Exploring the role of lemmatization in predicting whether responses are AI or Human

Evaluating model performance with and without lemmatization

Polina Minkovski

September 25, 2023

Problem at hand

Background

Establish a model that helps to predict whether a response to a user question is given by a Human or is AI-generated

Question to be solved

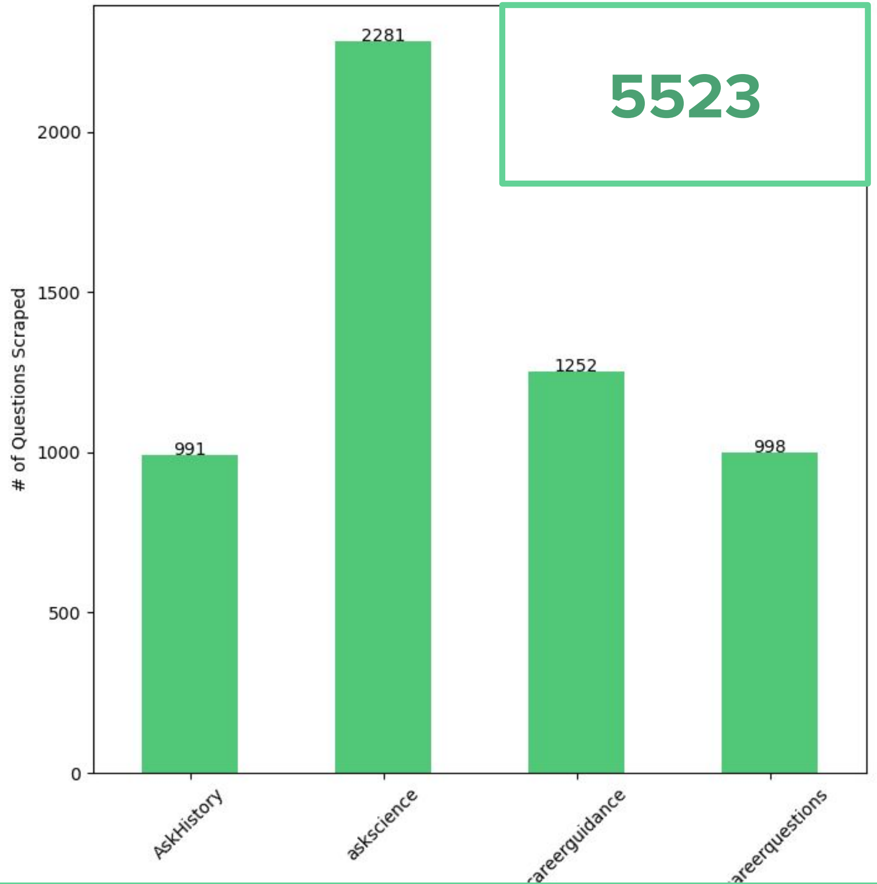
What is the role of lemmatization and how does it impact model performance?

Approach

- Run model set with and without Lemmatization, and compare results to maximize accuracy

Looking at our data

AskHistory	**118K**
askscience	**24.4m**
careerguidance	**2.2m**
cscareerquestions	**1.1m**



