



Funny Dads Council

EMERGENCY MEETING

Opening Joke

- Why should one never play poker in a casino?

Dad Joke #1

It's filled with too many cheetahs.

Problem Statement:

- ▶ Are the Dads Among us Playing Poker?

Data Science Workflow



1. Web-Scraping
2. Data Cleaning
3. Feature Engineering
4. Cleaning and Pre-Processing
5. Modeling
6. Conclusions and Next Steps

Web Scraping

- ▶ Scraping Tool Provided by Gwen Rathgeber and Ben Mathis
- ▶ Scraped Two Subreddits
 - ▶ Dad Jokes
 - ▶ 3,697 days worth of data collected
 - ▶ 50,102 rows and 90 columns in our Raw Data Frame
 - ▶ Poker
 - ▶ 5,395 days worth of data collected
 - ▶ 50,234 rows and 99 columns in our Raw Data Frame

Data Cleaning

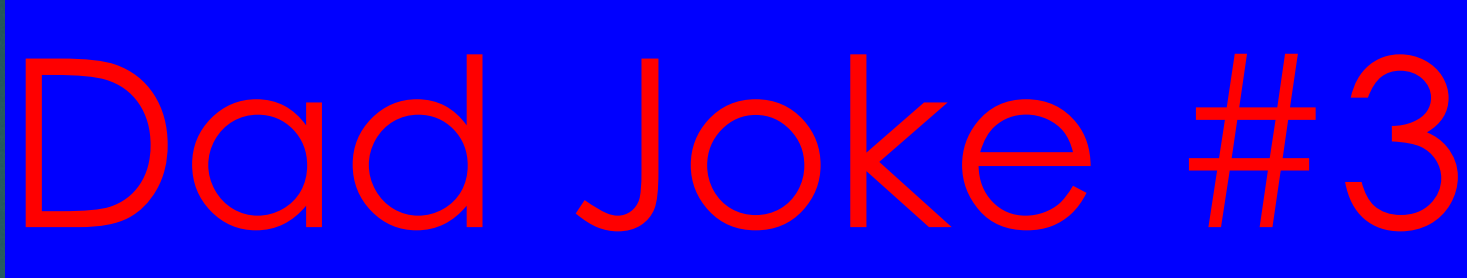
- ▶ Column Reduction
 - ▶ Reduced from 90+ columns to 3
 - ▶
- ▶ Row

