R/NARUTO VS. R/ANIME

A PRESENTATION ON CLASSIFICATION MODELS USING EARLY 21ST CENTURY MACHINE LEARNING TECHNIQUES

PRESENTED BY: DANIEL WANG (OF THE NEO-EARTH RESEARCH COUNCIL)

3023 Mars Conference on Pre-World War 3 Humanoid Anthropology

PRESENTATION OUTLINE

1. PROBLEM STATEMENT

Recovered Pre-WWIII data discs of popular 21st century website 'reddit'

2. BACKGROUND INFO

What is anime? Who is Naruto?

3. DISCUSSION OF METHODS

21st Century Machine Learning & Classification Methods

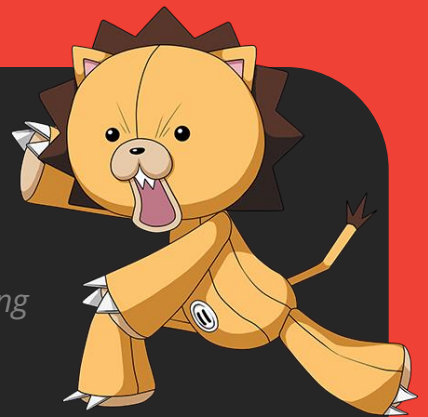
4. RESULTS & FINDINGS

Classification Model Results

5. CONCLUSIONS

Evaluating Findings

PROBLEM STATEMENT



- ***Recovered 21st century artifacts***

  During a recent expedition to the ruins of alpha-Earth, an exploration team recovered a semi-operable server containing data from the popular 21st century website Reddit.

- ***21st century data science techniques***

  Additionally, the team discovered a manual on early 21st century data science techniques for classification and machine learning.

- ***Using 21st century methods to analyze 21st century data***

  We explored the use of natural language processing and several classification algorithms to differentiate between posts from r/anime vs. those from r/Naruto.

ANIME

Anime is hand-drawn and computer-generated animation originating from Japan. Outside of Japan and in English, anime refers to Japanese animation, and refers specifically to animation produced in Japan. However, in Japan and in Japanese, anime (a term derived from a shortening of the English word animation) describes all animated works, regardless of style or origin. Animation produced outside of Japan with similar style to Japanese animation is commonly referred to as anime-influenced animation.

Naruto was an anime/manga series initially airing from 2002-2017. The series' popularity coincided with (& helped fuel) an explosion in popularity of anime/manga outside of Japan.*

*It's effects on the genre were significant enough that it can be considered the quintessential shounen** series of the early 21st century, upon which many subsequent anime tropes were based.*

** Naruto (2002–07)*

Naruto: Shippuden (2007–17)

*** shounen = "boys' comics", targeting an audience of adolescent boys*

NARUTO
-ナルト-

free printable invitation - drevio.com

Methods

1. Data scraping via API
2. Tokenizing
3. Lemmatizing
4. Stop Words
5. n-word grams
6. Count vectorization
7. Grid Search
8. Confusion Matrices

Classification Methods

1. Naïve Bayes (Using Standard CountVectorizer)
2. NB w/TfidfVectorizer (Term Frequency-Inverse Document Frequency Vectorizer)
3. Random Forest
4. K Nearest Neighbors

Baseline

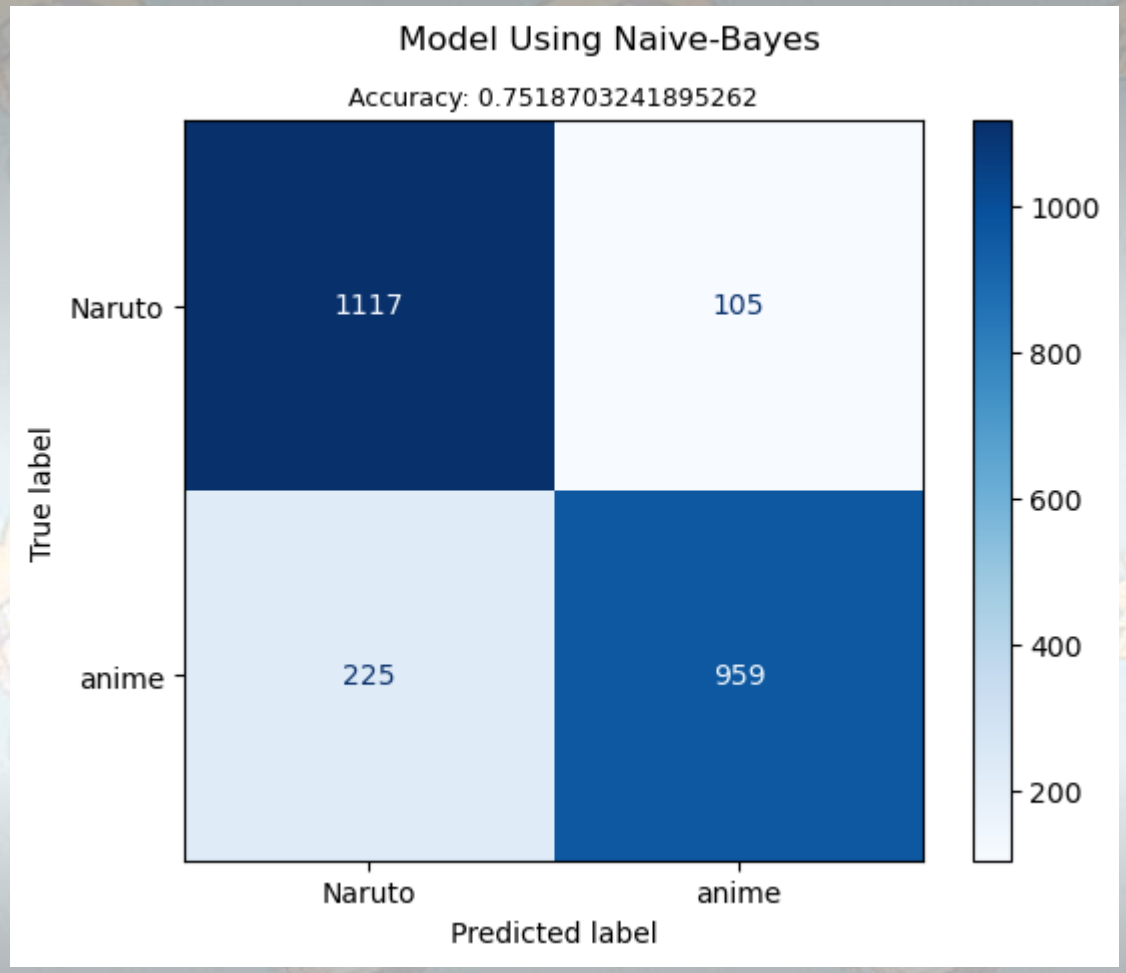
Dataset:

4887 posts from r/anime (50.8%)

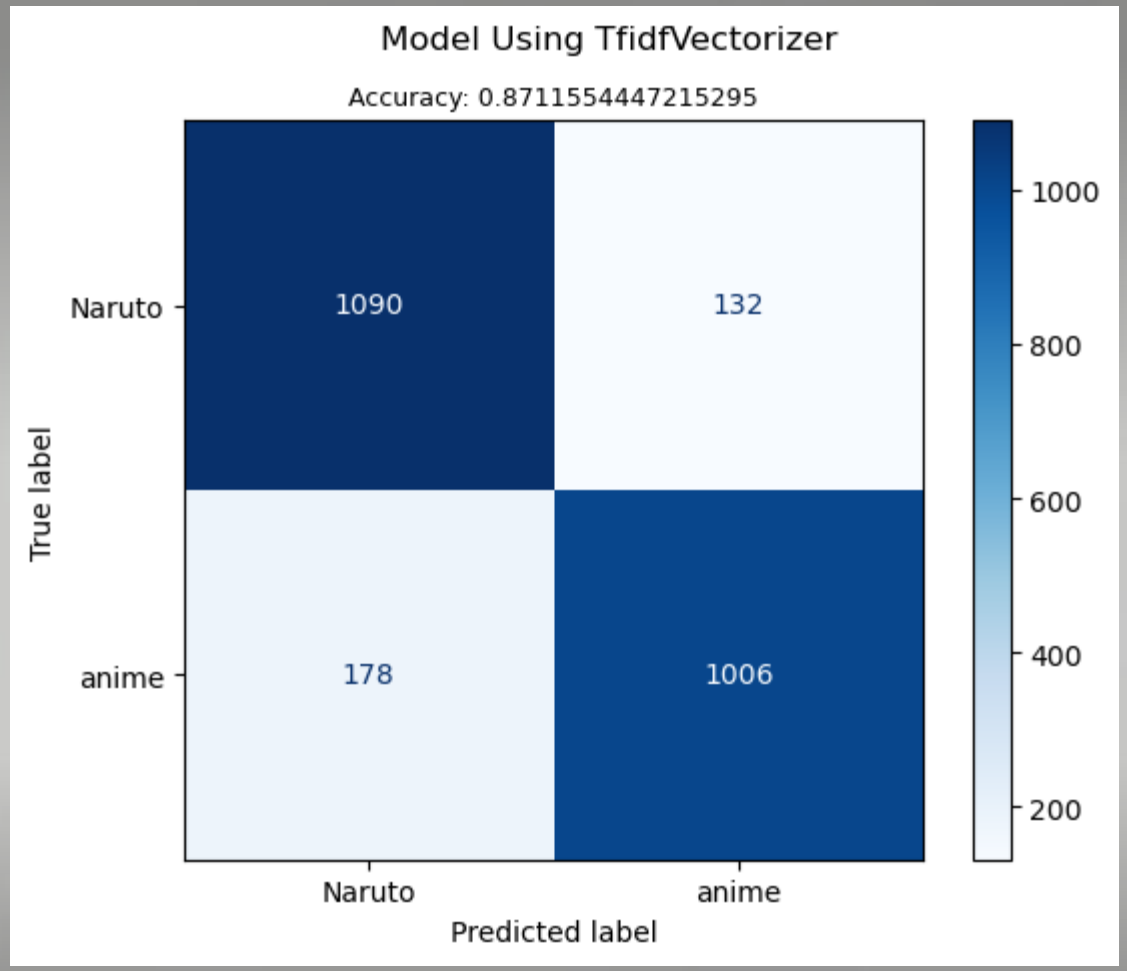
4734 posts from r/Naruto (49.2%)

