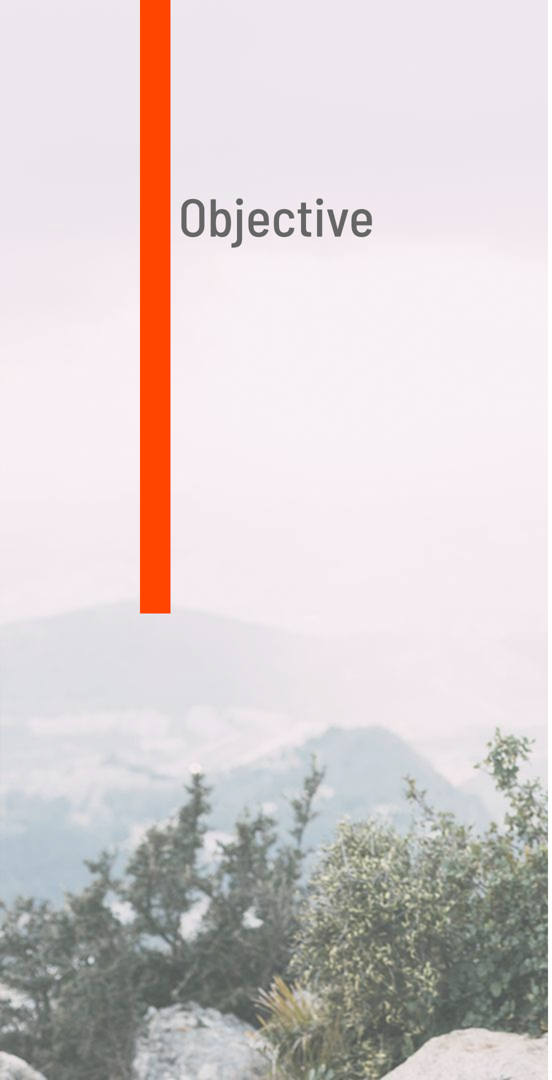
MENTAL HEALTH DIAGNOSES USING REDDIT

Objective



By using Natural Language Processing, train a machine learning model to accurately classify whether a reddit post belongs to the depression subreddit or the bipolar subreddit

