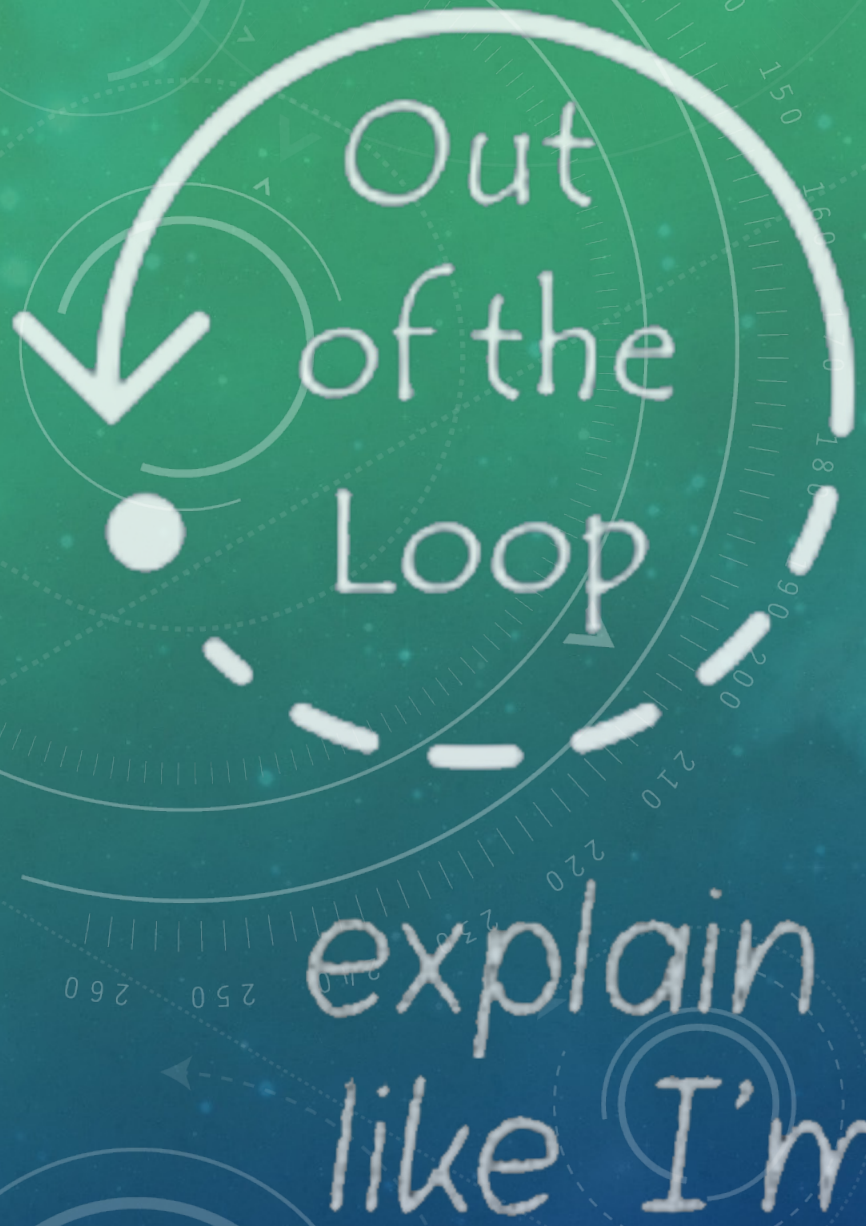



# SUBREDDITS CLASSIFICATION

PROJETC #3

# PROBLEM STATEMENT

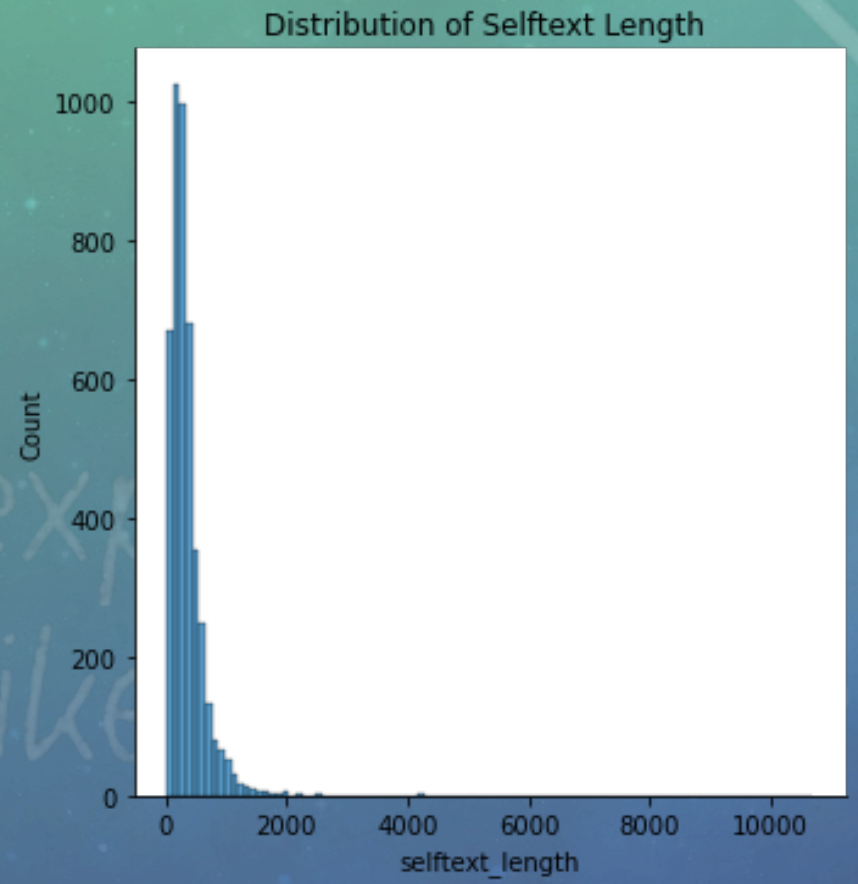
- **GOAL**: to develop a predictive model to classify posts from two different subreddits
- **USE CASE:**
  - I'm a data scientist at the reddit and In these specific scenario, I'm trying to build a model which could predict the posts from these two different subreddits.
