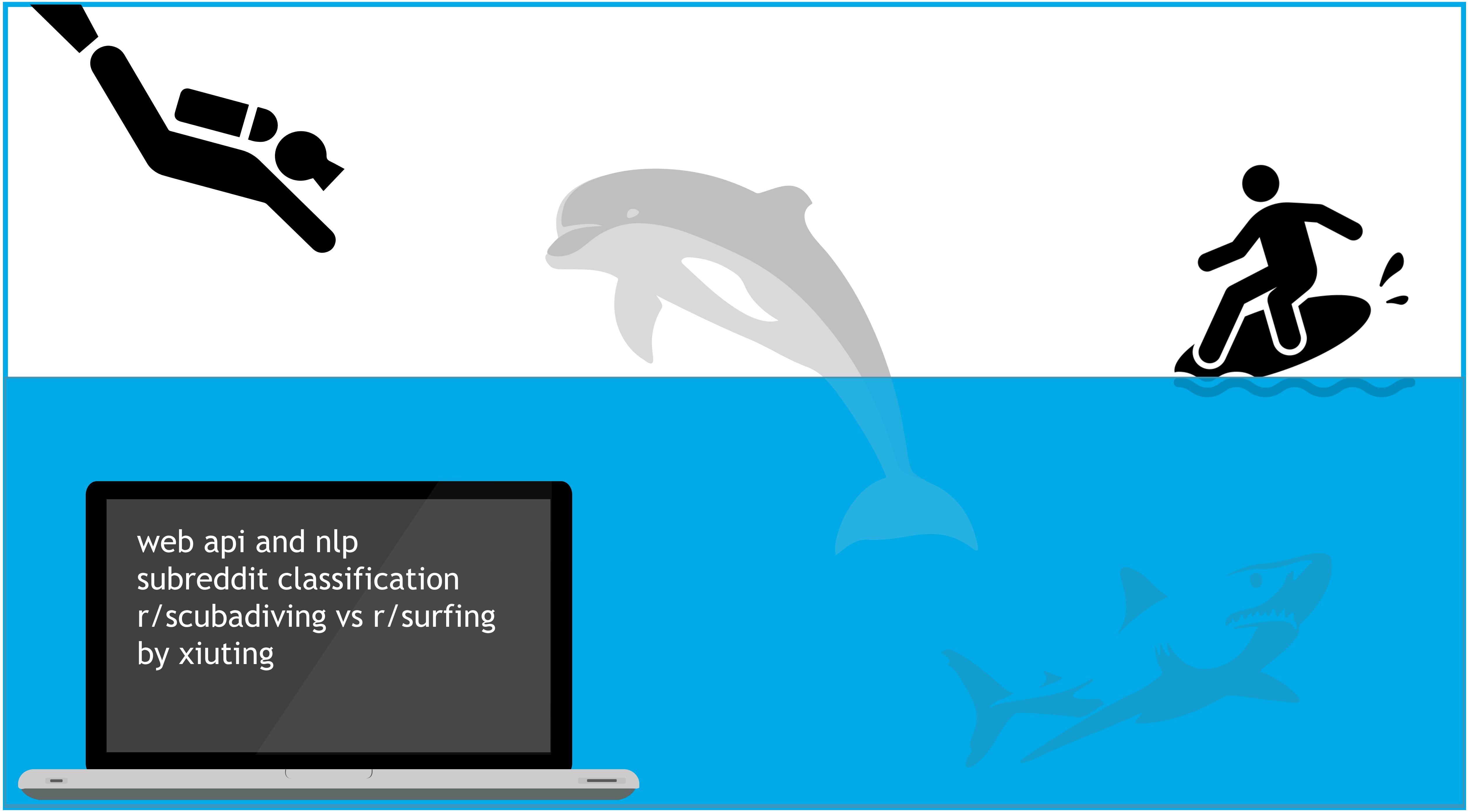
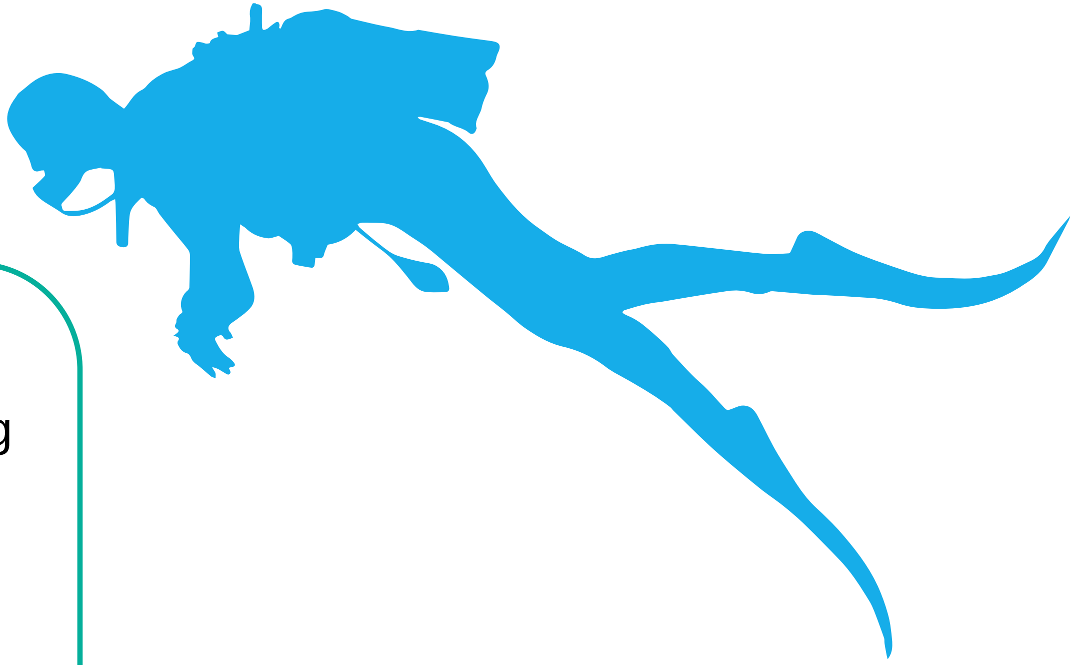


