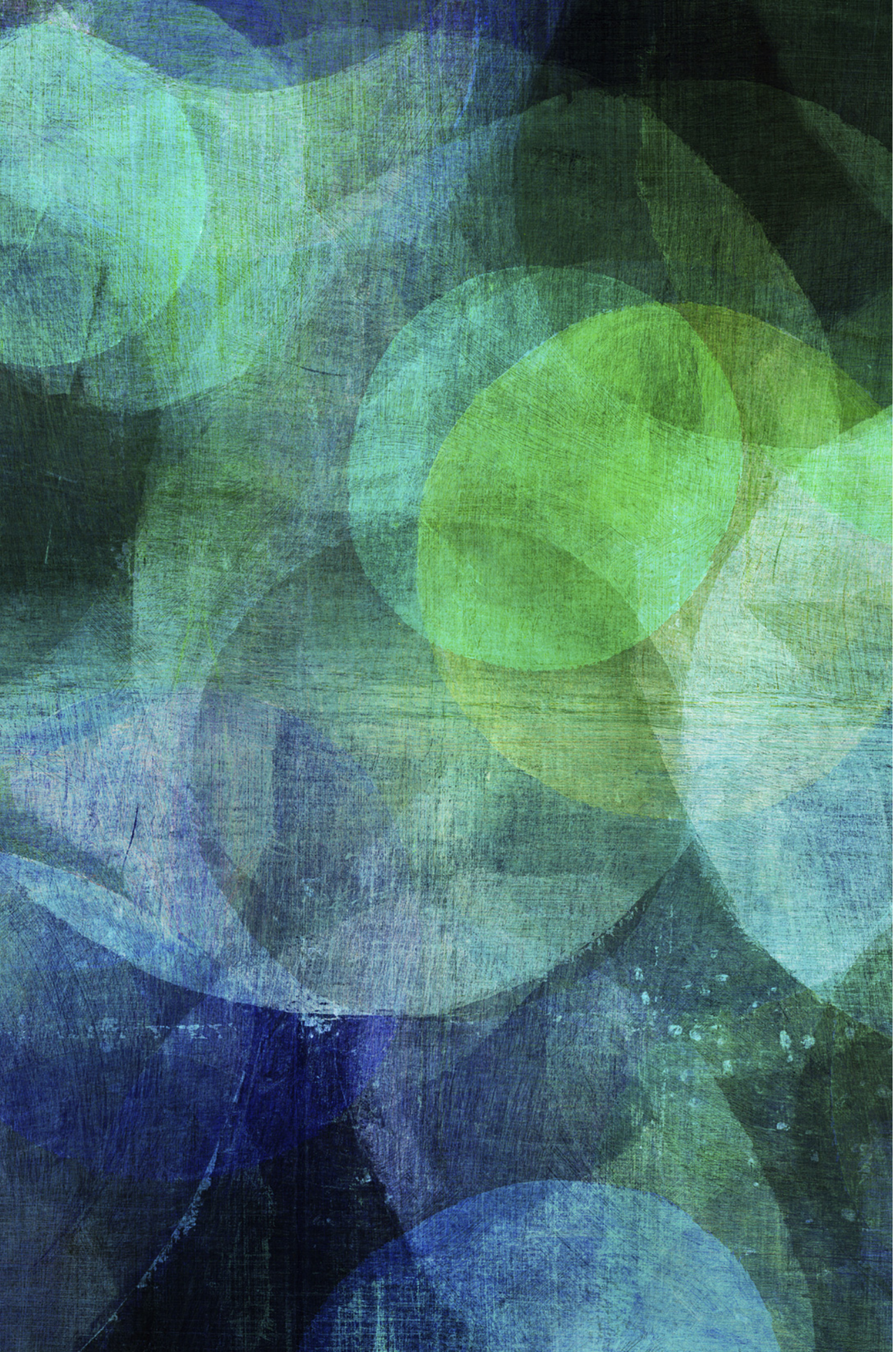
MARVEL VS. DC COMICS SUBREDDIT EDITION

Presented by Sharnique Beck

THE DATA SCIENCE PROCESS

- Problem Statement
- Data Collection
- Data Cleaning & Preprocessing
- Modeling
- Evaluation
- Conclusion and Recommendations

