# Science
# or
# Fake Facts?

Classification Model for Subreddit Posts

# Problem Statement

To build a classification model that predicts which subreddit: Fake Facts or Science, a post came from

Fake Facts

r/FakeFacts

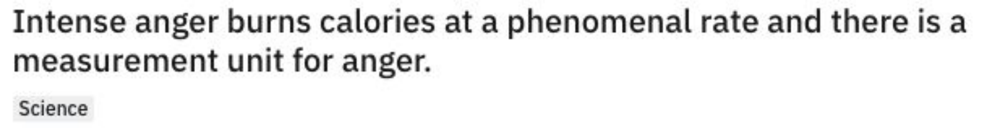
Reddit Science

r/science

# Reddit

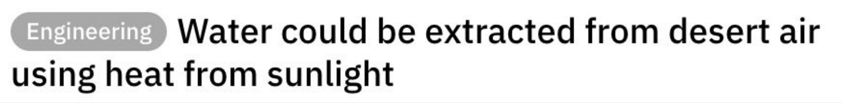
- Reddit is the 6th most visited site in the United States and the 7th most visited website in the world.
- Reddit has over 1.2 million different subreddits.
- Reddit has spent just $500 on ads in all its existence.

FAKE



Intense anger burns calories at a phenomenal rate and there is a measurement unit for anger.

SCIENCE



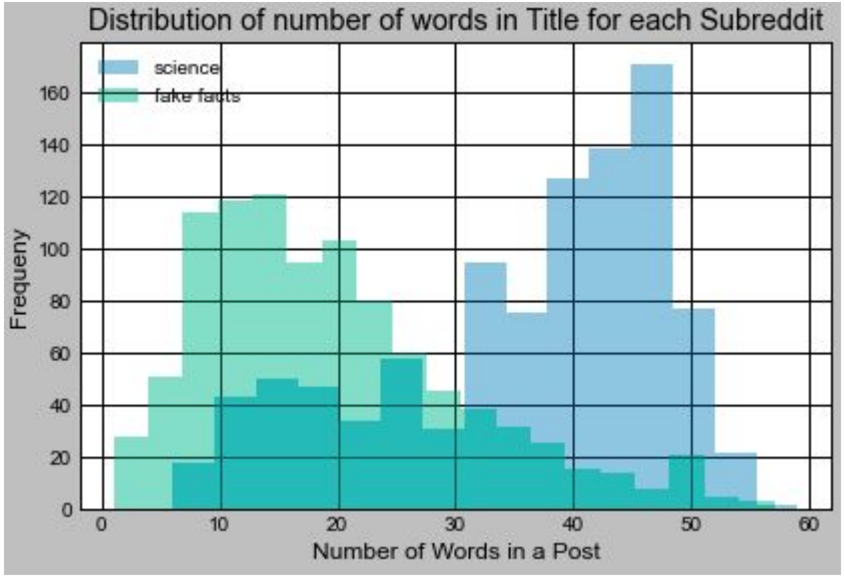
Water could be extracted from desert air using heat from sunlight

Source: https://websitebuilder.org/blog/reddit-statistics/>

# Data Collection

- Retrieved data from two subreddits
- Data collected for each post: Title, User ID, URL, Number of comments, Date Created, and Body Text.
- Used PRAW, Python Reddit API Wrapper, to collect this data from Reddit, transfer it to a Python format
- Result: 1971 “top” posts collected from 2017-2020
    - Science: 990,  Fake Facts: 981

# EDA Findings



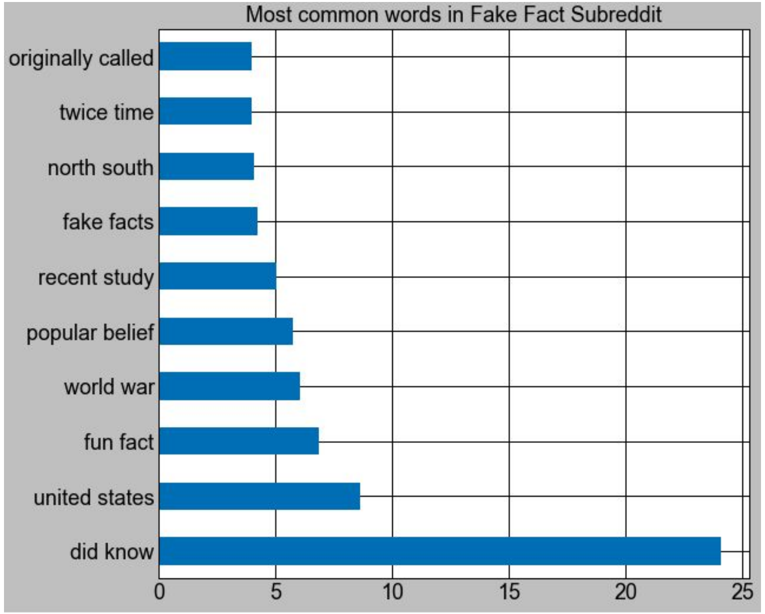
| Subreddit | Number of Comments | Score |
|---|---|---|
| Fake Facts | 3.5 | 62.4 |
| Science | 1983 | 50589.7 |