web api and nlp
subreddit classification
r/scubadiving vs r/surfing

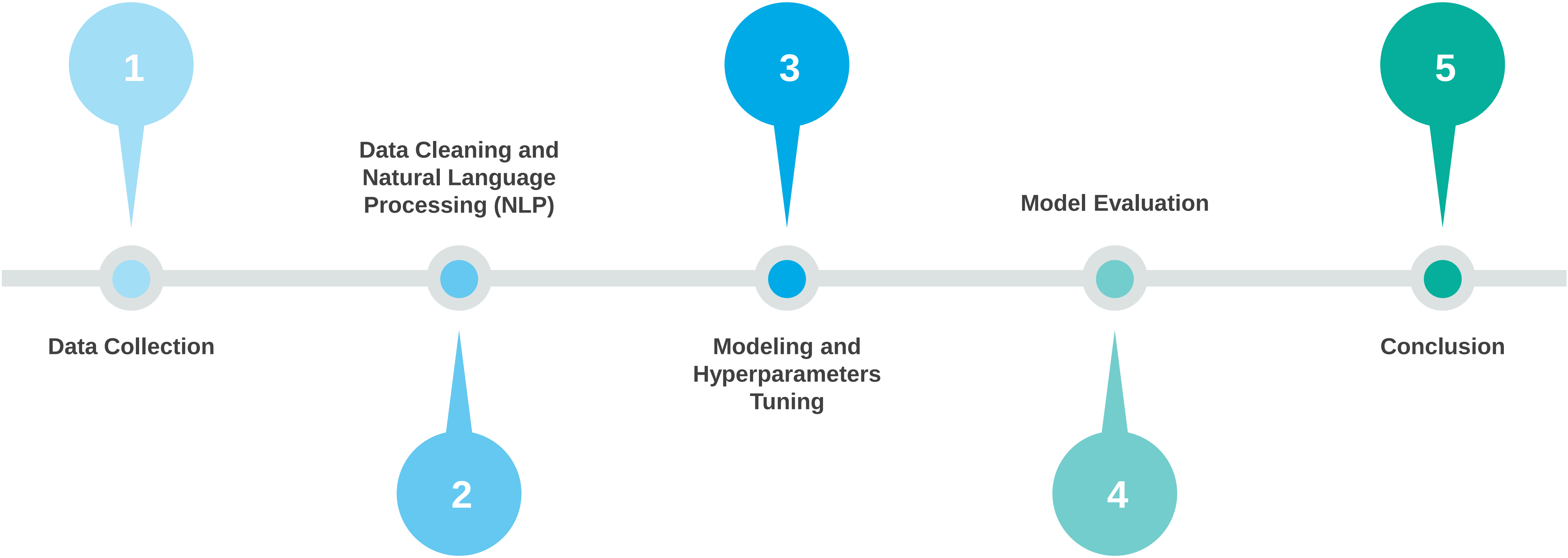
by xiuting

Problem Statement

With more intense competition from many sporting goods companies which are moving to e-commerce retail amidst the pandemic, an e-commerce water sports apparel and equipment company, which also operates a discussion forum on their website, is looking to improve their sales revenue by staying ahead of the e-commerce competition.



