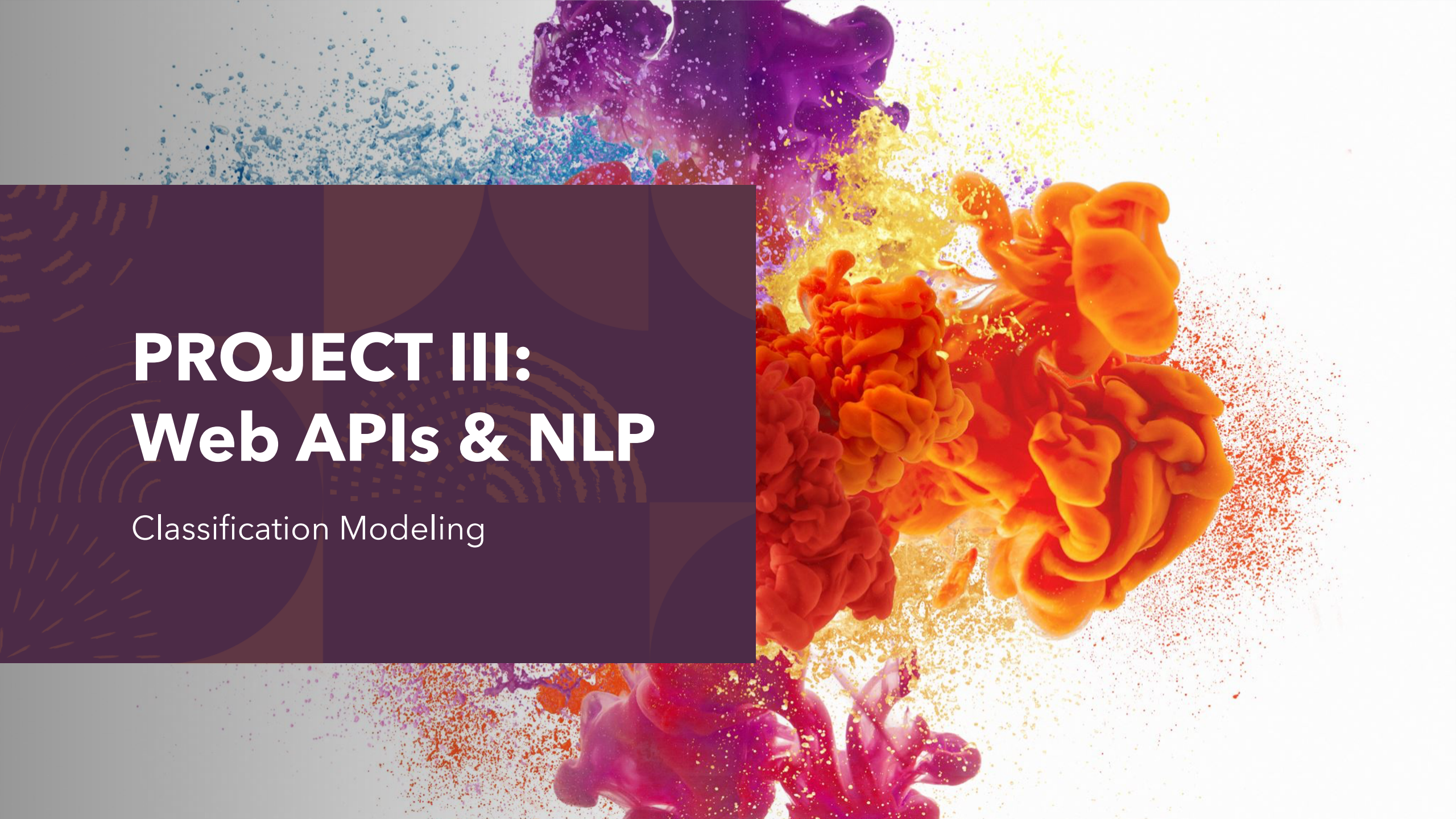


PROJECT III: Web APIs & NLP

Classification Modeling

Content

- Reddit
- Problem Statement
- Data Gathering
- Cleaning & EDA
- Modeling
- Conclusion
- Next Steps

Overview - Reddit?

What is Reddit?

- Reddit is a social media website where members can post comments (links, text and images). “Posts are organized by subject into user-created boards called "subreddits” which covers a wide variety of topics”
- As of October 2020, Reddit ranks as the 17th-most-visited website in the world and 7th most-visited website in the US (from https://en.wikipedia.org/wiki/Reddit)

Pull at least 1,000 posts each from 2 subreddits forums (startups & startup_ideas) and build a classification model that identifies which subreddit each post came from with an accuracy score of at least 90%

Problem Statement

Data Sets

For this project I decided to pull data from the following subreddits:

Startups	Startups Ideas

Initially the data were pulled (1,000 rows each) from the above subreddits using "praw", This was then used to investigate which classification model will best fit my dataset