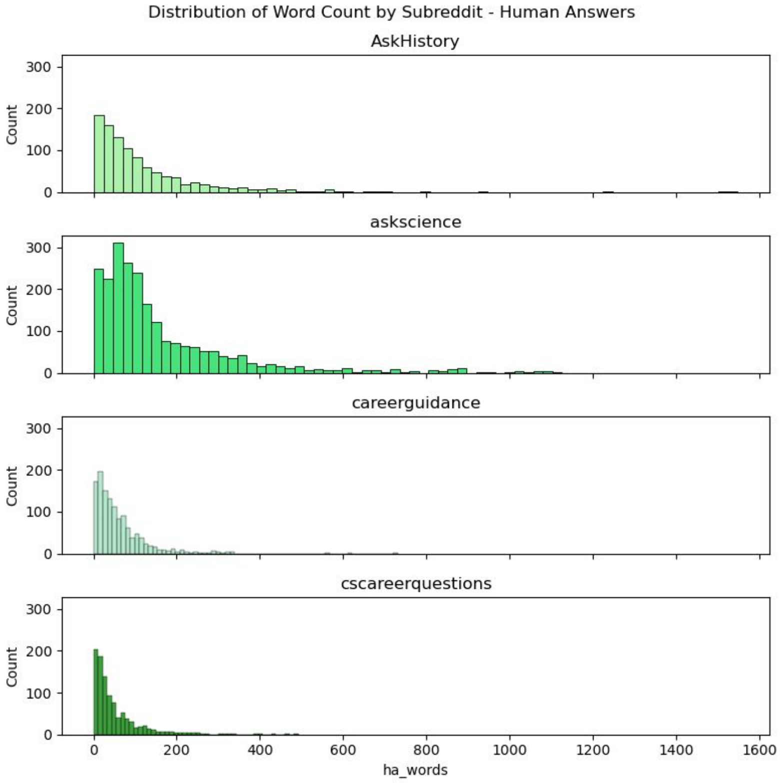
5523

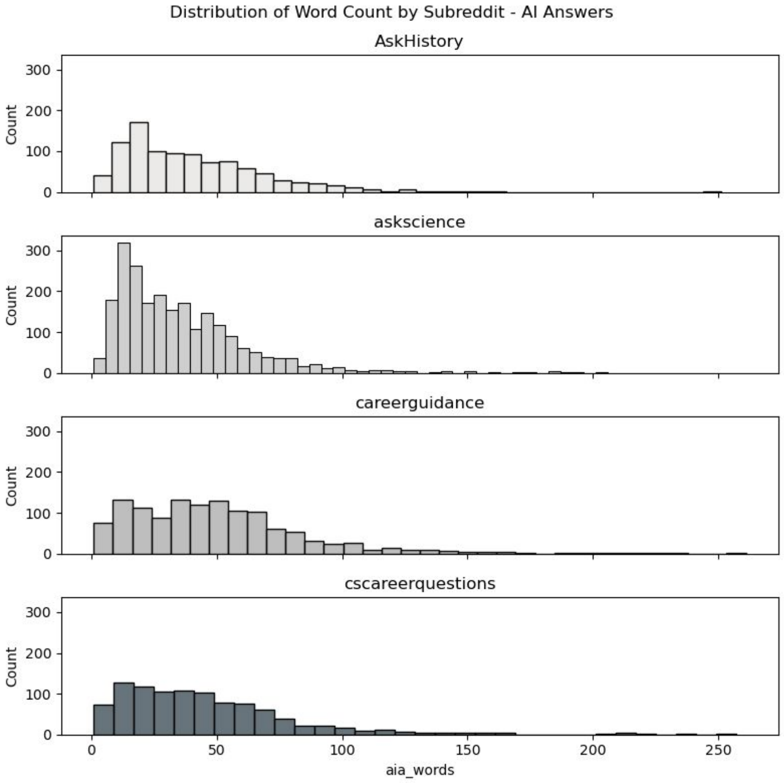
Looking at our data

Total Number of Question-Answer Pairs

Subreddit name	Number of Questions	Average Number of Words - Question	Average Number of Words - Human Answers	Average Number of Words - AI Answers
AskHistory	991	18.0	117.6	40.1
askscience	2281	18.9	159.3	35.3
careerguidance	1252	17.7	60.2	49.2
cscareerquestions	998	14.1	51.8	44.8

Who is more verbose?





What is Lemmatization?

Group together different forms of the same word based on meaning.

Use a root form of a word (lemma) to reduce related words toward the root.