Baseline prediction accuracy: 50.8%

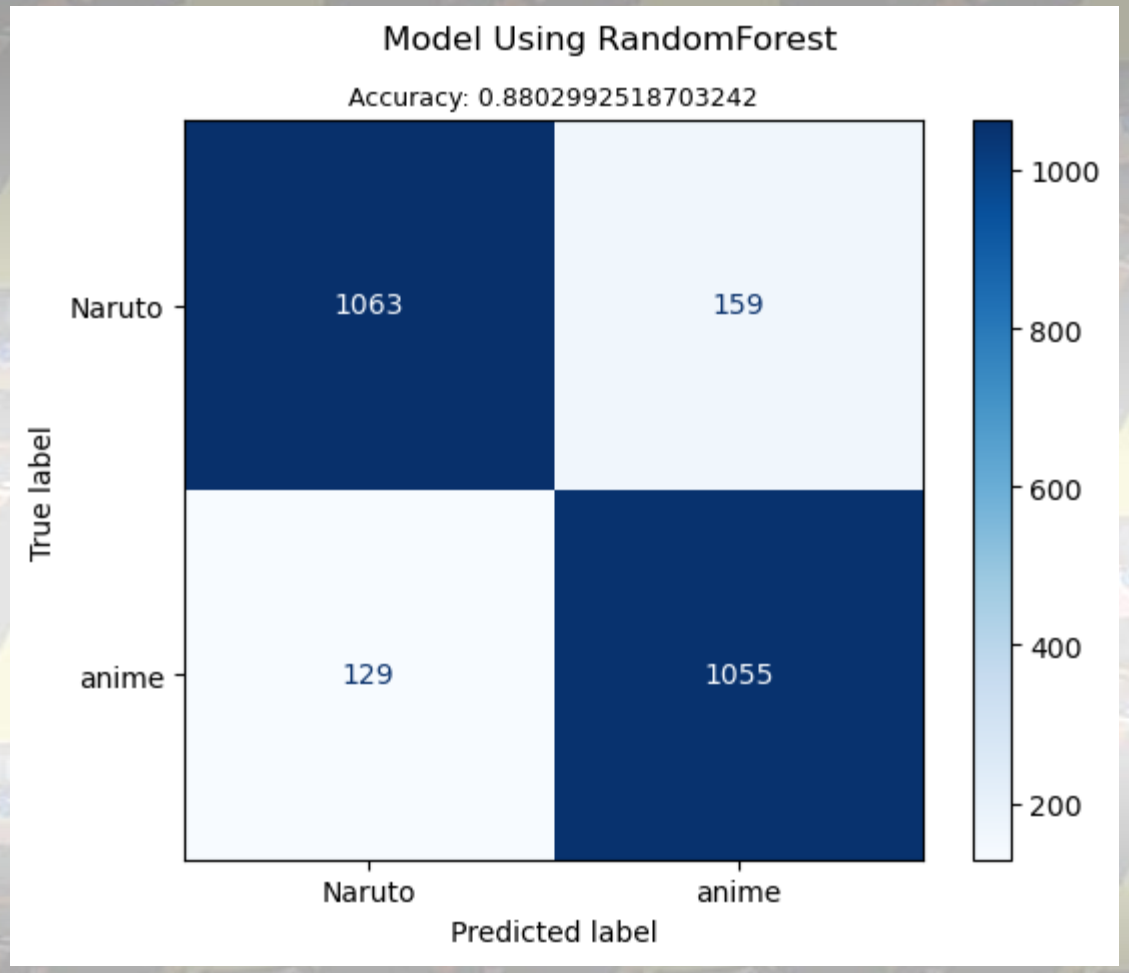
Naïve Bayes (w/ Standard CountVectorizer)