TABLE OF CONTENTS

01 BACKGROUND

02 DATA COLLECTION

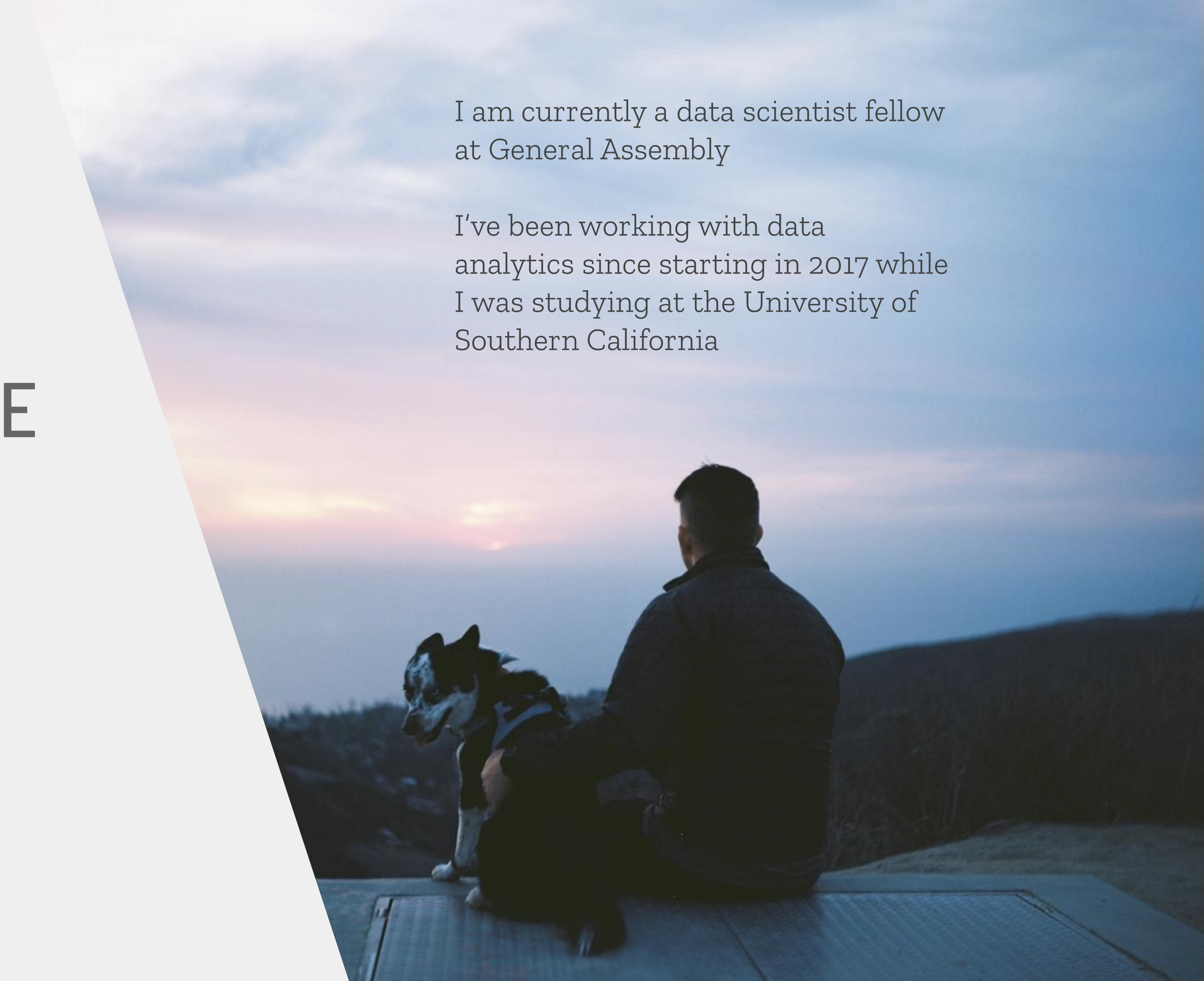
03 DATA EXPLORATION

04 BUILDING A MODEL

05 RESULTS

06 CONCLUSIONS

ABOUT ME



I am currently a data scientist fellow at General Assembly

I've been working with data analytics since starting in 2017 while I was studying at the University of Southern California

BACKGROUND

DEPRESSION VS. BIPOLAR DISORDER

DEPRESSION

A mental health disorder characterized by persistently depressed mood or loss of interest in activities, causing significant impairment in daily life.

BIPOLAR DISORDER

A disorder associated with episodes of mood swings ranging from depressive lows to manic highs.

WHY DIFFERENTIATE?

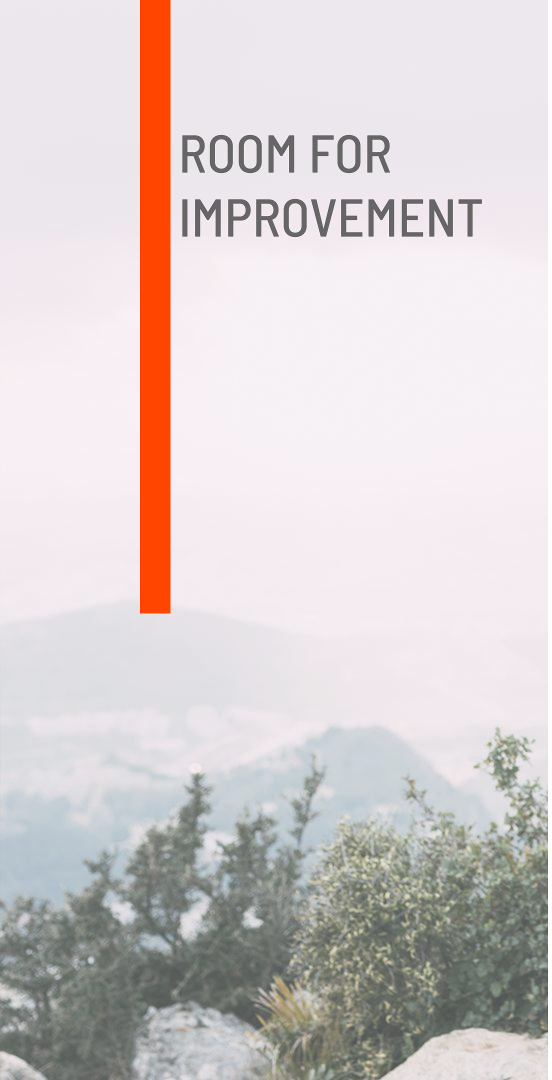
NON-DIAGNOSIS CAN BE DANGEROUS

Often bipolar patients are prescribed traditional depression medications with serious consequences, including increased suicide risk.