Approach



1

Data Collection

2

Data Cleaning and Natural Language Processing (NLP)

3

Modeling and Hyperparameters Tuning

4

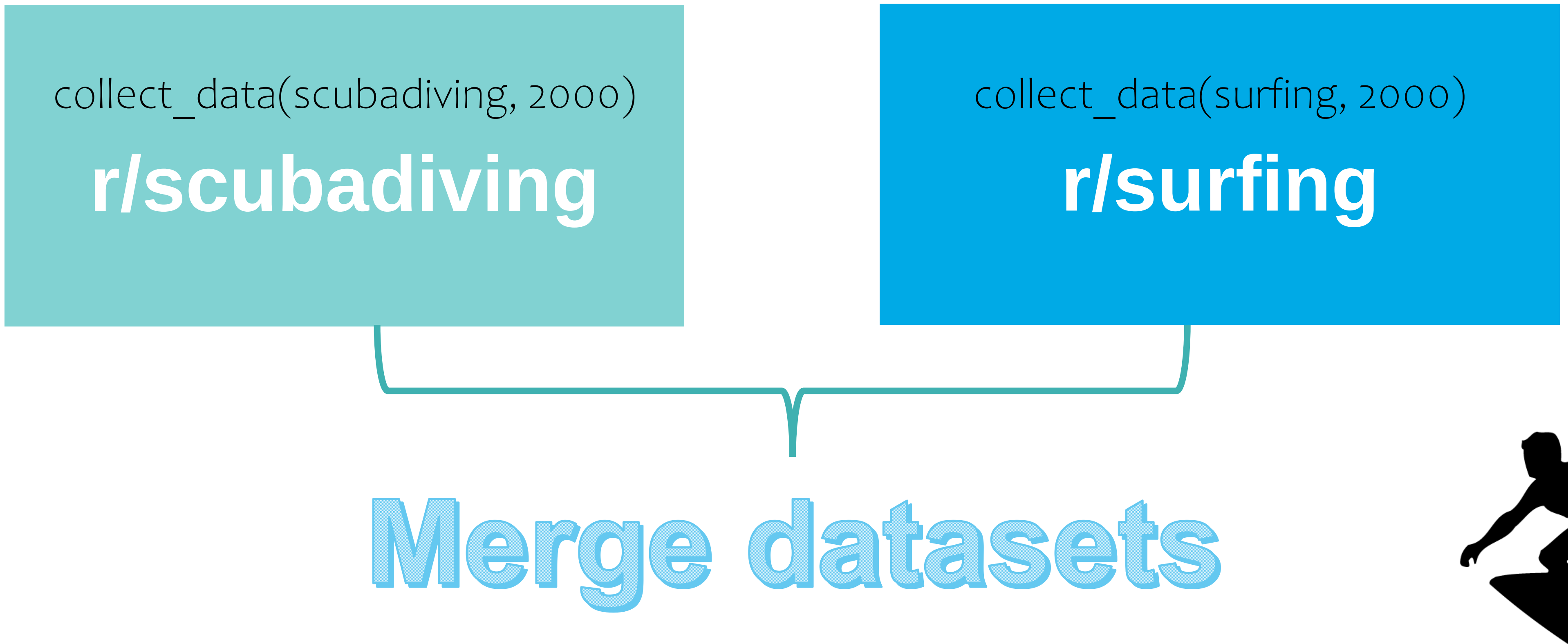
Model Evaluation

5

Conclusion

Data Collection

Pushshift API



collect_data(scubadiving, 2000)

