WEB API & NLP

Analyzing Reddit US &
International News Discussions

PROBLEM

THE PROBLEM

The United States has an expanding and immediate influence on international news and discussion.

Guiding Question: What topics in U.S. discourse carry the most influence on global political discussion?

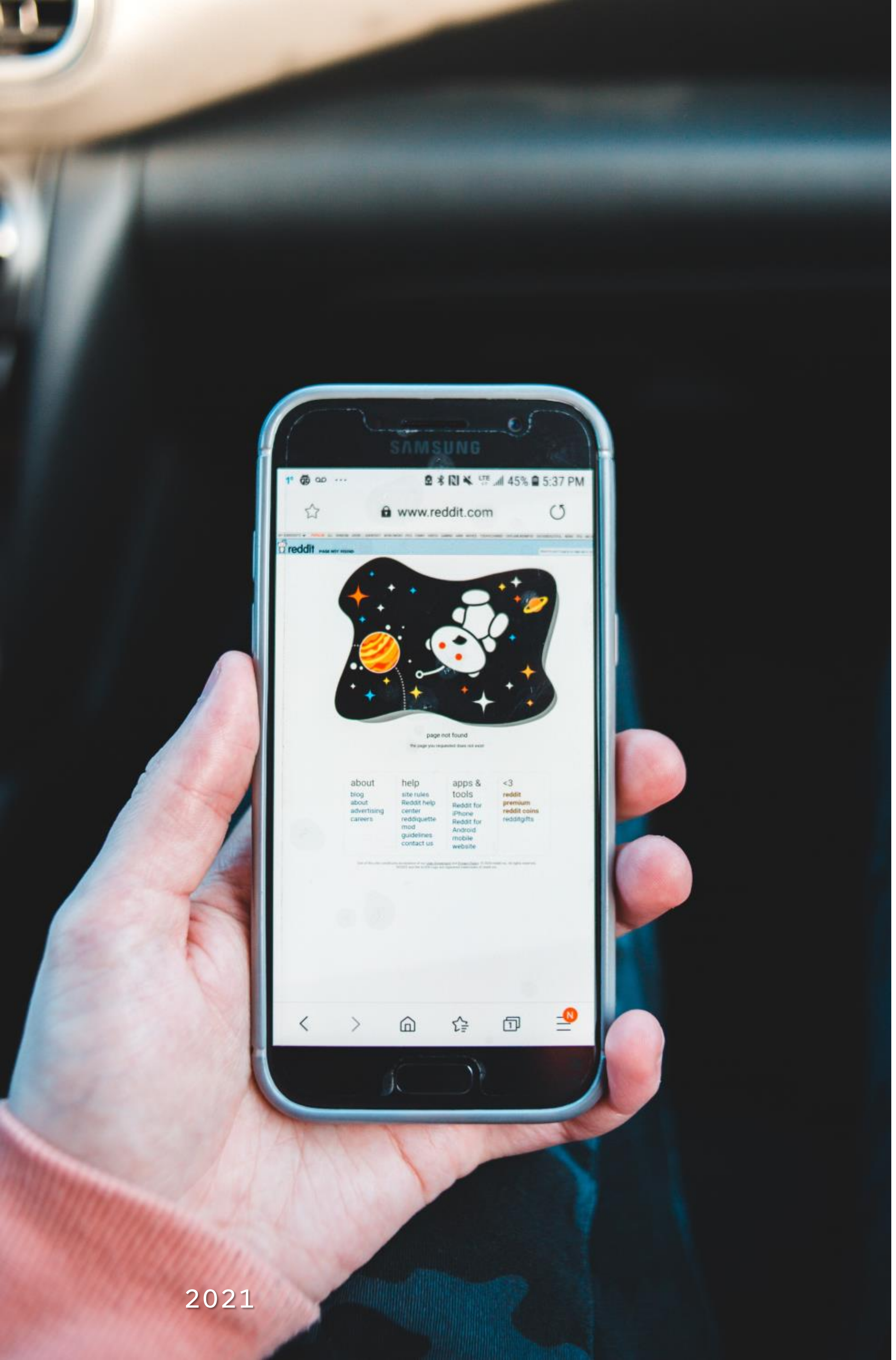
THE DATA PROBLEM

How can natural language processing (NLP) help guide our ability to understand and predict the influence of U.S. discourse on International interests?