MISDIAGNOSIS CAN ALSO BE NEGATIVE

Those incorrectly identified as bipolar may not receive adequate treatment and it may prolong their depression

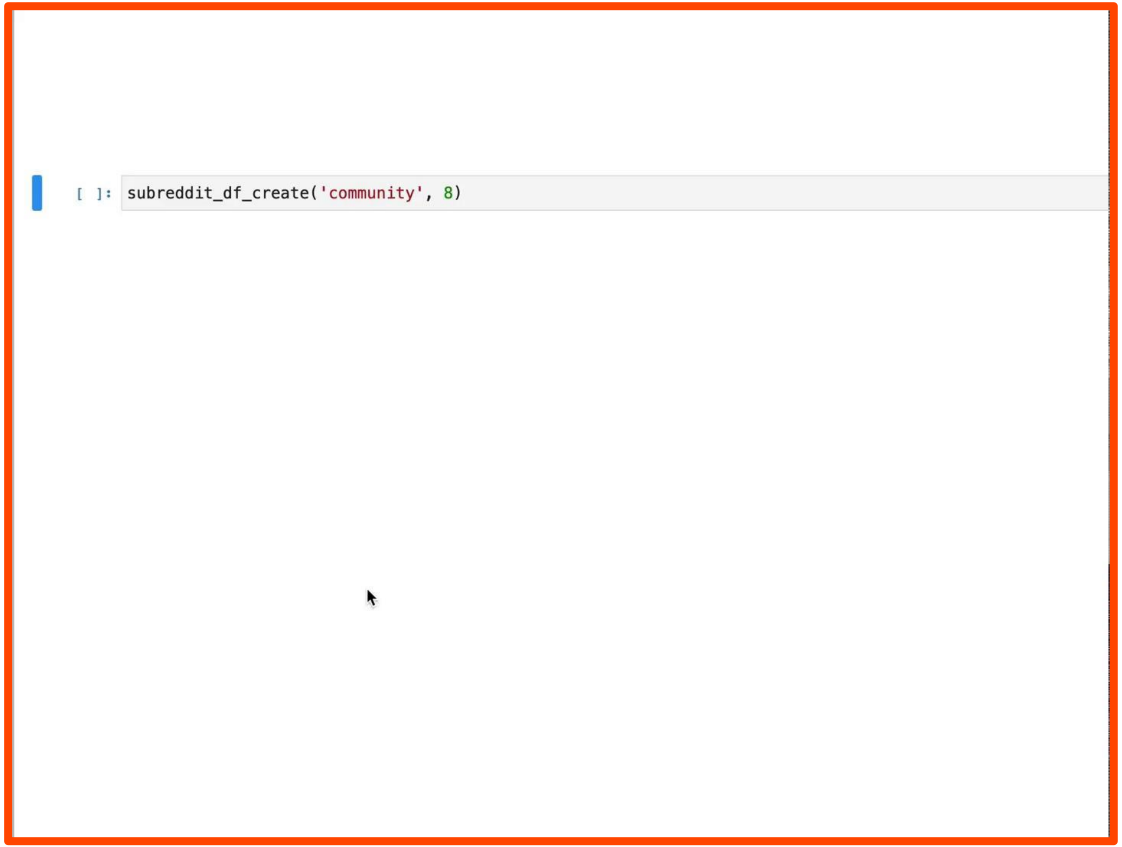
ROOM FOR IMPROVEMENT



The American Journal of Psychiatry studied the most common screening for bipolar disorder, the Mood Disorder Questionnaire, in November 2000 and correctly diagnosed Bipolar Disorder just 73% of the time (sensitivity).