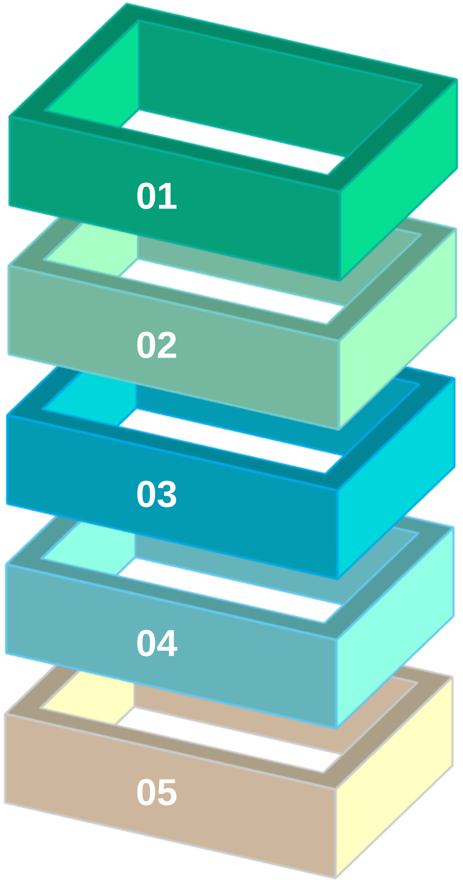
r/scubadiving

collect_data(surfing, 2000)

r/surfing

Merge datasets

Data Cleaning



MISSING VALUES

Missing values in ‘selftext’ and ‘selftext’ with values ‘[removed]’, ‘[deleted]’ are replaced empty strings ‘ ‘

BINARY ENCODING

Create new column called ‘diving’ and map 1 and 0 based on ‘subreddit’ values. 1: scubadiving, 0: surfing

COMBINE/DROP COLUMNS

‘title’ and ‘selftext’ are combined in a new column ‘text’. Columns ‘title’, ‘selftext’ and ‘subreddit’ are dropped

DUPLICATED ROWS

Checked and dropped duplicated rows.

TEXT CLEANING

String in ‘text’ is converted to lowercase, html and special characters, punctuations, etc. are removed.

Natural Language Processing

01 STEMMING VS LEMMATIZATION

TEXT BEFORE STEMMING / LEMMATIZATION

“**going diving** tomorrow for an open water cert and doc has told me i have **swimmers** ear today going diving tomorrow for an open water cert and doc has told me i have **swimmers** ear today i have no pain or **irritation** a few days ago i went swimming and **equalised** with no pain should i go or skip out and **leave** it for another day doc has **prescribed** me **antibiotics** if i did dive can i take these in conjunction with the dive”

TEXT AFTER STEMMING

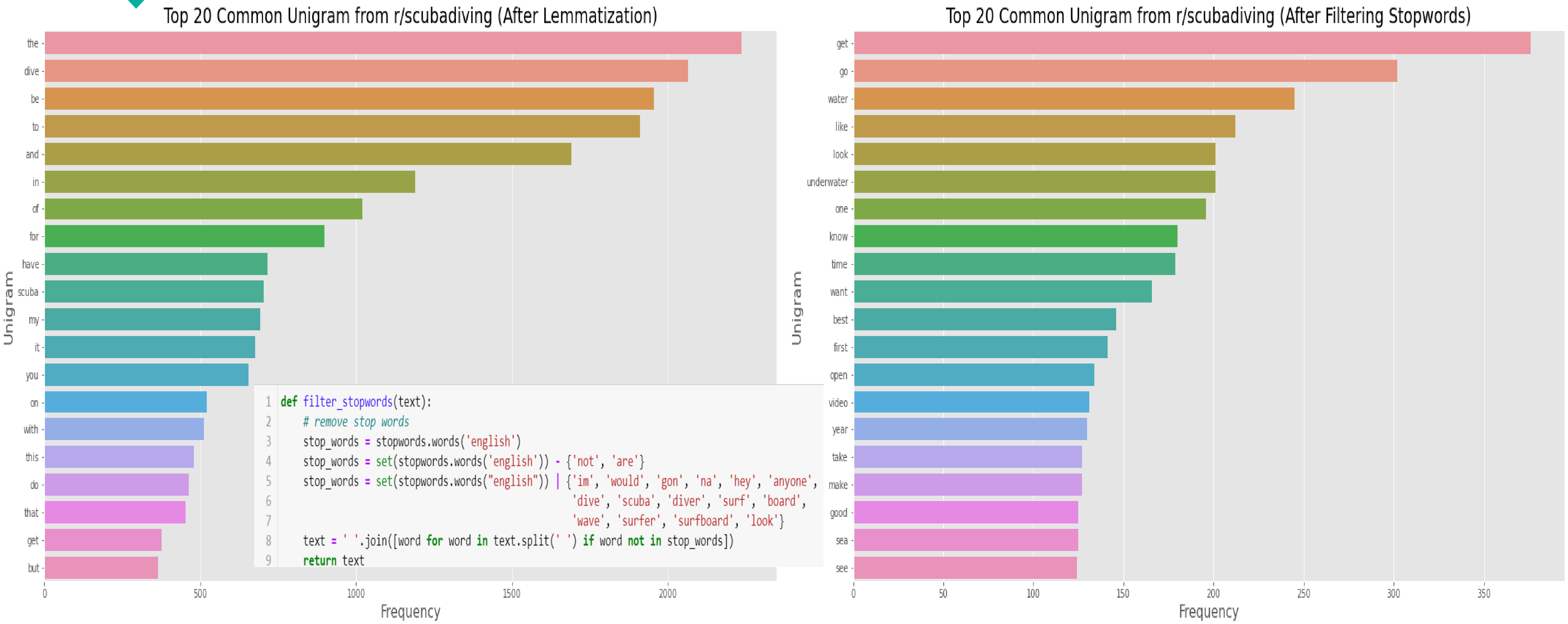
go dive tomorrow for an open water cert and doc has told me i have **swimmer** ear today go dive tomorrow for an open water cert and doc has told me i have **swimmer** ear today i have no pain or **irrit** a few day ago i went swim and **equalis** with no pain should i go or skip out and **leav** it for anoth day doc has **prescrib** me **antibiot** if i did dive can i take these in conjunct with the dive