THE DATA

r/politics: News and discussion about U.S. politics.
(7.8 million members)

r/worldnews: News and discussion about international political interests.
(27.2 million members)



REDDIT AND PUSHSHIFT API

REDDIT

An American social media company whose website/app fosters an active community of 330 millions users discussing interests across over 138,000 subreddits (communities).

PUSHSHIFT'S API

Application Programming Interface (API) developed by /r/dataset mod team to allow for intuitive access and use of data within Reddit's collection of communities

DATA SCIENCE PROCESS

DATA CLEANING & EDA

Data is appropriately acquired, cleaned and, investigated with thought given to problem statement.

PRE-PROCESSING

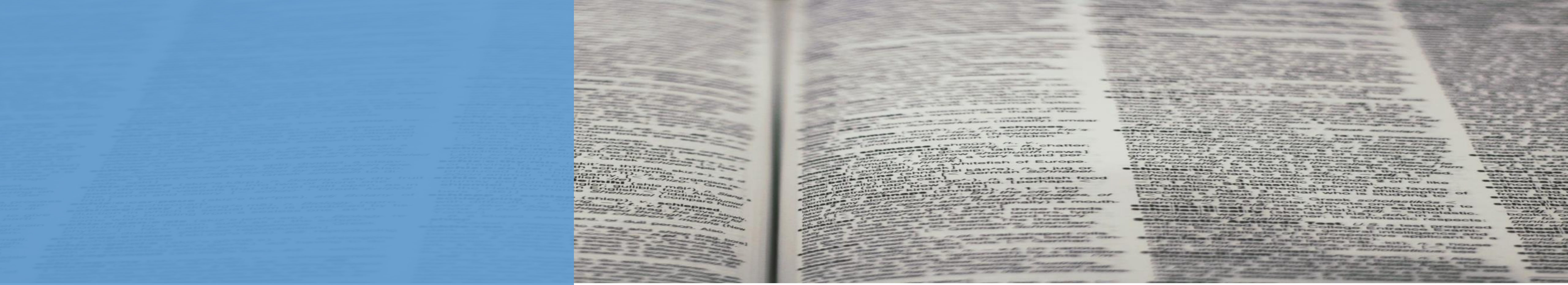
Stop words & lemmatization utilized to prepare data for successful modelling.

MODELLING

Test and evaluate a variety of models under select conditions to best identify the ideal product for solution.

EVALUATIONS & CONCLUSION

Analyze mode's measurements with thoughtful interpretation and consideration of final product.



CLASSIFICATION MODELS

VECTORIZERS:

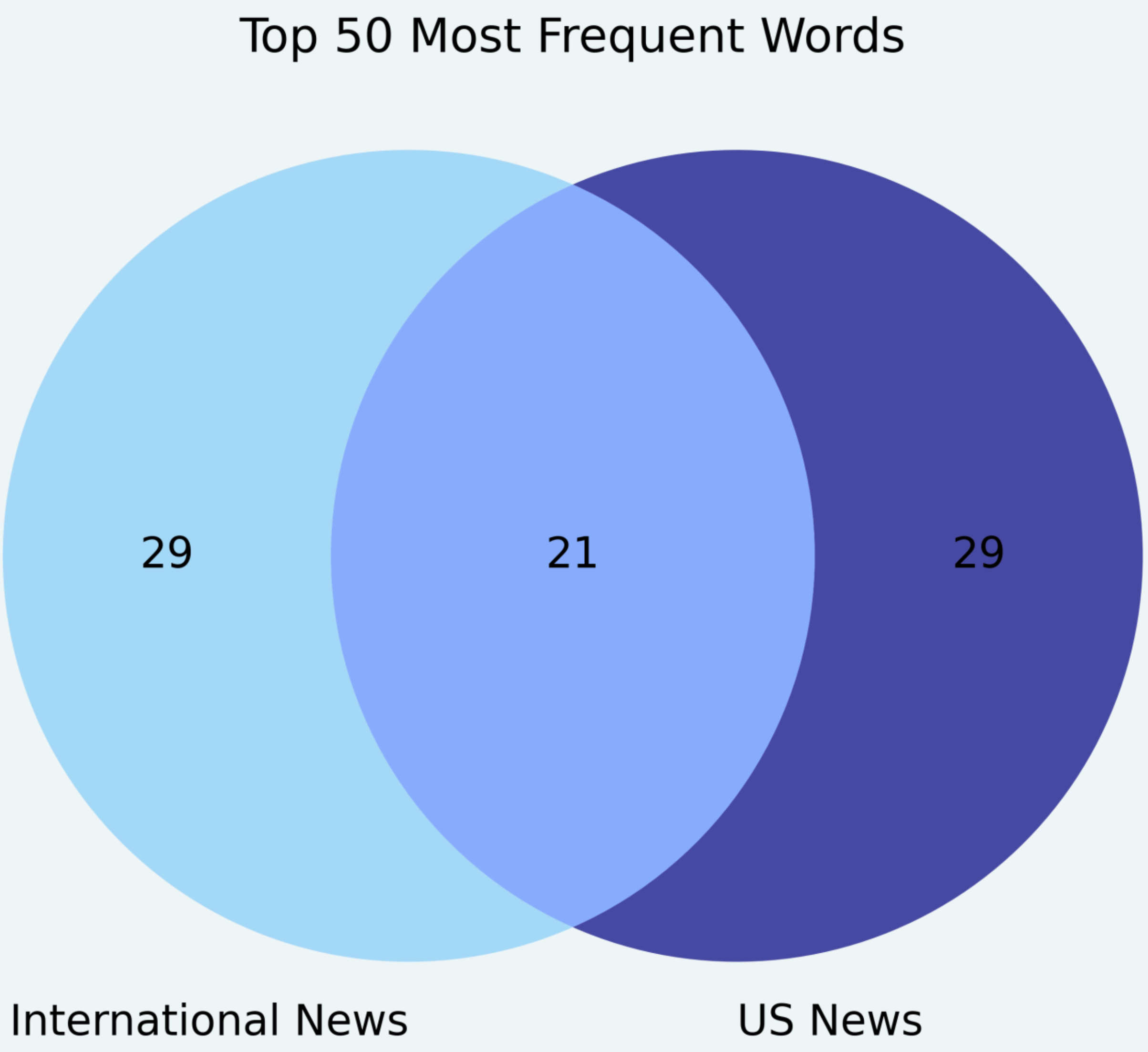
Count Vectorizer (cvec): Converts string of characters into matrix of integer values.

TF-IDF Vectorizer (tvec): Converts string of characters into matrix as float values.

PIPELINE & GRID SEARCH CV:

Pipeline: Enables easy and fluid direction of model pre-processors for application.

Grid Search CV: Evaluates and optimizes directed models by provided hyperparameters.