Data Collection - PRAW Method

After several attempts trying to pull 1,000 post in 1 go my best pull was

- Startup - 900 rows
- Startup_Ideas – 991 rows

At this point I decided to start my classification modeling with this dataset

Data Cleaning

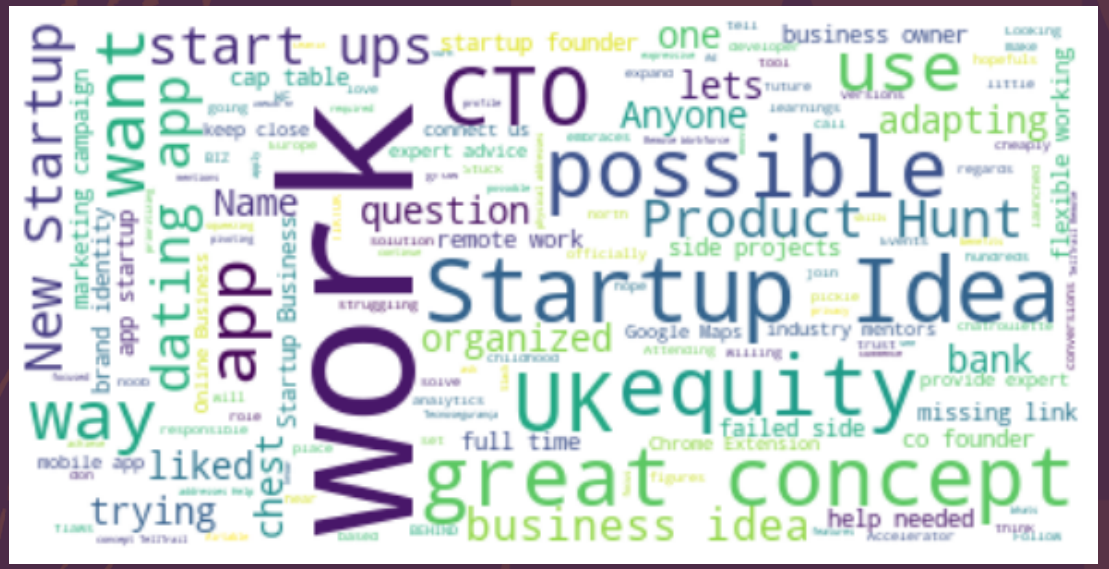
The 2 DataFrames were merged and the following cleaning were done on the merged dataset:

- **Used title and subreddit columns setting 1 for startup and 0 for startup_ideas**
- **Use Regex to take out hyperlinks**
- **Added 2 new columns word_count & title_length (for EDA)**
- **Removed punctuations & special characters**
- **Remove stop words**