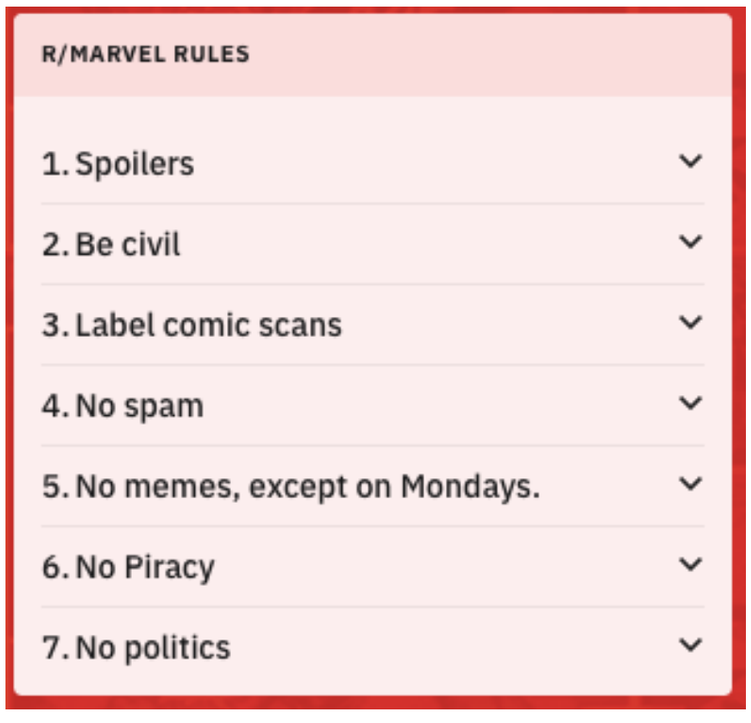


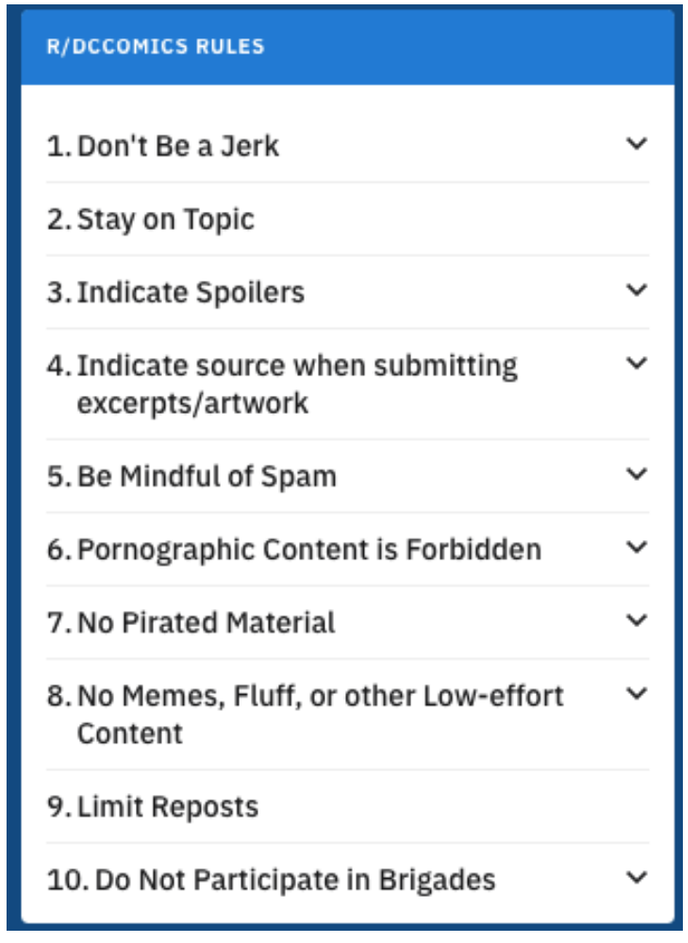
PROBLEM STATEMENT

- An online comic book store wants to determine where forum posts are originating from.
- Create and train a classifier model on which subreddit a given post came from, Marvel or DC.