TEXT AFTER LEMMATIZATION

go dive tomorrow for an open water cert and doc have told me i have **swimmer** ear today go dive tomorrow for an open water cert and doc have told me i have **swimmer** ear today i have no pain or **irritation** a few day ago i go swim and **equalise** with no pain should i go or skip out and **leave** it for another day doc have **prescribed** me **antibiotic** if i do dive can i take these in conjunction with the dive

Natural Language Processing

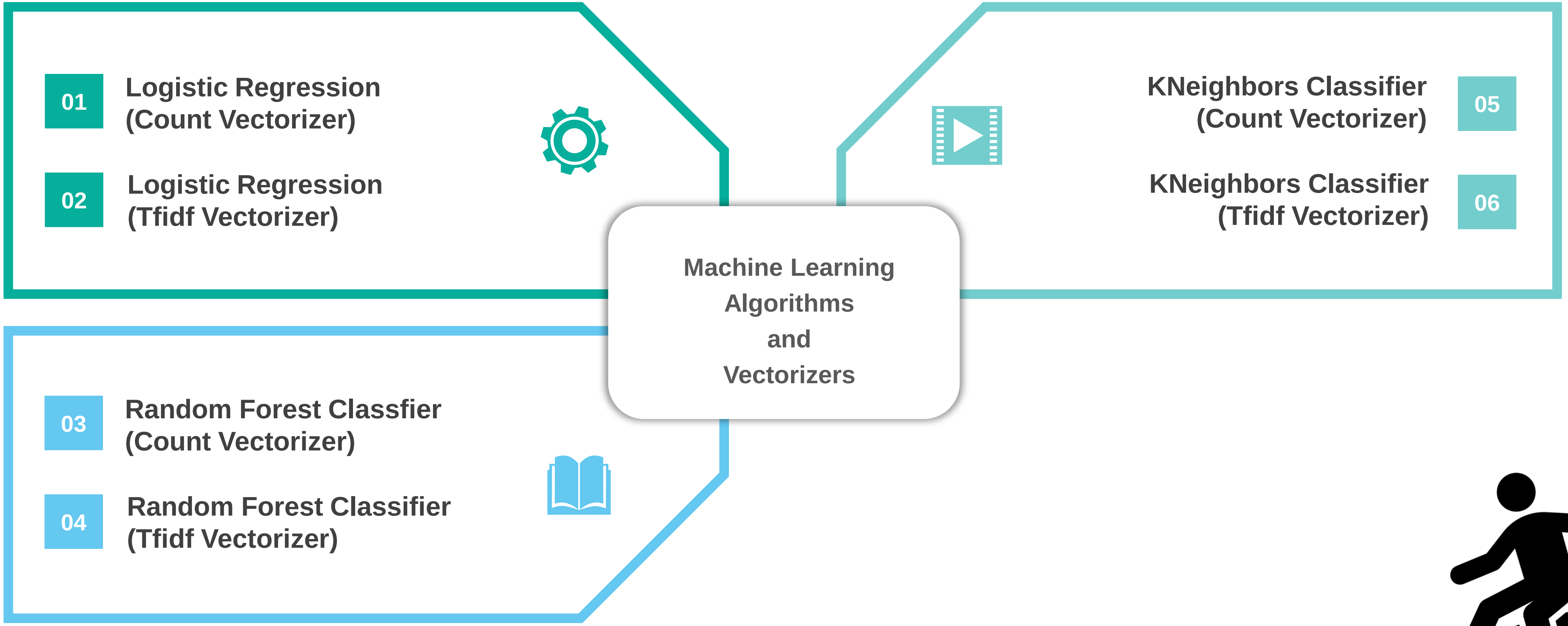
02 FILTER STOPWORDS


Top 20 Common Unigram from r/scubadiving (After Lemmatization)