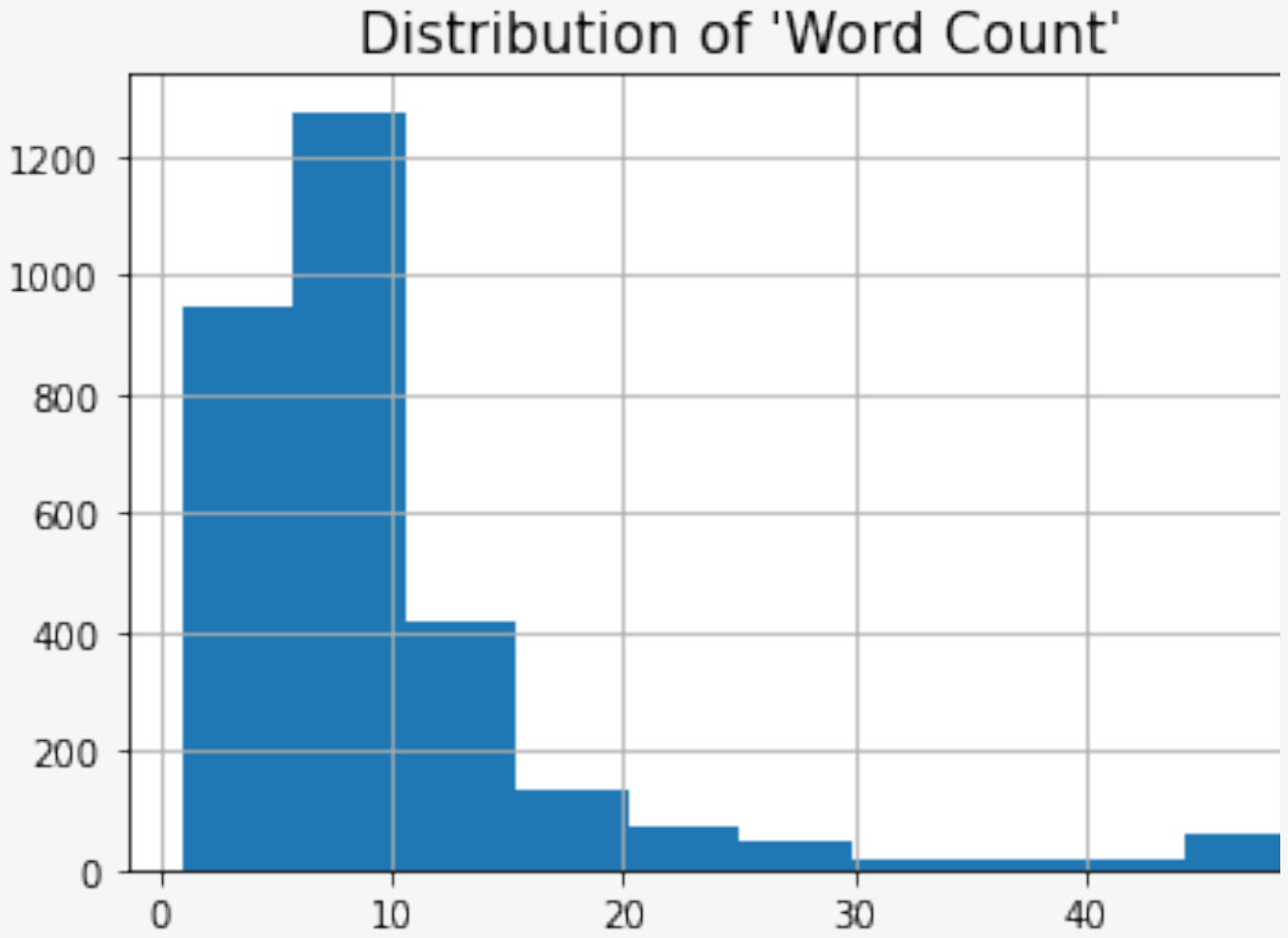
Data: First Glance

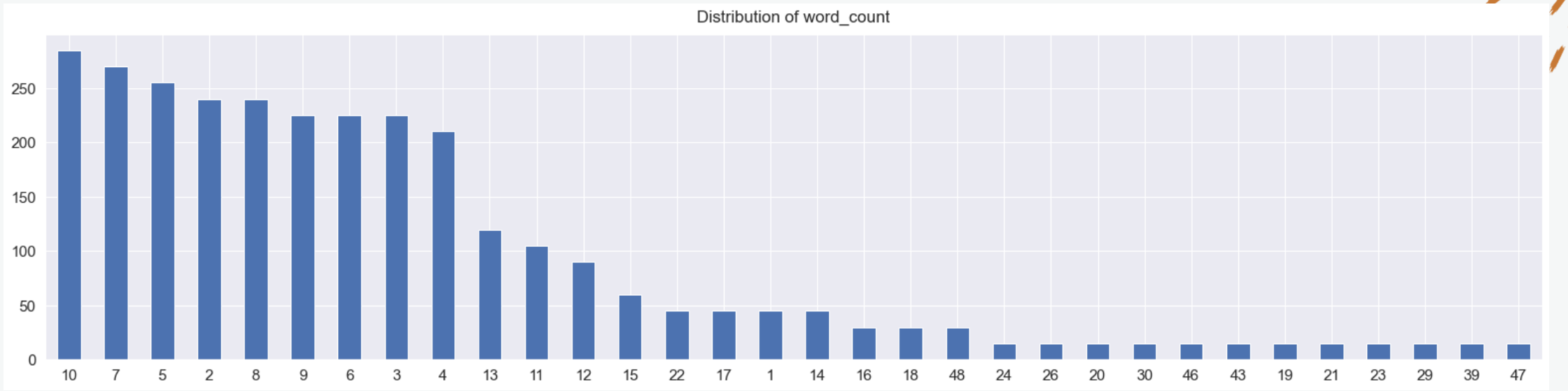
With stop words

Without stop words

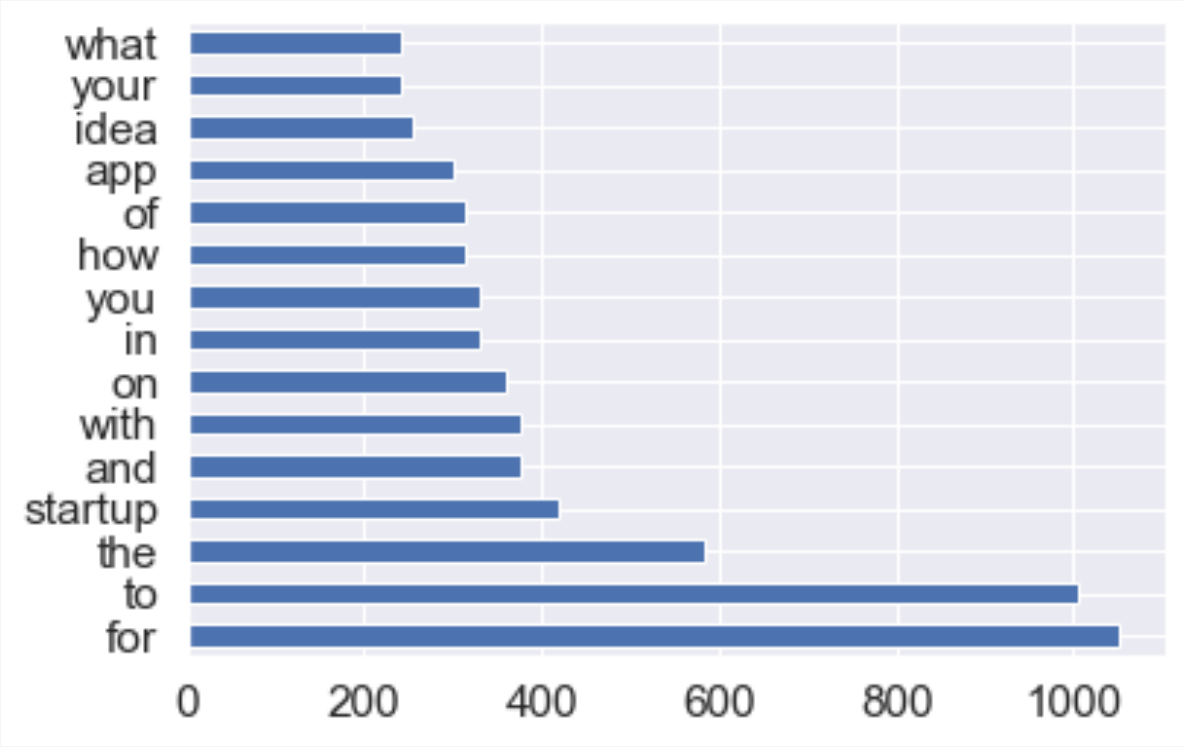


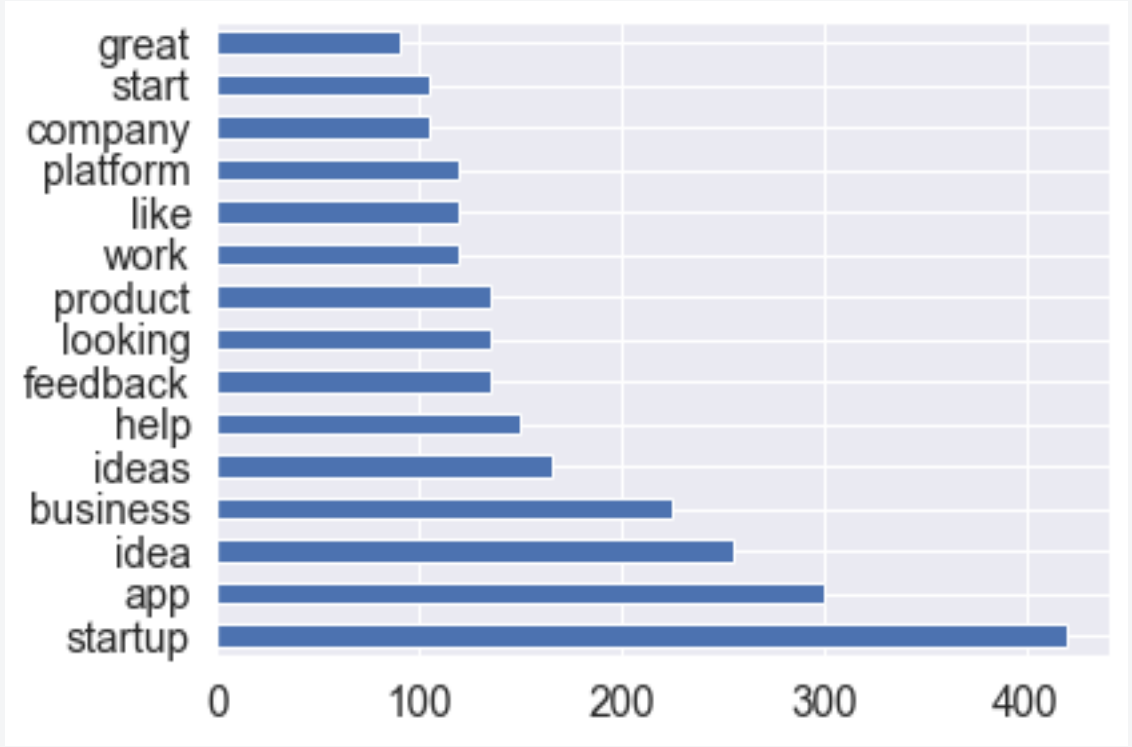
