ASKING or ANSWERING? NLP MODELING PROBLEM

Kathy Simon | November 5, 2021

TABLE OF CONTENTS

Problem Statement

Are you answering a question or responding to a statement stating information? The problem is to determine whether a written statement is answering a question or in response to a statement. I will create a model to determine whether a written statement is answering a question or commenting on the information provided.

- Collection of comments from Reddit.com
- Classification Model of Natural Language Processing
- Metrics of Success based on accuracy and f1-score.

ABOUT THE DIFFERENT SUBREDDITS

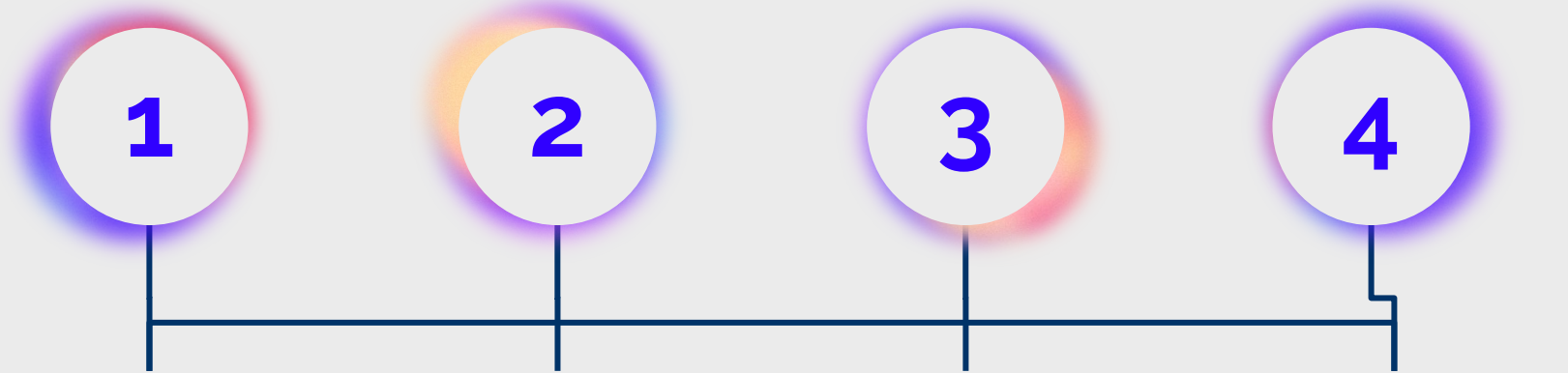
r/askscience

- Submissions ask a question about science.
- 21.4 million members.
- Collected comments
- Highly monitored

r/todayilearned

- Submissions state a fact
- 26.3 million subscribers.
- Collected comments
- Not highly monitored