DATA COLLECTION

- https://api.pushshift.io/reddit
  - r/Marvel
  - r/DCcomics
- Collected 5000 latest post from each





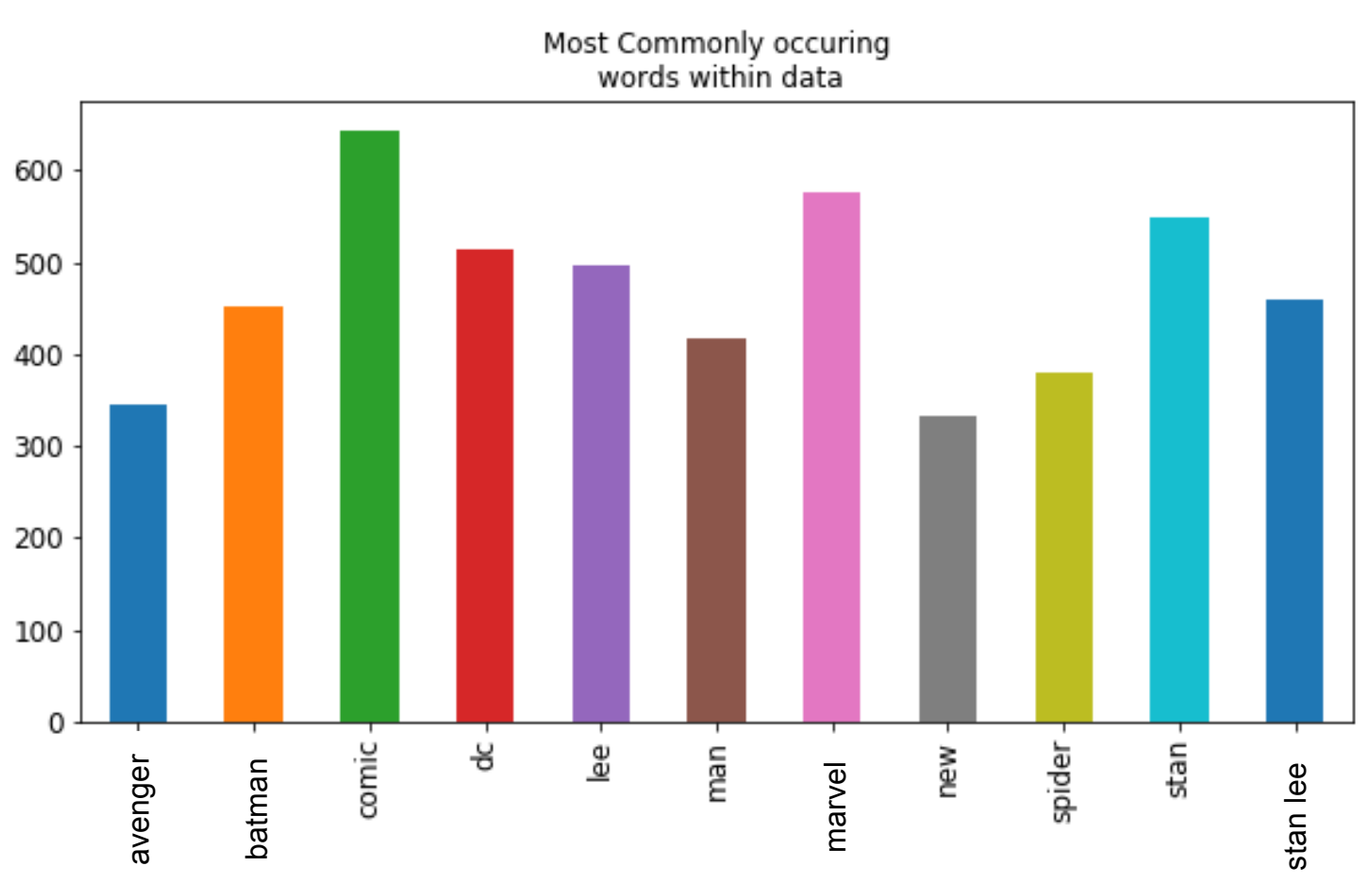
DATA CLEANING & PREPROCESSING

- Cleaned collected subreddit titles using:
  - Regular Expression
  - Removing stop words
  - Lemmatizing
- Resulted in some empty strings
  - 'Don’t do it'
  - '🤷‍♀\u200d♂....'
- Unbalanced data