Most titles have 10 or fewer words





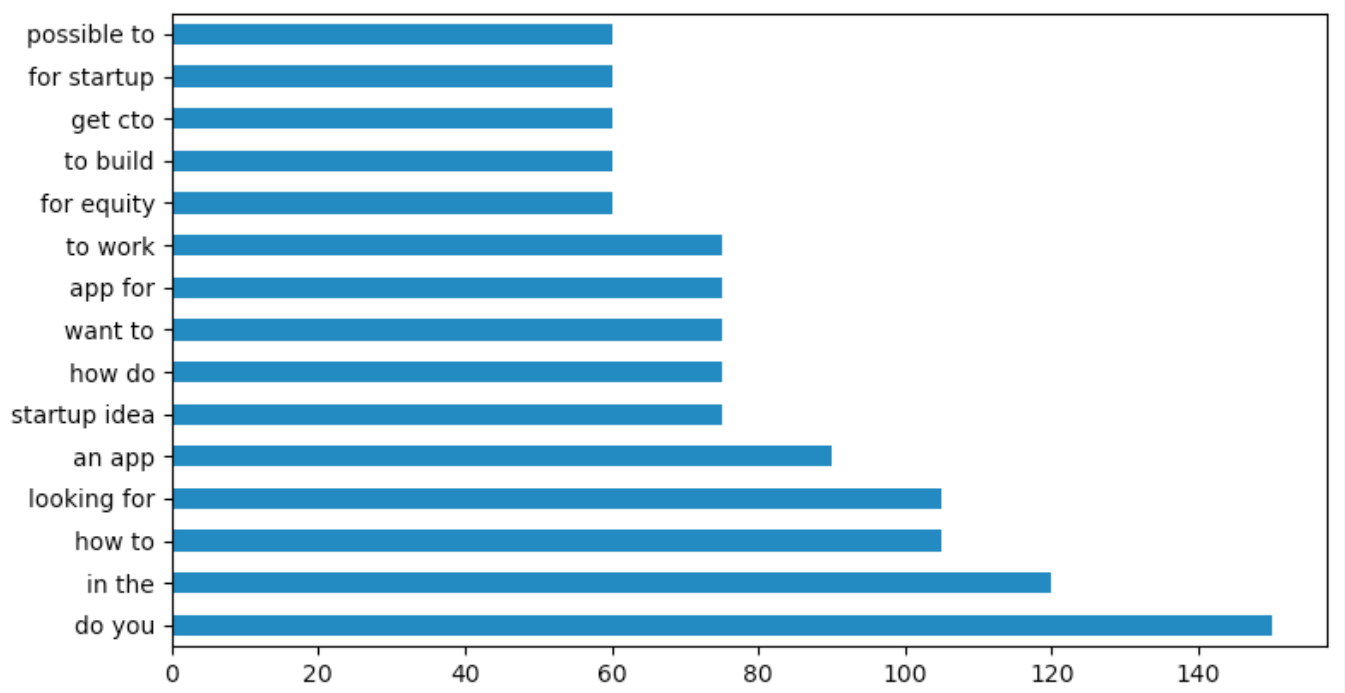
Distribution of words count / Title



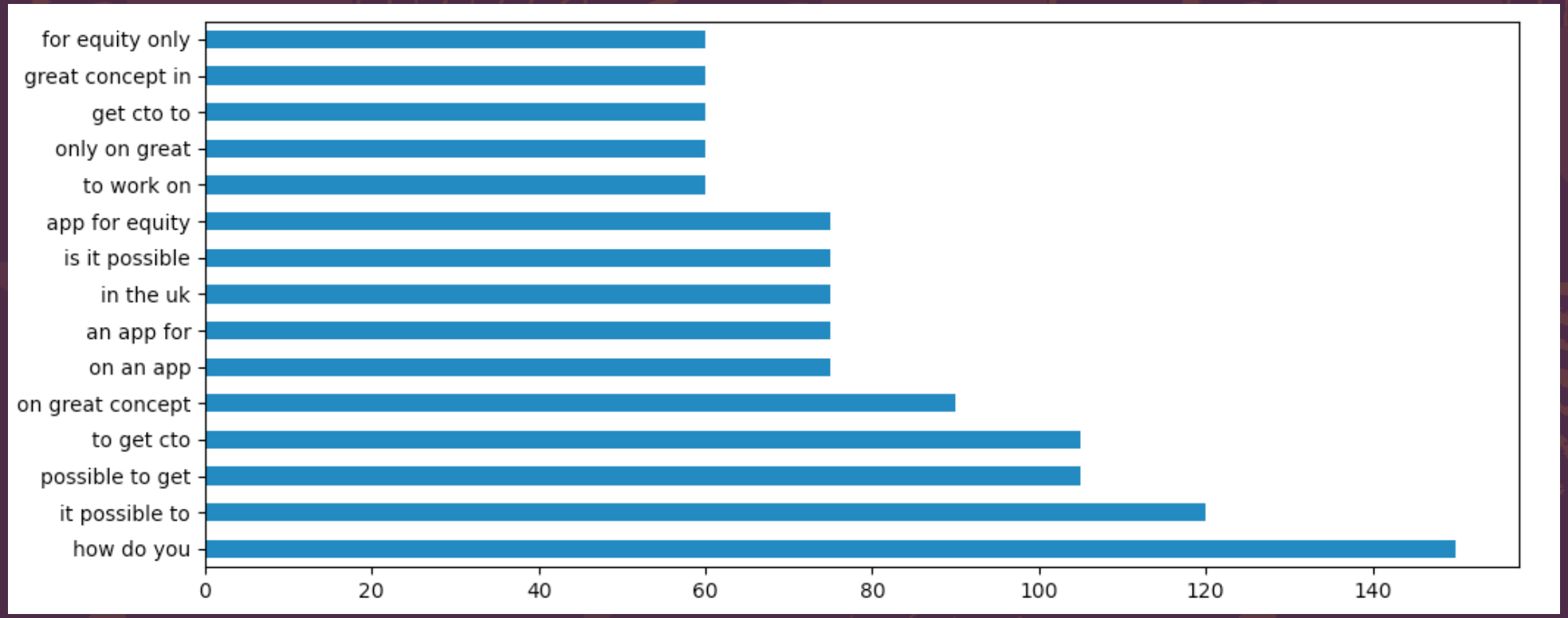


Unigram with / without Stop words

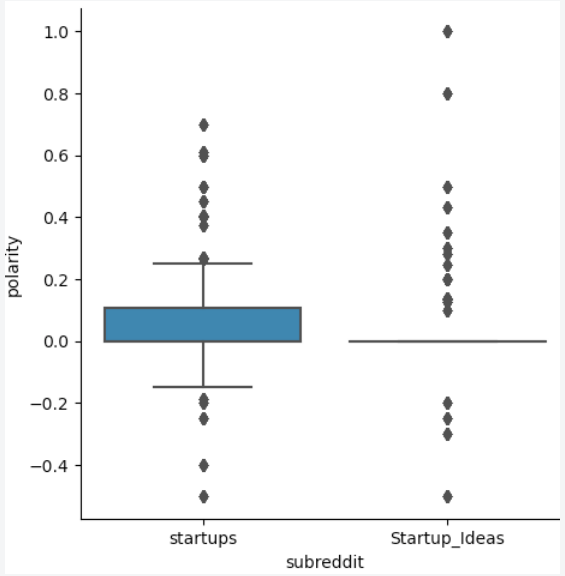
