SUPER HEROES
MEDIA

Digitalizing our Marketing and Branding Infringement Checks

SUPER HEROES MEDIA

Who Are We?

- Super Heroes Media Incorporated
- Provide marketing and branding services
- Two largest accounts: Disney & Warner Bros

Our Two Biggest Portfolios

SUPER HEROES MEDIA

Problem Statement

Infringement of material between our portfolios

SUPER HEROES MEDIA

Infringement Checks

- These portfolios are very similar yet distinct
- Easy for marketing/branding team to mix up information
- Mix up could result in loss of business, damage to our reputation/credibility or lawsuit

Infringement Checks

Infringement Checks

Infringement Checks

Infringement Checks

SUPER HEROES MEDIA

Why Digitalize our Infringement Checks?

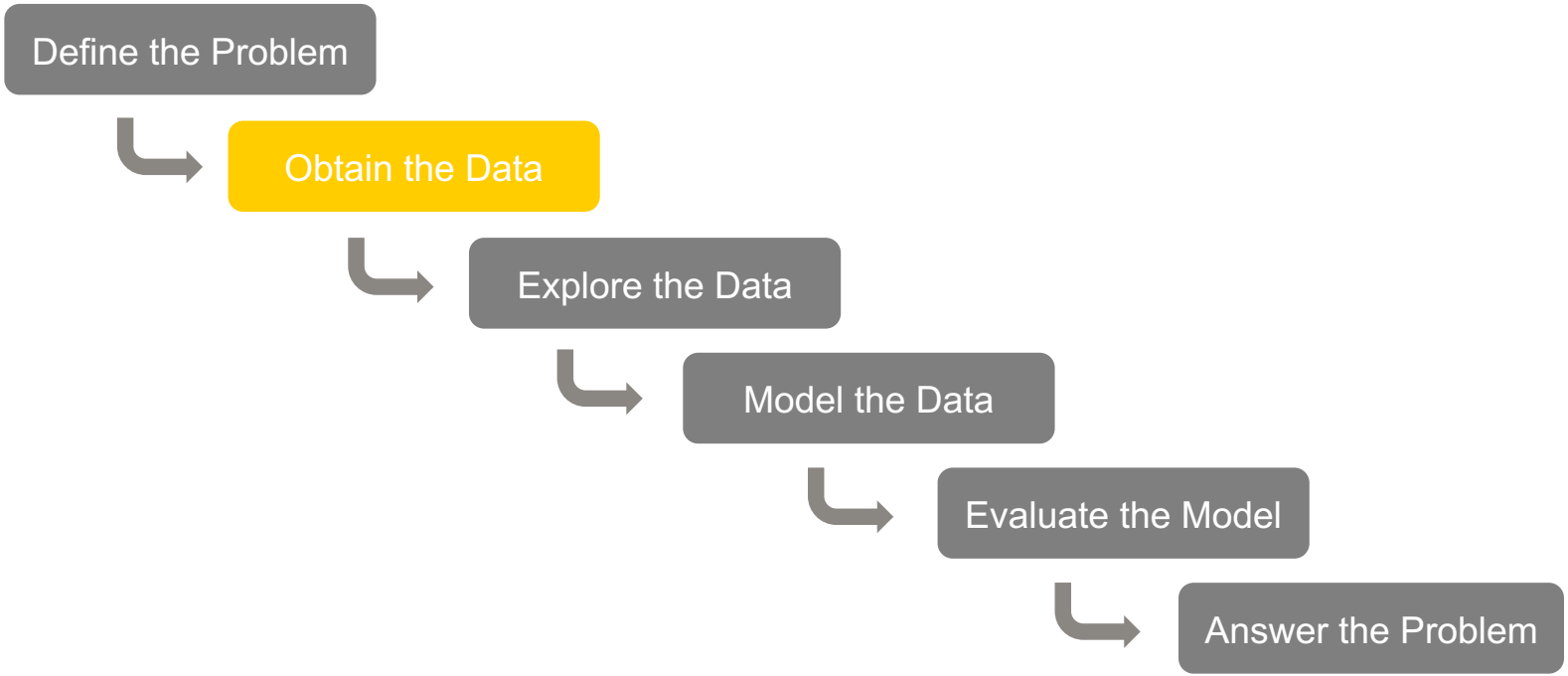
- More work is coming our way
  - MCU Phase 4 and 5
  - DCEU reboot
- Leverage on technology to manage our increased workload
  - Automate checks
  - Reduce mistakes
  - Increase efficiency

SUPER HEROES MEDIA

First Step in Our Digitalization Journey

- Trial implementation of a Natural Language Processing (NLP) Model to process text from
  - scripts,
  - concept write ups,
  - media publications and releases,
  - promotions,
  - etc

