

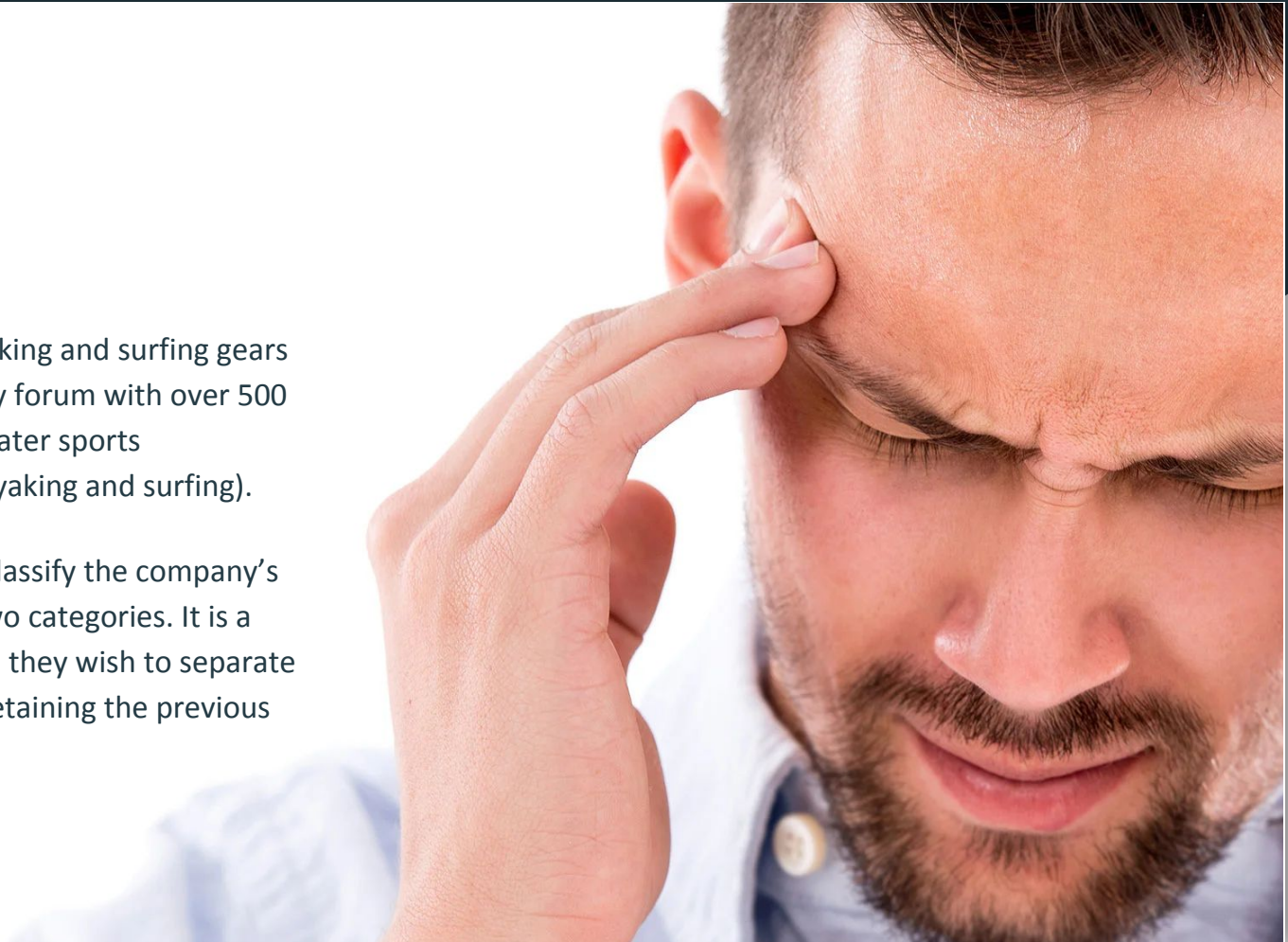
# Kayaking vs Surfing

By: Asyraf Taha

# Problem!

A company selling kayaking and surfing gears has a online community forum with over 500 posts regarding their water sports experiences (mainly kayaking and surfing).

I have been tasked to classify the company's forum posts into the two categories. It is a single channel post and they wish to separate posts by hobby while retaining the previous posts.