Bigram

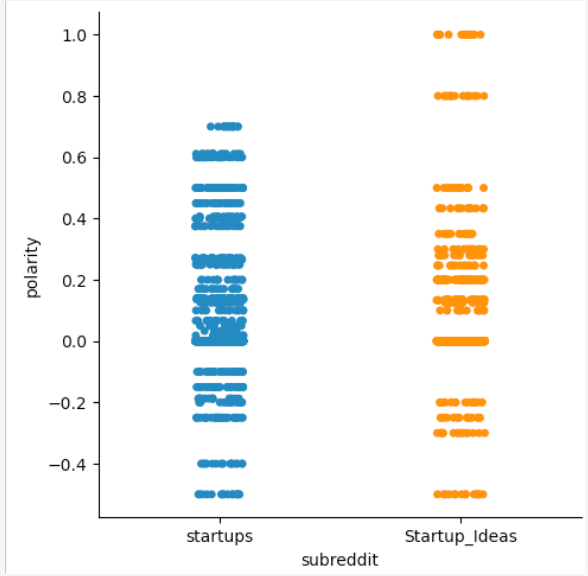

possible to
for startup
get cto
to build
for equity
to work
app for
want to
how do
startup idea
an app
looking for
how to
in the
do you
0
20
40
60
80
100
120
140