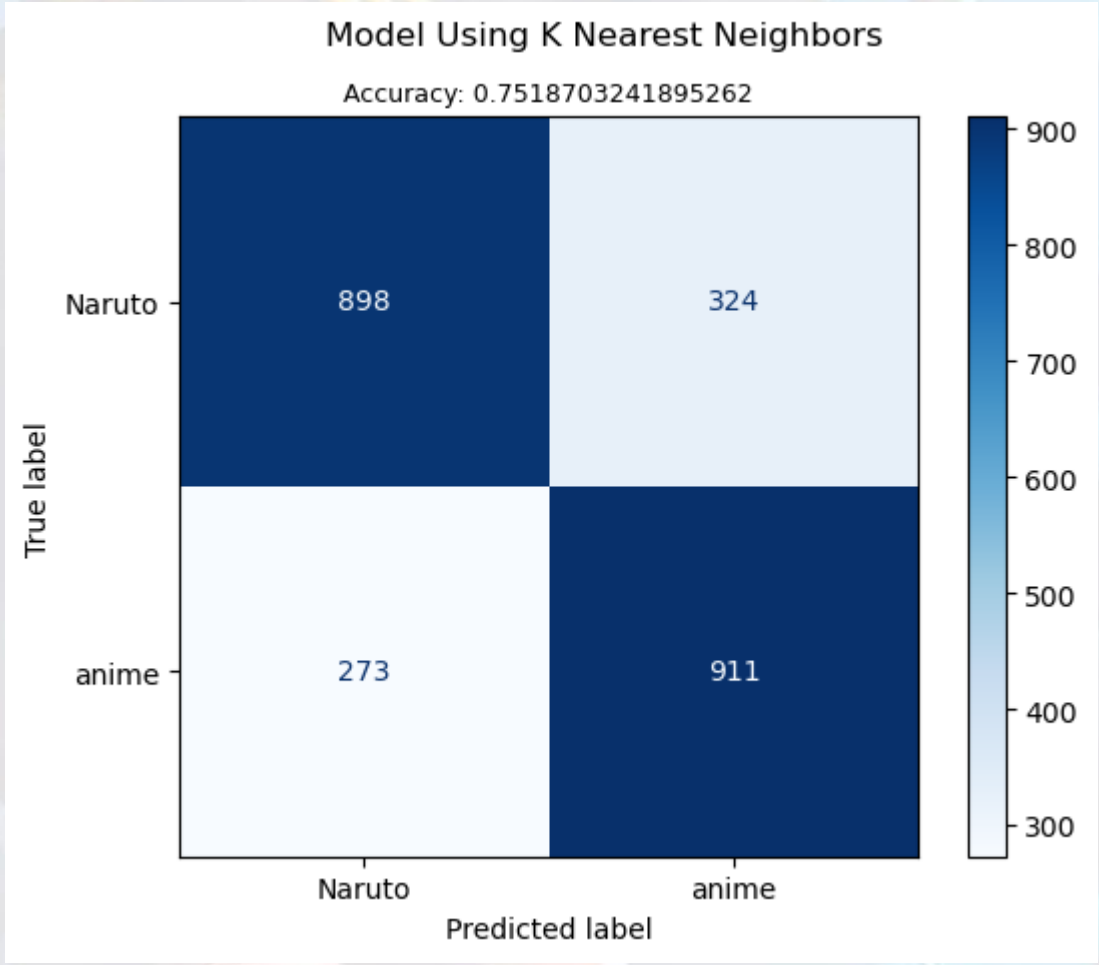
NB w/ TfidfVectorizer



Random Forest



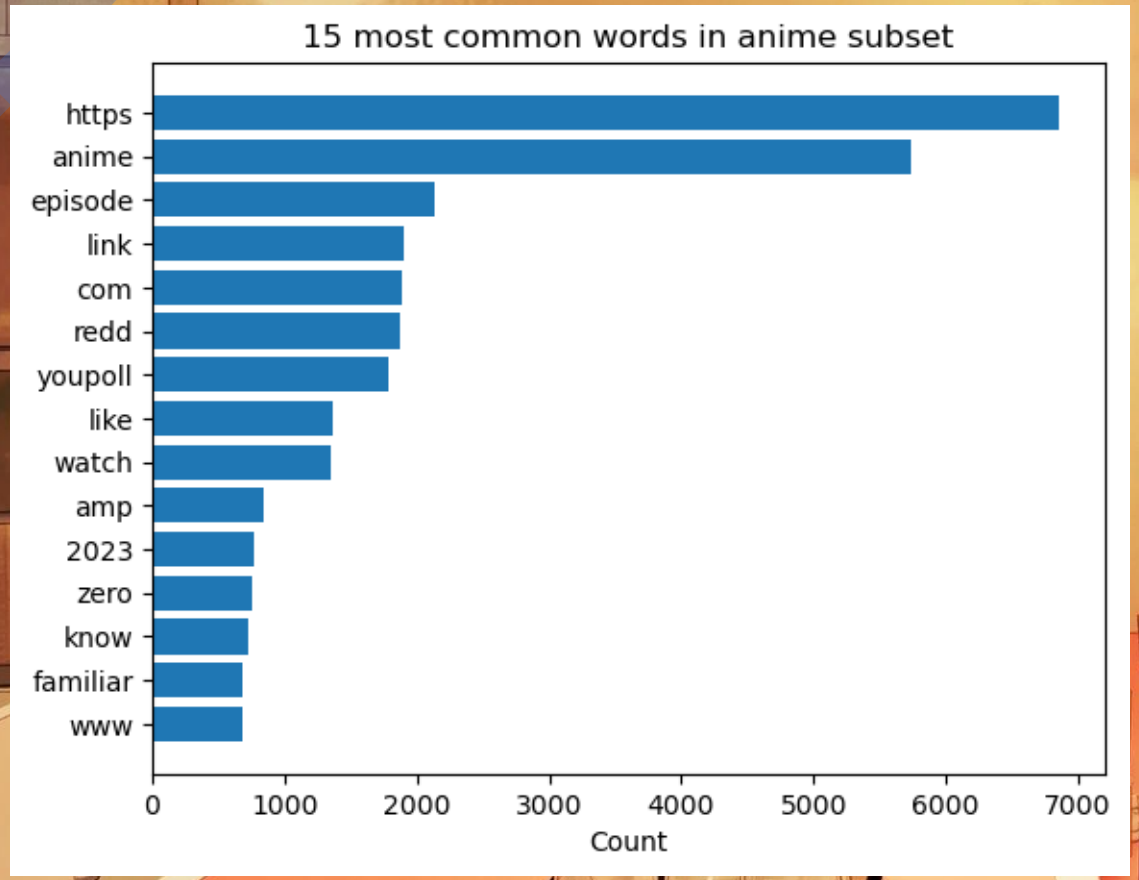
