



MARVEL VS. DC COMICS  
SUBREDDIT EDITION

Presented by Sharnique Beck

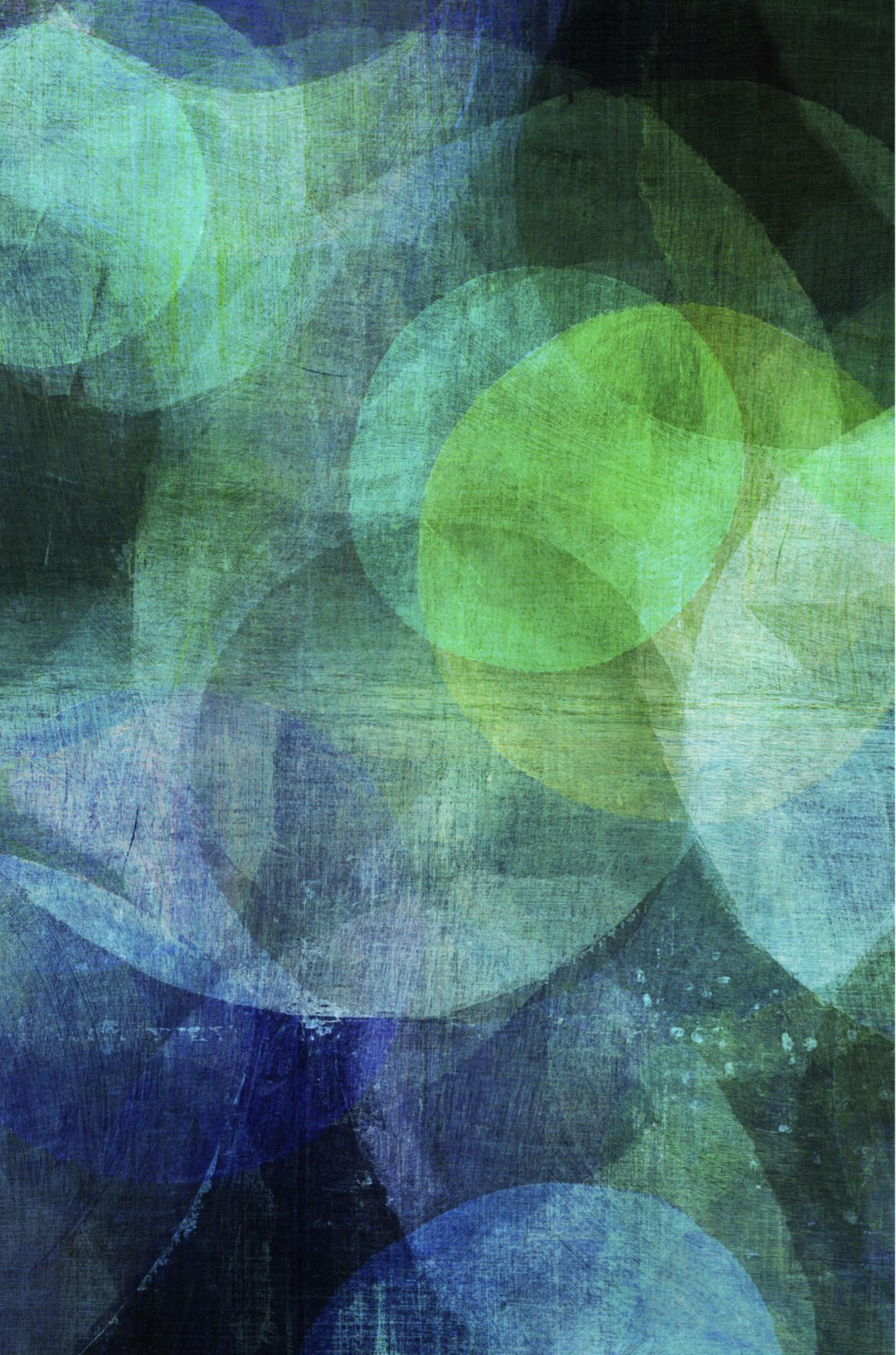




# THE DATA SCIENCE PROCESS

---

- 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

R/MARVEL RULES	
1. Spoilers	▼
2. Be civil	▼
3. Label comic scans	▼
4. No spam	▼
5. No memes, except on Mondays.	▼
6. No Piracy	▼
7. No politics	▼