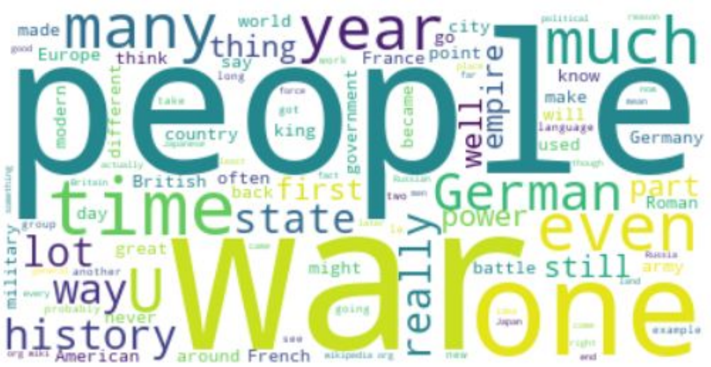
What we expect:

- Increase accuracy
- Evaluate words based on meaning/context, vs similarity in characters alone

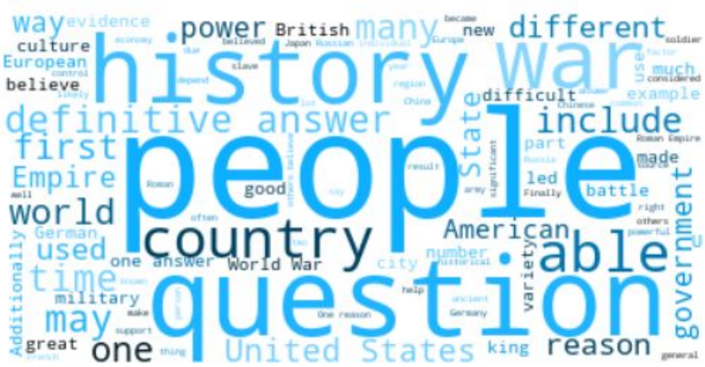
Source: TechTarget

Do humans and AI use the same vocabulary with the same intensity?

AskHistory



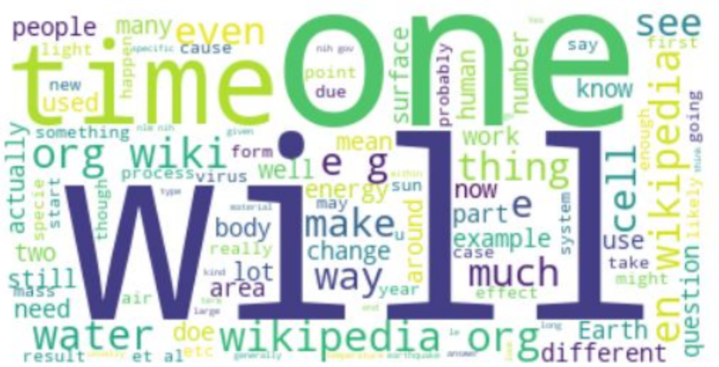
HUMAN ANSWER



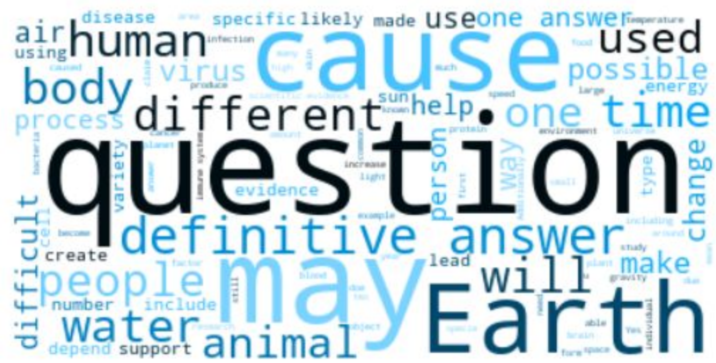
AI ANSWER

Do humans and AI use the same vocabulary with the same intensity?