K Nearest Neighbors

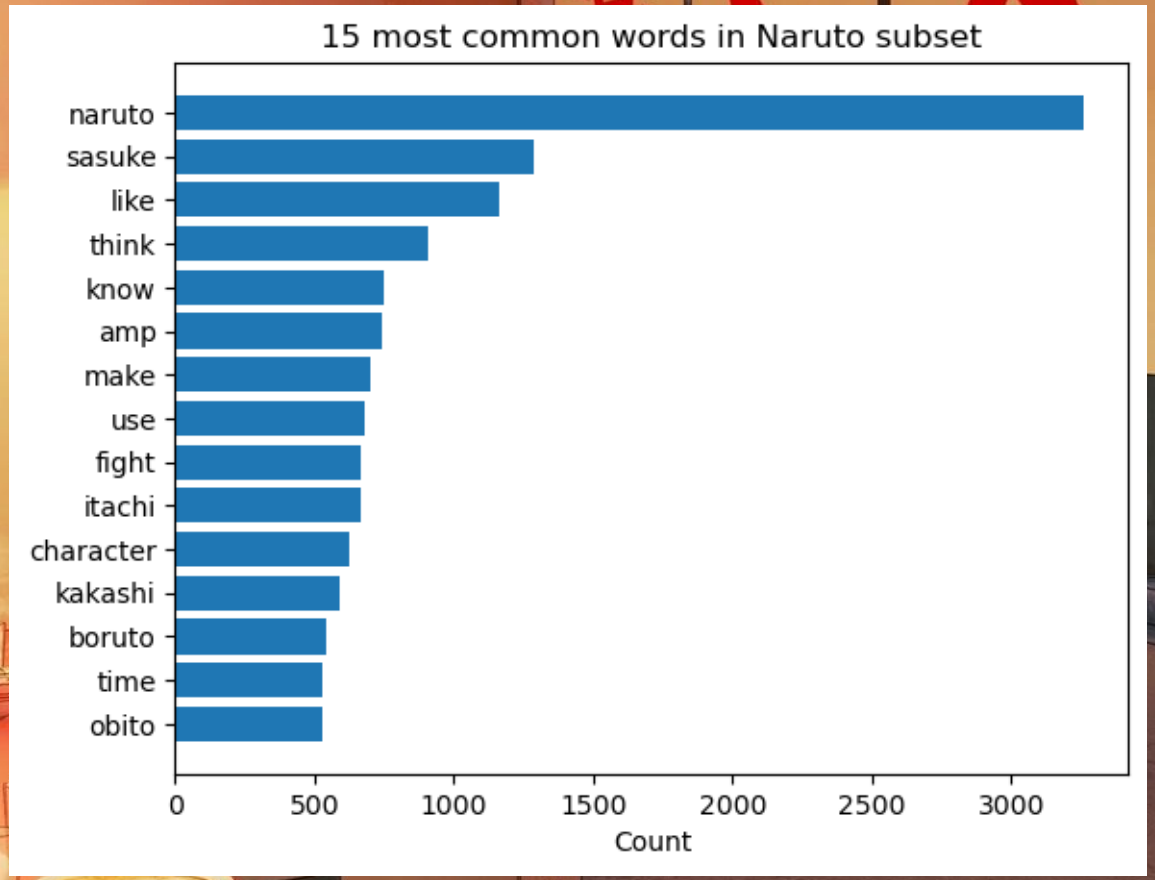


Final Rankings

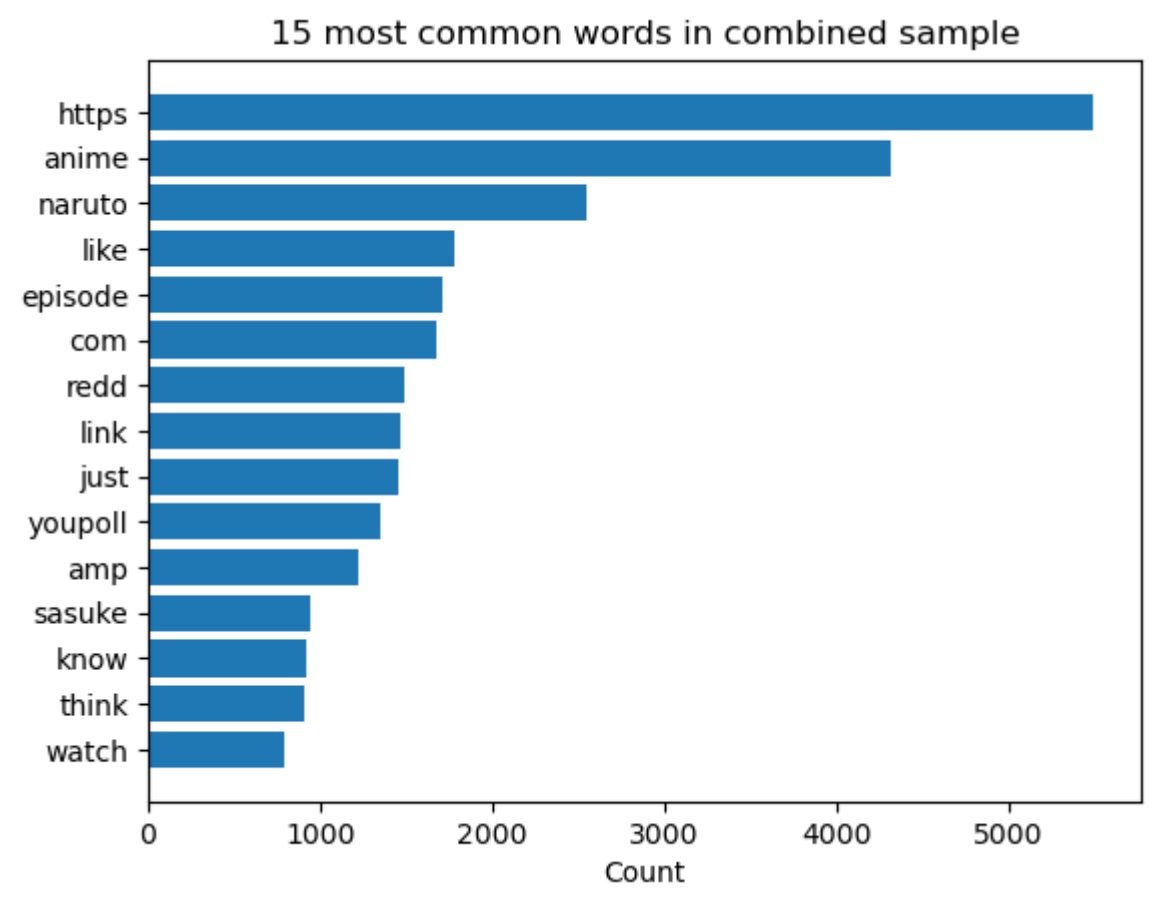
Rank	Algorithm	Accuracy
1.	Random Forest	88.03%
2.	TfidfVectorizer	87.12%
3.	Naive Bayes	86.28%
4.	K Nearest Neighbors	75.19%