Data Science Process



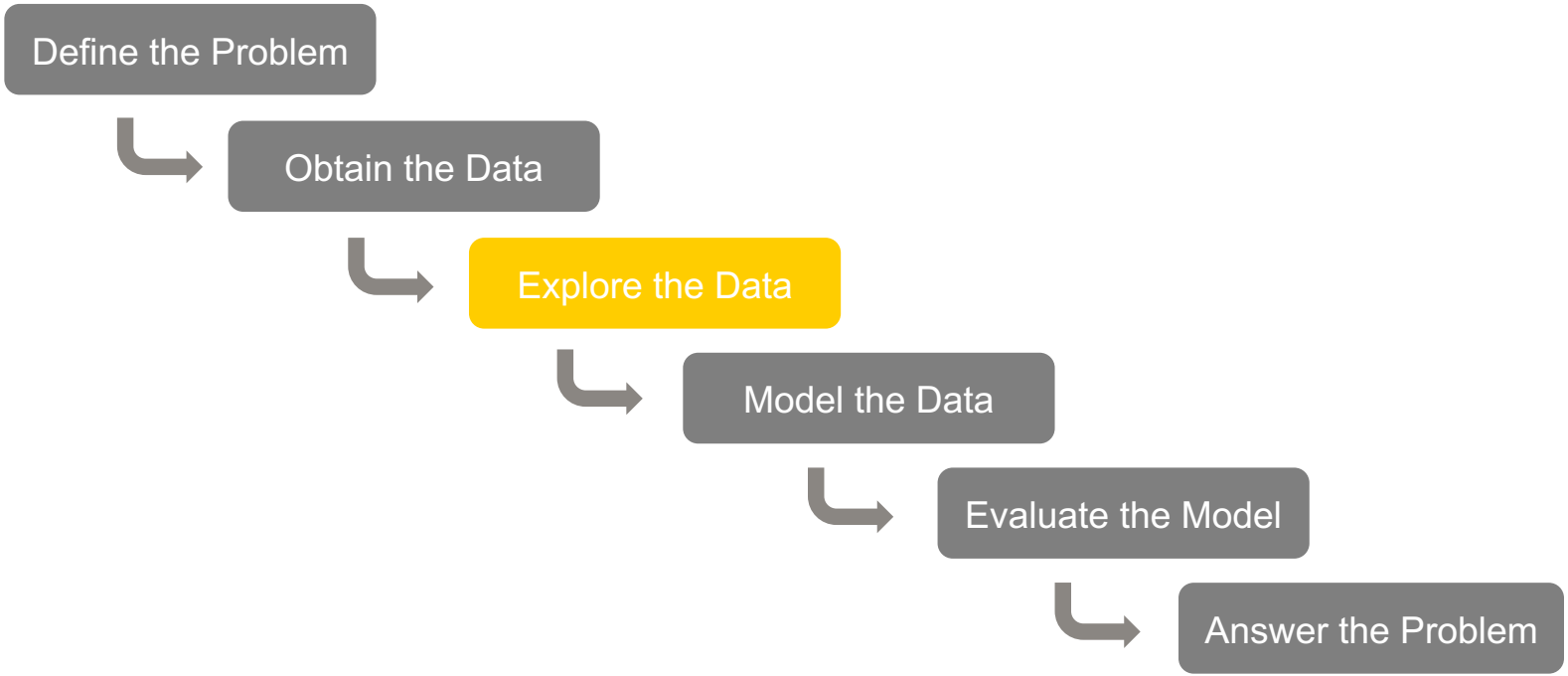
SUPER HEROES MEDIA

Scrapping from Reddit

- Subreddits
  - r/marvelstudios
  - r/DC_Cinematics
- Pushshift API
  - Post titles
  - Post text

SUPER HEROES MEDIA

Data Science Process

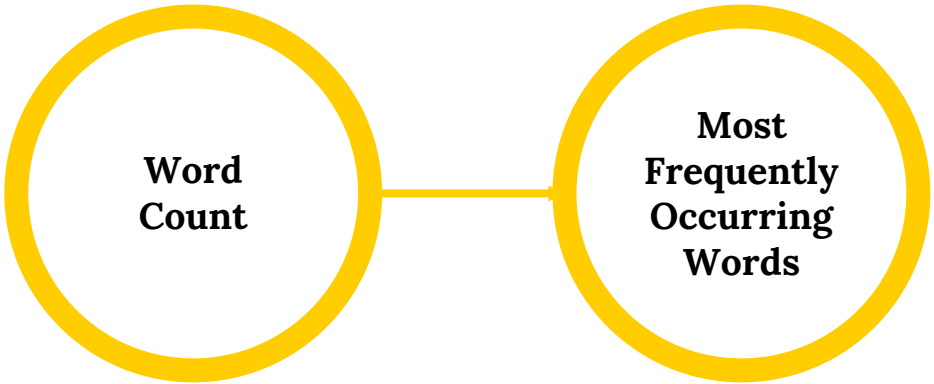


SUPER HEROES MEDIA

Data Cleaning & Processing

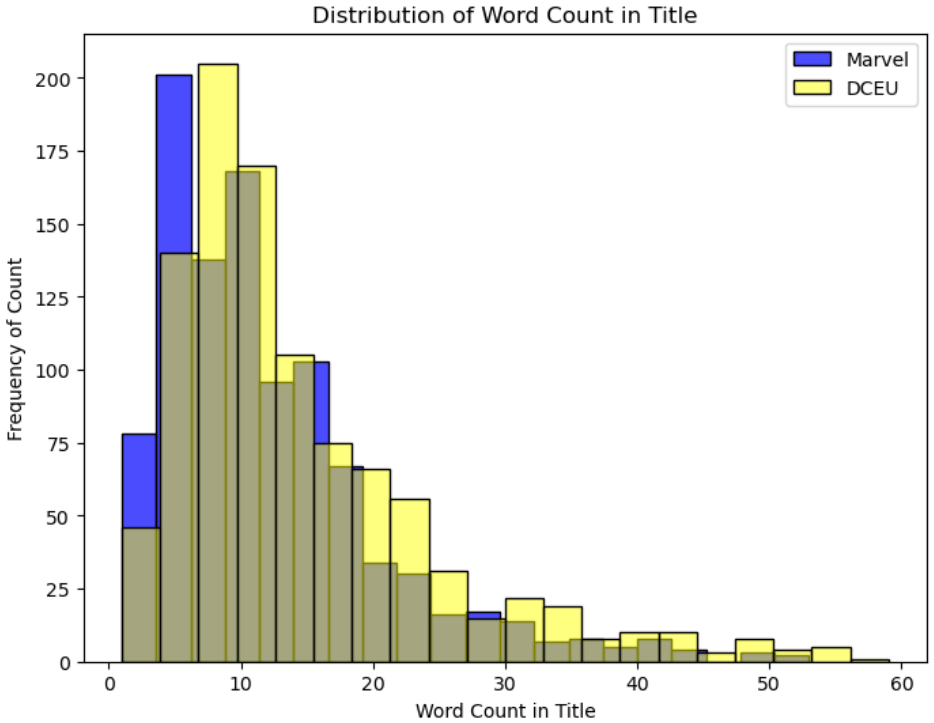
- Repeated posts
- Remove apostrophes, URLs, numbers, etc
- Convert sentences into individual texts
- Covert different words with the same meaning/intent into the same word
- Filler words that provide no context
- Words common to both data sets as well as giveaway words