# Where to learn and model my data?

Snoo to the  
rescue!



# Why Reddit?


r/Kayaking

**About Community**

All things paddling related! Kayaks, canoes, even SUPs are welcome -- this is your place to post your paddling photos, ask your gear questions, share your experiences, or just be a part of the paddling community!

75.7k	106
Paddlers	Online

---

 Created Nov 7, 2009

r/Surfing

**About Community**

Kooks on the internet

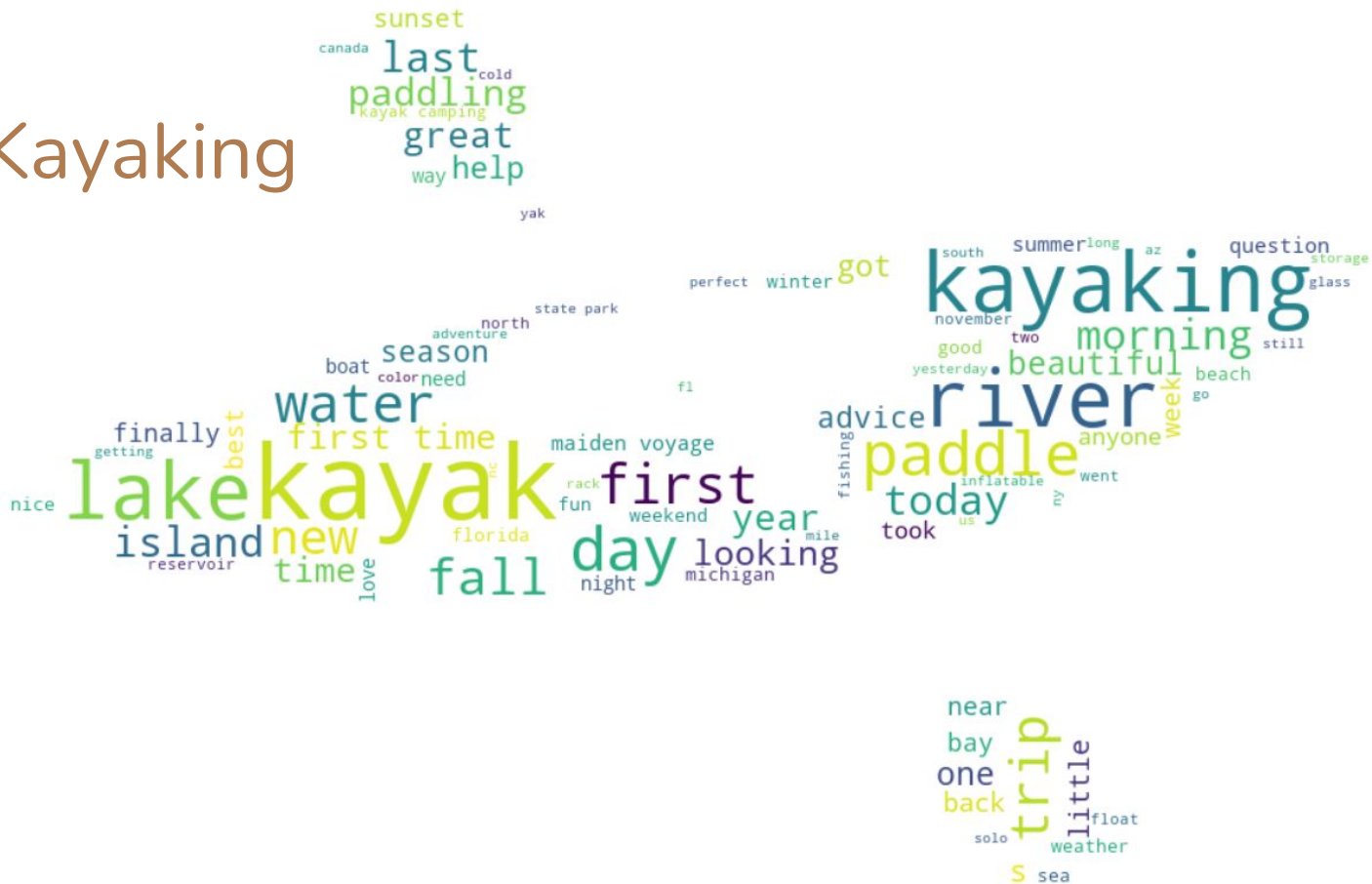
99.5k	400
Members	Online

---

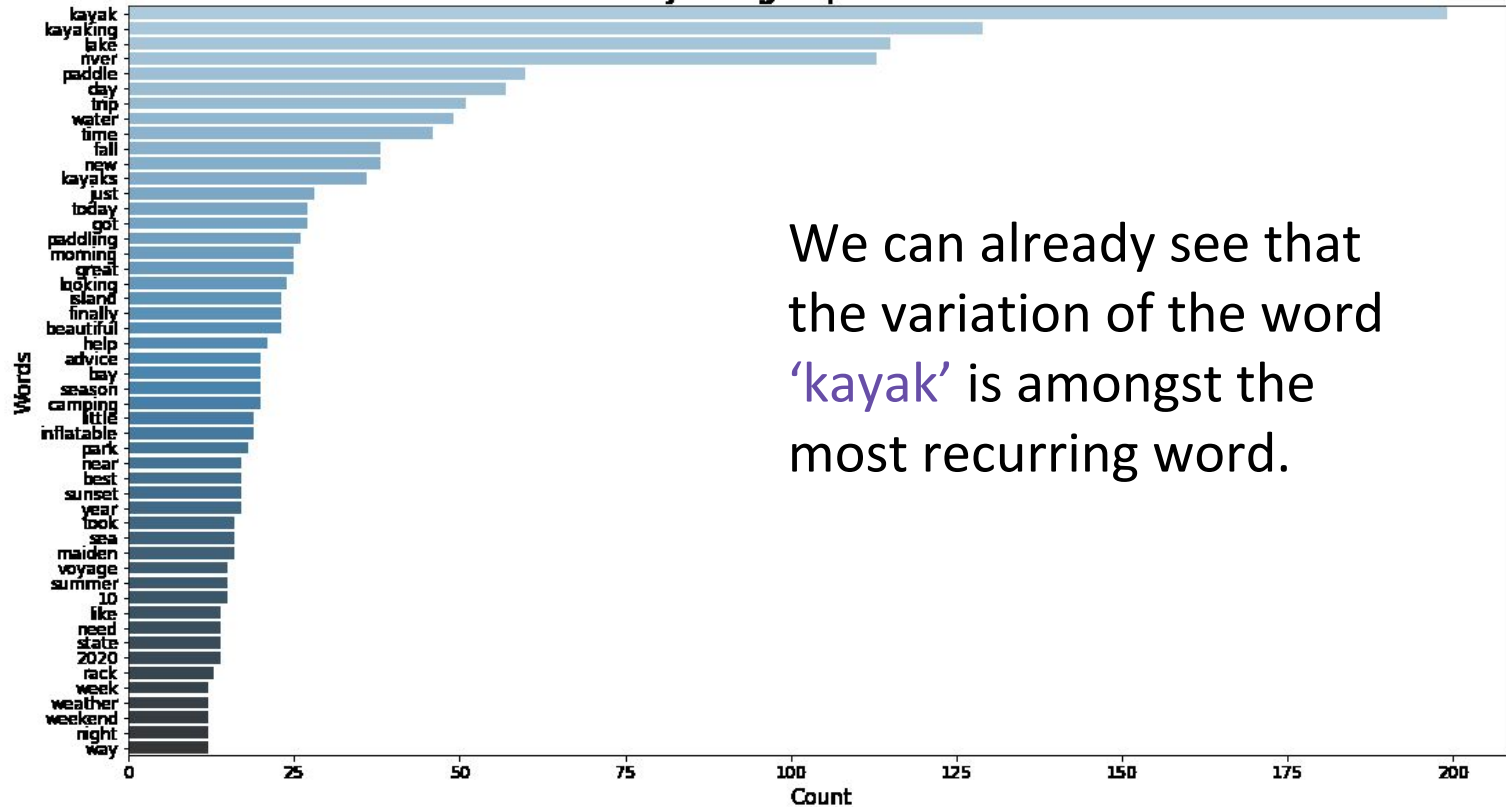
 Created Sep 11, 2008

Both have an extensive database of posts for us to learn!