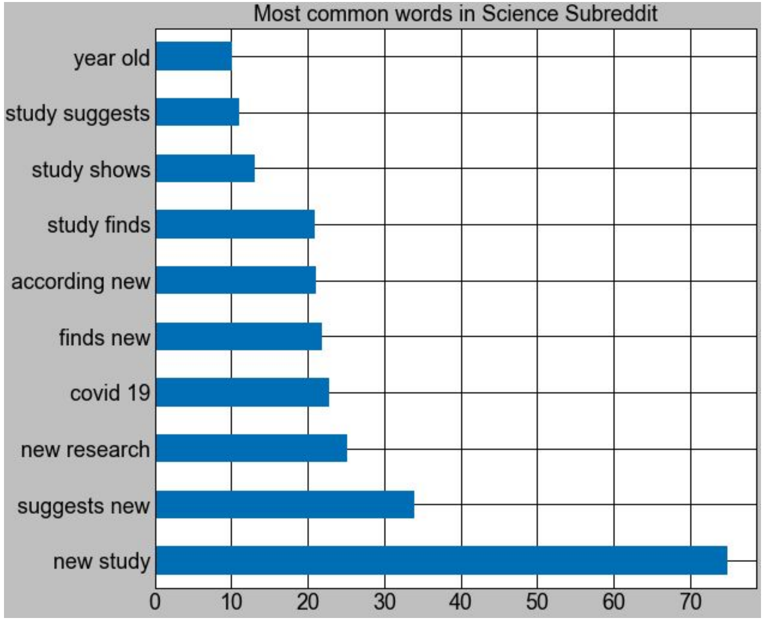
Left: most posts in the subreddit Fake Facts have fewer words than  the number of words in the subreddit Science
Top: the subreddit Science has higher scores and more comments than the subreddit Fake Facts.

# Feature Extraction - NLP

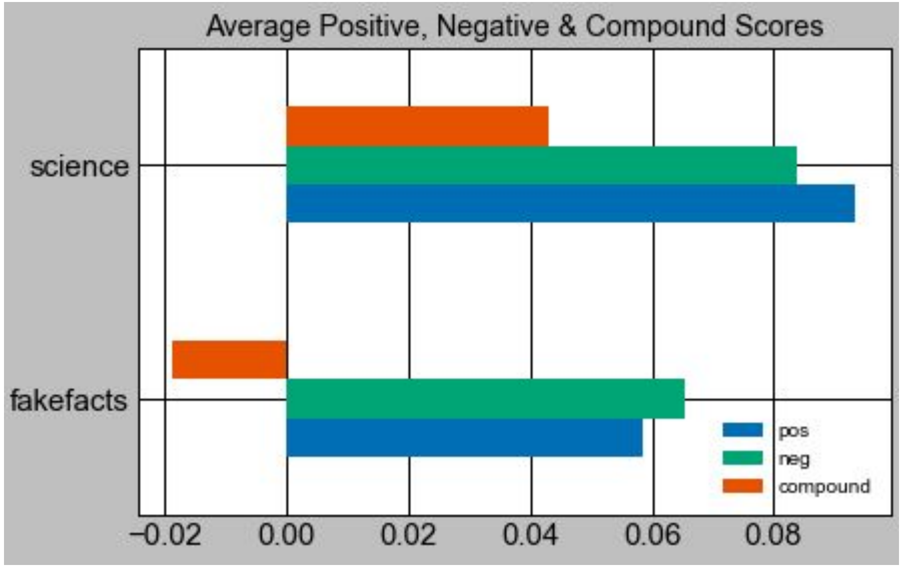
- Used Natural Language Processing to create features from the words in the Title of a subreddit.
- Transformed text data into numeric values using a vectorizer.
- Compared the results from CountVectorizer and Term frequency-inverse document frequency.
- Tuned in the hyperparameters for the vectorizer using gridsearch