DATA COLLECTION

Data Collection



```
[ ]: subreddit_df_create('community', 8)
```

Data Cleaning

2%

Increase in model accuracy by preprocessing language data

REMOVE CONTRACTIONS

Expand contractions like "can't" to "can" and "not"

LEMMATIZE WORDS

Return words to their base form

REMOVE PUNCTUATION

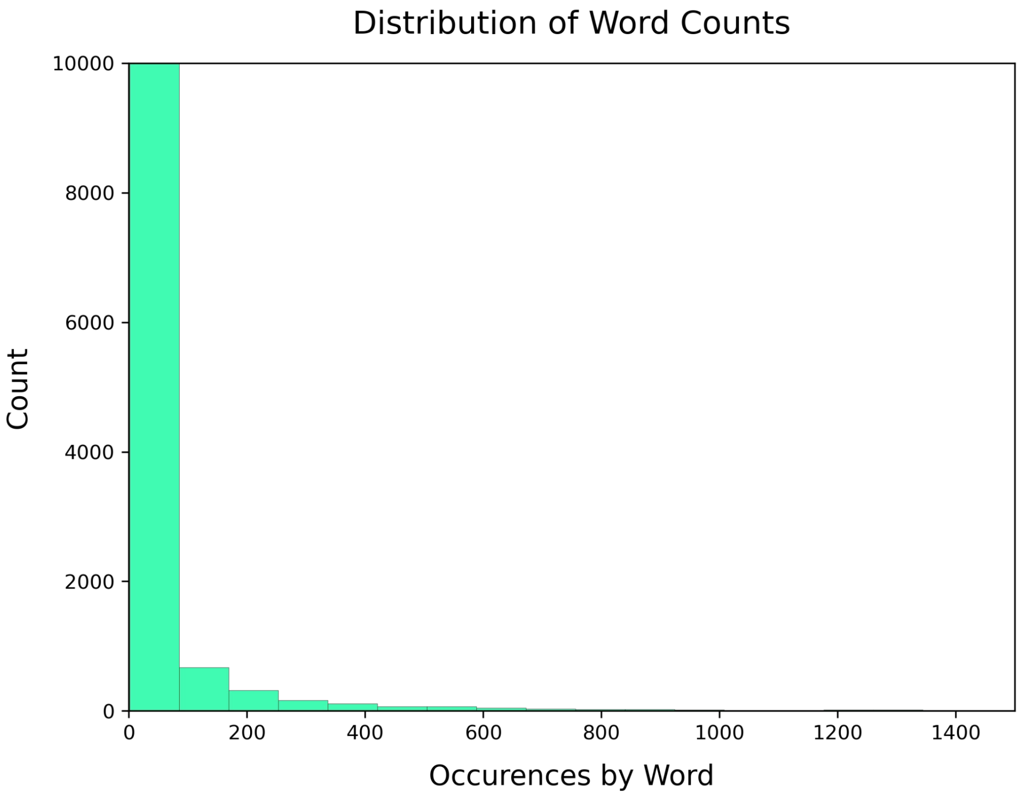
Punctuation is so commonplace and while it may have some meaning, machines have trouble with their meaning

REMOVE STOP WORDS

Stop words are very common words like "the" that will not help our analysis and maybe even hinder it