# r/Kayaking



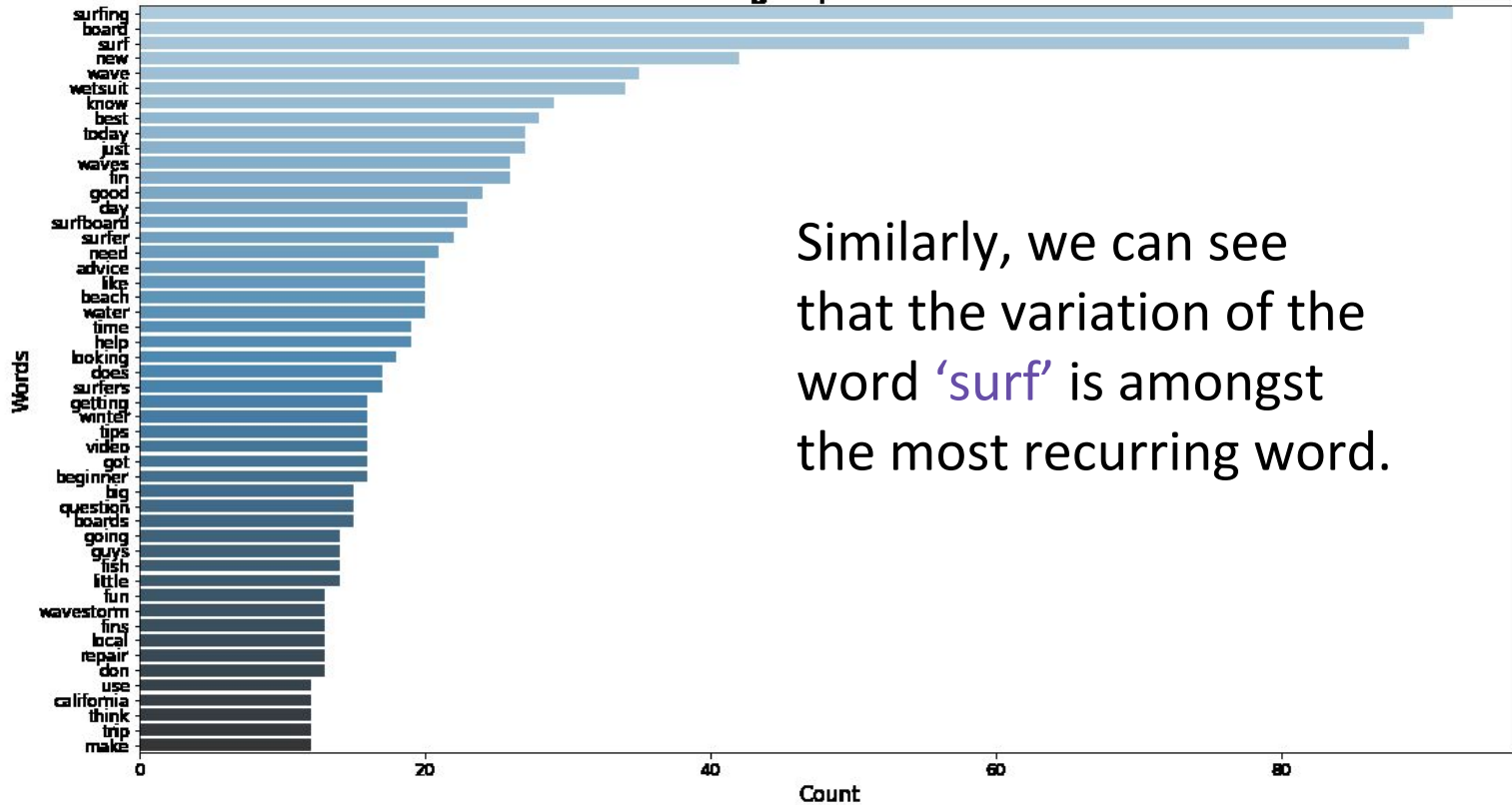
## Kayaking Top 50 Words



We can already see that the variation of the word 'kayak' is amongst the most recurring word.



## Surfing Top 50 Words



Similarly, we can see that the variation of the word 'surf' is amongst the most recurring word.