Top 20 Common Unigram from r/scubadiving (After Filtering Stopwords)

```python
def filter_stopwords(text):
    # remove stop words
    stop_words = stopwords.words('english')
    stop_words = set(stopwords.words('english')) - {'not', 'are'}
    stop_words = set(stopwords.words("english")) | {'im', 'would', 'gon', 'na', 'hey', 'anyone',
                                                    'dive', 'scuba', 'diver', 'surf', 'board',
                                                    'wave', 'surfer', 'surfboard', 'look'}
    text = ' '.join([word for word in text.split(' ') if word not in stop_words])
    return text
```

Modeling



Machine Learning Algorithms and Vectorizers

01 Logistic Regression (Count Vectorizer)

02 Logistic Regression (Tfidf Vectorizer)

03 Random Forest Classfier (Count Vectorizer)

04 Random Forest Classifier (Tfidf Vectorizer)

05 KNeighbors Classifier (Count Vectorizer)

06 KNeighbors Classifier (Tfidf Vectorizer)

Hyperparameter Tuning

01

LOGISTIC REGRESSION

'lr__C': np.linspace(0.7, 2, 10),
'lr__solver':['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
'lr__penalty' : ['l1', 'l2', 'elasticnet', None]

02

RANDOM FOREST CLASSIFIER

'rf__n_estimators' : [800, 1000, 1200]
'rf__max_depth’: [3, 7, 9]}

03

KNEIGHBORS CLASSIFIER

'knn__p': [1, 2],
'knn__weights' : ['uniform', 'distance']
'knn__metric': ['minkowski', 'euclidean']
'knn__n_neighbors': np.arange(1, 7, 2)

04

Count Vectorizer

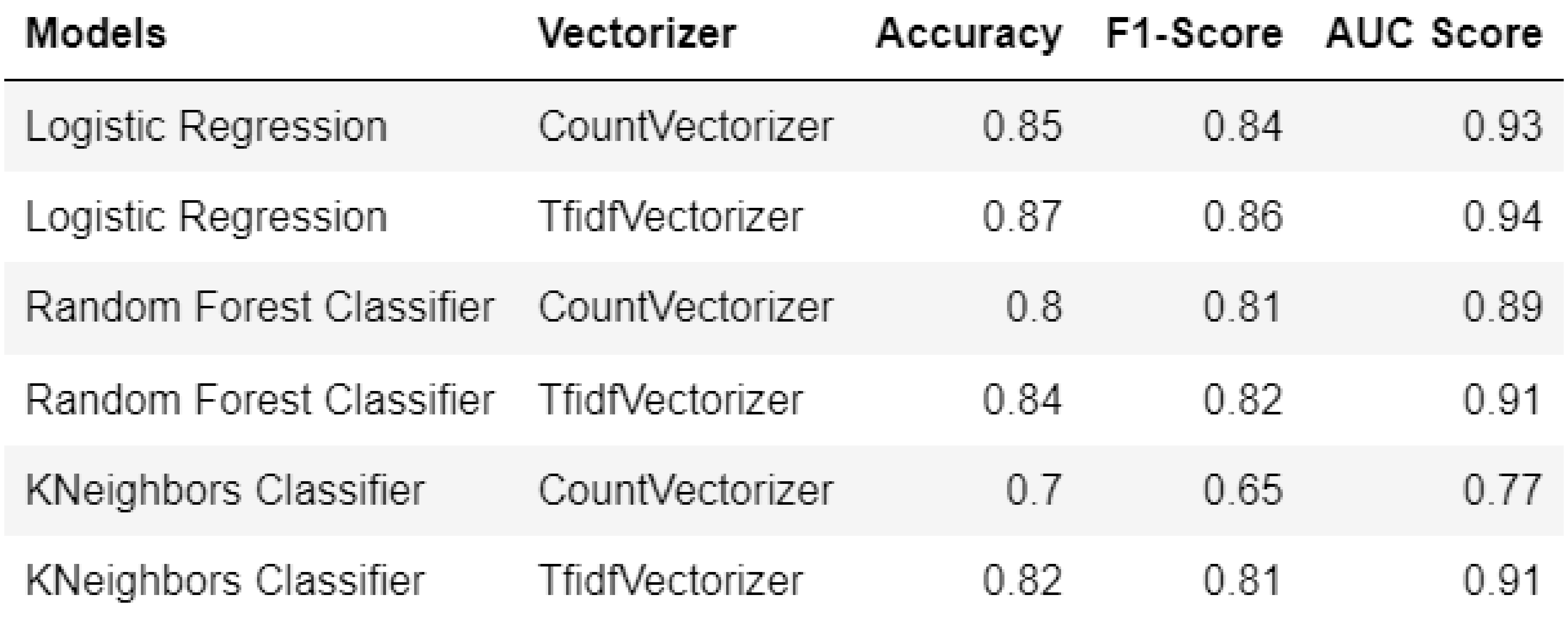
'cvec__ngram_range' : [(1, 1), (1, 2)],
'cvec__max_features' : [6500, 6700, 6900],
'cvec__max_df' : [.9, .95]’
'cvec__min_df' : [0.5, 1]'

