The background is a deep blue gradient with a subtle pattern of white dots. Overlaid on the left side are several concentric circles and arcs, some with degree markings (140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260) and arrows indicating a clockwise direction. The text is positioned on the right side of the image.

THE FUTURE, THE PAST & NATURAL LANGUAGE PROCESSING

NATE BUKOWSKI

THE PROCESS



Problem Identification



Data Collection



Modeling

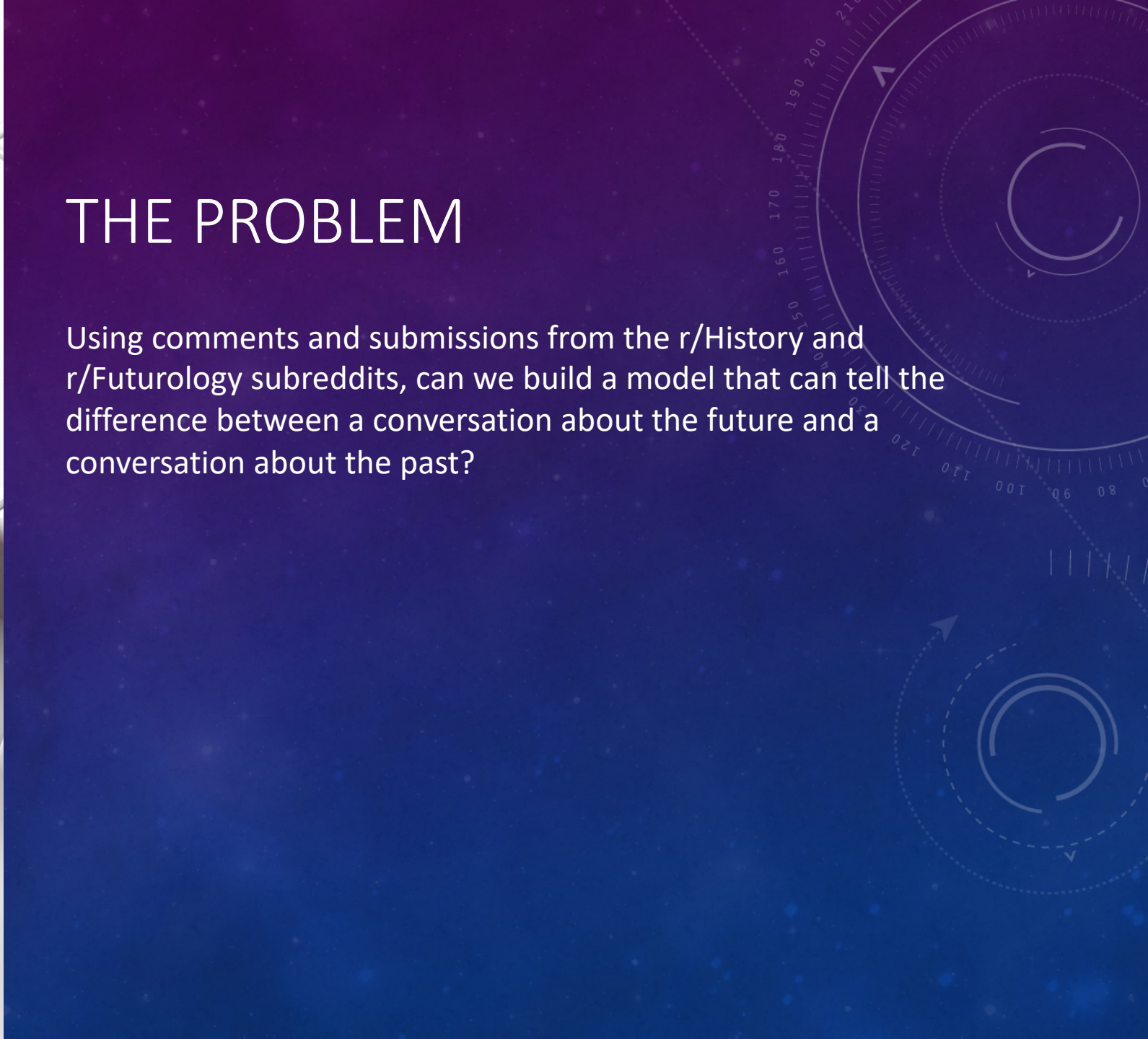


Result Interpretation



THE PROBLEM

Using comments and submissions from the r/History and r/Futurology subreddits, can we build a model that can tell the difference between a conversation about the future and a conversation about the past?