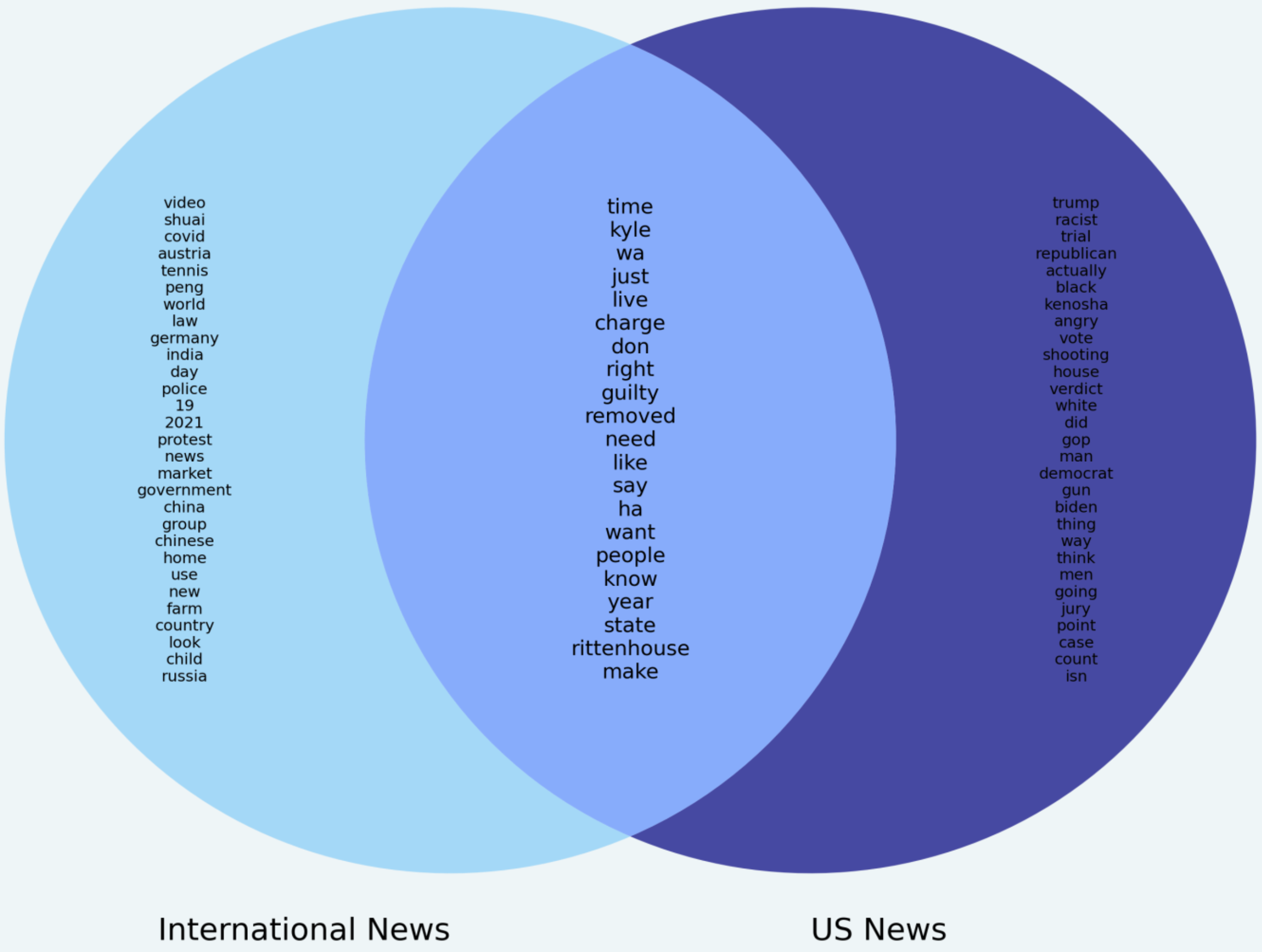


SELECTED WORDS FOR ANALYSIS

- 7,994 entries between selected US political and international discussion communities
- Top 50 most frequent words* within each discussion forum

*not including words addressed by Natural Language Tool-Kit's 'stop_words'.

Top 50 Most Frequent Words



TRENDS OF DISCUSSION

- Made evident by shared text is the influence of U.S. domestic legal proceedings.
- International text frequencies highlight interest in injustice and exterior political influence.
- U.S. text frequencies highlight interest in domestic interpersonal matters.