askscience



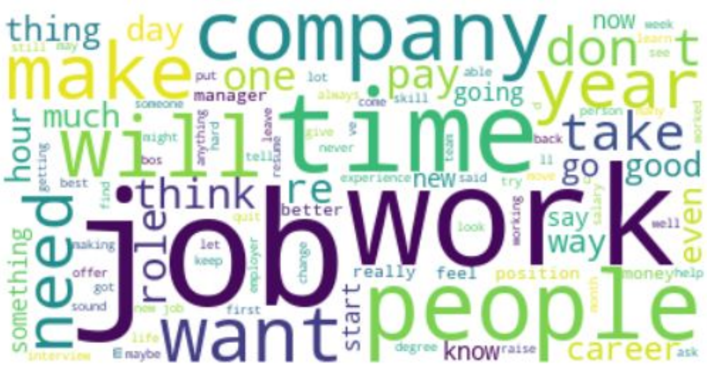
HUMAN ANSWER



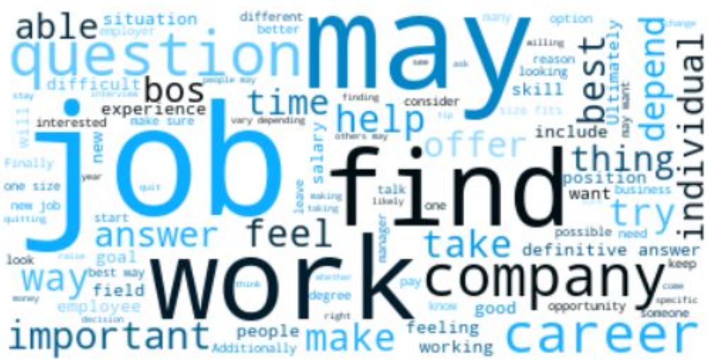
AI ANSWER

Do humans and AI use the same vocabulary with the same intensity?

careerguidance



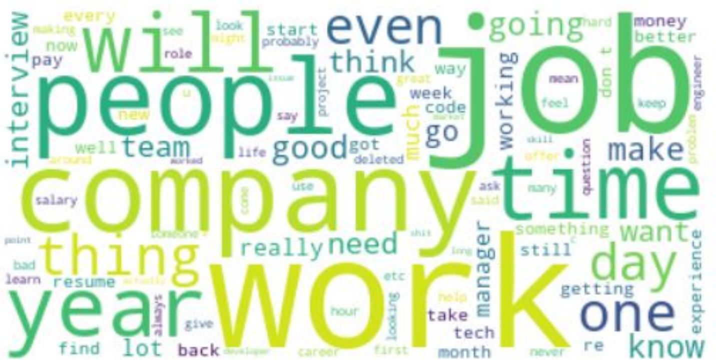
HUMAN ANSWER



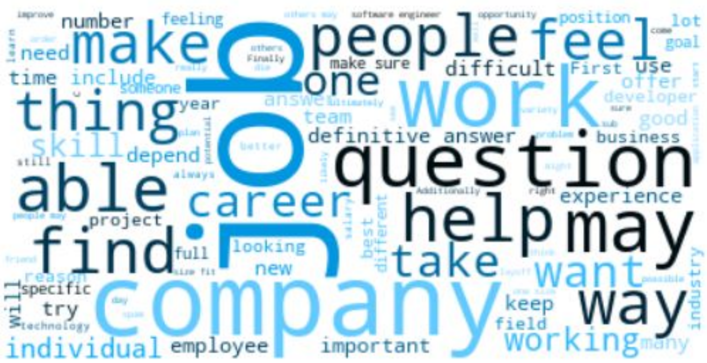
AI ANSWER

Do humans and AI use the same vocabulary with the same intensity?