  - The intention of the board, afterwards, is directing ads to a specific user profile of each subreddits.
- **SUBREDDITS**: Out of the Loop vs. Explain Like I'm Five
  - not so different from each other / aware of not overlapping much terms;

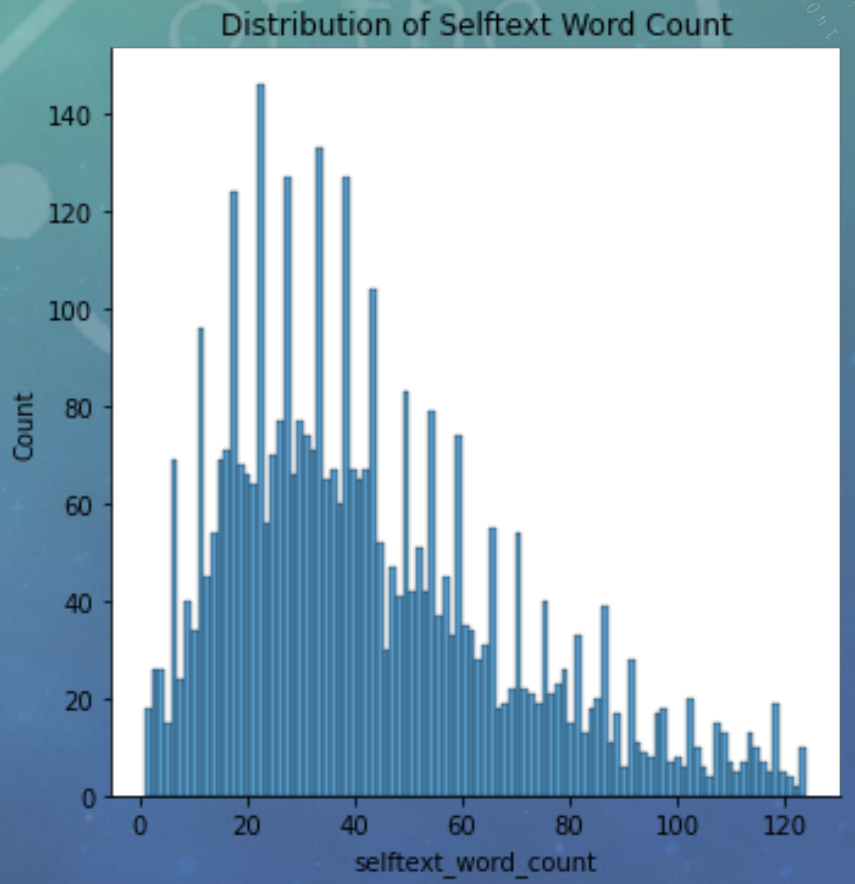
# PROBLEM STATEMENT

- **AUDIENCE:**
  - In this case, we would have as primary and secondary audience, owners of reddit sites and language course companies, respectively.
- **METRICS:**
  - F1 SCORE / ACCURRACY