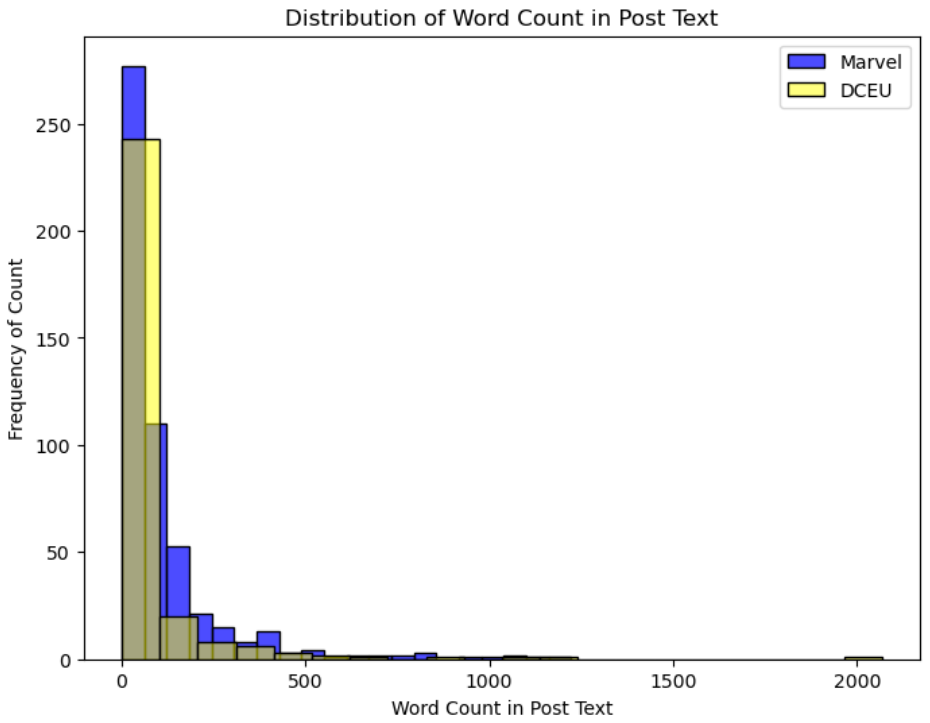
Data Exploration



SUPER HEROES MEDIA

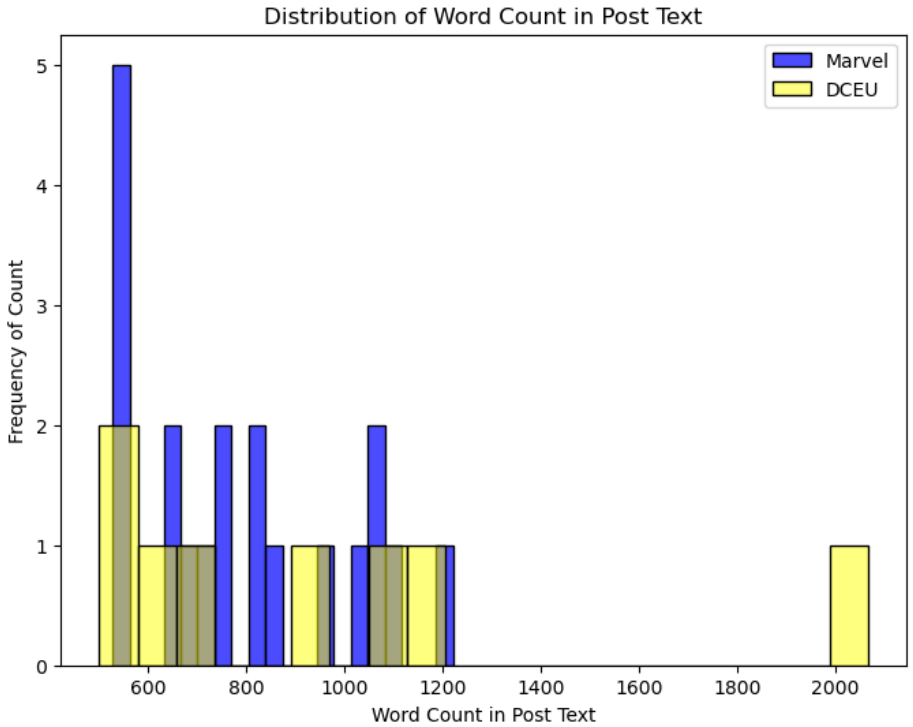
Title and Post Length





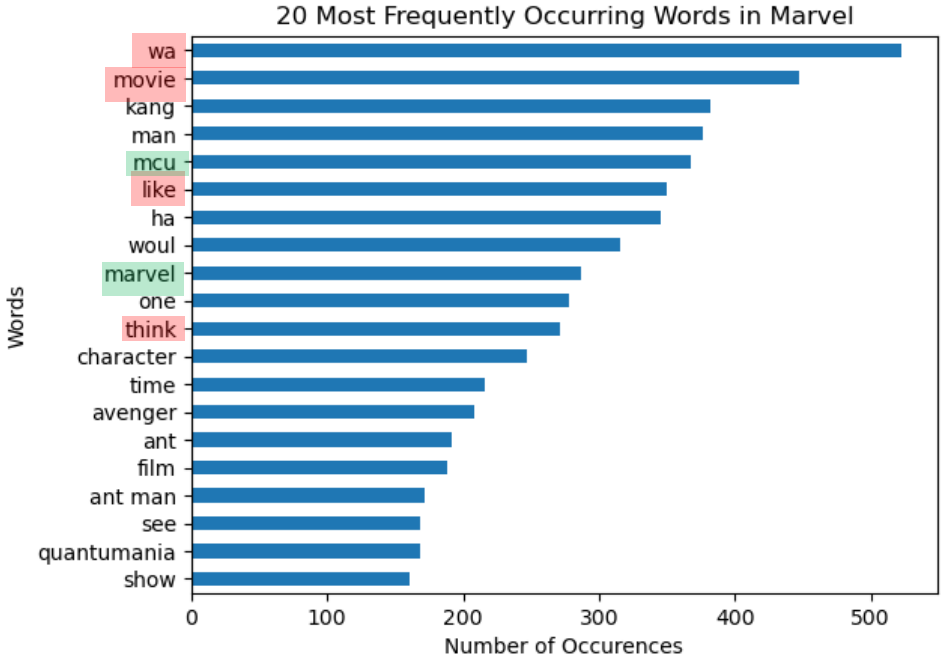
SUPER HEROES MEDIA

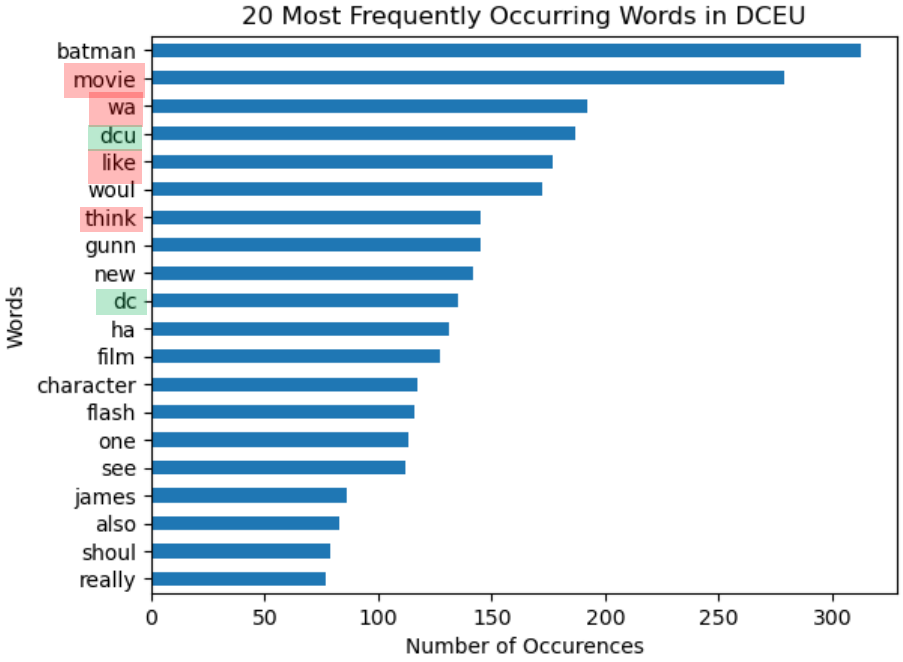