DATA COLLECTION: USING THE PUSHSHIFT API

- The `get_reddit_data` function takes in three arguments:
 - `subreddit`: The subreddit from which the data is to be scraped. We will be using the 'history' and 'Futurology' subreddits.
 - `endpoint`: The type of data to scrape; either 'comment' or 'submission'.
 - `n_iter`: The number of times the API will run. Because we are limited to 1,000 posts per scrape, `n_iter` allows us to scrape $n_iter * 1,000$ posts at a time. We used `n_iter = 10` for each endpoint & subreddit.

THE DATA

Using Pushshift's API, 10,000 comments and 10,000 submissions were collected from each subreddit for a total of 40,000 rows of data.

r/History

- /r/History is a place for discussions about history. Feel free to submit interesting articles, tell us about this cool book you just read, or start a discussion about who everyone's favorite figure of minor French nobility is!

reddit.com/r/history



r/Futurology

- Welcome to r/Futurology, a subreddit devoted to the field of Future(s) Studies and speculation about the development of humanity, technology, and civilization.

reddit.com/r/Futurology



Base

Logistic
Regression

Naïve Bayes

K-Nearest
Neighbors

Support
Vector
Machine

THE MODELS

The background is a deep blue gradient with a starry space texture. On the left side, there are several concentric circular patterns, some with tick marks and numbers, resembling a compass or a clock face. These patterns are rendered in a lighter blue color. The text 'BASE MODEL' is centered in the upper half of the image in a large, white, sans-serif font. Below it, the text '0.50 ACCURACY' is centered in a smaller, white, sans-serif font.

BASE MODEL

0.50 ACCURACY

LOGISTIC REGRESSION

Hyperparameters: Count Vectorizer & Tfidf-Vectorizer

- Max Feature Limit: None, 5,000, 10,000
- With & without stopwords
- 1 & 2 n-gram range

Best Count Vectorized Model

- No max feature limit
- No stopwords
- 2 n-gram range
- Training Accuracy: 0.8942
- Testing Accuracy: 0.8976

Best Tfidf-Vectorized Model

- No max feature limit
- With stopwords
- 1 gram range
- Training Accuracy: 0.8963
- Testing Accuracy: 0.8975

NAÏVE BAYES

Multinomial Hyperparameters

- Max Feature Limit: None, 5,000, 10,000
- With & without stopwords
- 1 & 2 gram range Best Count Vectorized Model
- No max feature limit

Best Multinomial Model

- No Max feature limit
- No stopwords
- 2 gram range
- Training Accuracy: 0.9081
- Testing Accuracy: 0.9081