MODELS

KNN

K-Nearest Neighbor

Utilizes ‘feature similarity’ to predict the values of new datapoints.

LOGREG

Logistic Regression

Regression classification model that evaluates variety of variables to produce predicted binary outcome.

DT

Decision Tree

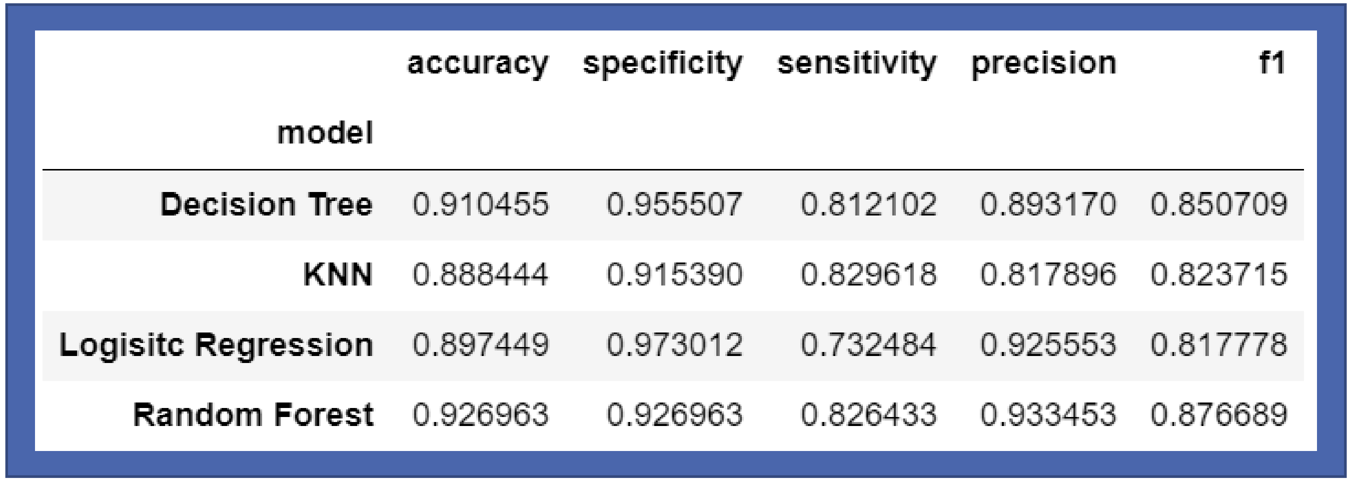
Evaluates and splits data in binary nodes to produce predictions

RF

Random Forest

Produces and combines multiple decision trees by random subspace method.

CHOICE PRODUCTION MODEL



	accuracy	specificity	sensitivity	precision	f1
model					
Decision Tree	0.910455	0.955507	0.812102	0.893170	0.850709
KNN	0.888444	0.915390	0.829618	0.817896	0.823715
Logisitc Regression	0.897449	0.973012	0.732484	0.925553	0.817778
Random Forest	0.926963	0.926963	0.826433	0.933453	0.876689

CLASSIFICATION METRICS

- In relation to the problem statement, accuracy is the primary metric of analysis.
- Specificity = predictions of U.S. discussions text
- Sensitivity = predictions of international discussion text.

RANDOM FOREST CLASSIFIER

Of the supervised classification models that were studied, Random Forest produced the most optimal metrics.

Random Forest algorithms are consistently accurate in relation to other non-linear classifiers.

Slower processing time does not challenge ability to address the problem statement

GOING FORWARD

CONCLUSION

Of all political topics, U.S. domestic instability most directly influences international discussion, and behavior.

The U.S. has a formidable online presence that further stimulates this evolving trend.

RECOMMENDATIONS

Expanding this analysis by means of data collection can help identify, predict, and study the U.S. online presence within the world wide web.

THANK YOU

William Englehart

DSI 1011

Project 3

November 22, 2021