Most common words

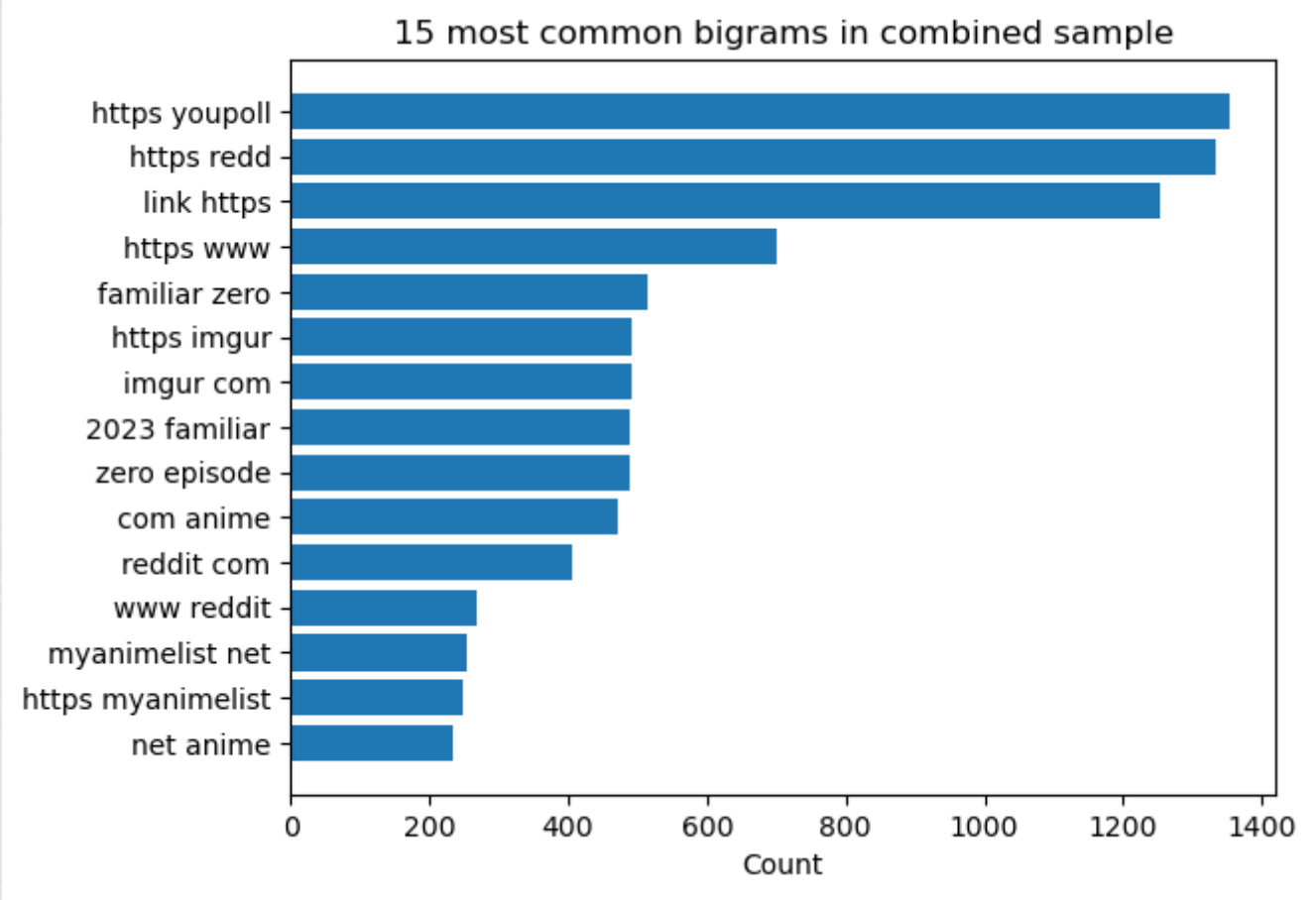




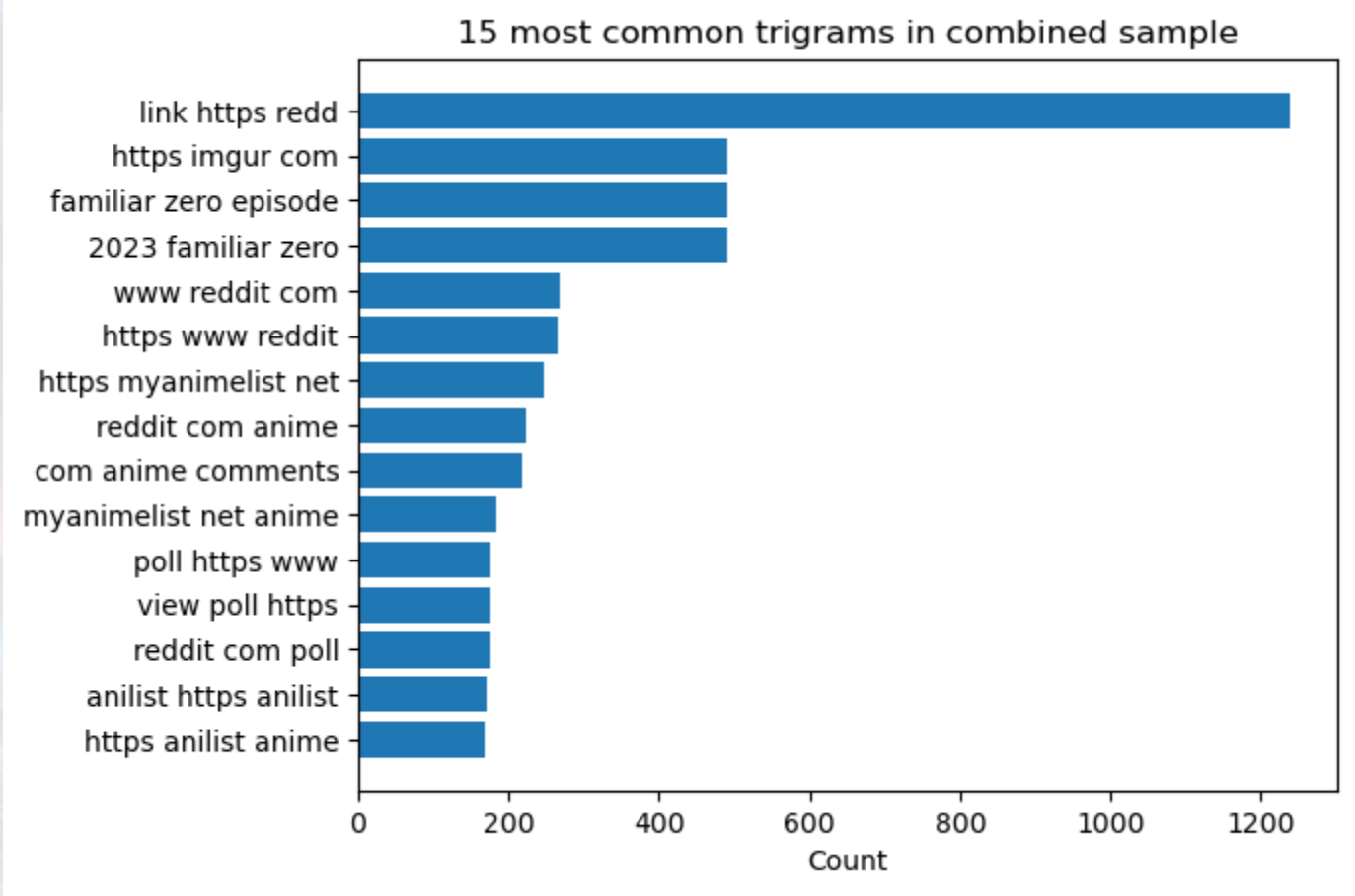
Most common words

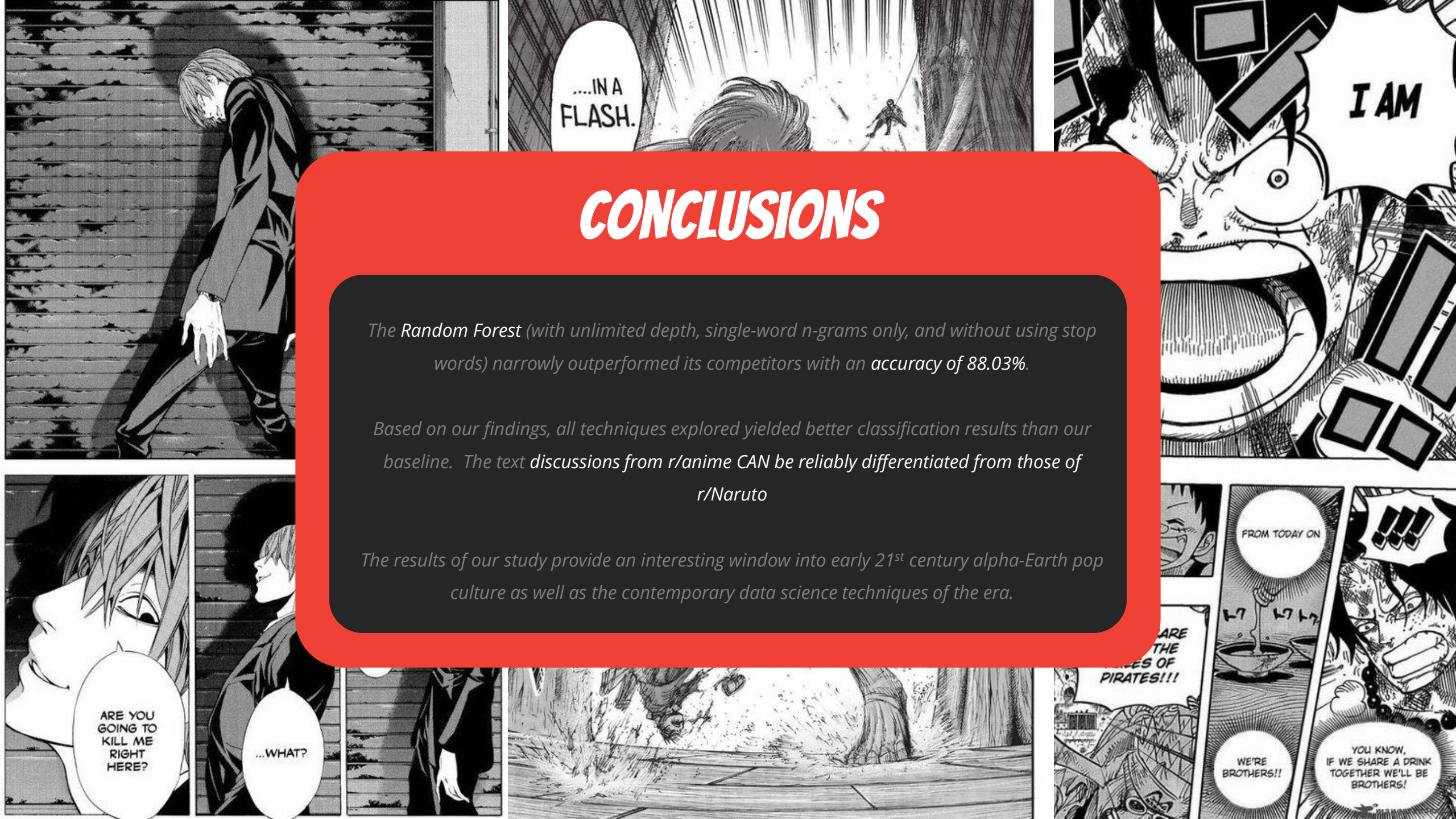


Most common bigrams



Most common trigrams





CONCLUSIONS

The **Random Forest** *(with unlimited depth, single-word n-grams only, and without using stop words) narrowly outperformed its competitors with an* **accuracy of 88.03%**.

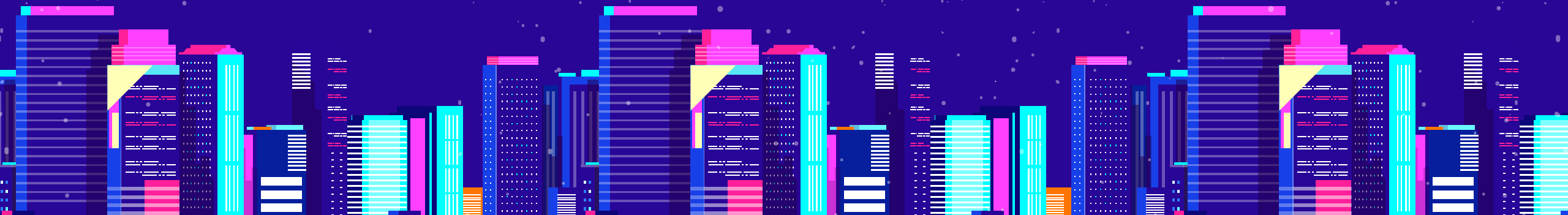
Based on our findings, all techniques explored yielded better classification results than our baseline. The text **discussions from r/anime CAN be reliably differentiated from those of r/Naruto**

The results of our study provide an interesting window into early 21st century alpha-Earth pop culture as well as the contemporary data science techniques of the era.

THANK YOU!
QUESTIONS/COMMENTS?

CREDITS

Data Scraping: Pushshift Reddit API

Main Slides Template: Slide Chef

Additonal Graphics: Drevio [1]
imagemlegal.com.br [1, 2, 3]
fazendoanossafesta.com.br [1]