R/DCCOMICS RULES	
1. Don't Be a Jerk	▼
2. Stay on Topic	
3. Indicate Spoilers	▼
4. Indicate source when submitting excerpts/artwork	▼
5. Be Mindful of Spam	▼
6. Pornographic Content is Forbidden	▼
7. No Pirated Material	▼
8. No Memes, Fluff, or other Low-effort Content	▼
9. Limit Reposts	
10. Do Not Participate in Brigades	▼



## 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

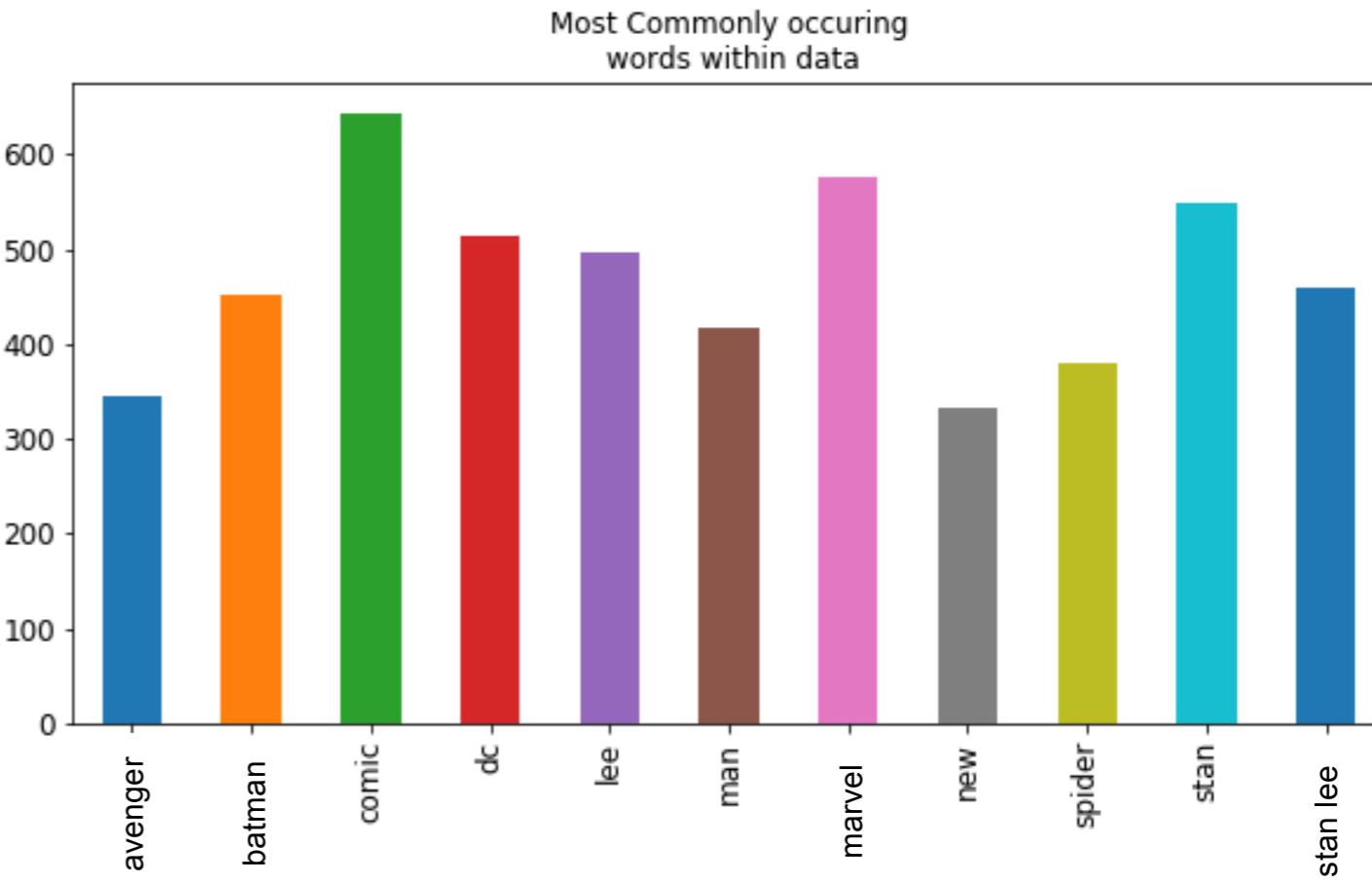
Test Score: 0.9037304452466908

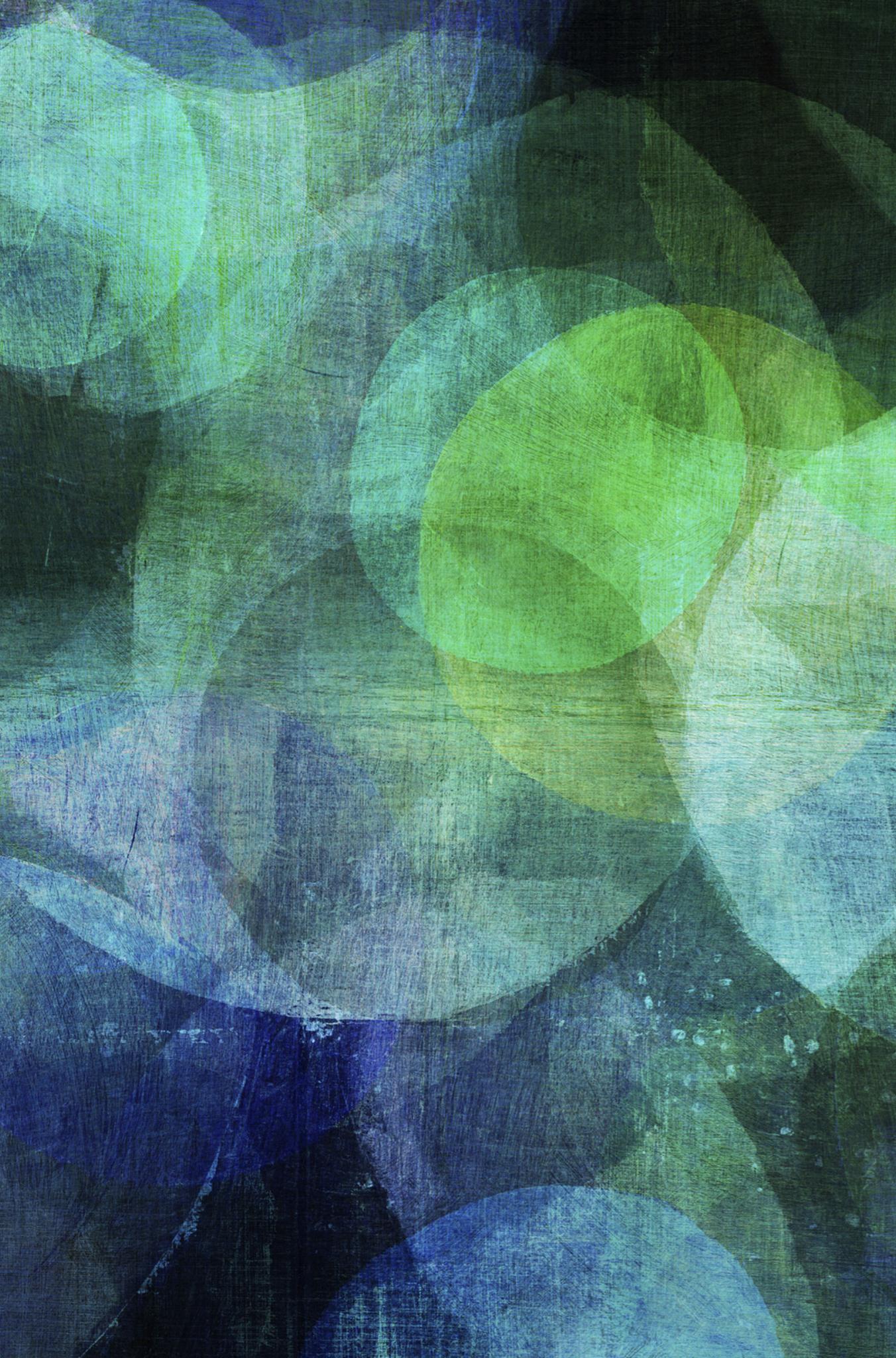
## ► Extra Trees

Best parameters: {'max\_depth': None, 'min\_samples\_split': 4, 'n\_estimators': 30}