DATA EXPLORATION

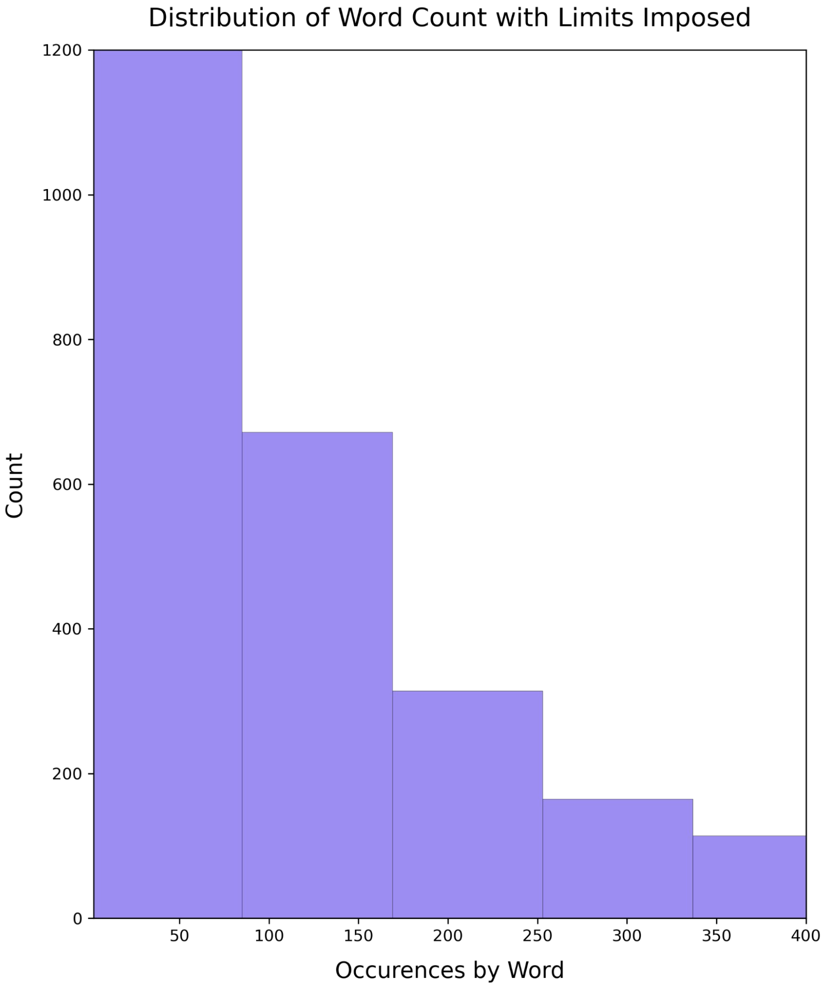
ANALYZING WORD COUNTS



Most Words Occur Very Few Times

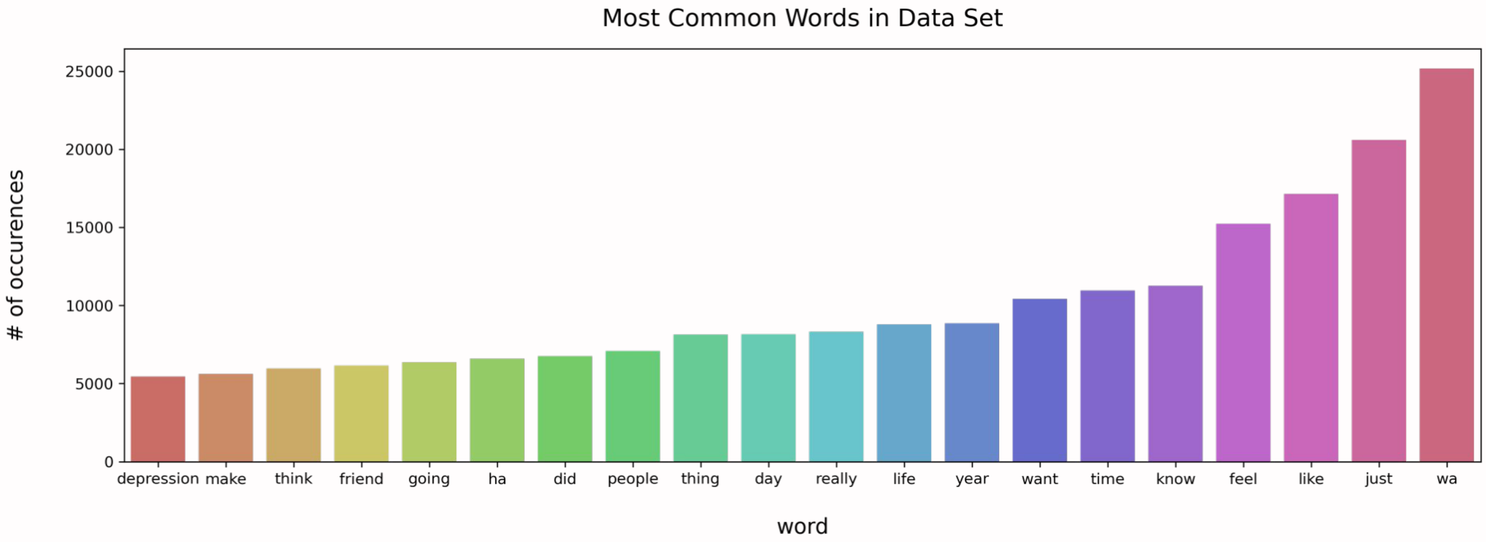
If a word occurs only a few times in a dataset of millions, how important is it?