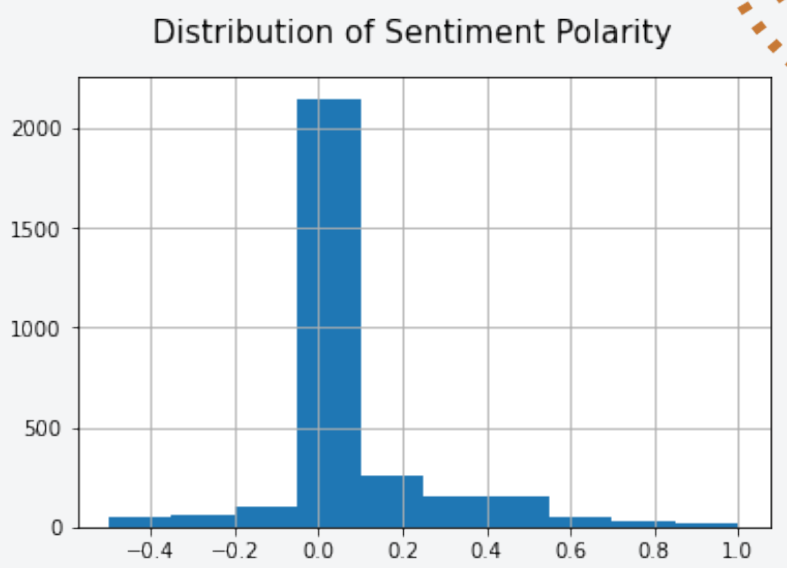
Trigram

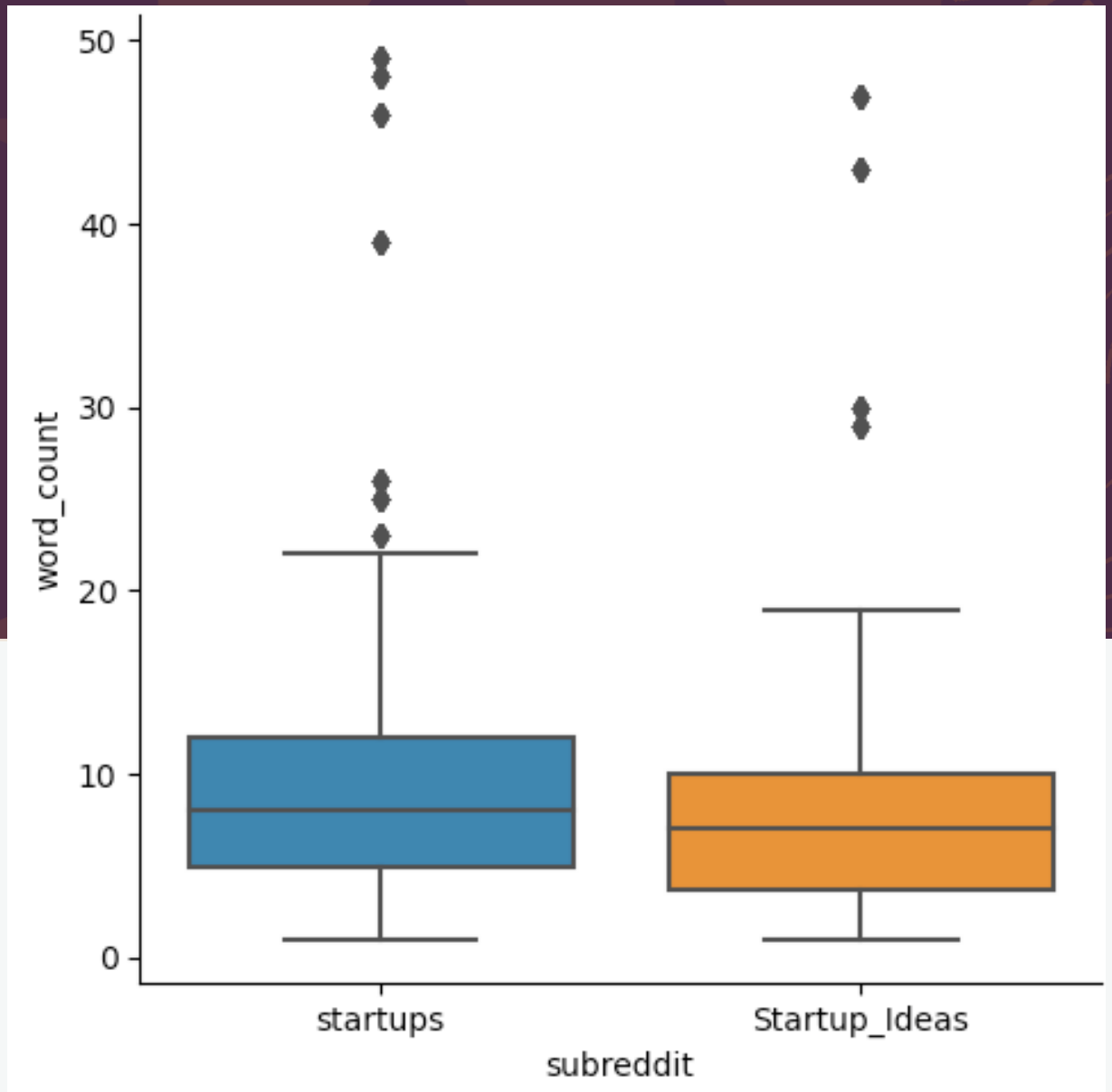

for equity only
great concept in
get cto to
only on great
to work on
app for equity
is it possible
in the uk
an app for
on an app
on great concept
to get cto
possible to get
it possible to
how do you
0
20
40
60
80
100
120
140