Outliers in Post Length



SUPER HEROES MEDIA

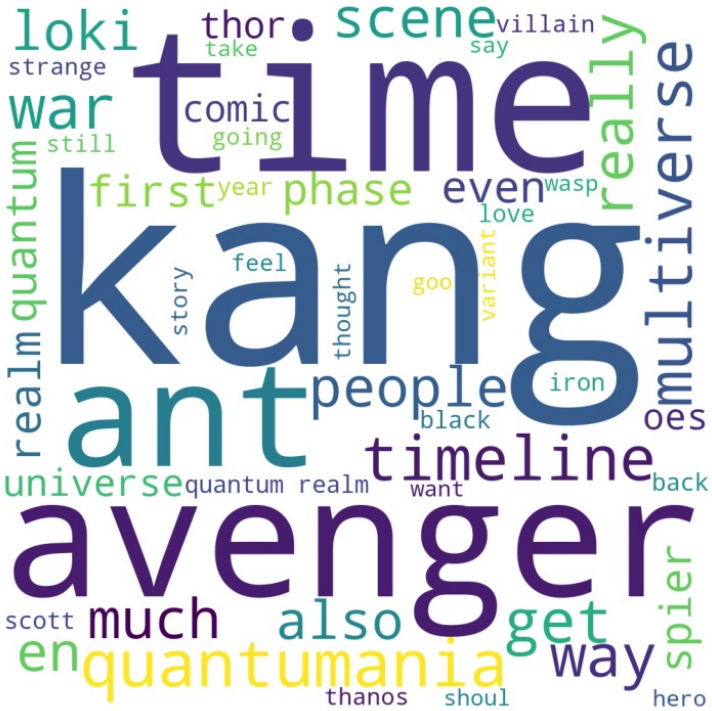
Most Frequently Occurring Words





Removing Common Words

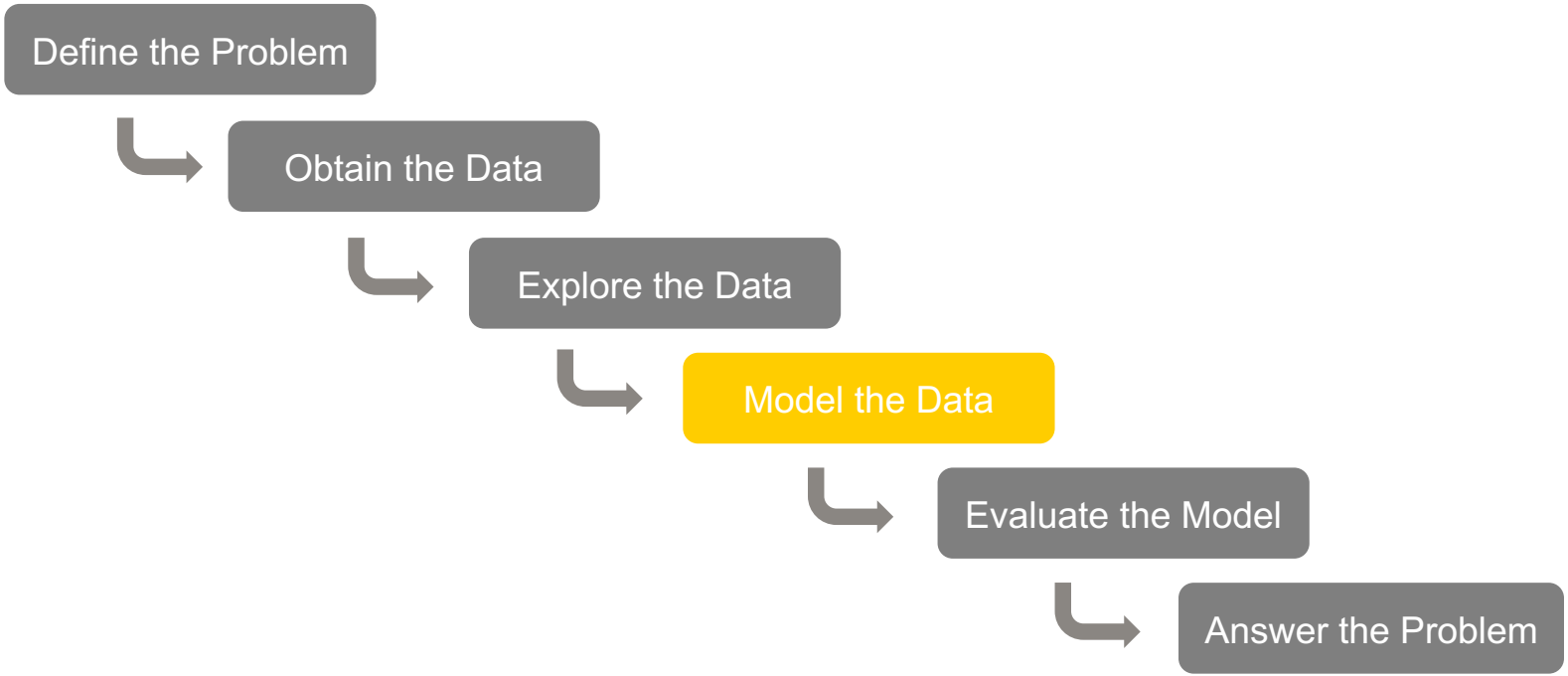
MCU



DCEU

SUPER HEROES MEDIA

Data Science Process



Modelling Process