Dad Joke #2

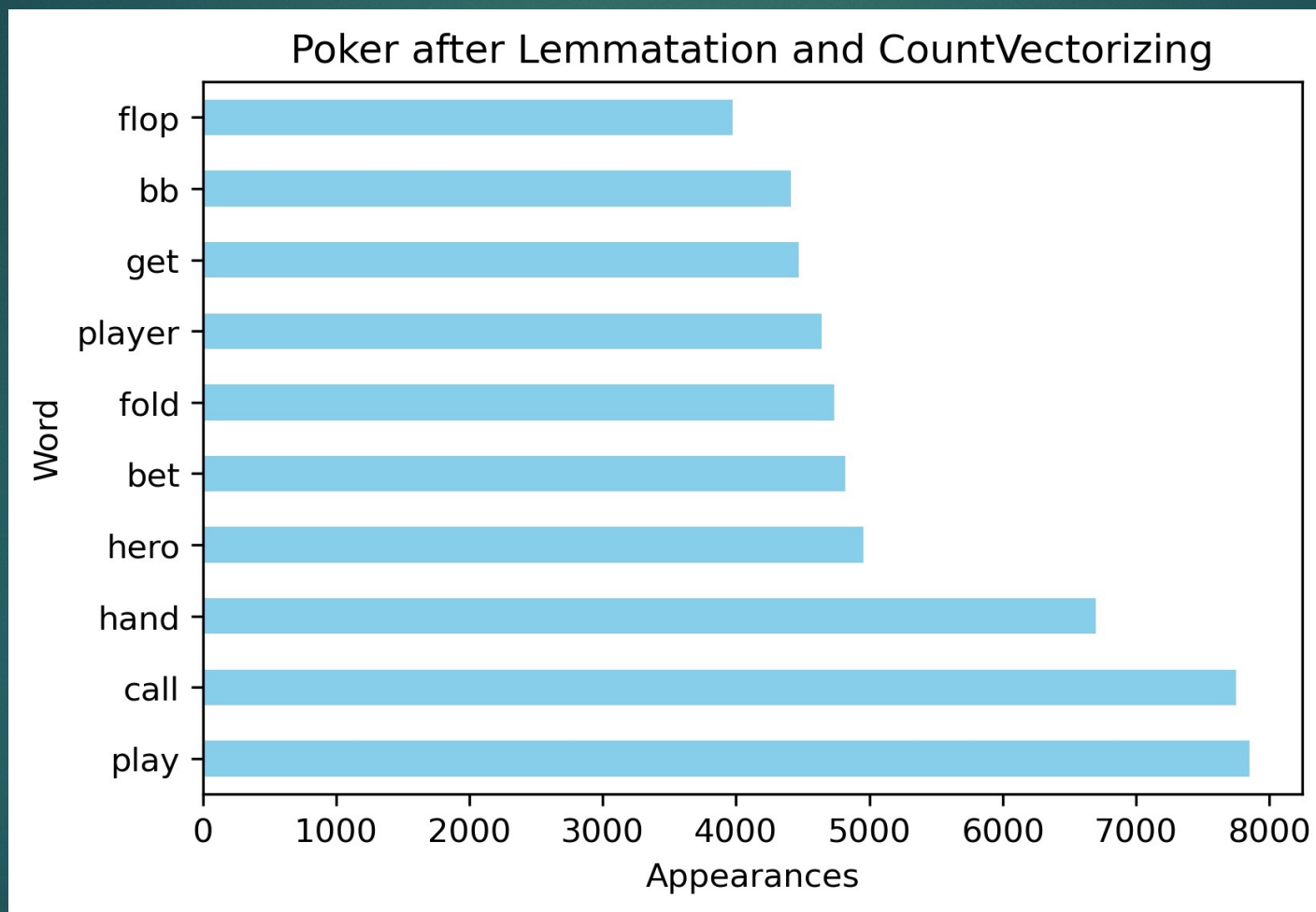
Reduction Step	Poker	Dad Jokes
Initial	50,234	50,102
Nulls	29,948	48,613
High Value Counts	27,405	47,378