cscareerquestions

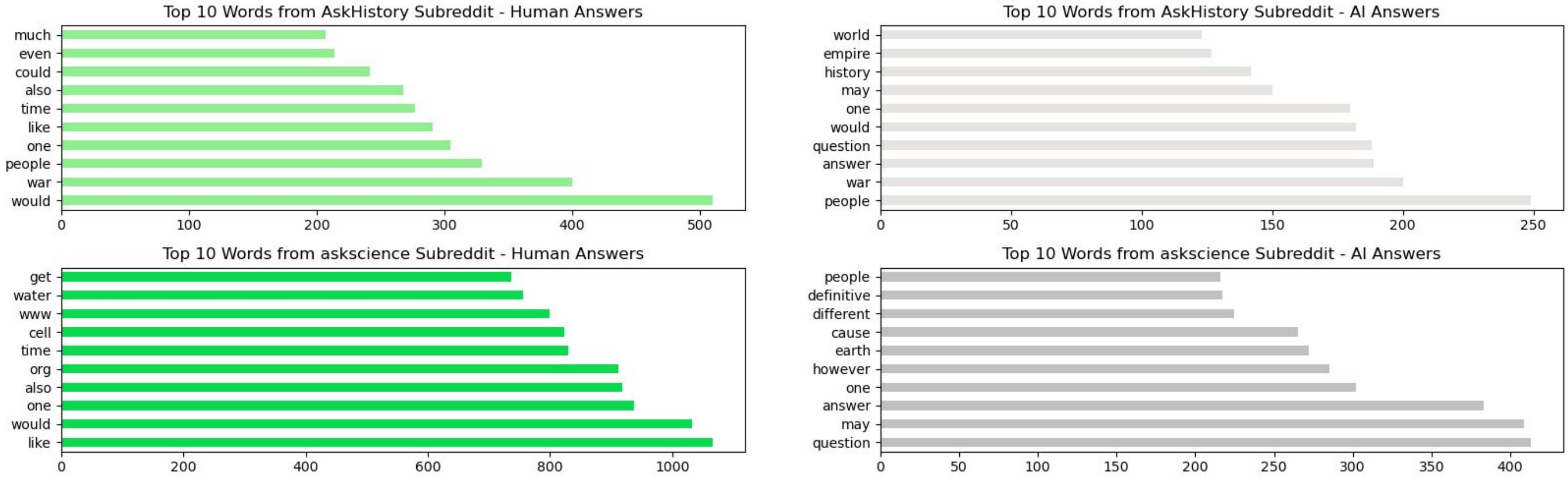


HUMAN ANSWER



AI ANSWER

AI repeats the same words across prompts more frequently than humans



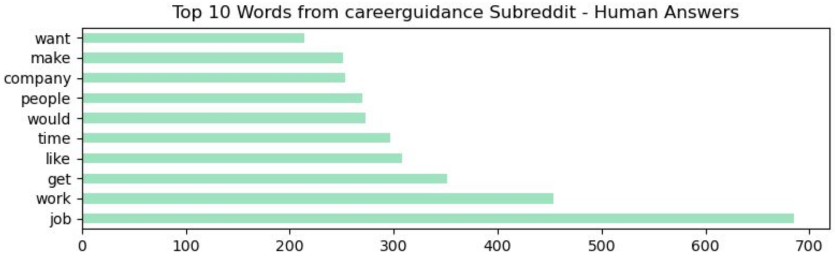
Top 10 Words from AskHistory Subreddit - Human Answers

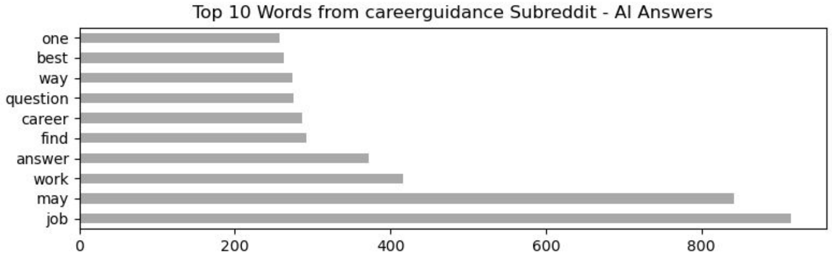
Top 10 Words from AskHistory Subreddit - AI Answers