MODELING

- Baseline accuracy

```
0    0.501003
1    0.498997
```

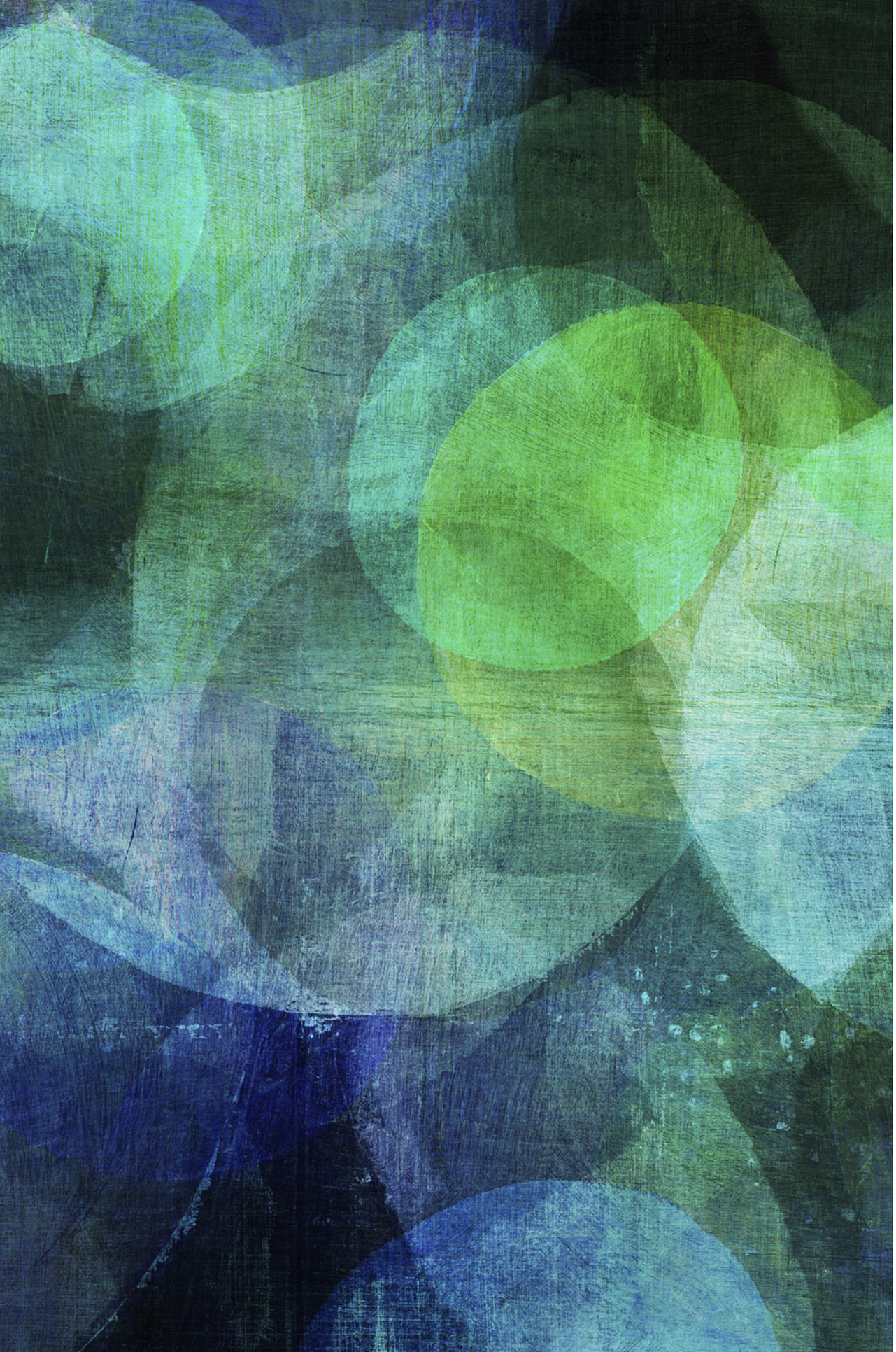


- **Binomial Naive Bayes**

```
Best parameters: {'tvect__max_features': None, 'tvect__ngram_range': (1, 2)}
Train Score: 0.8967776440700629
Test Score: 0.9037304452466908
```

- **Extra Trees**

```
Best parameters: {'max_depth': None, 'min_samples_split': 4, 'n_estimators': 30}
train score: 0.9949191068324642
test score: 0.8728439630966707
```



EVALUATION

Naive Bayes Model

- Confusion Matrix

	predicted neg	predicted pos
actual neg	1096	153
actual pos	87	1157

- 90.37% accuracy
- 153 Type I errors
- 87 Type II errors

CONCLUSION AND RECOMMENDATIONS

- The model will be able to predict a posts subreddit with very good accuracy
- If I had more time
  - Explore random forest and extra trees parameters further
  - Collect more data, include comments
- Use for sentiment analysis of combined comic boards, weight comment biases
  - Constructive vs critical

SOURCES

- https://comicbook.com/marvel/2018/02/01/marvel-vs-dc-twitter-trending-2018/
- https://api.pushshift.io/reddit/search/submission/?subreddit=Marvel
- https://api.pushshift.io/reddit/search/submission/?subreddit=DCcomics