



Project 3: **Reddit, APIs, and a Whole Lot of Modeling**

Presenter: Thomas Prich

TABLE OF CONTENTS

01

INTRODUCTION

What are we doing?

02

The Process

1. Gather Posts
2. ???
3. Profit

03

EDA

Not what I was expecting.

04

Models

The best of the best.

Introduction



Problem Statement

Can we use Natural Language Processing to determine a post's subreddit?



r/science

The positive in the model. Focuses on science and technology, most posts being links to articles.



r/todayilearned

The negative in the model. Covers a wide variety of fun facts people learn. More influenced by world events.

The Process



Gather Data

Use Pushshift's Reddit API to gather 1,500 posts for each subreddit



EDA

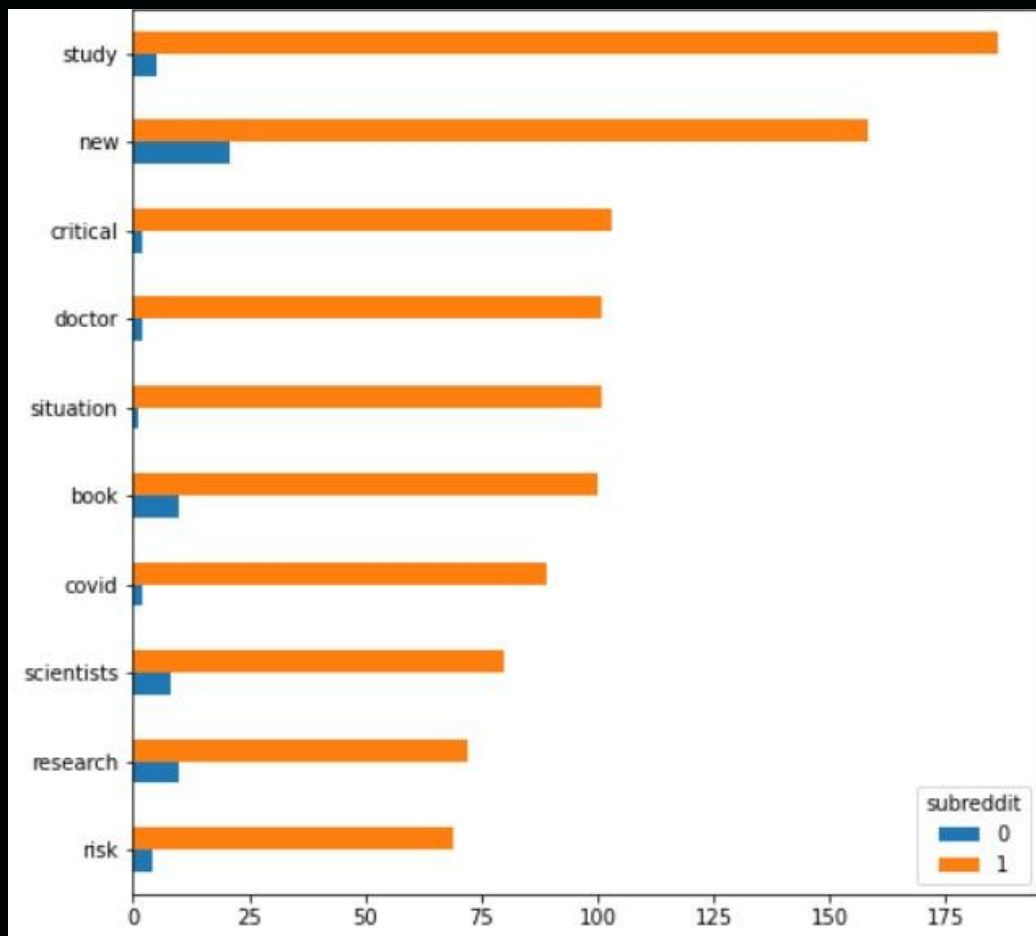
Explore the data to find the most common words and other features