Gaussian Model

- Training Accuracy: 0.87
- Testing Accuracy: 0.7549

K-NEAREST NEIGHBORS

Hyperparameters: Count Vectorizer & Tfidf-Vectorizer

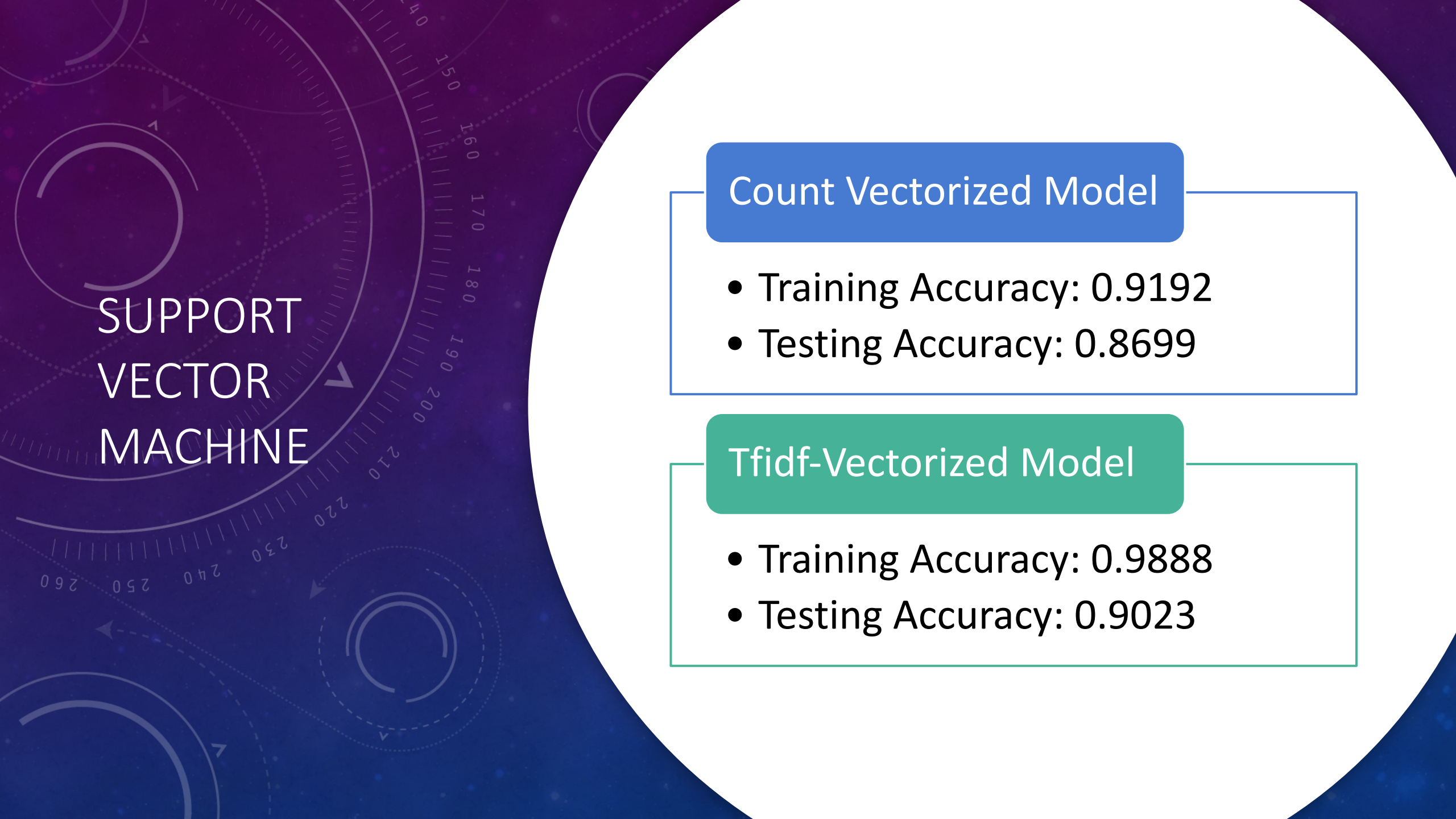
- 5, 15 & 25 nearest neighbors

Best Count Vectorized Model

- 5 Nearest Neighbors
- Training Accuracy: 0.6658
- Testing Accuracy: 0.6781

Best Tfidf-Vectorized Model

- 25 Nearest Neighbors
- Training Accuracy: 0.8141
- Testing Accuracy: 0.7569



SUPPORT VECTOR MACHINE

Count Vectorized Model

- Training Accuracy: 0.9192
- Testing Accuracy: 0.8699

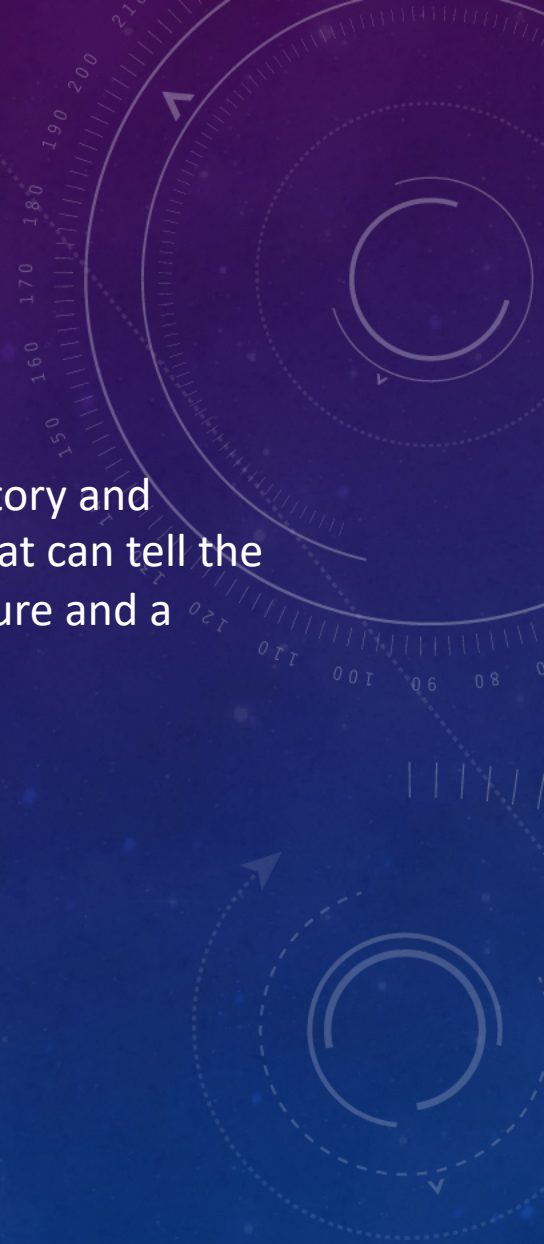
Tfidf-Vectorized Model

- Training Accuracy: 0.9888
- Testing Accuracy: 0.9023



THE PROBLEM

Using comments and submissions from the r/History and r/Futurology subreddits, can we build a model that can tell the difference between a conversation about the future and a conversation about the past?