Vectorization

1. Count Vectorizer
2. TF IDF Vectorizer

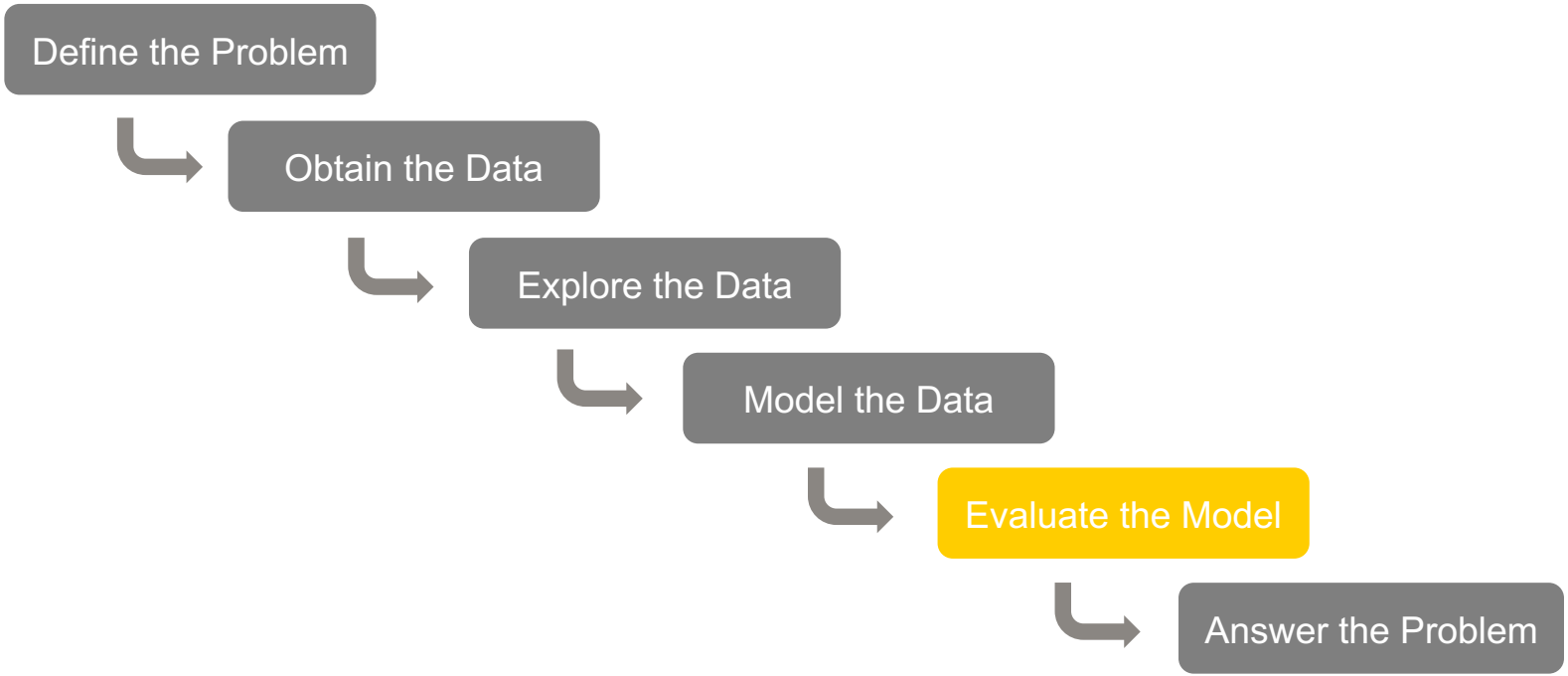
Apply Classification Algorithms

1. Base Model
2. Multinomial Naïve Bayes
3. Logistic Regression
4. KNN Classifier
5. Random Forest
6. Support Vector Machine