# Most common words



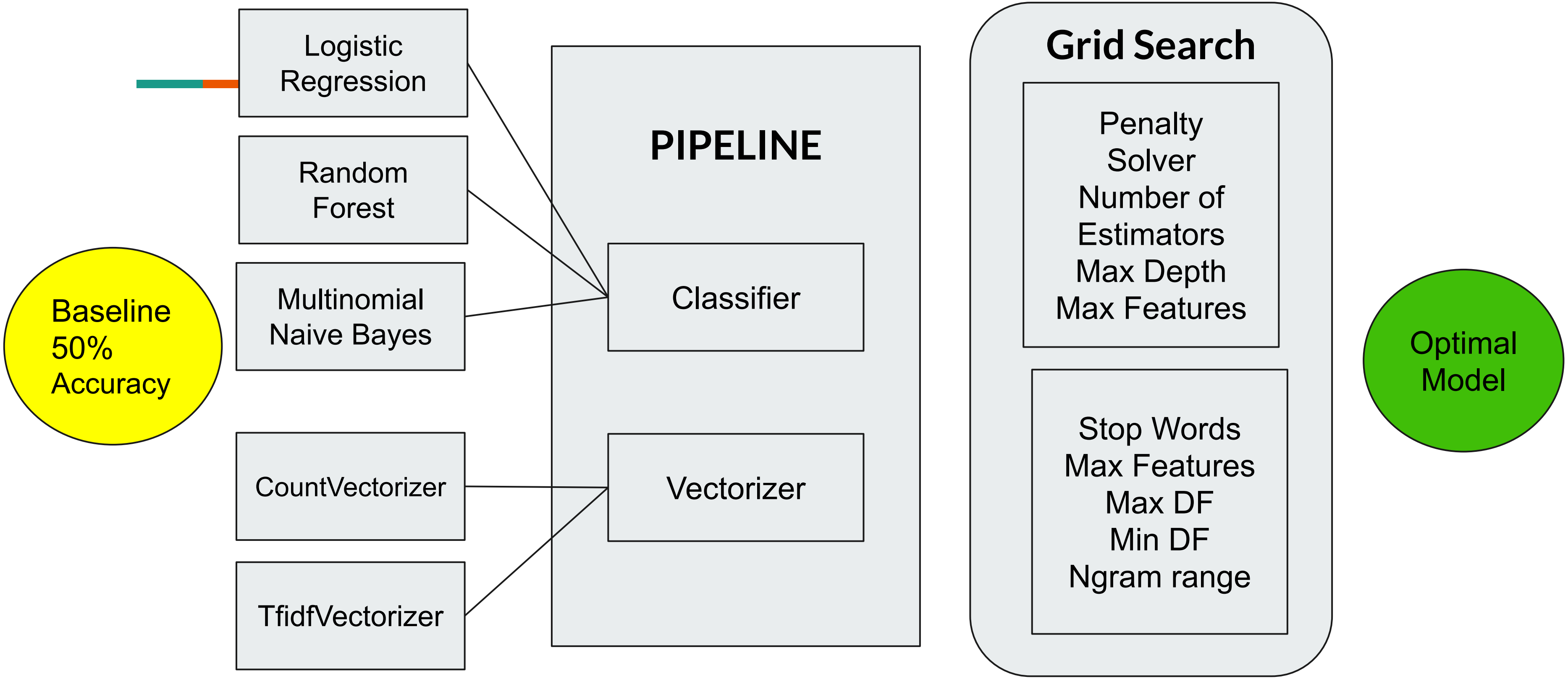


# Sentiment Analysis using VADER



As seen by the orange bar, the text in Fake Facts expresses a negative opinion as opposed to the text in Science which has a positive opinion

# Classification Modeling



Best Estimators: TfidfVectorizer: max_features=1000, ngram_range=(1, 2),use_idf=False)), RandomForestClassifier(n_estimators=90))

# Model Evaluation

| Model | Accuracy | Sensitivity | Specificity | Precision |
|---|---|---|---|---|
| Logistic Regression | 0.90 | 0.87 | 0.93 | 0.93 |
| Multinomial Naive Bayes | 0.92 | 0.94 | 0.90 | 0.91 |
| Random Forest | 0.92 | 0.88 | 0.96 | 0.96 |

Precision: can I trust my model?

# Conclusion and Recommendations

- A Random Forest classification model was built with 96% precision to predict if based on the words in a subreddit title, a post comes from Fake Facts or Science
- Evaluate further the pros and cons of using a Random Forest vs a Naive Bayes Classifier
- To try this model for subreddits in other languages.
- To validate this model over time and evaluate its accuracy with new posts.

# Sources

1. 109 Ridiculous Reddit Statistics & Facts to Know in 2020 <https://websitebuilder.org/blog/reddit-statistics/>
2. What is an API Wrapper <https://rapidapi.com/blog/api-glossary/api-wrapper/>