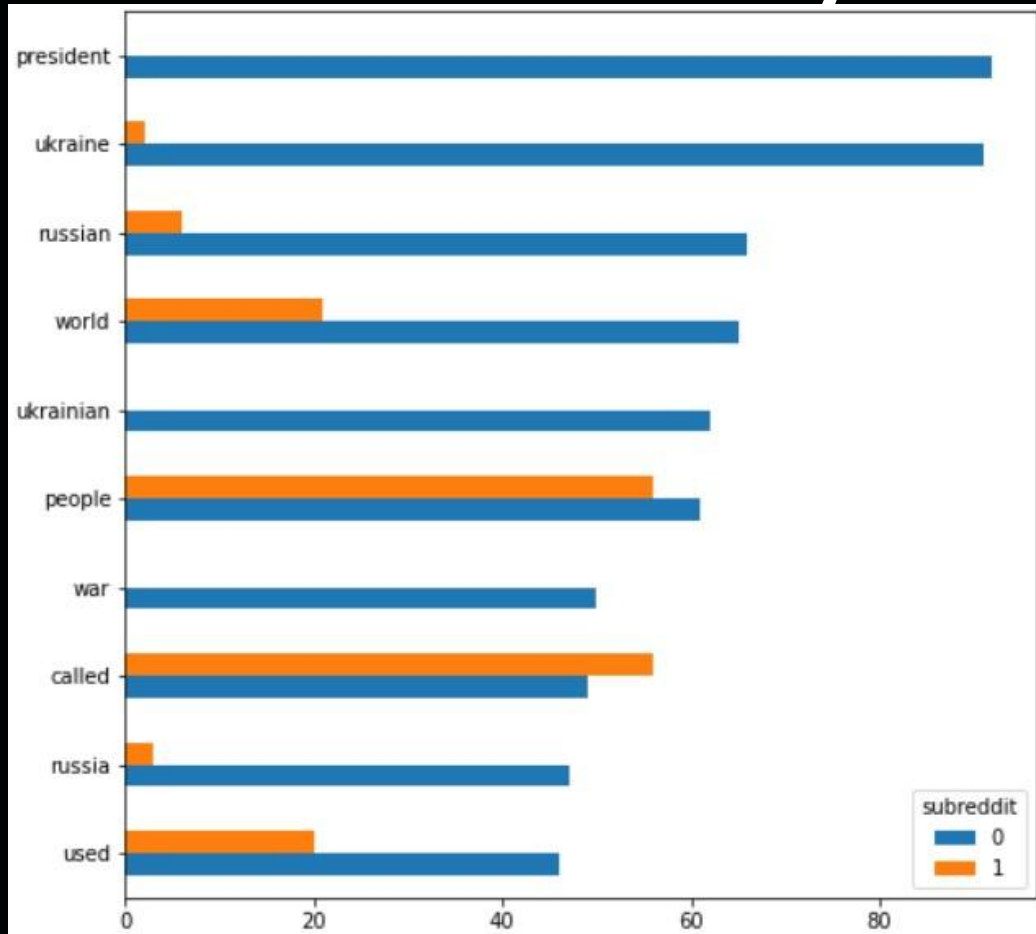
Model

Create a model to predict the subreddit of posts in the test dataset.

