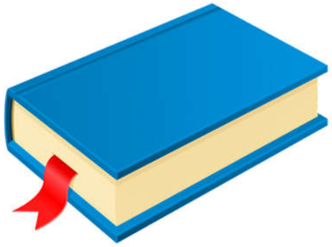
r/Basketball · Posted by u/basementmeth 1 day ago

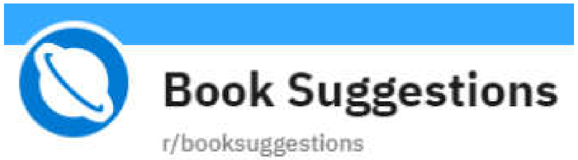
Are Handles Important?

GENERAL QUESTION

For reference, I am a sophomore in HS and plan on trying out (Season starts now due to COVID). I mostly rely on defense and general play knowledge with court vision. My jump shot is okay though my finishing is terrible but I think I can improve that by learning to play under pressure. Problem is, lots of the kids my age make their crossovers and moves look so smooth, while I can barely make it through the legs consistently without losing tempo on my dribbling. I've been working on improving my feel for it, but how important are they in a real game, and will a coach look for them at tryouts?

Exploratory data analysis



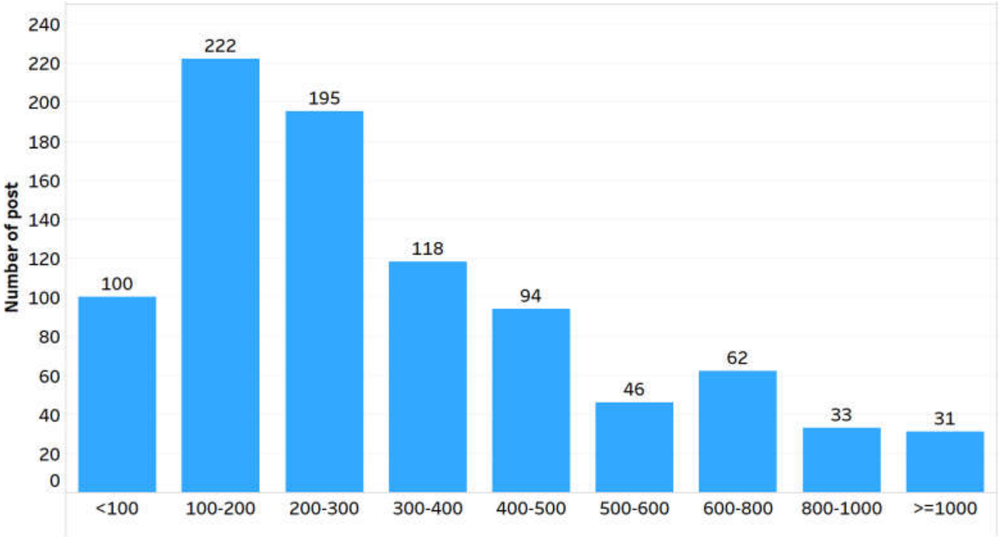


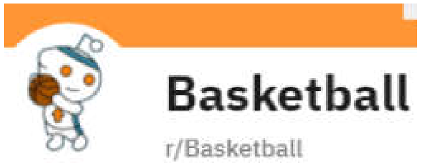
Book Suggestions

r/booksuggestions

900 posts

About **348 characters** per post



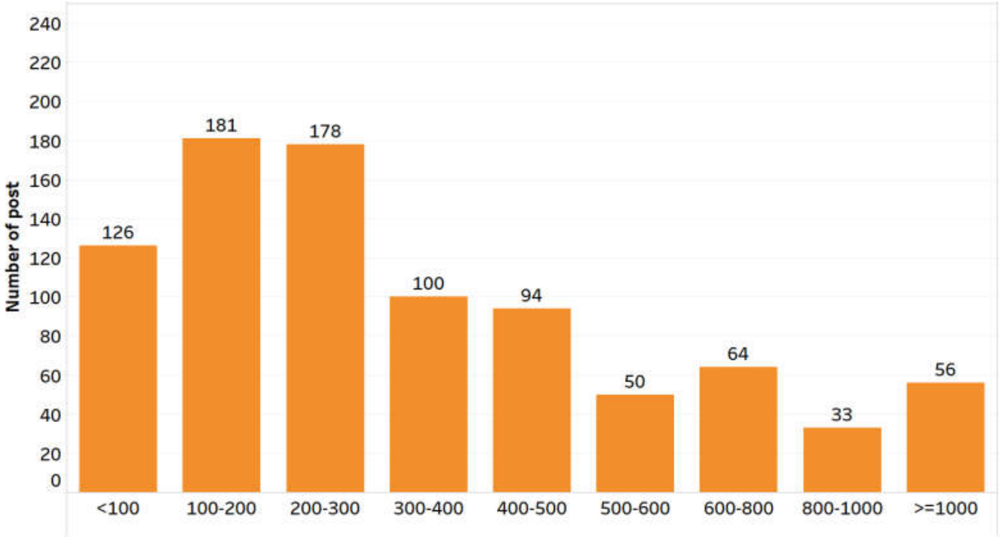


Basketball

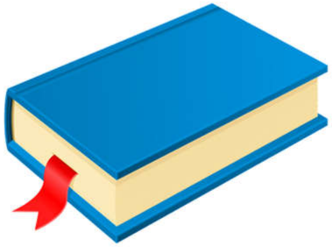
r/Basketball

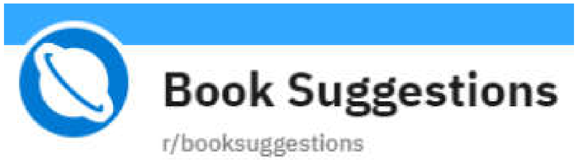
882 posts

About **393 characters** per post



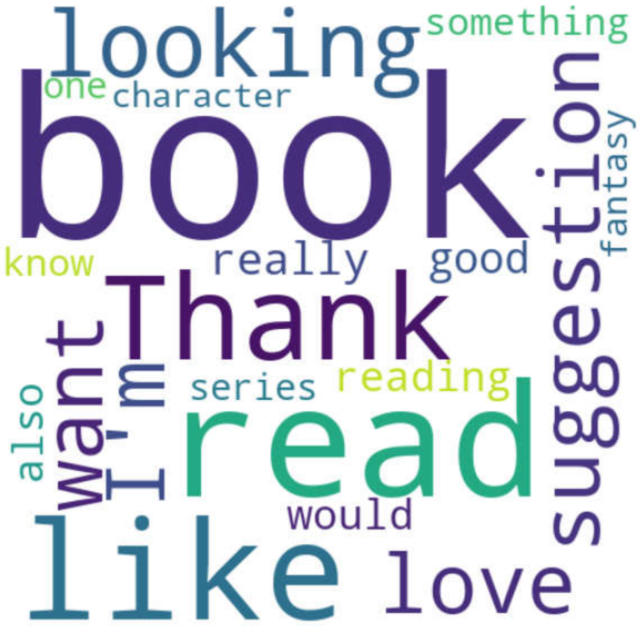