Duplicates	6,778	9,713

Feature Engineering

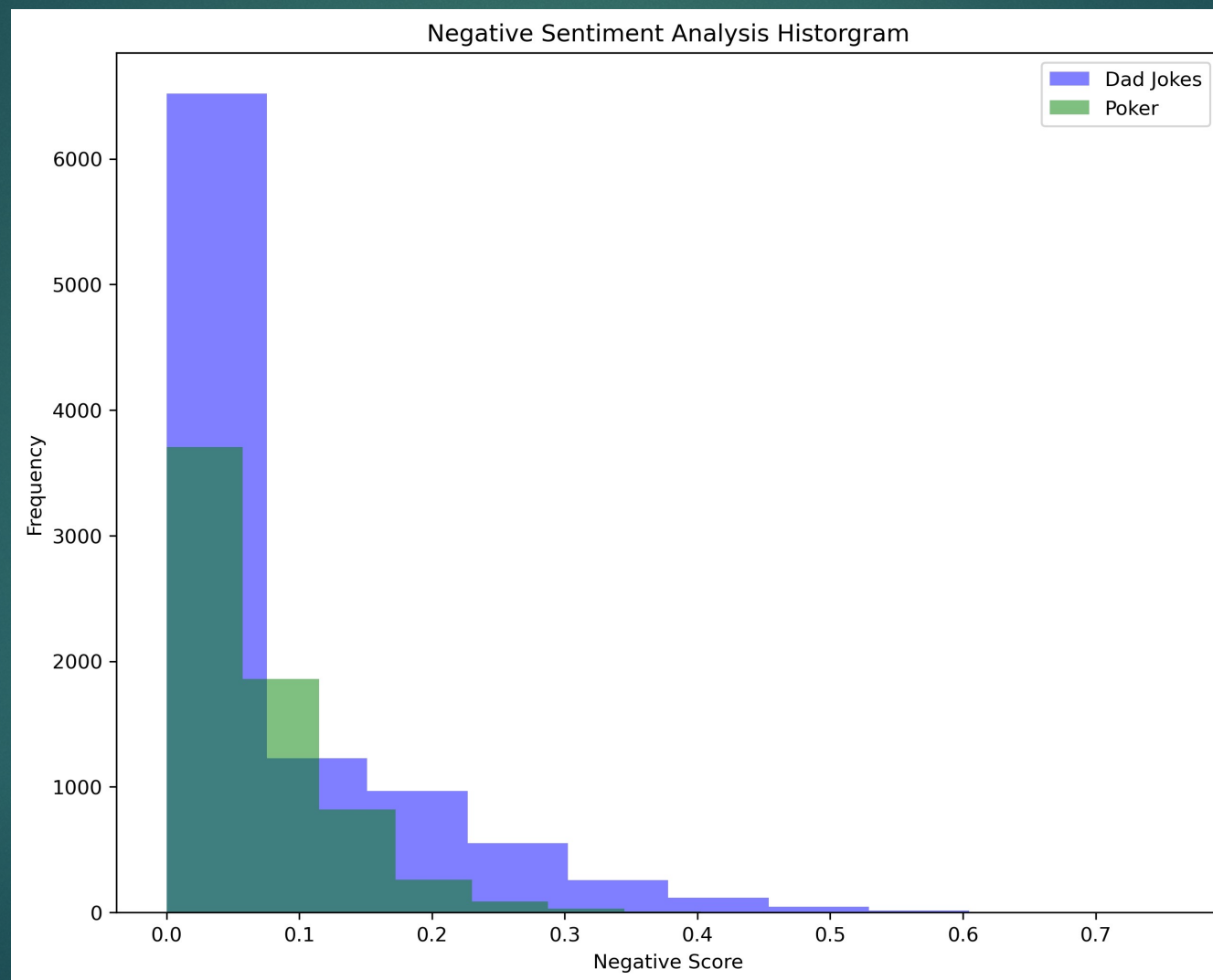
- ▶ Tokenization
- ▶ Sentiment Analysis
- ▶ Lemmatization
- ▶ Count Vectorization