# Some preprocessing....

Looking at the top 50 words, we can see there are some overlaps in those words, such as 'water', 'day', etc.

- List them and add them to the list of stopwords.
- Remove variation of word 'kayak' and 'surf'
- Remove numbers and words with 2 or less letters.

.....Done! Let's Model!

# Baseline Model Accuracy

r/Kayaking

**51.3%**

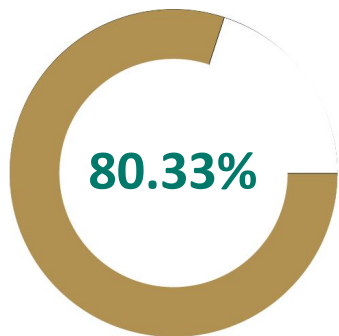
r/Surfing

**48.7%**

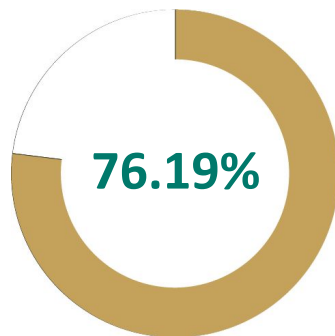
This means that when the model is unable to classify a post, there is a slightly higher probability it will predict it as a Kayaking-related post.

# CountVectorizer + LogisticRegression

Train Score



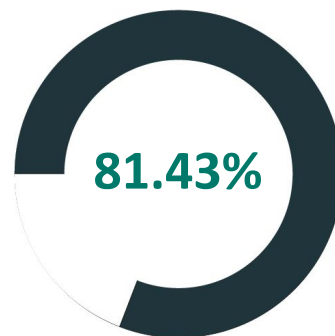
Accuracy Score



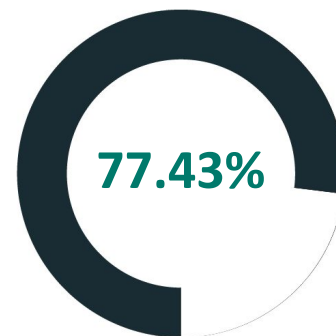
	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	120	115
Actual r/Kayaking	0	248

# TfidfVectorizer + LogisticRegression

Train Score



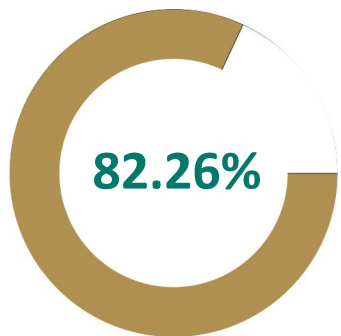
Accuracy Score



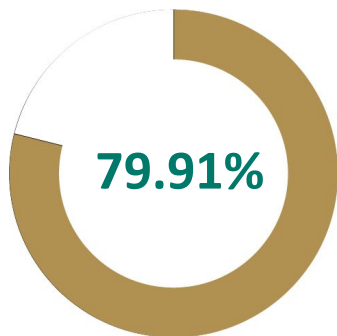
	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	126	109
Actual r/Kayaking	0	248

## CountVectorizer + MultinomialNB

Train Score



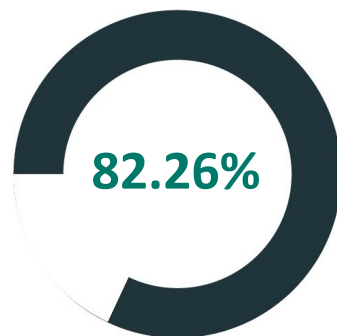
Accuracy Score



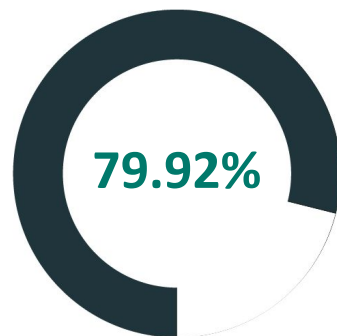
	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	138	97
Actual r/Kayaking	0	248