Most Common: r/science



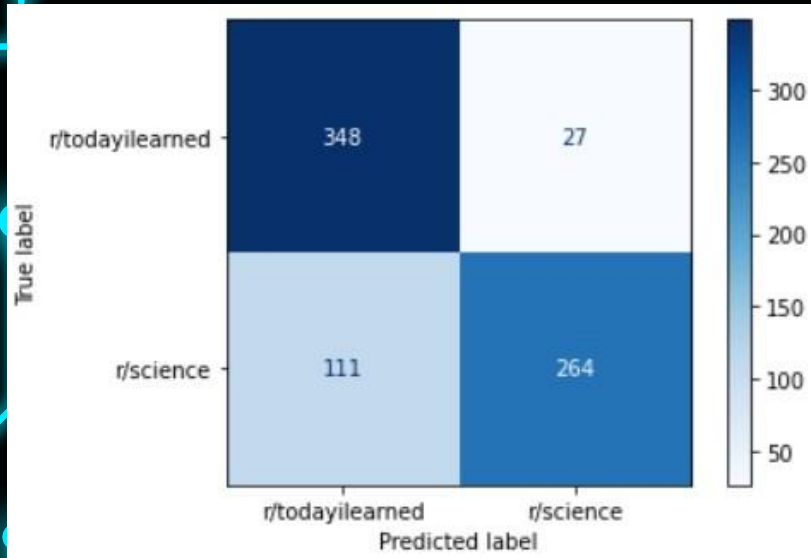
Most Common: r/todayilearned



The Models

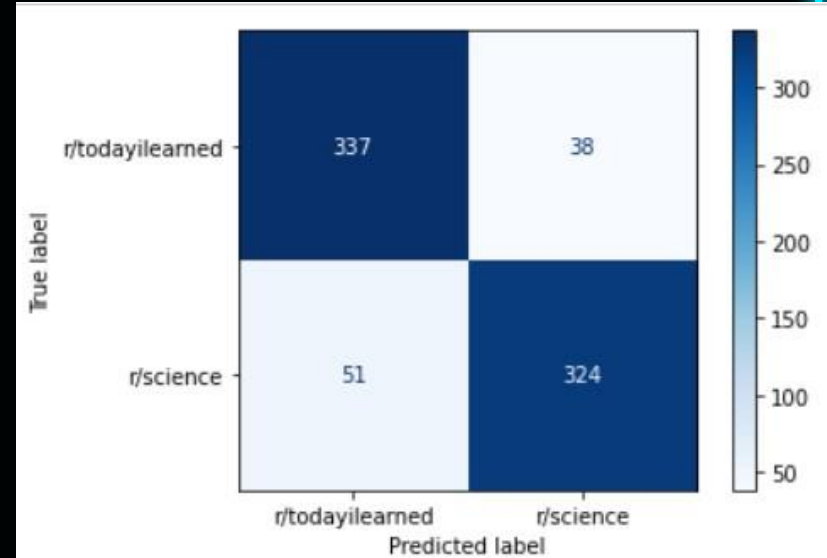
K-Nearest Neighbors

Best Accuracy: 81.6%



Logistic Regression

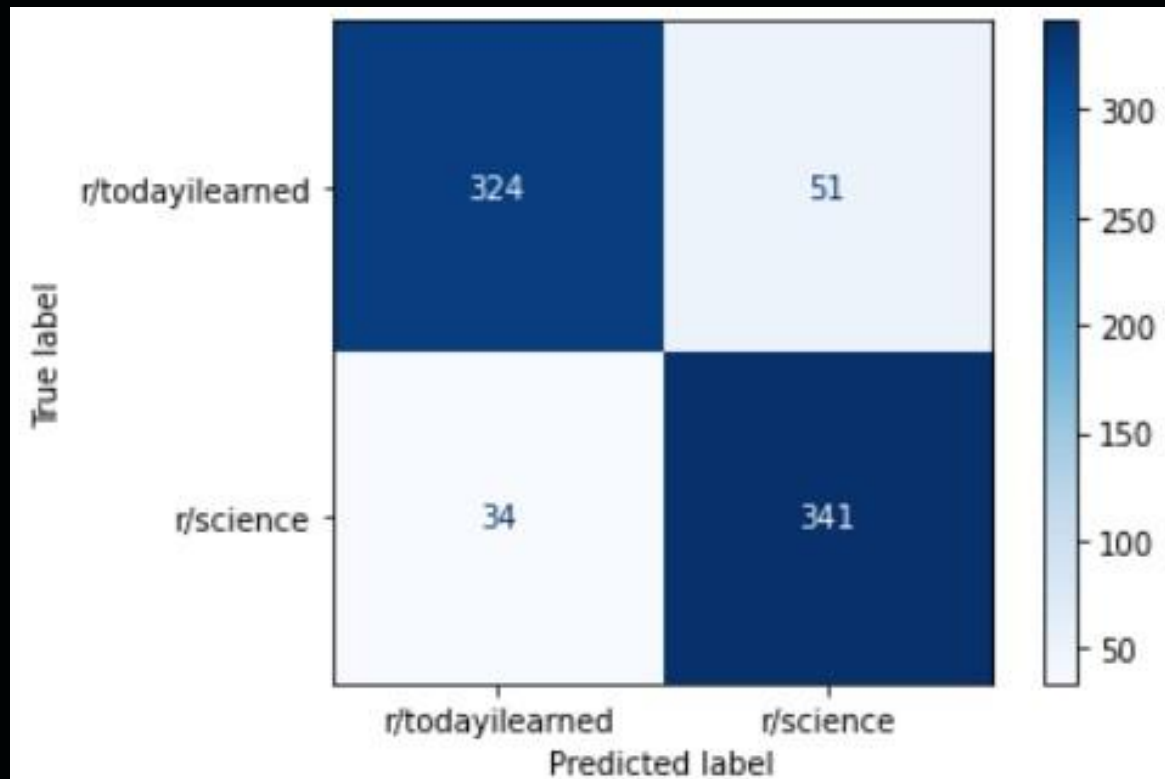
Best Accuracy: 88.1%



The Best Model

Logistic Regression (w/ word_count) with TF-IDF Vectorizer

88.67%



THANKS!



Do you have any questions?

Credits: This presentation template was created by
Slidesgo, including icons by **Flaticon**, and
infographics & images by **Freepik**