# DATA CLEANING AND EDA

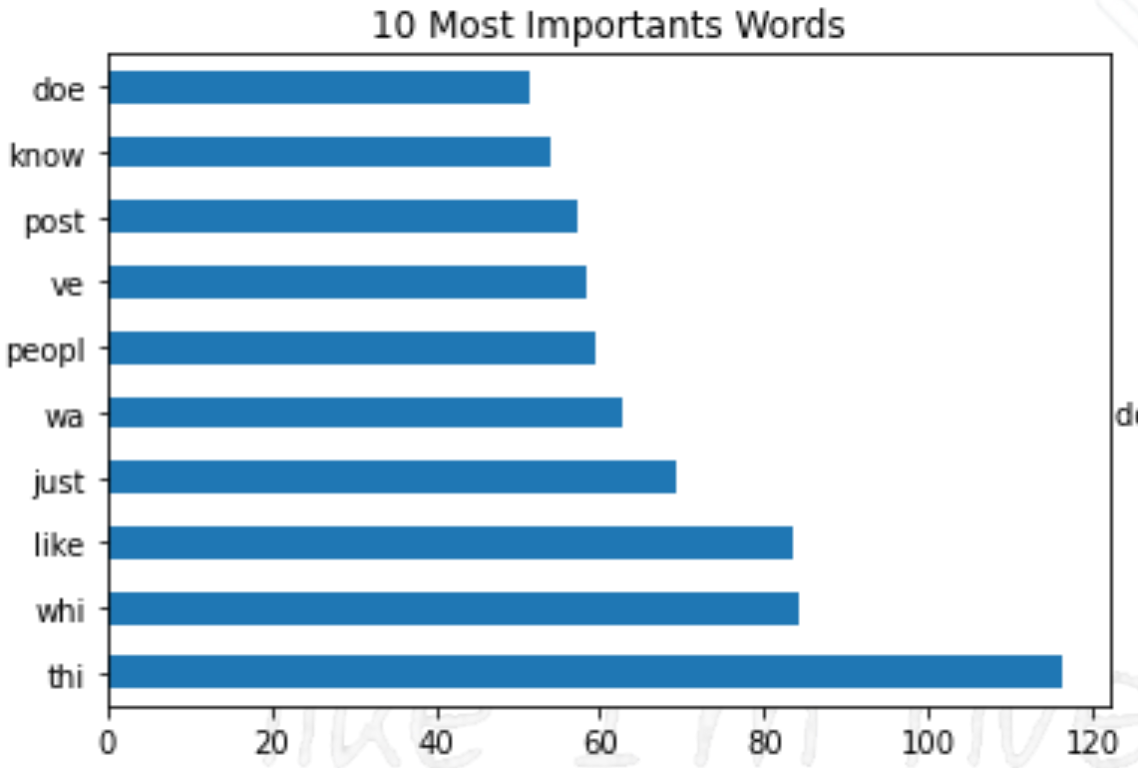
## DISTRIBUTIOS OF DOCUMENTS AFTER FEATURE ENGINEERING