ANALYZING WORD COUNTS

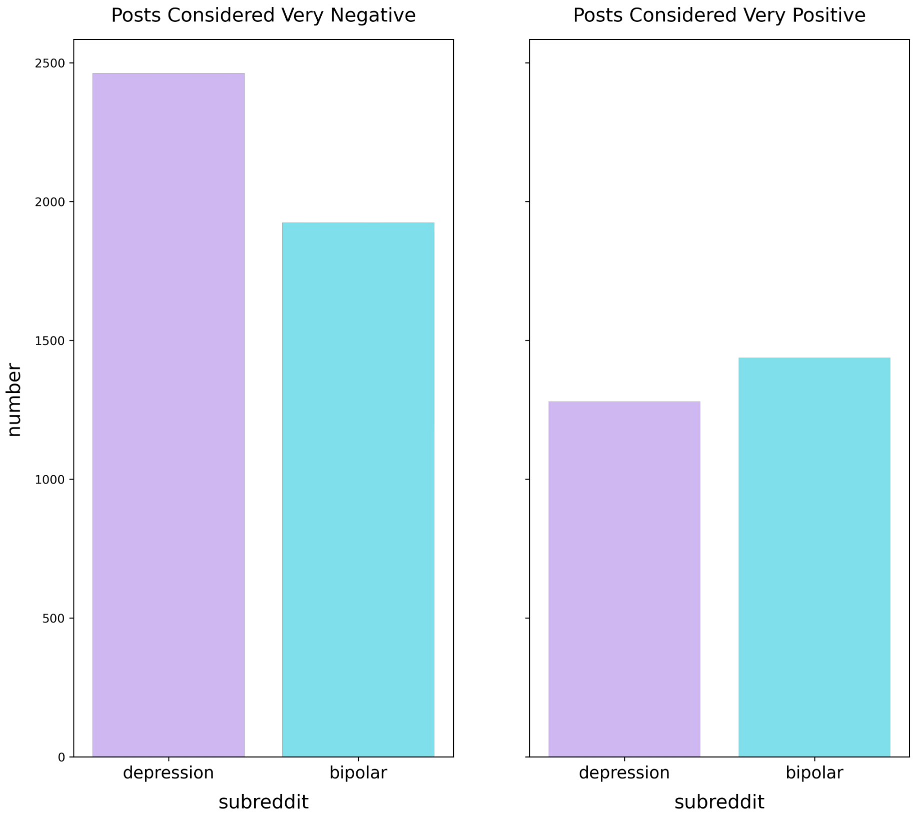


This is now the distribution of our words when limits are imposed

We will keep this in mind when we start modeling


Most Common Words in Data Set
# of occurences
25000
20000
15000
10000
5000
0
depression
make
think
friend
going
ha
did
people
thing
day
really
life
year
want
time
know
feel
like
just
wa
word

SENTIMENT ANALYSIS



The general distribution of the sentiment scores had a large amount of positives

Looking closer, there might be a genuine reason for this and it may provide predictive value

BUILDING A MODEL

STEP 1: TFIDF Vectorization

Term frequency–inverse document frequency, or **TFIDF**, is a way to weight a word depending on how important it is to our data

TFIDF ADVANTAGES

Analyze which words are being weighted heavily

Test what the optimal amount of features should be

Analyzes combinations of words without using too much memory

STEP 2: Model Selection

Select models that make sense for language processing

Get a baseline accuracy for our data: ours is 0.5 or 50%

Use "grid searches" to try thousands of different model builds and find the highest performing ones