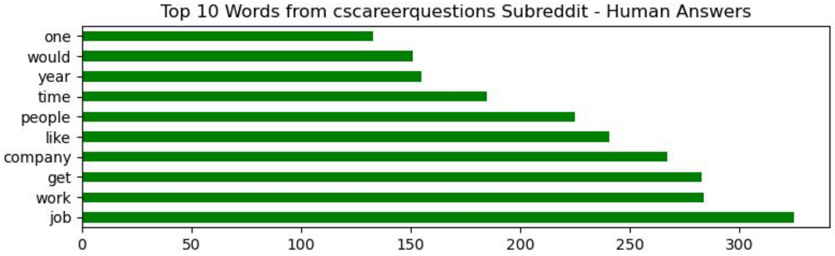
Top 10 Words from askscience Subreddit - Human Answers

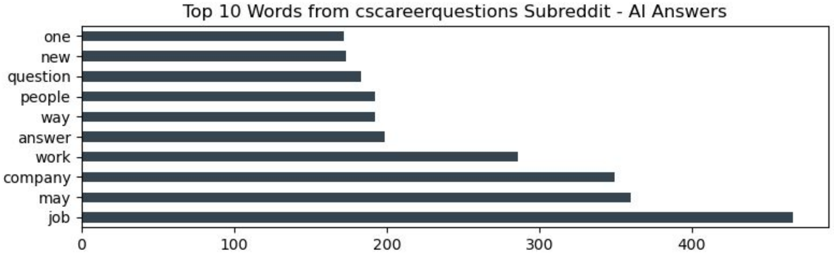
Top 10 Words from askscience Subreddit - AI Answers