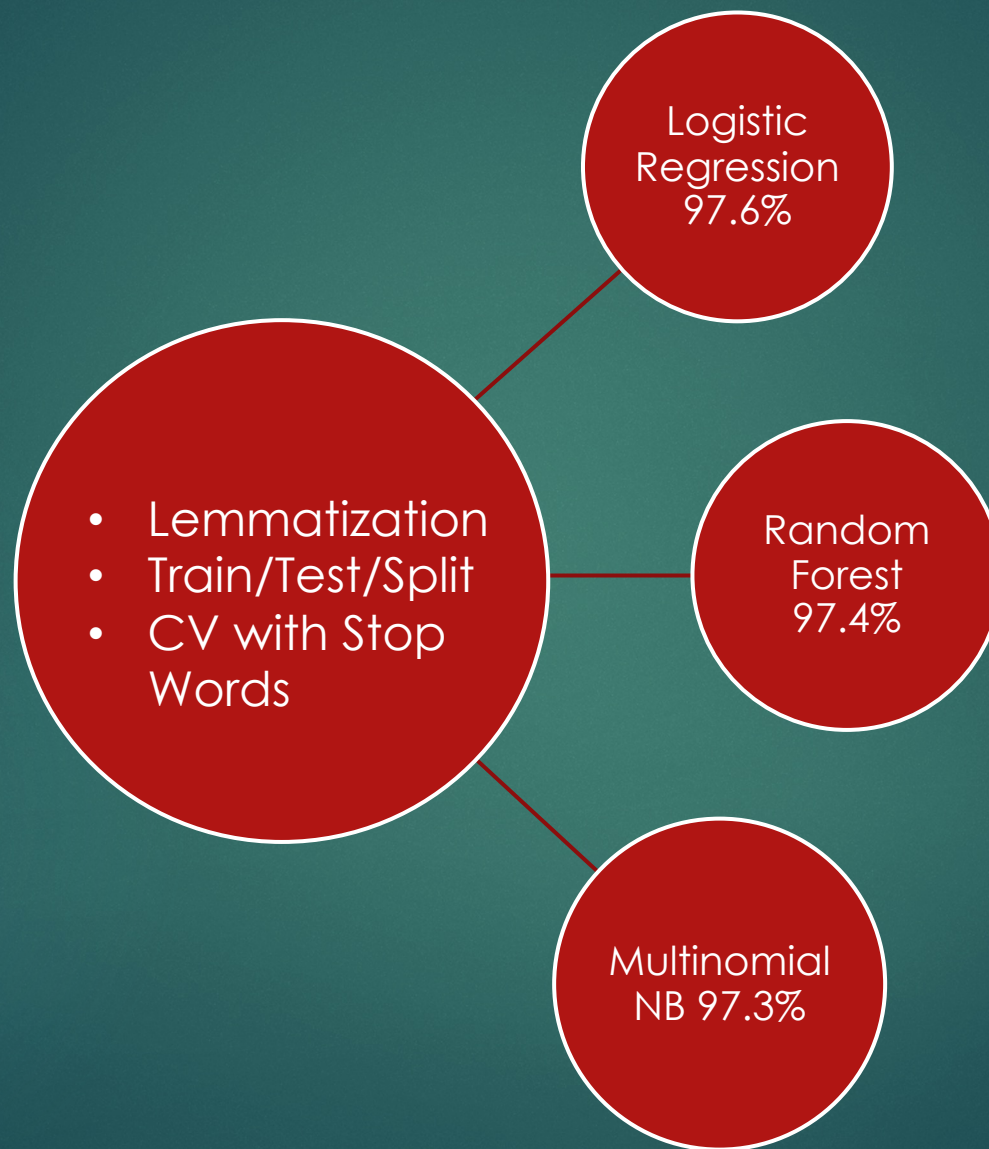
EDA – Feature Engineering



EDA – Feature Engineering

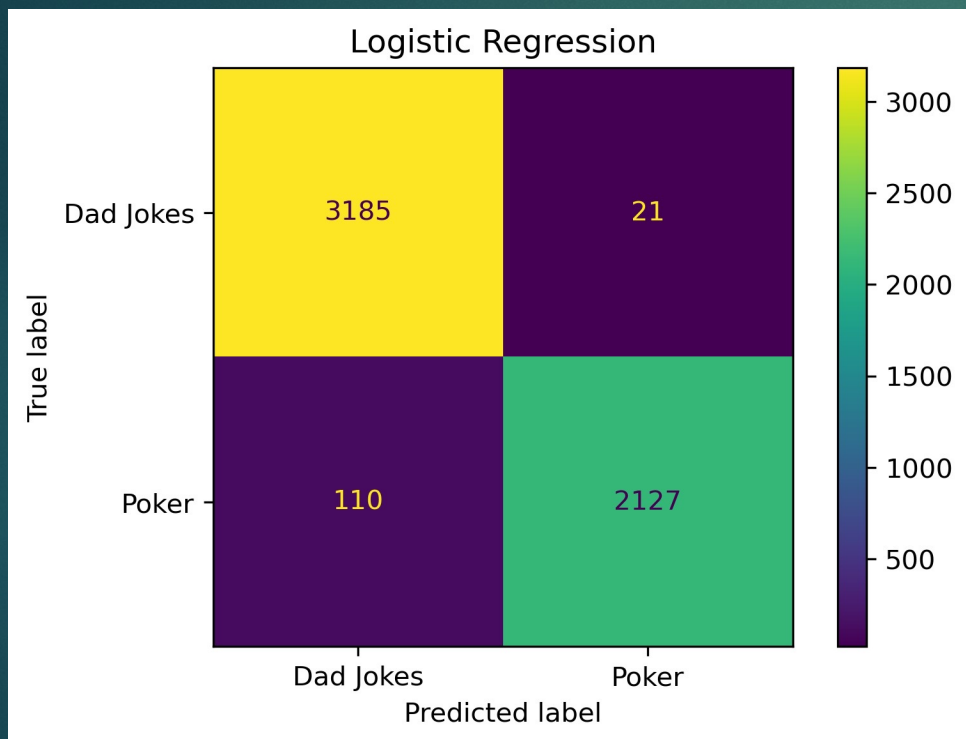


Modeling



*Overfitting Likely

Conclusions – LogReg is Best



	Baseline	Model Precision
Dad Jokes	58.9%	97%
Poker	41.1%	99%

In Conclusion

- ▶ We should be Safe
- ▶ Grab a flier of our Lemmatized, CountVectorized Poker Most Common Words
- ▶ Stay Vigilant