05

Tfidf Vectorizer

'tfidf__max_features': [6000, 6500, 7000],
'tfidf__ngram_range': [(1,1), (1,2), (2,2)]Solver: {‘newton-cg’, ‘lbfgs’, ‘liblinear’,
'cvec__max_df' : [.9, .95]’
'cvec__min_df' : [0.5, 1]'

Model Evaluation



Models	Vectorizer	Accuracy	F1-Score	AUC Score
Logistic Regression	CountVectorizer	0.85	0.84	0.93
Logistic Regression	TfidfVectorizer	0.87	0.86	0.94
Random Forest Classifier	CountVectorizer	0.8	0.81	0.89
Random Forest Classifier	TfidfVectorizer	0.84	0.82	0.91
KNeighbors Classifier	CountVectorizer	0.7	0.65	0.77
KNeighbors Classifier	TfidfVectorizer	0.82	0.81	0.91

01

Best Model: Logistic Regression (Tfidf Vectorizer)

You can simply impress your audience and add a unique zing and appeal to your Presentations.

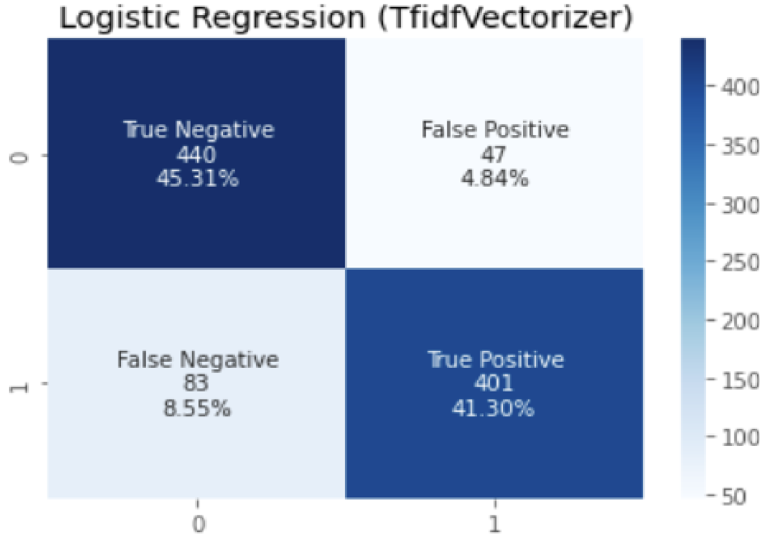
02

Worst Model: KNeighbors Classifier (Count Vectorizer)