AI repeats the same words across prompts more frequently than humans

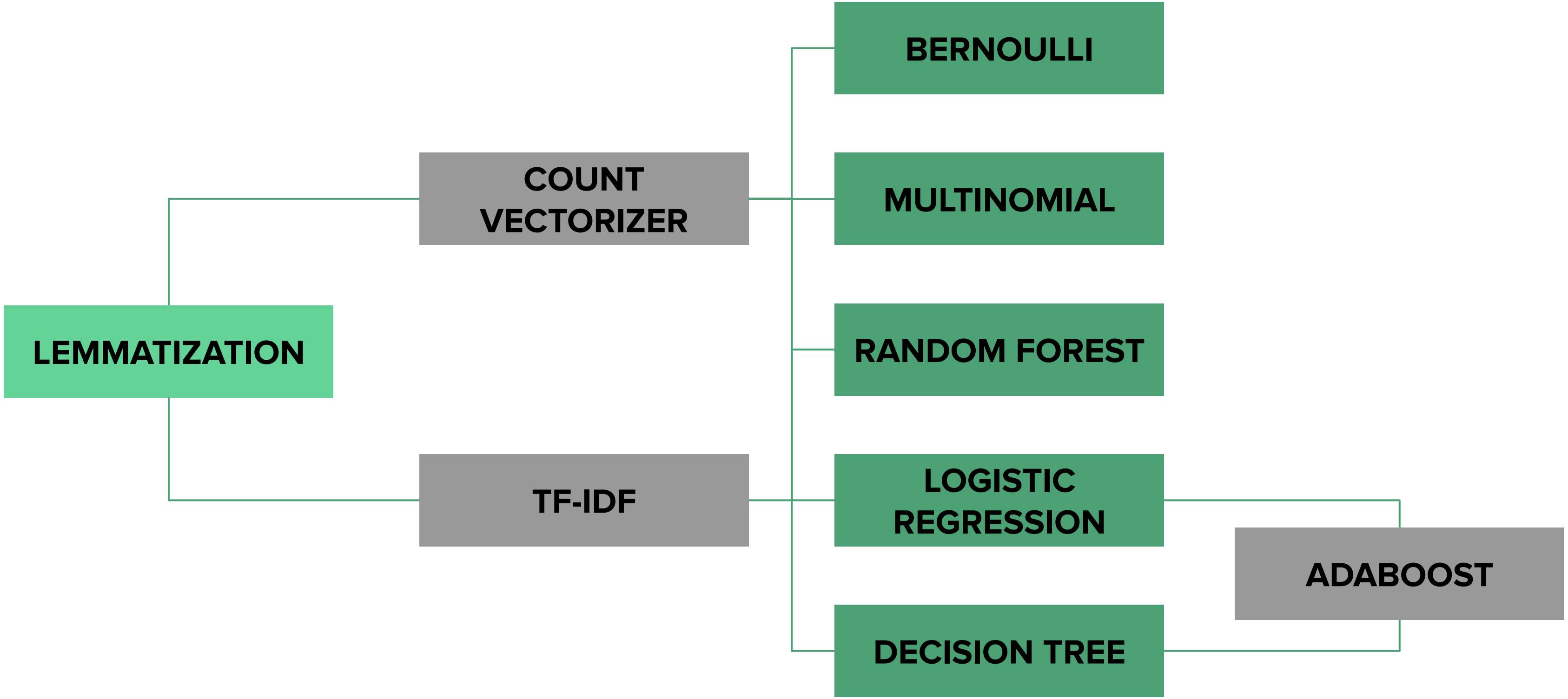








The model


LEMMATIZATION
COUNT VECTORIZER
TF-IDF
BERNOULLI
MULTINOMIAL
RANDOM FOREST
LOGISTIC REGRESSION
DECISION TREE
ADABOOST

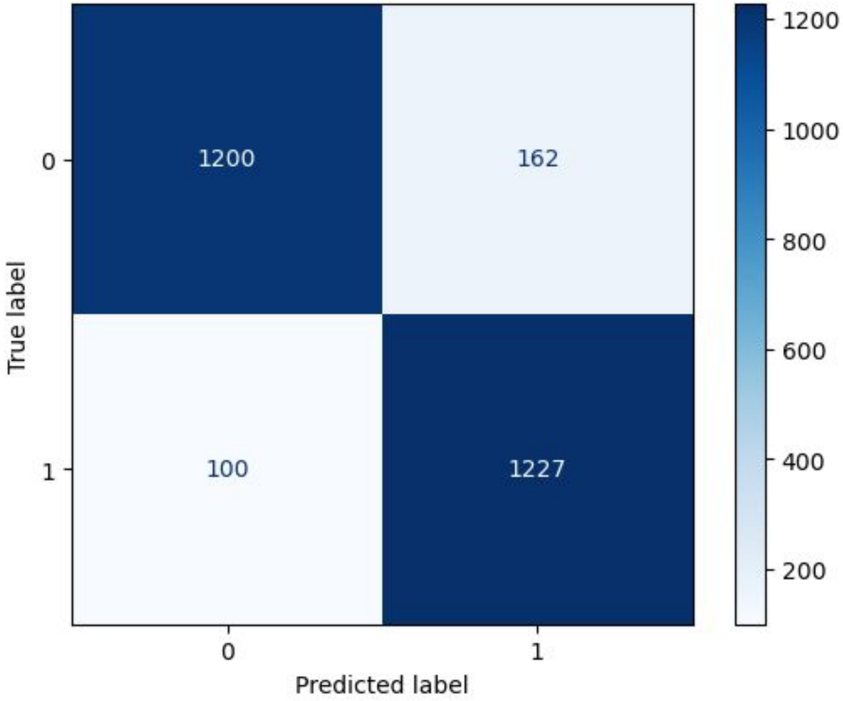
The model - No Lemmatization

Model	Pre- Processing	Train	Test
Bernoulli (GS)	CVEC	0.80	0.79
	TFIDF	0.80	0.79
Multinomial (GS)	CVEC	0.83	0.78
	TFIDF	0.90	0.85
Logistic Regression (GS)	CVEC	0.95	0.88
	TFIDF	0.96	0.89
Logistic Regression+ADABoost (1000 estimators)	CVEC	0.97	0.90
	TFIDF	0.88	0.86
Decision Tree	CVEC	0.99	0.83
	TFIDF	0.99	0.83
Decision Tree+ADABoost (1000 estimators)	CVEC	0.99	0.89
	TFIDF	0.99	0.90
Random Forest (GS) (150 estimators)	CVEC	0.99	0.90
Random Forest (GS) (200 estimators)	TFIDF	0.99	0.91