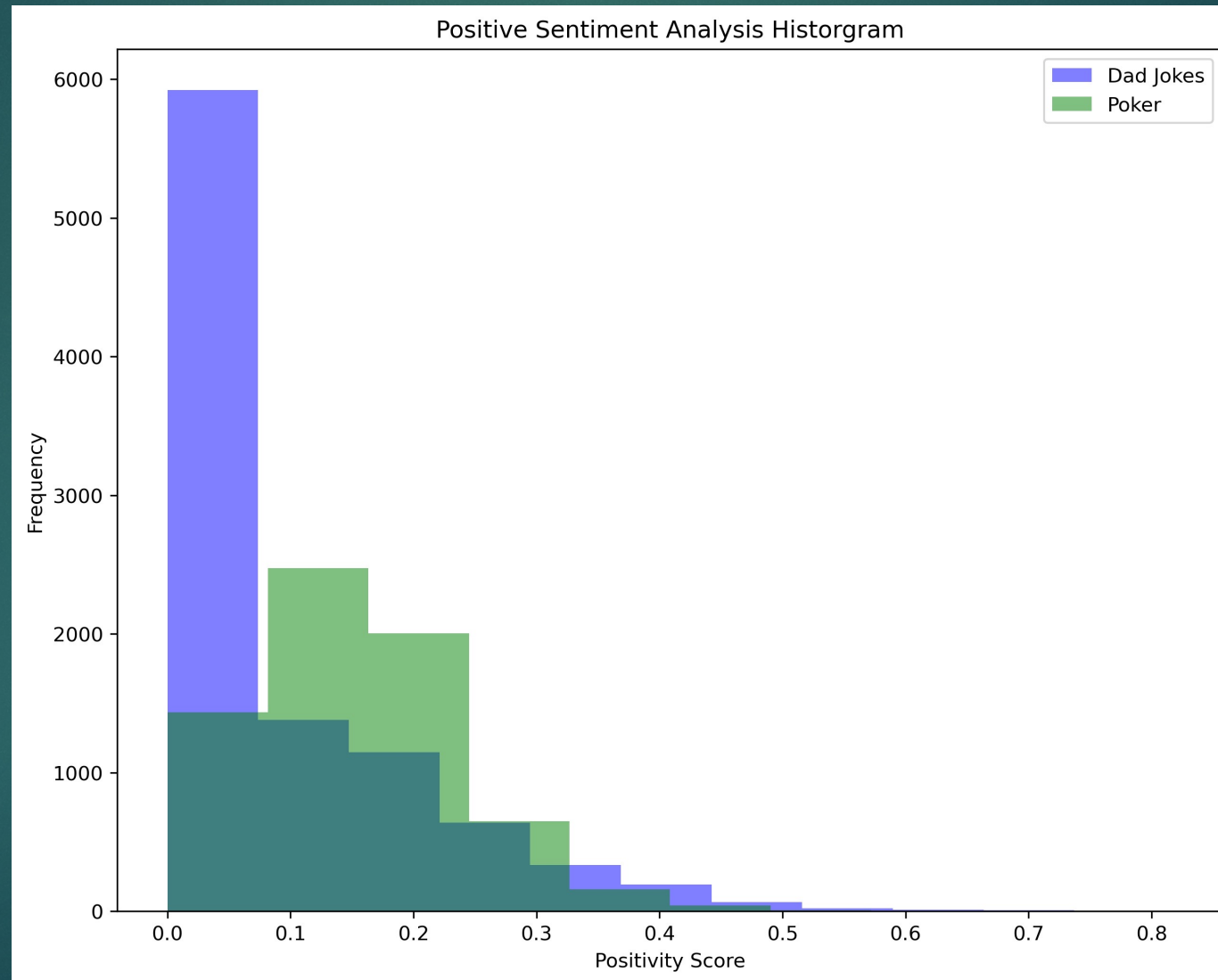


Thank you for your Time!