train score: 0.9949191068324642

test score: 0.8728439630966707





# EVALUATION

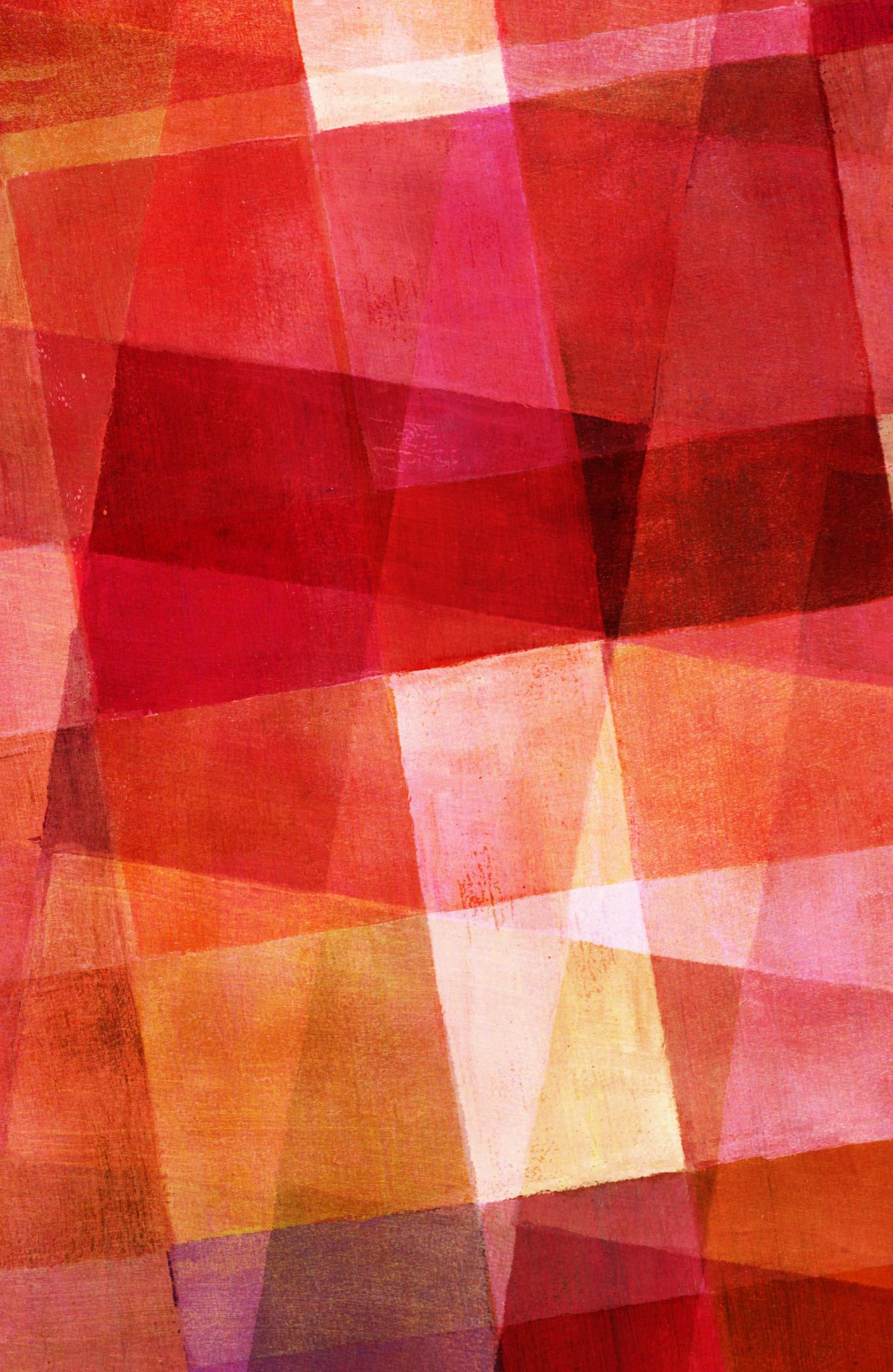
---

## Naive Bayes Model

- Confusion Matrix

	<b>predicted neg</b>	<b>predicted pos</b>
<b>actual neg</b>	1096	153
<b>actual pos</b>	87	1157

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



## CONCLUSION AND RECOMMENDATIONS

---

- The model will be able to predict a post's 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>