EDA: Frequency Word

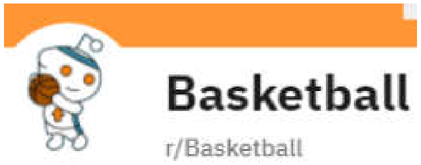




Book Suggestions

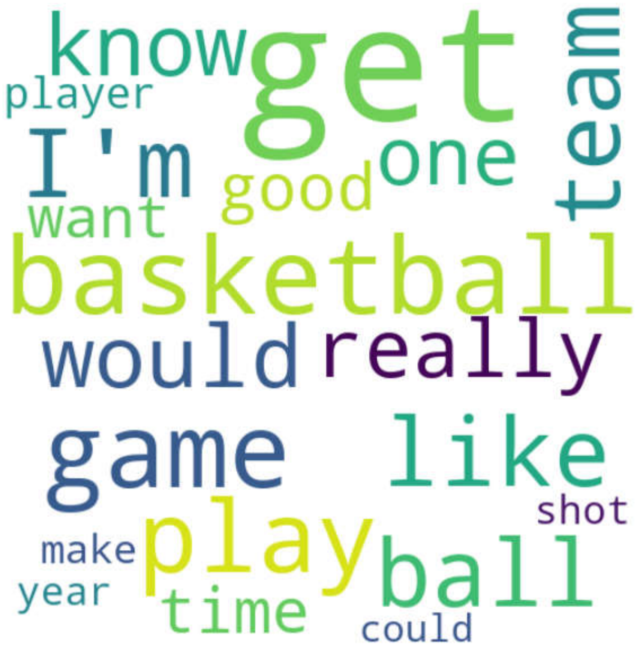
r/booksuggestions



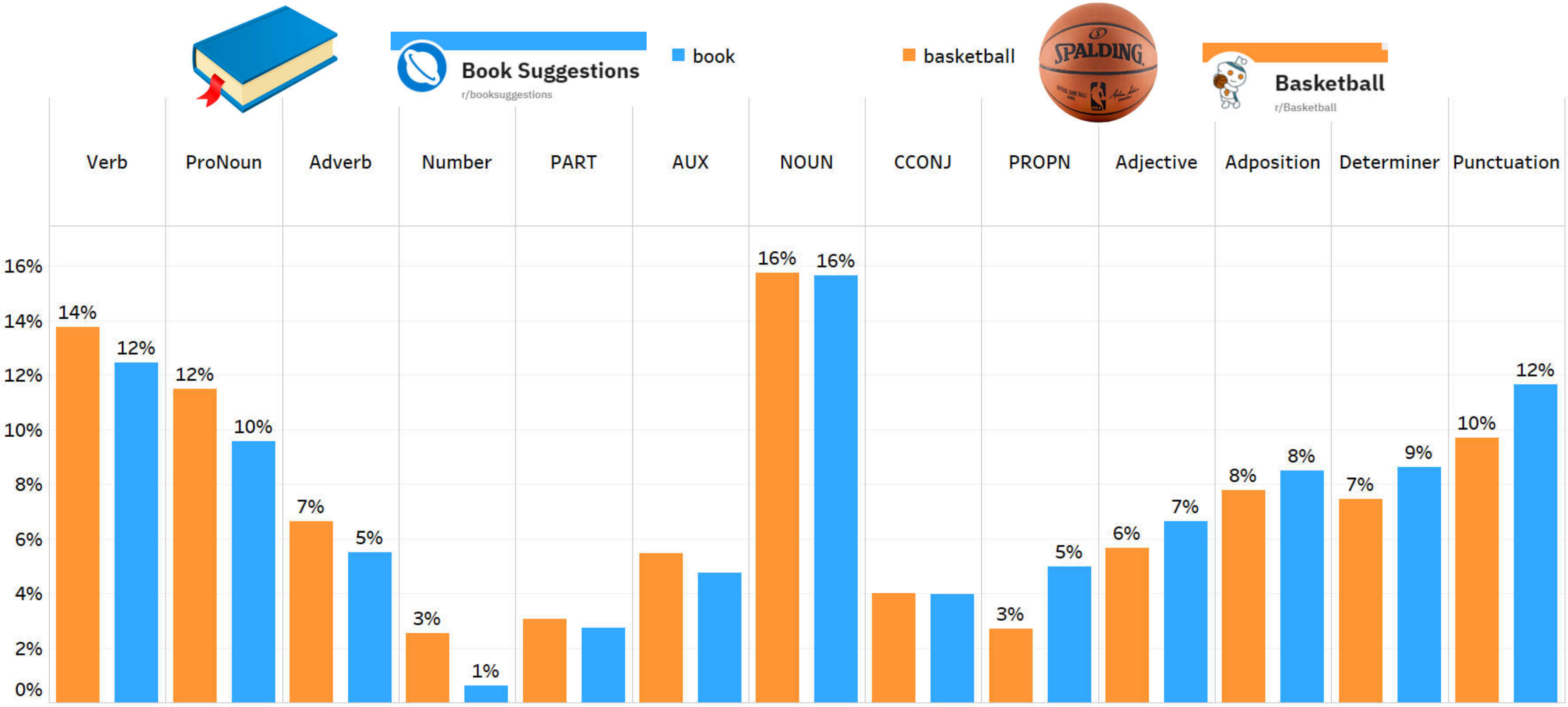


Basketball

r/Basketball



EDA: Proportion of part of speech in posts



Data preprocessing (text in the posts)

Step 1. **Remove** non-letters. (such number, #!/()[]&$~`+=)

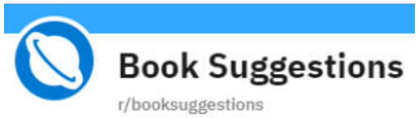
Step 2. **Convert to lower case**.

Step 3. **Remove stopwords**. (such as a, an, the, in, on, at, has, have)

Step 4. **Lemmatization**, grouping words. (such as “cats” and “cat”)