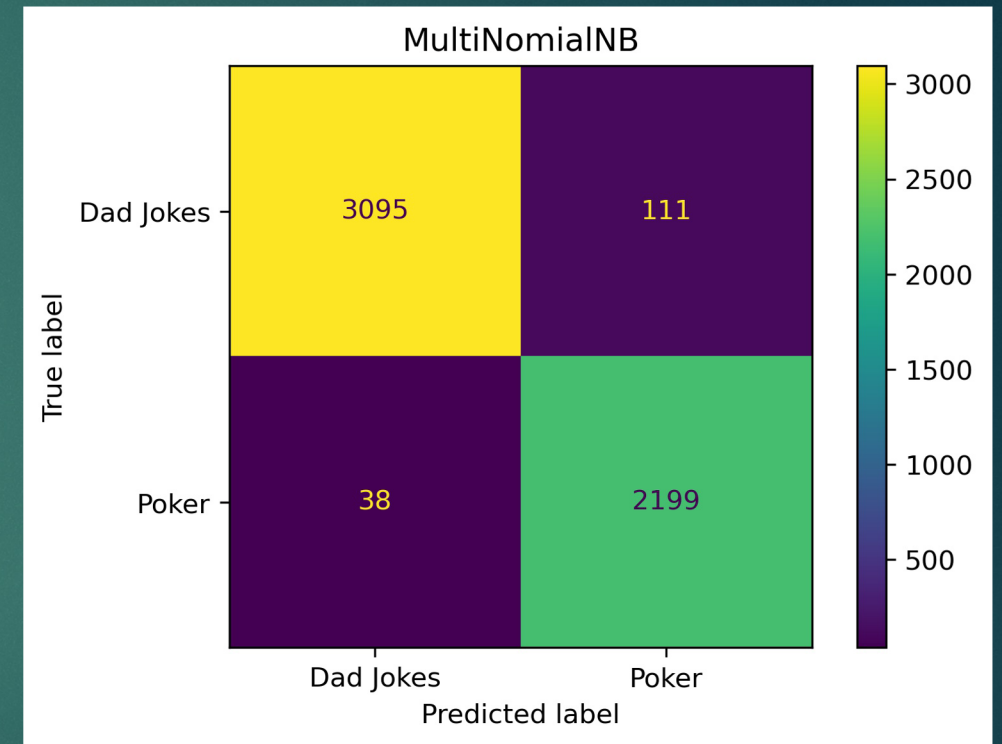
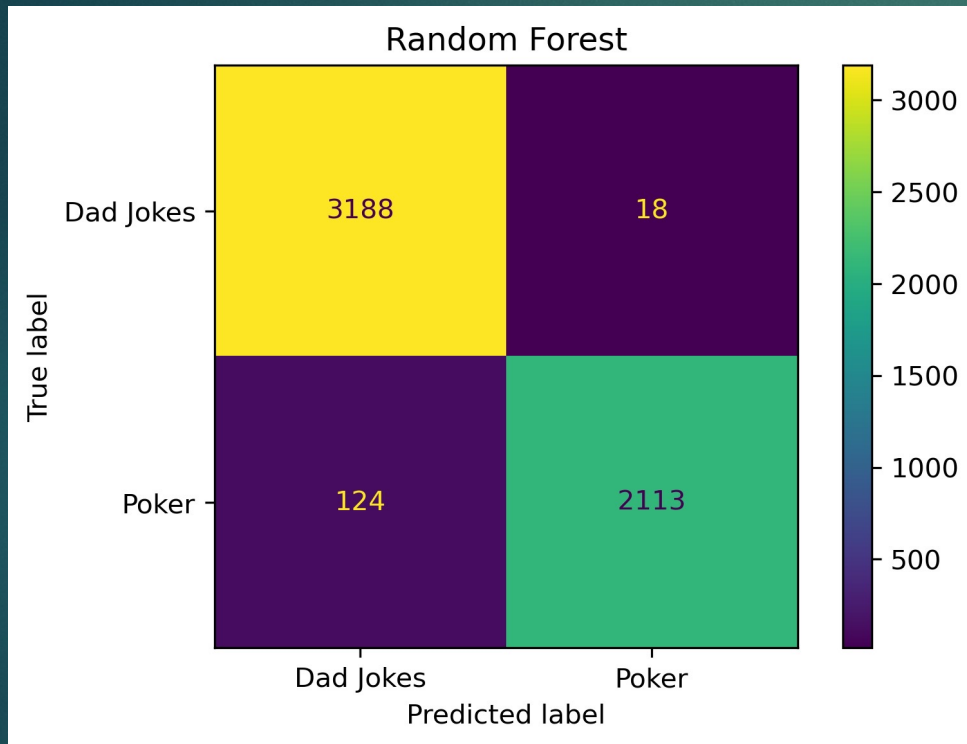
Questions or Comments?