Distribution of Sentiment Polarity



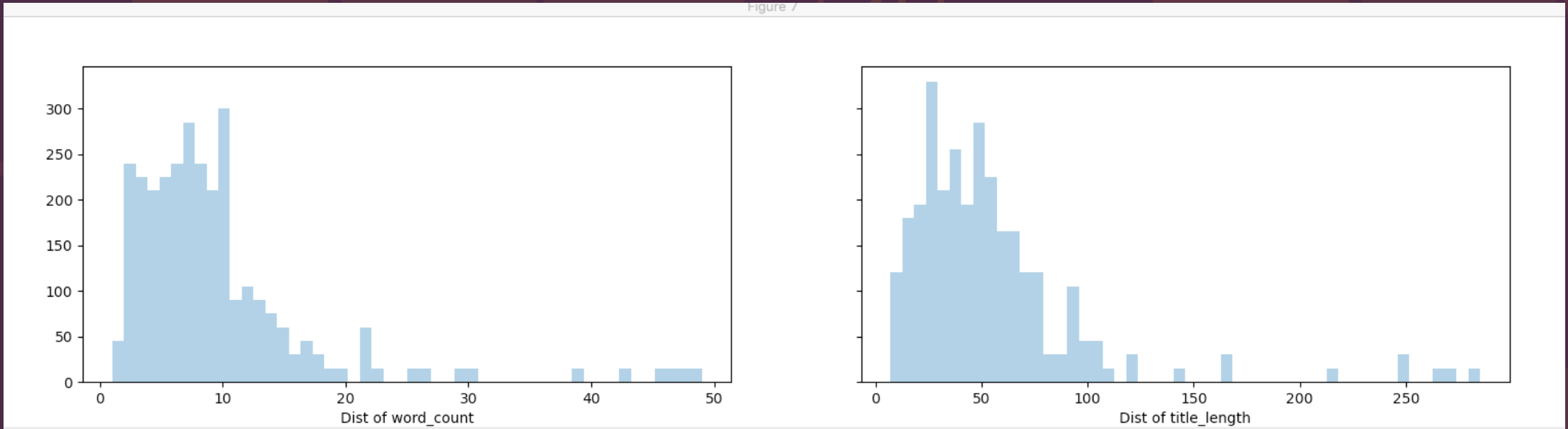






Box plot of subreddit vs word count

Distribution of word count vs title length



Results are in

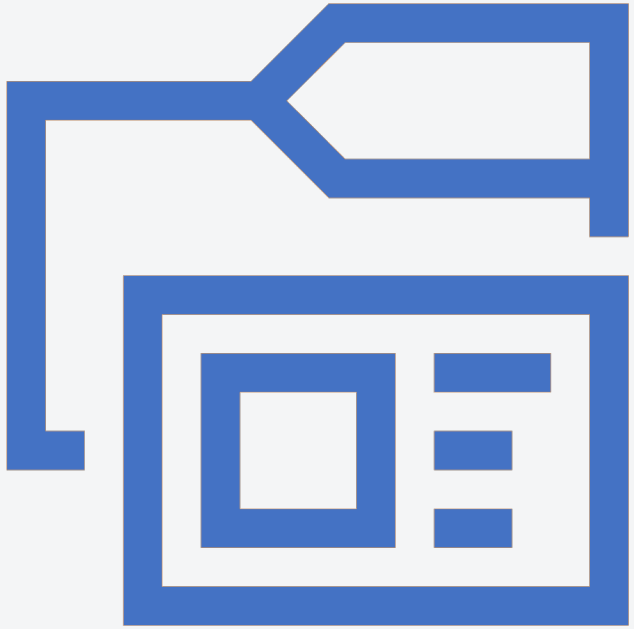
Modeling

Model	Train Score	Test Score
Logistic Regression	0.96759	0.77309
Logistic Regression - with Scaling	1	0.7124
Decision Tree	1	0.72559
Random Forest	0.79496	0.78901
ExtraTreesClassifier	0.78306	0.77511

What’s Next!

More Data collection

At this point I felt that I should gather a complete dataset of at least 2000 total records before moving on to complete additional models. This time I will use pushshift API to collect 1,500 post from each of my 2 subreddit