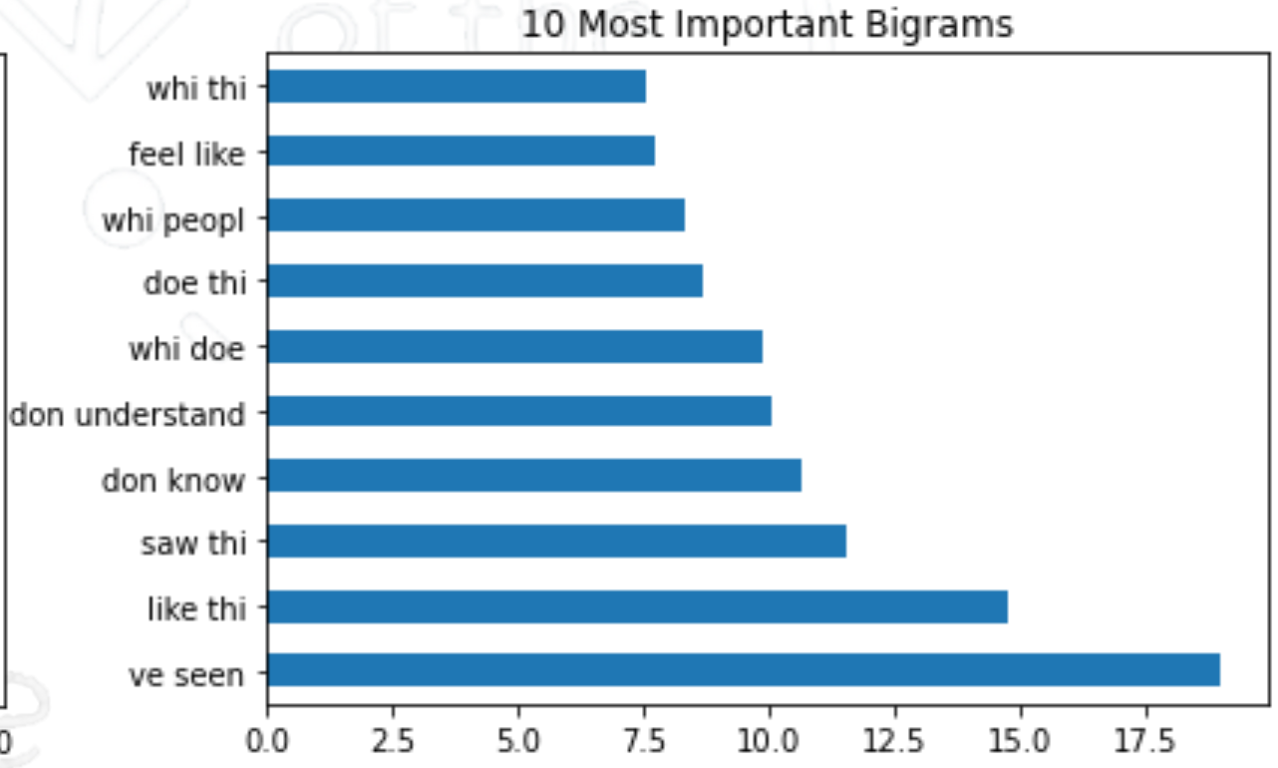


# PREPROCESSING