Classification Model

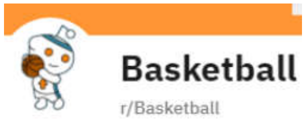
Class = 1



Book Suggestions

r/booksuggestions

Class = 0



Basketball

r/Basketball

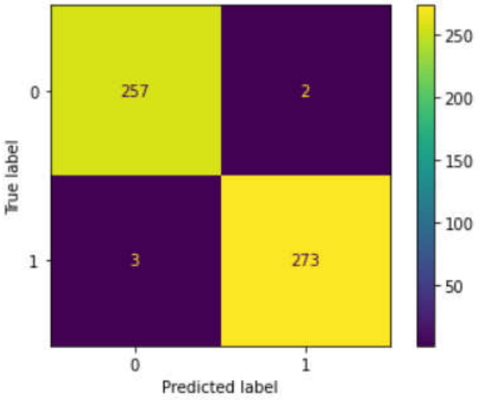
- 1. Bernoulli Naive Bayes Model
  - BernoulliNB()
- 2. Logistic Regression Model
  - LogisticRegression()
- 3. Voting Classifier Model
  - LogisticRegression()
  - BernoulliNB()
  - DecisionTreeClassifier()
  - AdaBoostClassifier()
  - GradientBoostingClassifier()

1. Bernoulli Naive Bayes Model

- GridSearch & Pipeline

```
{'cvec__max_df': 0.9,
 'cvec__max_features': 4000,
 'cvec__min_df': 2,
 'cvec__ngram_range': (1, 2),
 'nb__alpha': 0.2}
```

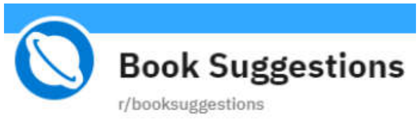
- **Train Score = 0.99919**
- **Cross Validation Score = 0.99679**
- **Test Score = 0.99065**



Accuracy	**=**	**99.06%**
Missclassification	**=**	**0.94%**
Precision	**=**	**99.23%**
Sensitivity	**=**	**99.85%**
Specificity	**=**	**99.27%**

- **Model Evaluation**
  - Performance of model is good in unseen data and testing data, misclassification is very low (less than 1%)

Main Key words



book
read
like
looking
would
suggestion
good
love
really
something

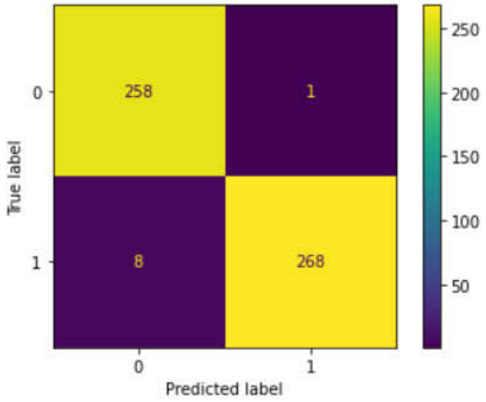
aau
able dunk
able jump
able play
access
adidas
advantage
aggressive
airball
anger

2. Logistic Regression Model

- GridSearch & Pipeline

```
{'cvec__max_df': 0.9,
 'cvec__max_features': 2000,
 'cvec__min_df': 2,
 'cvec__ngram_range': (1, 1),
 'lr__C': 2.5,
 'lr__penalty': 'l2'}
```

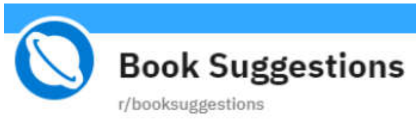
- **Train Score = 1.0**
- **Cross Validation Score = 0.98555**
- **Test Score = 0.98317**



Accuracy = 98.32%
Missclassification = 1.68%
Precision = 99.61%
Sensitivity = 96.99%
Specificity = 99.63%

- **Model Evaluation**
  - Performance of model is good but slightly over fit and misclassification is higher than NB

Main Key words



Book Suggestions
r/booksuggestions

Basketball
r/Basketball

book
novel
read
reading
something
suggestion
similar
looking
recommendation
life

basketball
play
team
game
ball
kg
nba
player
poll
shot

3. Voting Classifier Model

LogisticRegression()
BernoulliNB()
DecisionTreeClassifier()
AdaBoostClassifier()
GradientBoostingClassifier()

- GridSearch & Pipeline

```
{'ada__n_estimators': 120,
 'gb__n_estimators': 125,
 'lr__C': 2,
 'nb__alpha': 0.1,
 'tree__max_depth': None}
```

- **Train Score = 0.98636**
- **Test Score = 0.97943**

Model Evaluation

Performance of model is not good as **Naive Bayes** and slightly over fit

Conclusion

- The **Bernoulli Naive Bayes model** is best performance on training, unseen, testing data and % miss classification is very low when compare with Logistic Regression Model which is slightly over fit with the training data.

- Here are the list of **main keywords** which can use to identify the posts

book, read, like, looking, would, suggestion, good, love, really, something, novel, reading, similar, recommendation, life

basketball, play, team, game, ball, nba, player, poll, shot, aau, able dunk, able jump able play, access