## TfidfVectorizer + MultinomialNB

Train Score



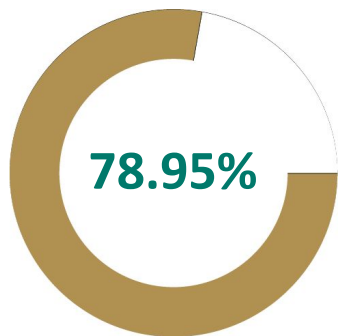
Accuracy Score



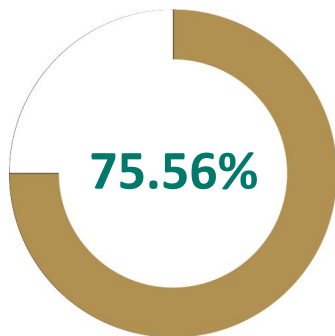
	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	138	97
Actual r/Kayaking	0	248

## CountVectorizer + RandomForest

Train Score



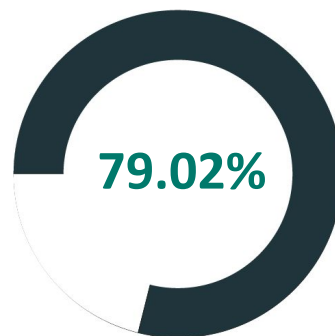
Accuracy Score



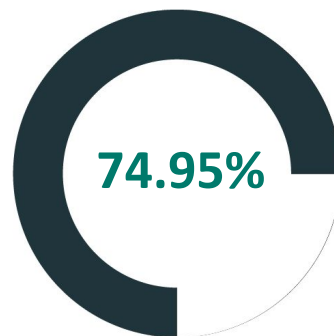
	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	117	118
Actual r/Kayaking	0	248

## TfidfVectorizer + RandomForest

Train Score



Accuracy Score



	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	114	121
Actual r/Kayaking	0	248



TfidfVectorizer + MultinomialNB  
Performed the best!

# Misclassification Analysis

	Predicted r/Surfing	Predicted r/Kayaking
Actual r/Surfing	138	97
Actual r/Kayaking	0	248

- High Type 1 Error (False Postive)
- Very Low Type 2 Error (False Negative)

# Type 1 Error

- 2 or less word(s) titles - in orange
- Generic titles (non related to hobby - in blue
- Location specific posts - in red
- Name specific posts - in green
- Words that have failed to be classified together with their root word (eg. 'waves' and 'wave') - in yellow
- Animal specific post - in black
- Other surfing technique related posts (carving) - not boxed

	title	actual	predicted
1759	Ireland going off ' nut yesterday 🙋	0	1
1788	Good waves and bad technique from Morocco	0	1
1380	Ribs	0	1
1483	Beginner spots the bay area	0	1
1482	People who film themselves what you use and what mount holders you with? Looking mainly...	0	1
1129	Fun waves today 🙋	0	1
1630	STOP	0	1
1182	Pelican tries drop , gets worked	0	1
1062	This one felt good	0	1
1103	Felt great, looked good for once.	0	1
1565	Yesterday hossegor, from the webcam, lockdown back	0	1
1500	Indo Trip	0	1
1809	Lot variety Ireland today	0	1
1215	Ooh yay...	0	1
1665	Falling while doing carve?	0	1
1180	How are the conditions Brighton right now? worth hour trip this weekend?	0	1
1761	Conor Maguire Mullaghmore yesterday. Massive.	0	1



Pelican Drop



# Conclusion

## Findings


- **TfidfVectorizer + MultinomialNB Model** performed the best amongst the other models in terms of reducing overfitting and accuracy.
- The best model is able to predict at a **79.92% accuracy**.
- **Prevalent Type 1 Error** (Wrongly Predicting Kayaking Post) throughout all 4 models.
- Generally **TfidfVectorizer tends to improve the model** by increasing score or reducing overfitting (except for RandomForest)

## Limitations

- Many of the errors, were due to posts having root words that have been identified with strong coefficients. The machine was unable to distinguish the words.
- At 79.92% accuracy, the model may not be able to effectively predict (with low error) as the number of posts increases.
- Machine is also unable to classify or predict usage of special characters or emojis that may be of a prevalent use in the future.



# Recommendations

- The accuracy suggests that you would require additional manpower to manually filter misclassified posts for your forum if the total number of posts is large.
  - Train machine to be able to identify variations of a root word via lemmatization.
  - Continue to learn and identify other water sports that may be in the company's forum post.
- 



Thank You