10 Models Trained and Tested

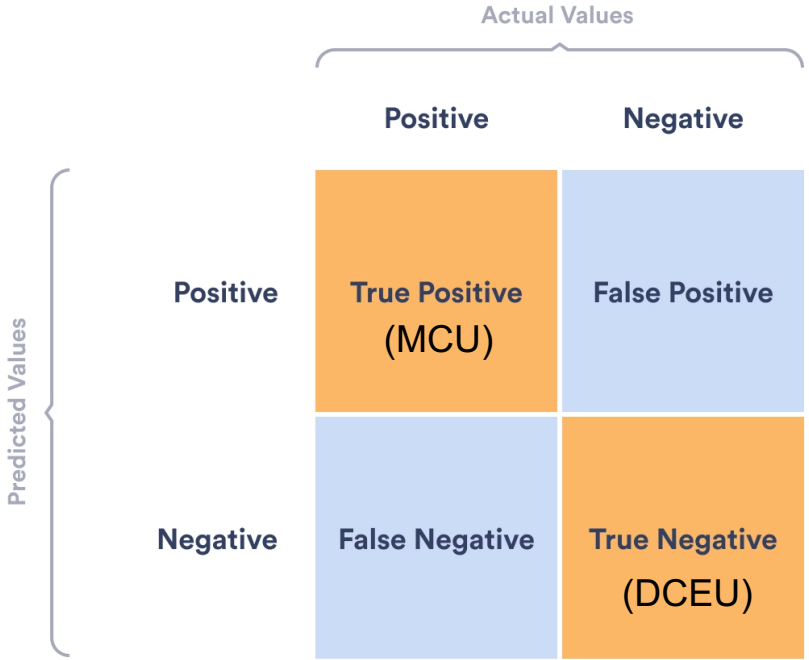
SUPER HEROES MEDIA

Data Science Process



SUPER HEROES MEDIA

Confusion Matrix



Model Evaluation Metrics

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Specificity

$$\frac{TN}{TN + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

Receiving Operating Characteristic (ROC) AUC

Measure of ability to distinguish between classes

F1

Harmonic man of Precision and Recall.