Additional Slides

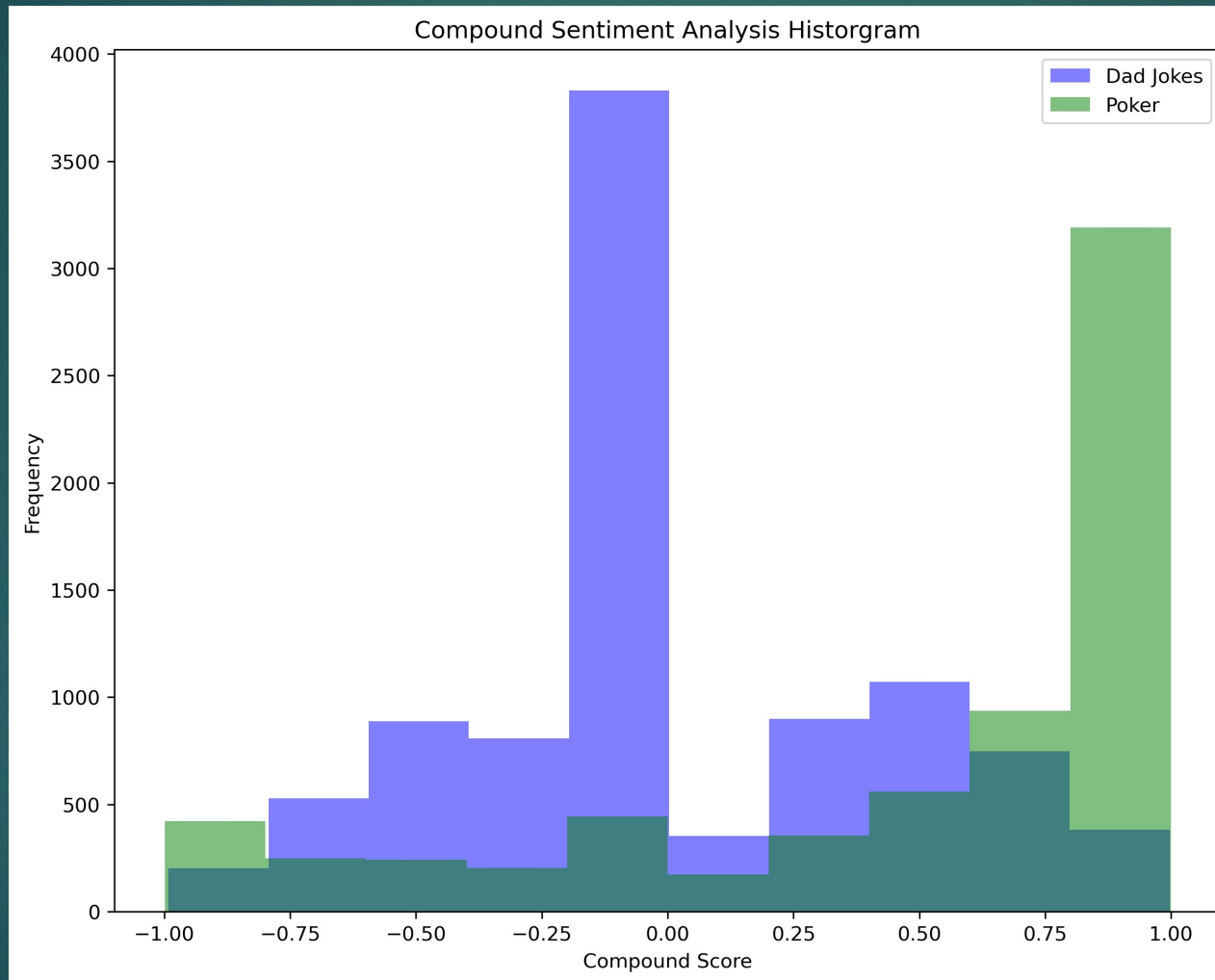
EDA – Feature Engineering



Conclusions



EDA – Feature Engineering



EDA – Feature Engineering