MODELING PROCESS



DATA COLLECTION

Scrap data using Pushshift API

EDA

Remove duplicates, clean data, explore the data

MODELING

Consider different models, parameters and hyperparments

CONCLUSIONS

and future recommendations

DATA FACTS

r/askscience

Number of unique cleaned comments

r/todayilearned

Number of unique cleaned comments

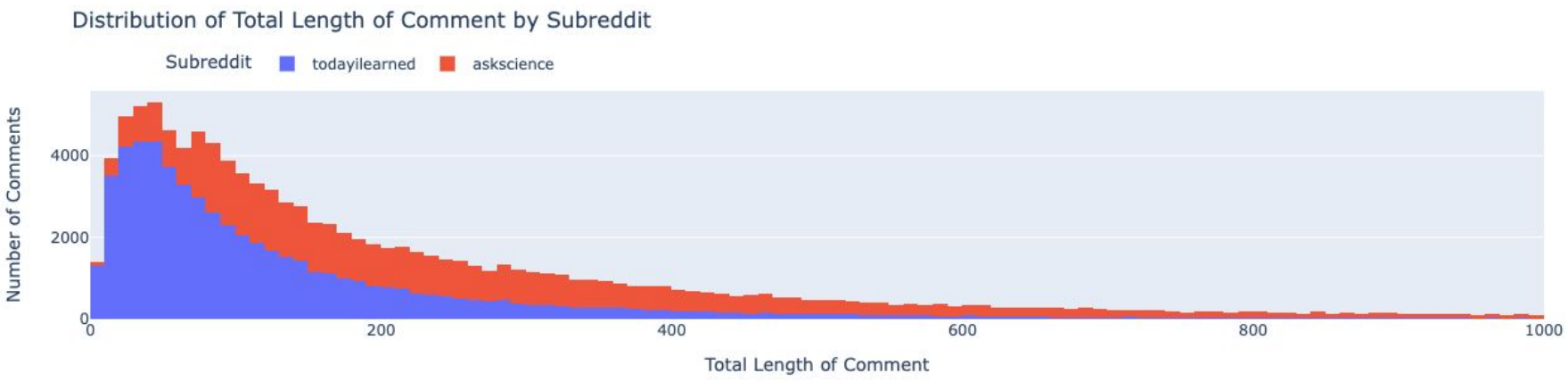
Sentiment Intensity Analyzer

Average Neutral Score

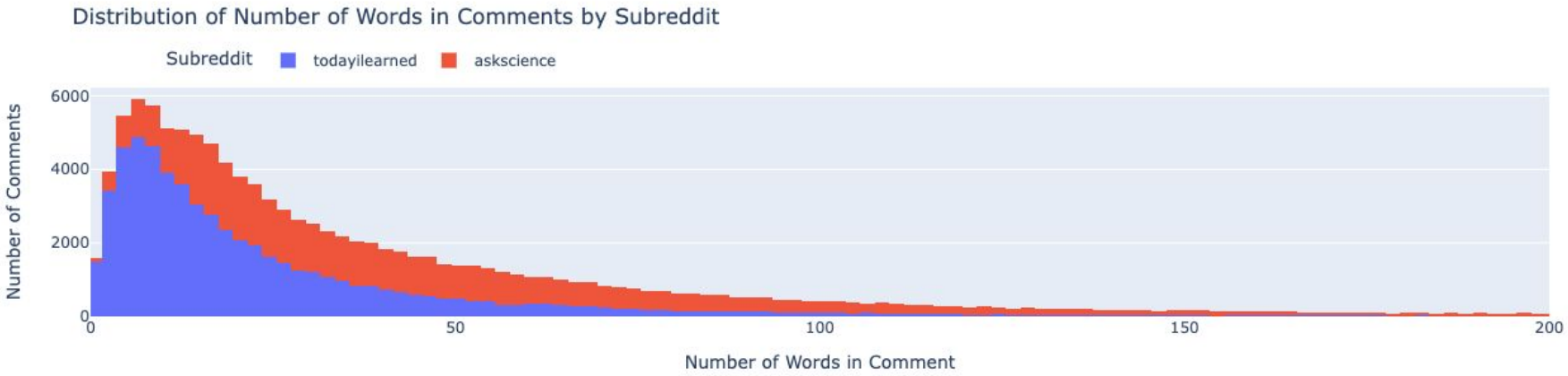
Sentiment Intensity Analyzer

Average Neutral Score

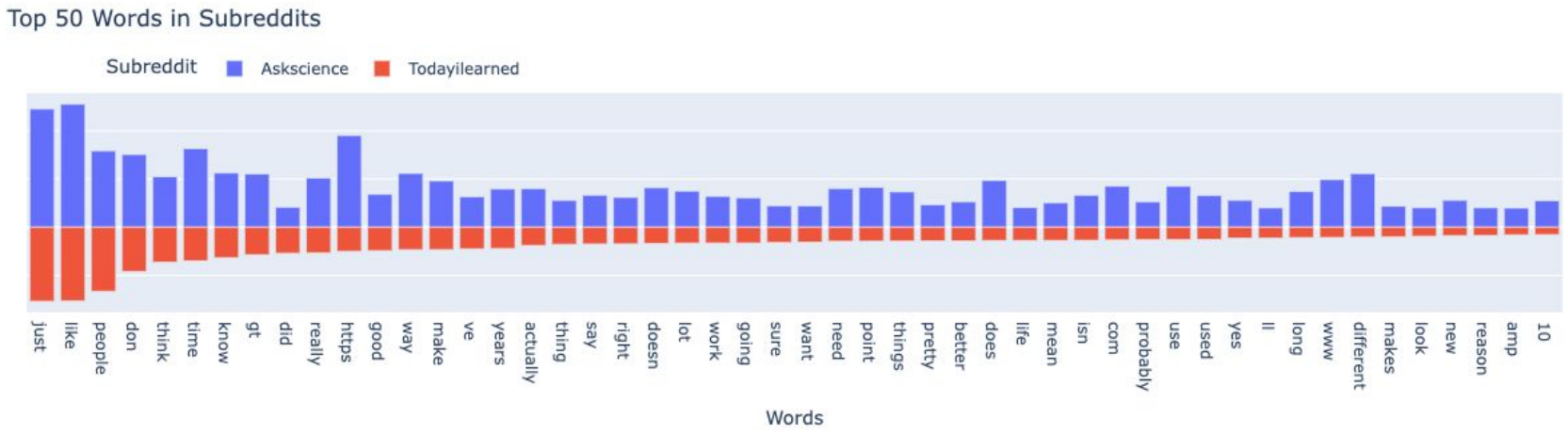
Total Length of Comments by Subreddit



Number of Words in Comment by Subreddit



Top 50 Words by Subreddit

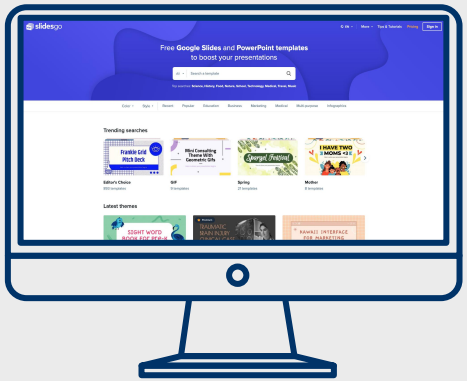


ABOUT MODELING

- Baseline Accuracy 50%
- CountVectorizer for preprocessing
- Logistic Regression
- Support Vector Machines
- Random Forest Classifier

CLASSIFICATION MODELS

	LOGISTIC REGRESSION	SUPPORT VECTOR MACHINES	RANDOM FOREST CLASSIFIER
Stop Words	Top 50 Words	Top 50 Words	English
Maximum Features	3,000	3,000	None
Accuracy Score Training Testing	 87.9% 86.4%	 88.7% 85.7%	 99.8% 84.1%
F1 SCORE	87.3%	86.6%	84.8%



CHECK OUT MY MODEL ON STREAMLIT!

CONCLUSIONS

BEST MODEL

- Logistic Regression, with Countvectorizer preprocessor
- Stop words - top 50 words in each subreddit
- Maximum Features - 3,000

NEXT STEPS

- Faster computer to run n_grams.
- Explore the top predicting words.
- Expand modeling to other subreddits
- Create a website to classify comment based on subreddit

THANKS!

DO YOU HAVE ANY QUESTIONS?

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