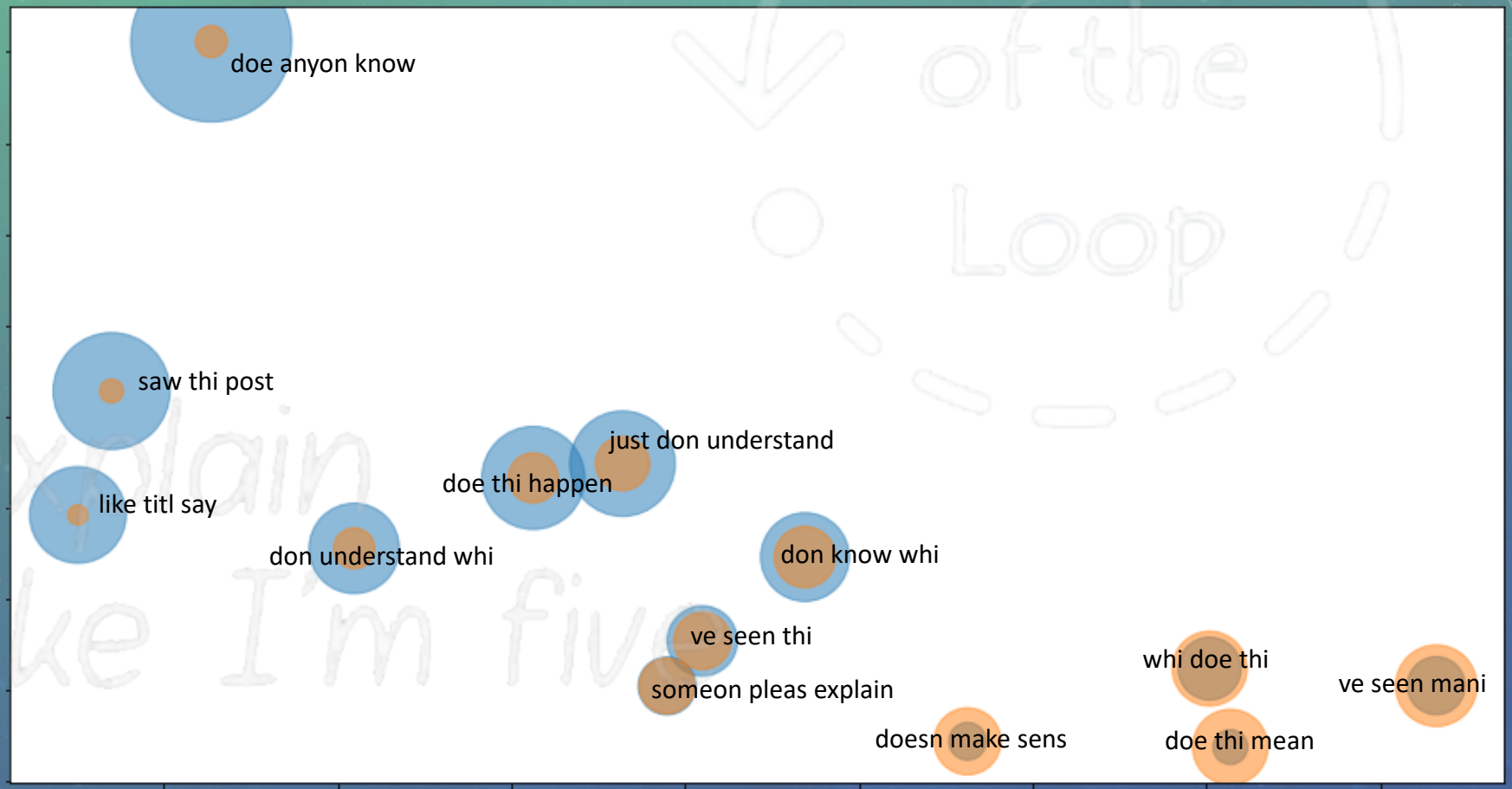
## MOST IMPORTANTS WORDS AFTER VECTORING CORPUS