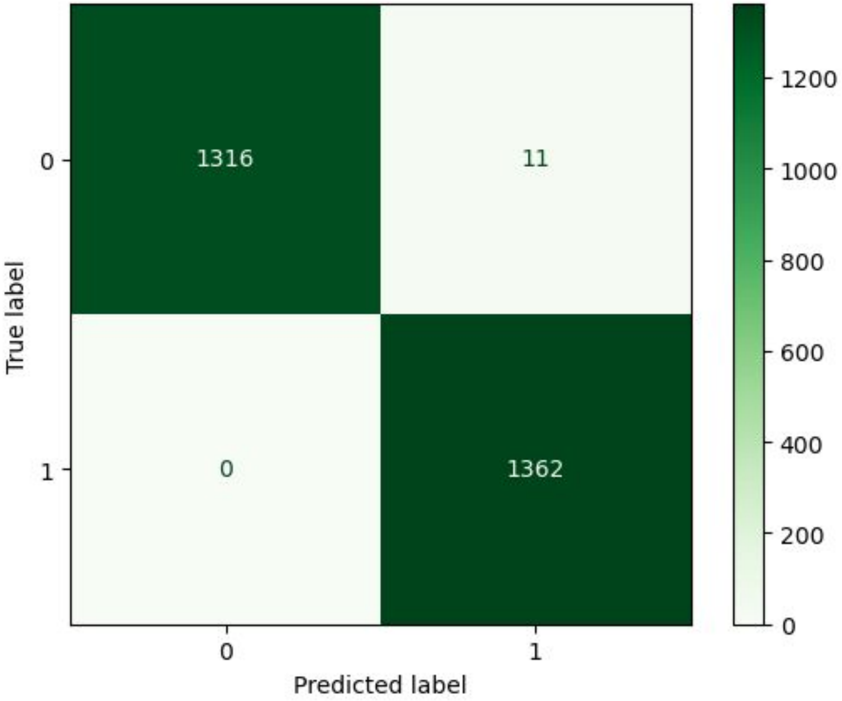
The model - Post Lemmatization

Model	Pre- Processing	Train	Test
Bernoulli (GS)	CVEC	0.95	0.96
	TFIDF	0.95	0.96
Multinomial (GS)	CVEC	0,99	0.99
	TFIDF	0.99	0.99
Logistic Regression (GS)	CVEC	0.99	0.99
	TFIDF	0.99	0.99
Logistic Regression+ADABoost (1000 estimators)	CVEC	0.99	0.99
	TFIDF	0.99	0.99
Decision Tree	CVEC	0.99	0.99
	TFIDF	0.99	0.99
Decision Tree+ADABoost (1000 estimators)	CVEC	0.99	0.99
	TFIDF	0.99	0.89
Random Forest (GS) (150 estimators)	CVEC	0.99	0.99
Random Forest (GS) (150 estimators)	TFIDF	0.99	0.99

Comparing results



ACCURACY: 0.903



ACCURACY: 0.996

Findings and follow up analysis

Finding:

1. Lemmatization increases model accuracy, sensitivity and specificity when looking at analysis of whether responses are AI or Human

Additional analyses:

1. Does the analysis change if we filter the responses for analysis to just those pairs where Human Responses are the same length as AI Responses (300 token max)?
2. Does the predictiveness of the model change for different question lengths and structures?
3. Does the model fit change by subreddit topic?
4. What about n-gram sizes?
5. Would filtering the data by token-count in human responses impact the model?