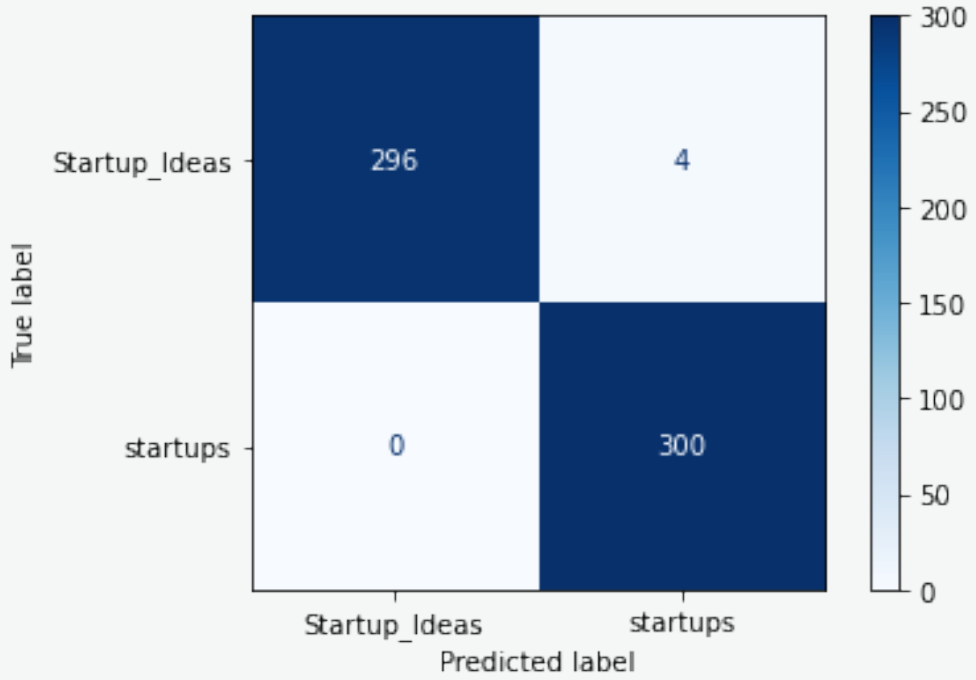
Complete all step as mentioned above

CLEAN THE DATA

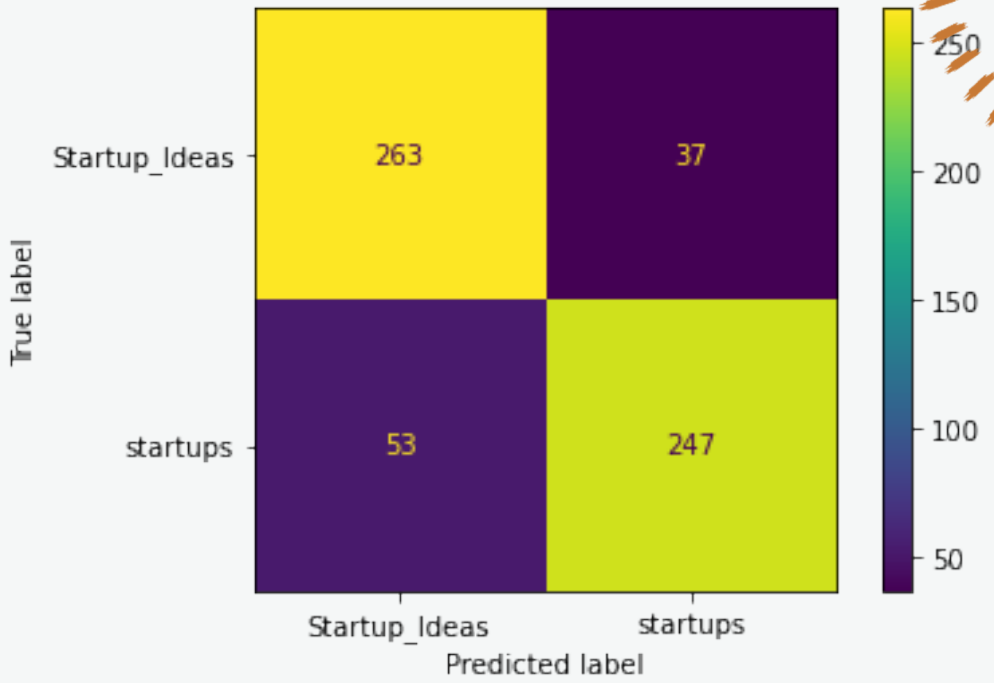
UPDATE MY EDA

RUN THE MODEL AGAIN USING MY NEW DATA

Logistic Regression



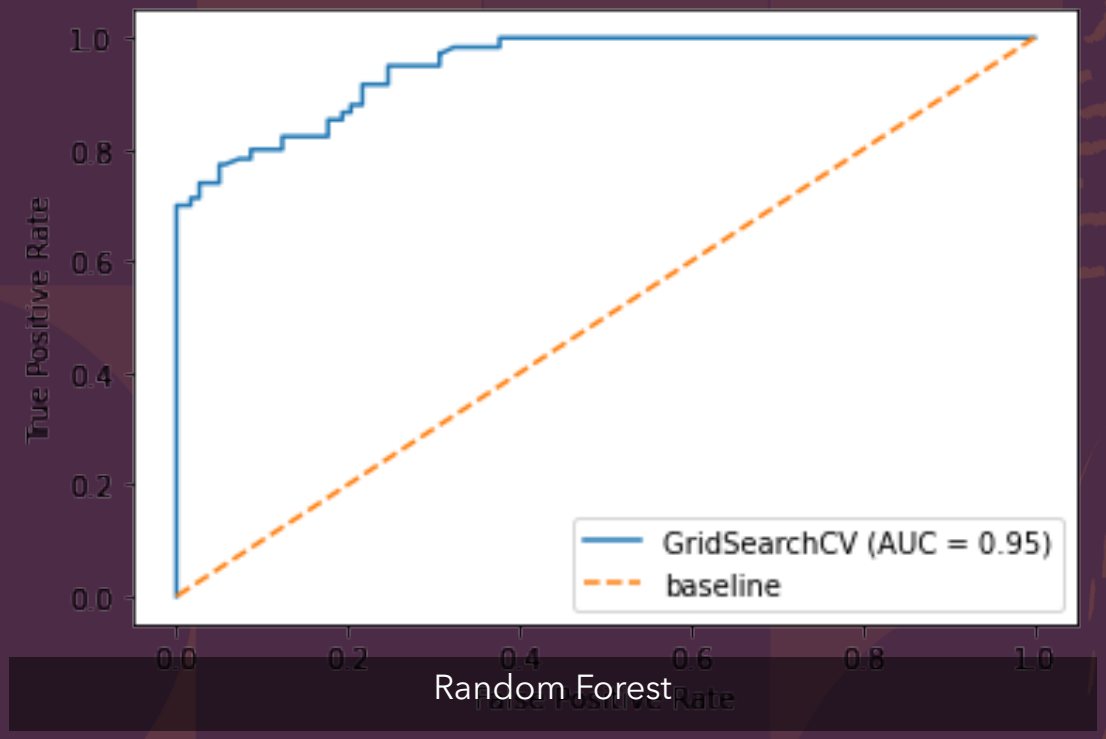
Random Forest Classification



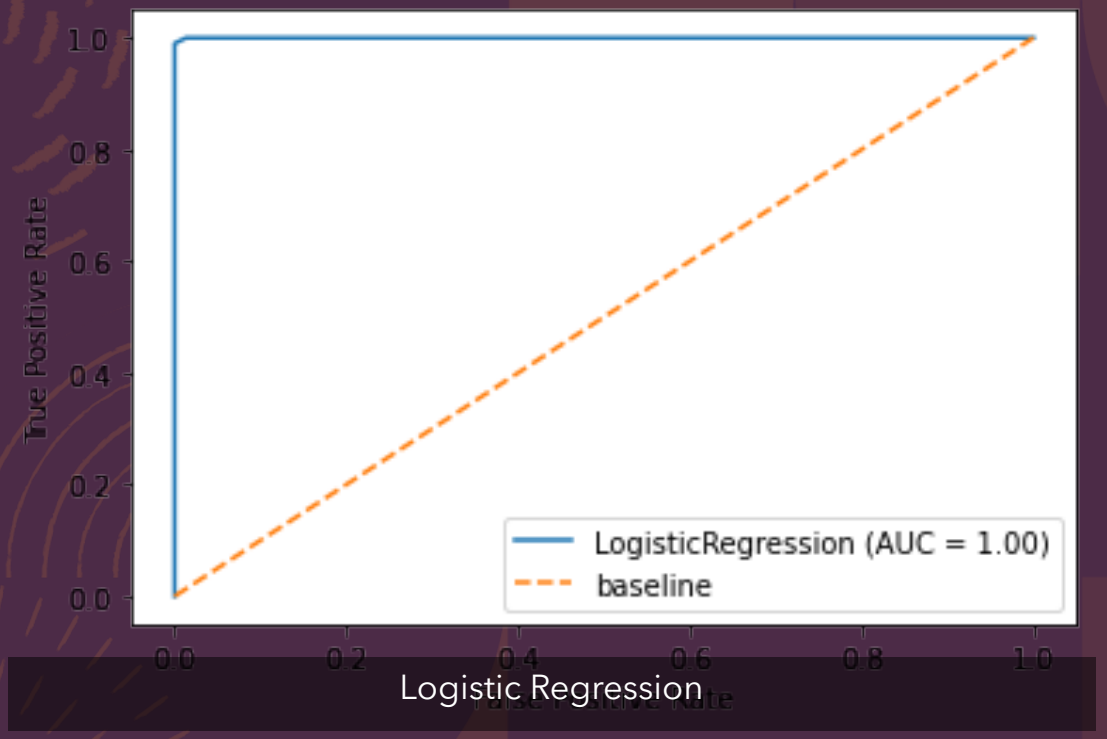
CONFUSION MATRIX

Logistic Regression Vs Random Forest

ROC Curves:



Random Forest



Logistic Regression

Results

On first try on the Logistic Regression Model I was able to do much better than my goal

Train Score: 0.99541

Test Score: 0.99333

Conclusion

- No matter how great your model is it can not cater for all possible data sets as I am seeing 99% accuracy on my final dataset while I saw an 80% on the previous set of data.
- Although I got my desired results, I am unsure how my model will perform with another set of data
- My Production model will be a toss up between my Logistic Regression and Random Forest Model. Initially the Logistic Model was performing better but after changing the params they are both performing equally as well

I would like to closely review both sets of my data to analysis what caused huge difference in the scores

- Is it that previously I did not have enough data to generate a good model
- Or was it the words used in the 2nd dataset?
- I would Merged the dataset to see what results I would get to came up with a Model that will work well on both datasets

Next Steps

STAY TUNE !

TO BE CONTINUED……