Effective metric when FP and FN are equally costly

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision

$$\frac{TP}{TP + FP}$$

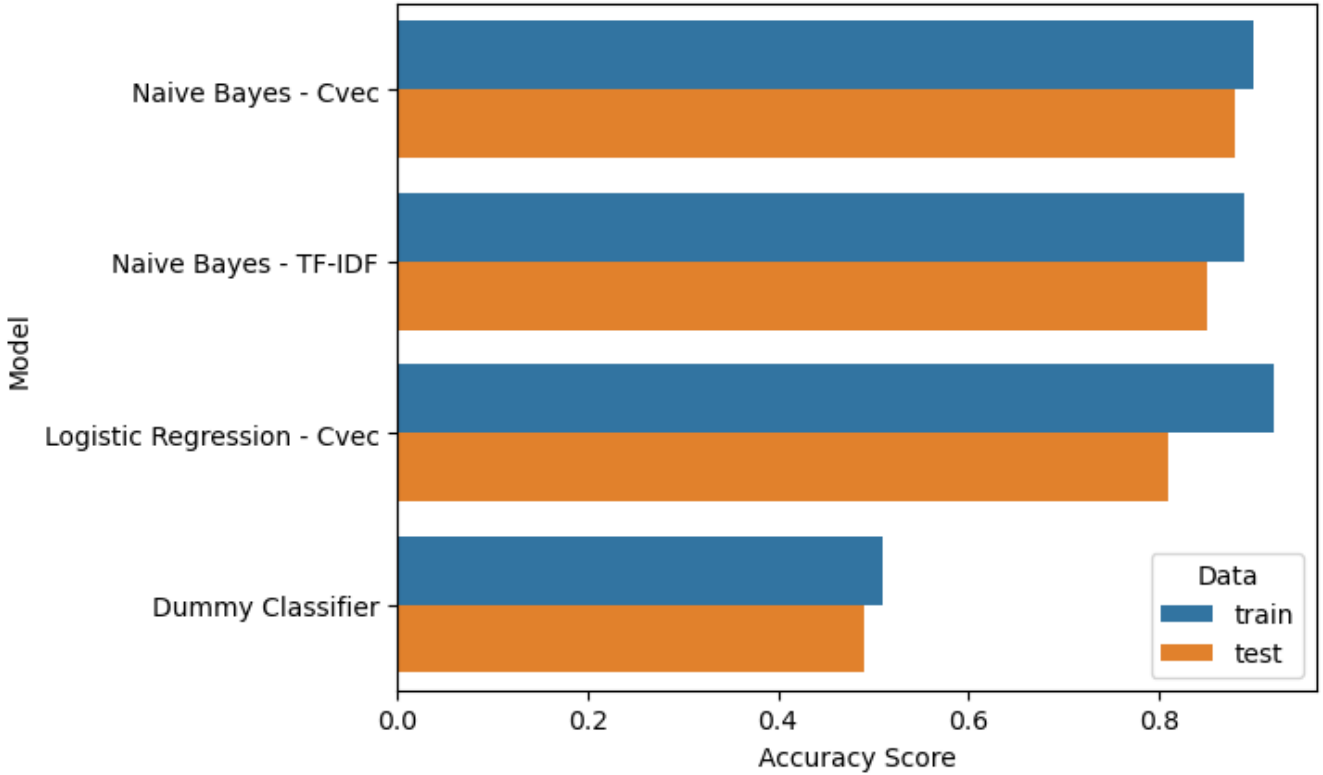
SUPER HEROES MEDIA

Results

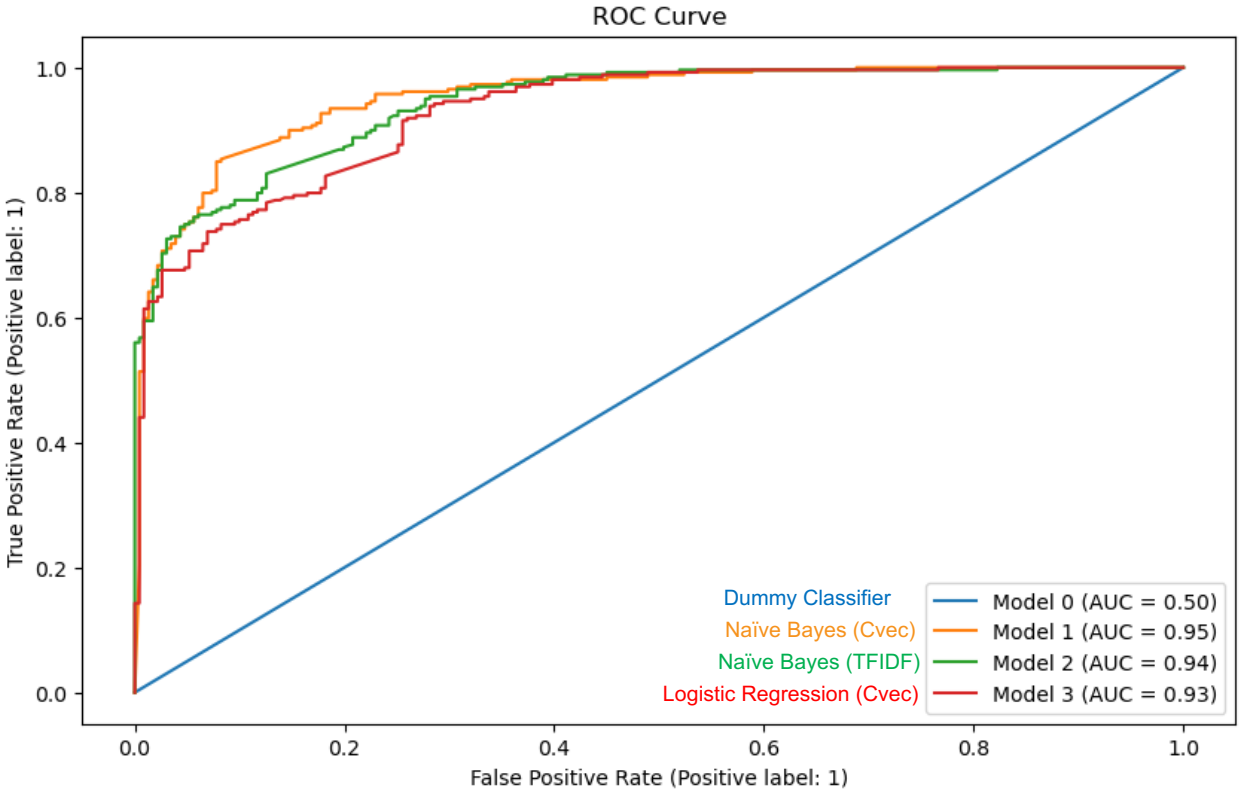
Model	Model Type	Training Accuracy	Test Accurary	Specificity	Recall	Precision	F1	ROC AUC
Model 0	Dummy Classifier	0.5133	0.4898	0.5152	0.4672	0.5193	0.4919	0.5
Model 1	Naive Bayes (Cvec)	0.8998	0.8837	0.9177	0.8533	0.9208	0.8858	0.95
Model 2	Naive Bayes (TF-IDF)	0.893	0.8510	0.8745	0.8301	0.8811	0.8545	0.94
Model 3	Logistic Regression (Cvec)	0.9264	0.8122	0.8182	0.8069	0.8327	0.8196	0.93
Model 4	Logistic Regression (TF-IDF)	0.8684	0.802	0.708	0.8842	0.7736	0.8252	0.91
Model 5	KNN Classifier (Cvec)	0.7778	0.6959	0.8571	0.5521	0.8125	0.6575	0.82
Model 6	KNN Classifier (TF-IDF)	0.7117	0.6612	0.71	0.6178	0.7049	0.6584	0.77
Model 7	Random Forest (Cvec)	0.8187	0.7939	0.9351	0.668	0.9202	0.774	0.91
Model 8	Random Forest (TF-IDF)	0.8098	0.8061	0.9697	0.6602	0.9607	0.7826	0.92
Model 9	Support Vector Machine (Cvec)	0.9298	0.7959	0.8225	0.7722	0.8299	0.8	0.9
Model 10	Support Vector Machine (TF-IDF)	0.9652	0.8347	0.8442	0.8263	0.856	0.8409	0.91