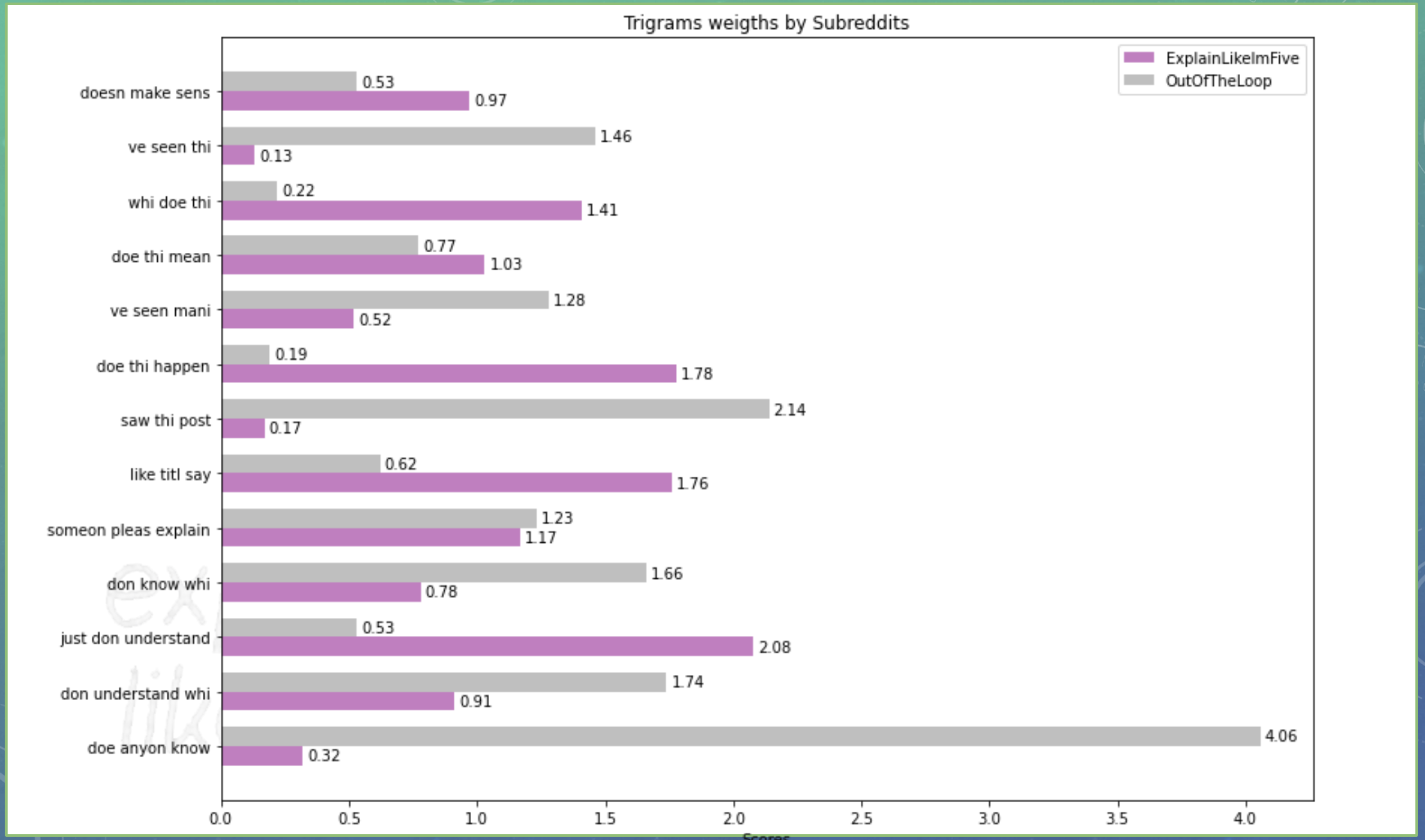


# PREPROCESSING

## 13 TRIGRAMS OCCURING IN BOTH SUBREDDITS




Trigrams weigths by Subreddits
ExplainLikeImFive
OutOfTheLoop
doesn make sens
0.53
0.97
ve seen thi
1.46
0.13
whi doe thi
0.22
1.41
doe thi mean
0.77
1.03
ve seen mani
1.28
0.52
doe thi happen
0.19
1.78
saw thi post
2.14
0.17
like titl say
0.62
1.76
someon pleas explain
1.23
1.17
don know whi
1.66
0.78
just don understand
0.53
2.08
don understand whi
1.74
0.91
doe anyon know
4.06
0.32
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Scores

# MODELS AND EVALUATION

| MODEL EVALUATION | TRAIN | TEST | BASELINE | F1 SCORE | ACCURACY |
|---|---|---|---|---|---|
| **MULTINOMIALNB** | **0.9460** | **0.9020** | **0.5386** | **0.9111** | **0.9020** |
| **LOGISTIC REGRESSION** | **0.9496** | **0.9000** | **0.5386** | **0.9076** | **0.9000** |
| ADA BOOST NB | 0.9573 | 0.8820 | 0.5386 | 0.8907 | 0.8820 |
| ADABOOST ET | 0.9350 | 0.8750 | 0.5386 | 0.8817 | 0.8750 |

## MULTINOMIAL NAÏVE BAEYS

```
Best Parameters: {
'tvec__max_df': 0.9,
'tvec__max_features': 5000,
'tvec__min_df': 3,
'tvec__ngram_range': (1, 2),
'tvec__stop_words': None}

Train Score: 0.946
Test Score: 0.902
F1 Score: 0.9111
Accuracy: 0.902
```

## LOGISTIC REGRESSION:

```
Best Parameters:
{
'lr__penalty': 'l2',
'tvec__max_features': 4000,
'tvec__min_df': 4,
'tvec__ngram_range': (1, 2),
'tvec__stop_words': None
}

Train Score: 0.9496
Test Score: 0.9
F1 Score: 0.9076
Accuracy: 0.9
```

# CONCLUSIONS AND RECOMMENDATIONS

- Looking at the metrics in the training and test sets, we can verify that both models are **not overfitting** or have high bias
- and based on the **F1 score metric** (and also accuracy), we can say that both models perform very well in predict  these two subreddits.
- Another conclusion is that there doesn't seem to be much **overlap in terms**, otherwise it wouldn't perform as well.

# FUTURE WORKS

**TEST VALIDATION :**

- Get more data from each subreddit and placement a test  in unseen data;

**ADD NEW INFORMATION :**

- Adding information  to the model from links extracted from the documents (YouTube links);

**DIFFERENT PROBLEM STATEMENT:**

- a scenario where it makes sense to rank one class more than the other. For example, consider that the ELI5 profile is a person who is learning the English language and needs explanations about a certain subject in a simple and lay language.