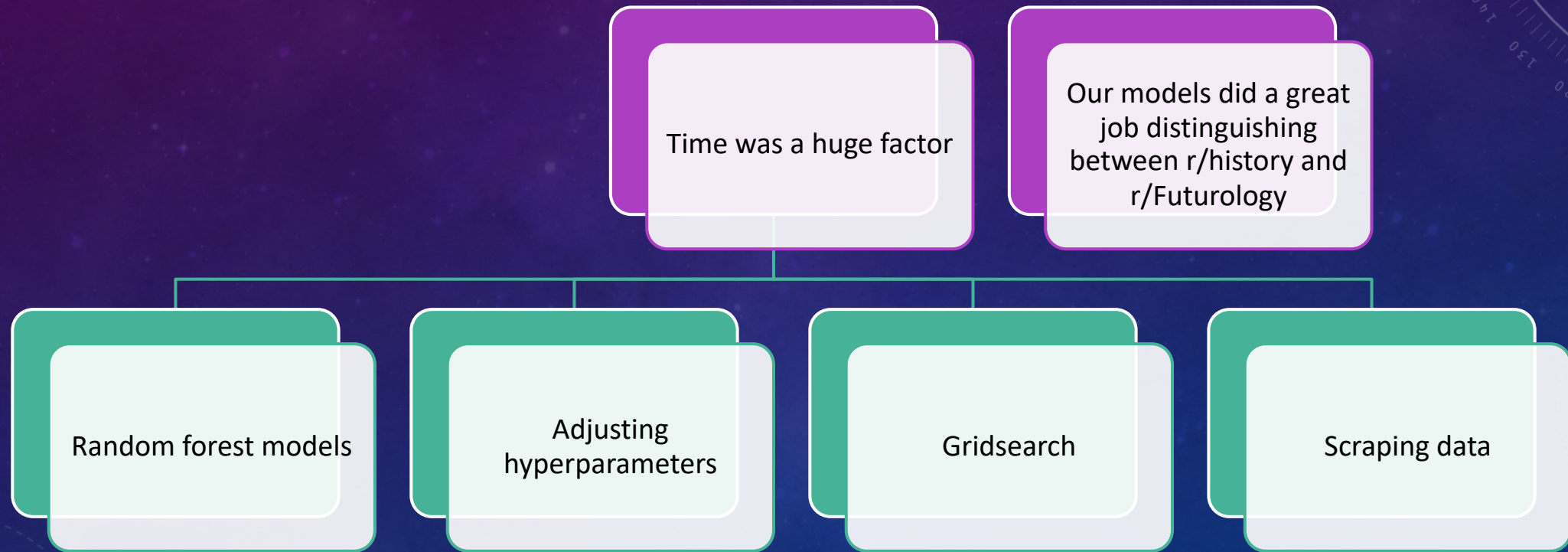
THE PROBLEM

Using comments and submissions from the r/History and r/Futurology subreddits, can we build a model that can tell the difference between a conversation about the future and a conversation about the past?

BEST MODELS

- Multinomial Naïve Bayes: 90.8% Accuracy
- Tfidf-Vectorized Support Vector Machine: 90.23% Accuracy

CONCLUSIONS



The background is a dark blue-grey gradient. On the left side, there are several faint, light blue circular patterns, some with dashed lines and arrows, resembling technical diagrams or clock faces. On the right side, there is a dense cluster of 3D question marks in a dark grey color, creating a sense of depth and complexity.

QUESTIONS?