SUPER HEROES MEDIA

Accuracy Score

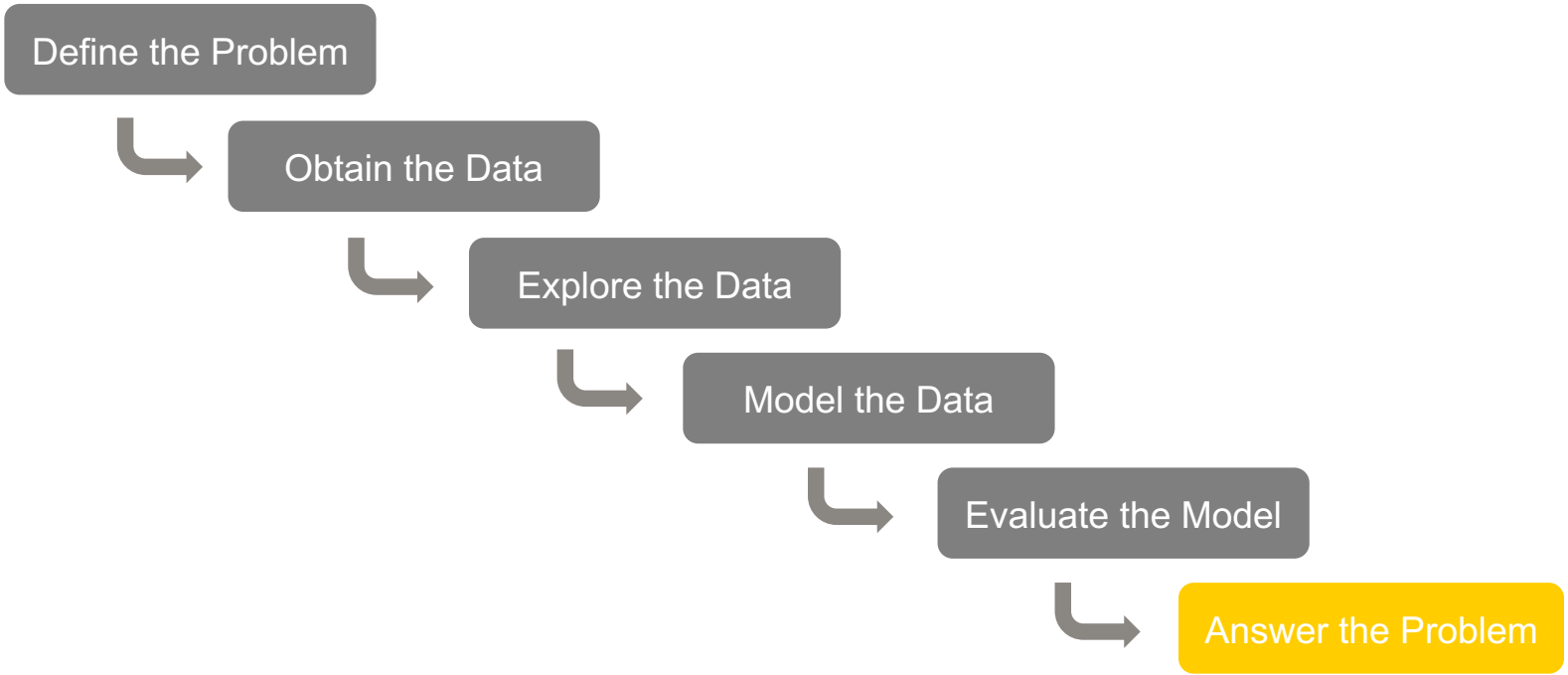


ROC AUC


ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
Dummy Classifier
Naïve Bayes (Cvec)
Naïve Bayes (TFIDF)
Logistic Regression (Cvec)
Model 0 (AUC = 0.50)
Model 1 (AUC = 0.95)
Model 2 (AUC = 0.94)
Model 3 (AUC = 0.93)

SUPER HEROES MEDIA

Data Science Process



SUPER HEROES MEDIA

Recap on Problem Statement

- Trial implementation of a Natural Language Processing (NLP) Model to prevent information infringement onto other client domains

SUPER HEROES MEDIA

Production Model

- Naïve Bayes

Count Vectorizer

Accuracy: 0.884

ROC AUC: 0.95

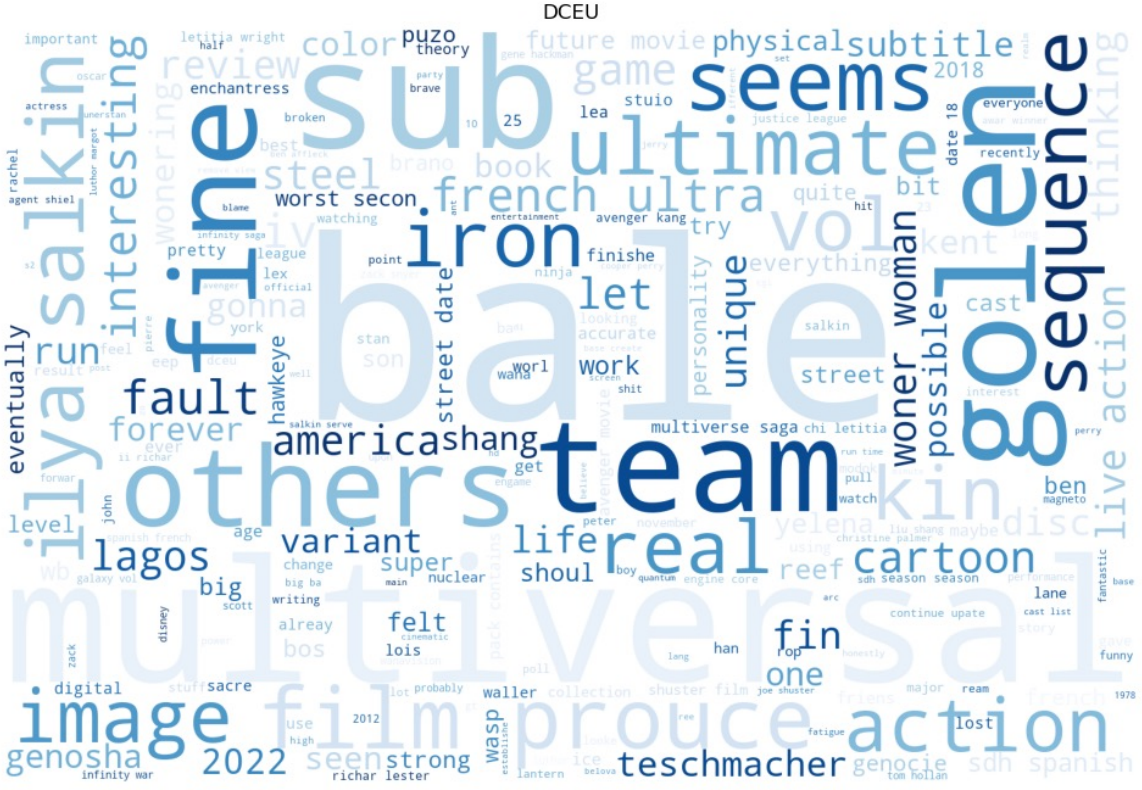
TF-IDF Vectorizer

Accuracy: 0.85

ROC AUC: 0.94

Explore TF-IDF Vectorizer further as it could help with word importance when a larger dataset is introduced

Feature Importance (DCEU)



Feature Importance (MCU)

SUPER HEROES MEDIA

Conclusion

- The model can differentiate between the two brands quite well
- To adopt as a secondary tool for automated infringement checks

SUPER HEROES MEDIA

Recommendations

Increase Model Accuracy

- Try other classification algorithms (e.g. decision trees, bagging, boosting, other complex models)
- Remove more common words

Increase Model Useability

- Include older data, going back to MCU and DCEU origins
- Use data from other sources (e.g Facebook, twitter)
- Consider data from Marvel and DC comics

Thanks!

Any questions ?