SELECTED MODELS

NAIVE BAYES

LOGISTIC REGRESSION

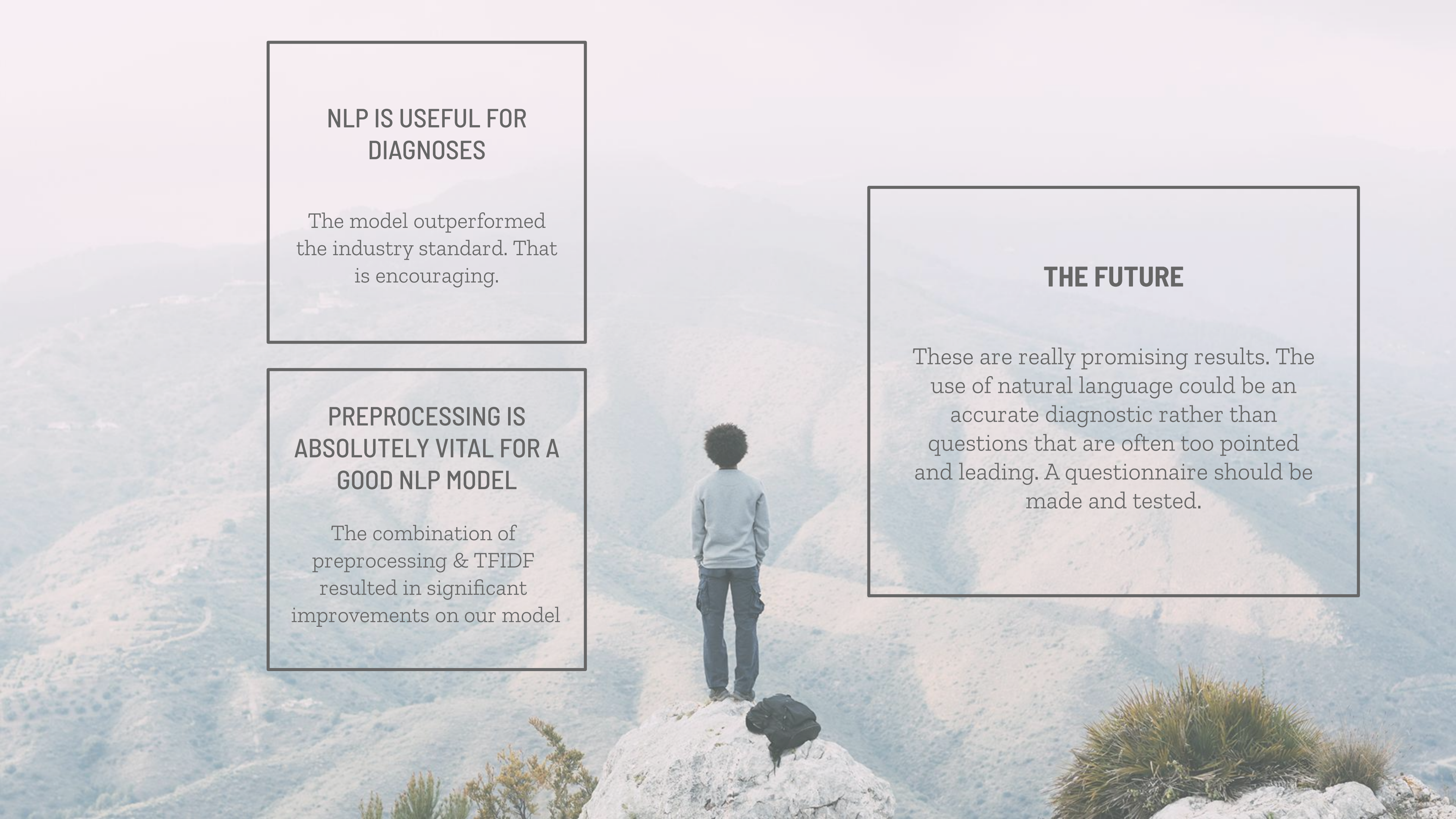
RANDOM FOREST

RESULTS

Model Performance

Score	Naïve Bayes	Logistic Regression	Random Forest
Accuracy (Baseline: 50%)	82.2%	84.8%	84.4%
Sensitivity (Recall)	78.5%	77.9%	80.2%
Precision	84.7%	80.2%	87.6%

CONCLUSIONS



NLP IS USEFUL FOR DIAGNOSES

The model outperformed the industry standard. That is encouraging.

PREPROCESSING IS ABSOLUTELY VITAL FOR A GOOD NLP MODEL

The combination of preprocessing & TFIDF resulted in significant improvements on our model

THE FUTURE

These are really promising results. The use of natural language could be an accurate diagnostic rather than questions that are often too pointed and leading. A questionnaire should be made and tested.

THANKS