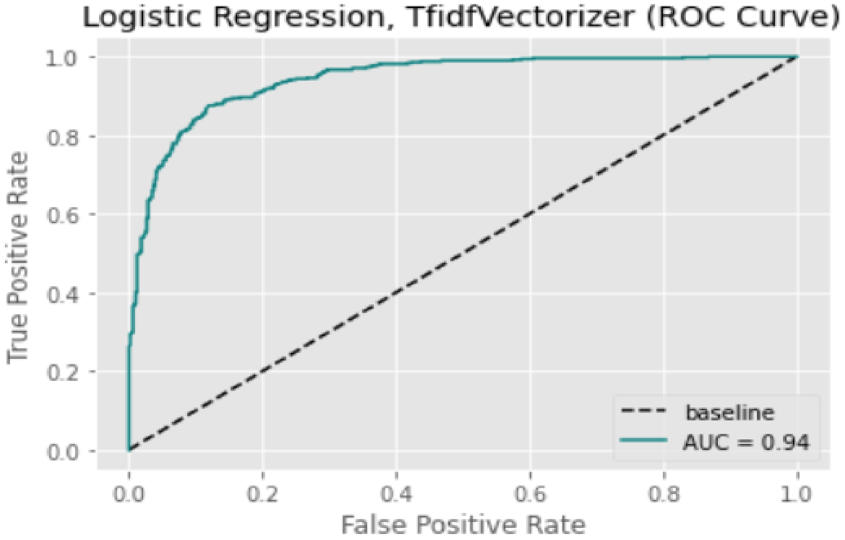
You can simply impress your audience and add a unique zing and appeal to your Presentations.

Conclusion

Logistic Regression with Tfidf Vectorizer will be recommended to the water sporting apparel and equipment company to help them identify their potential customers' interest; and provide them with highly personalized product recommendations.

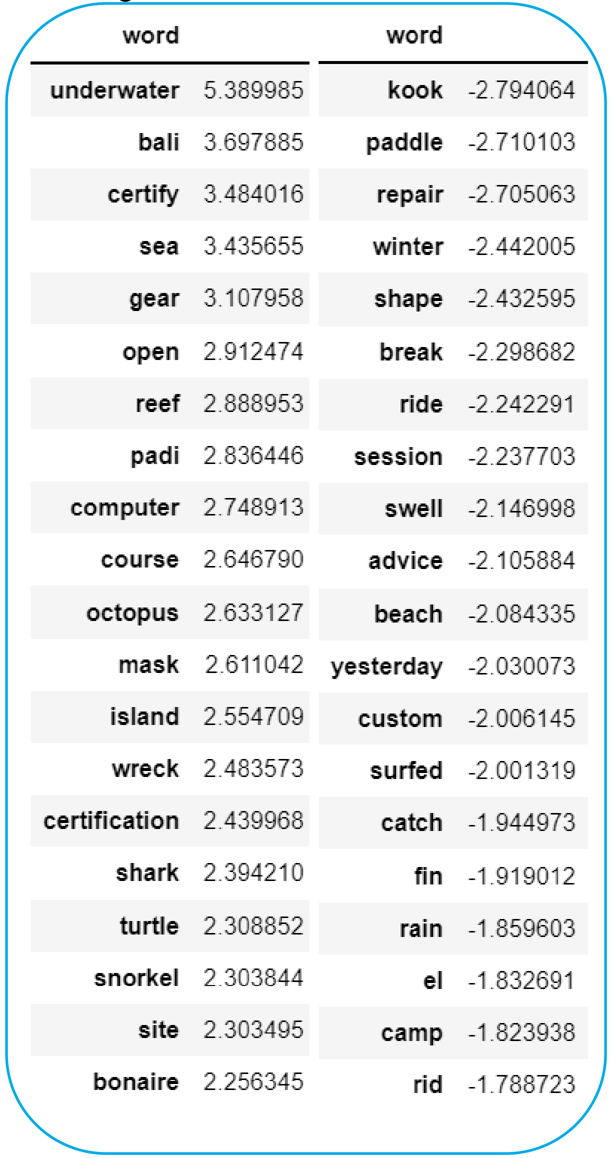


Logistic Regression (TfidfVectorizer)



Logistic Regression, TfidfVectorizer (ROC Curve)

word importance



word		word	
underwater	5.389985	kook	-2.794064
bali	3.697885	paddle	-2.710103
certify	3.484016	repair	-2.705063
sea	3.435655	winter	-2.442005
gear	3.107958	shape	-2.432595
open	2.912474	break	-2.298682
reef	2.888953	ride	-2.242291
padi	2.836446	session	-2.237703
computer	2.748913	swell	-2.146998
course	2.646790	advice	-2.105884
octopus	2.633127	beach	-2.084335
mask	2.611042	yesterday	-2.030073
island	2.554709	custom	-2.006145
wreck	2.483573	surfed	-2.001319
certification	2.439968	catch	-1.944973
shark	2.394210	fin	-1.919012
turtle	2.308852	rain	-1.859603
snorkel	2.303844	el	-1.832691
site	2.303495	camp	-1.823938
bonaire	2.256345	rid	-1.788723