# Reddit listing classification

DSI course

Jan 2024

**By: Masoud Alfi**

# Problem statement

- **Human language into numbers**
- **Interpret language and predict the context (subreddit classification)**

- **Challenges?**
- **Why is it important?**
- **Who benefits?**

# Outlines

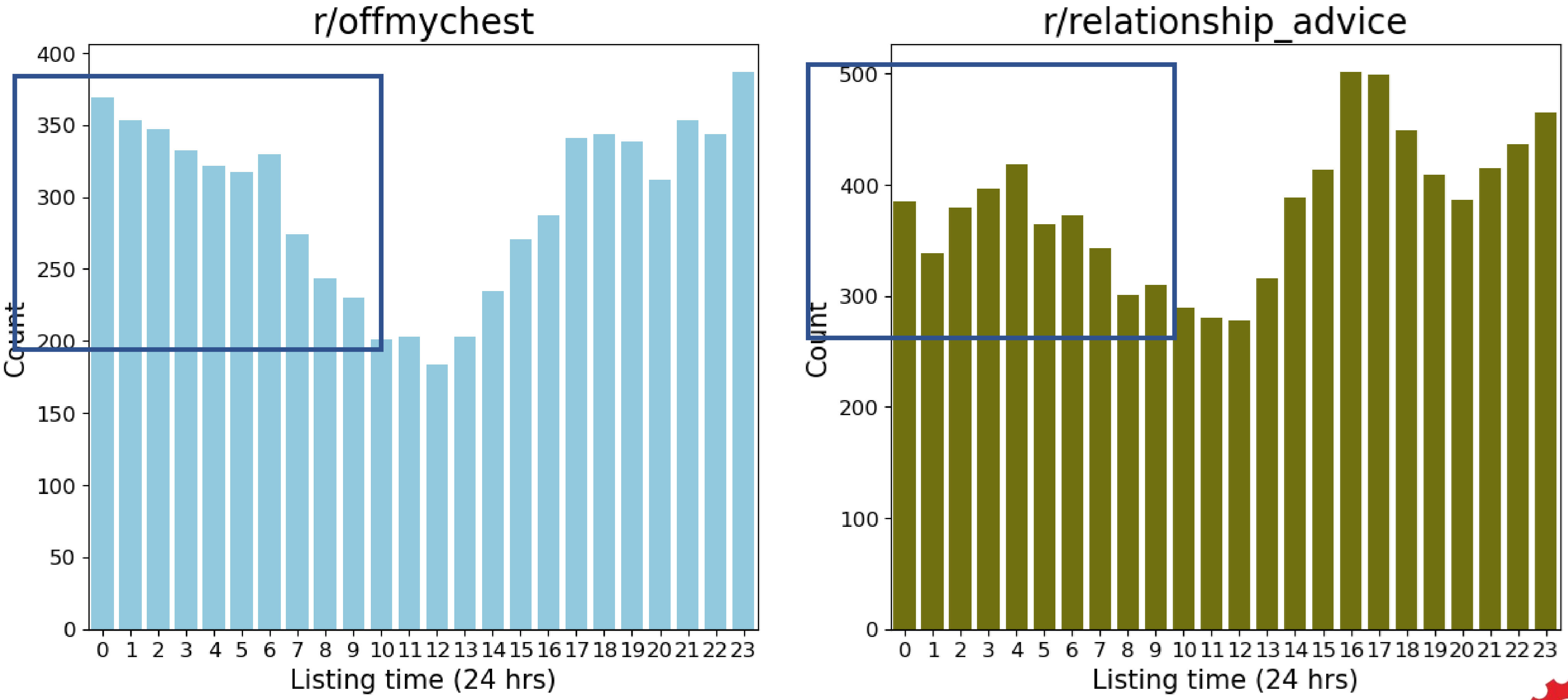
- Out data
- Exploratory data analysis
- Model benchmarking
- Model comparison
- Balanced vs. imbalanced data
- Conclusions

# Our data

- Reddit API's used
- 16,000+ listings
- Information extracted:
  - Listing and title
  - urls and media
  - Date and time

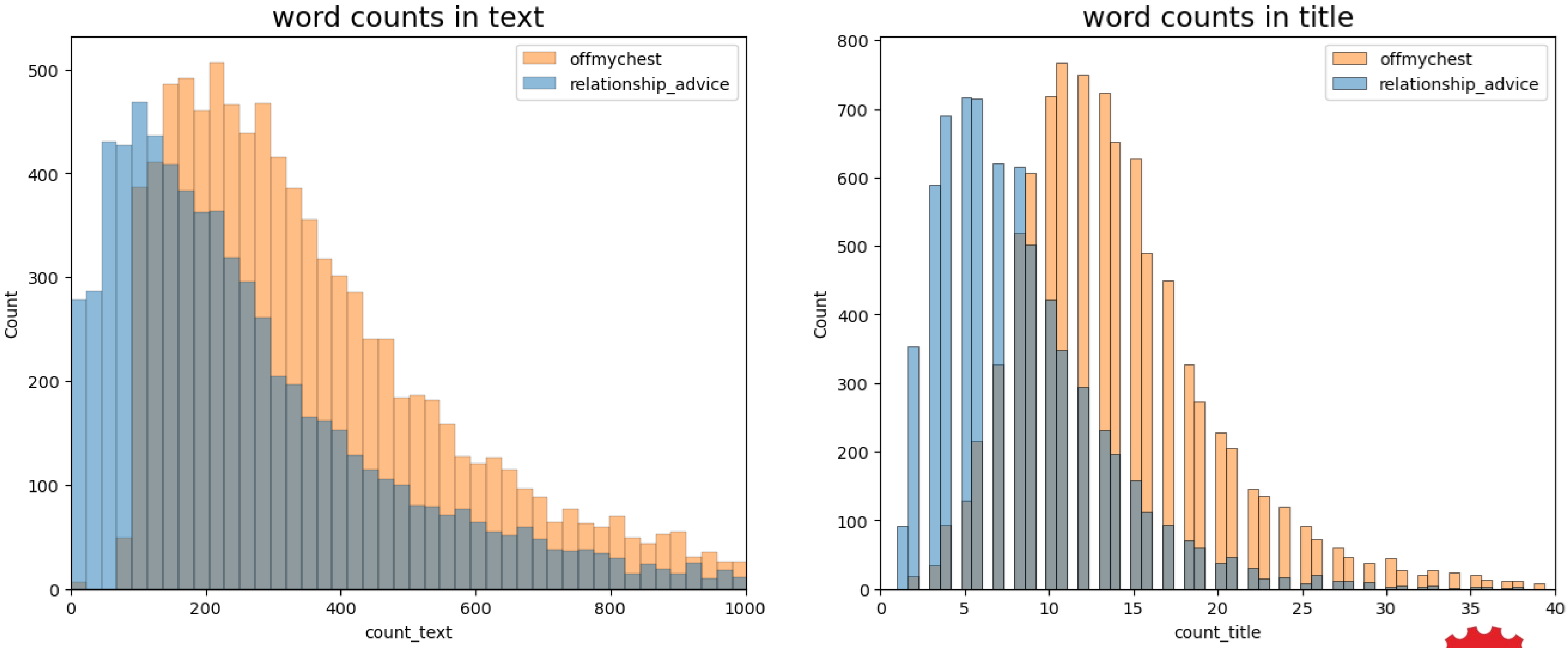
- **Subreddits:**
  - r/offmychest
  - r/relationship_advice
- No severe imbalance
- **Why these options?**

# EDA-Listing Time



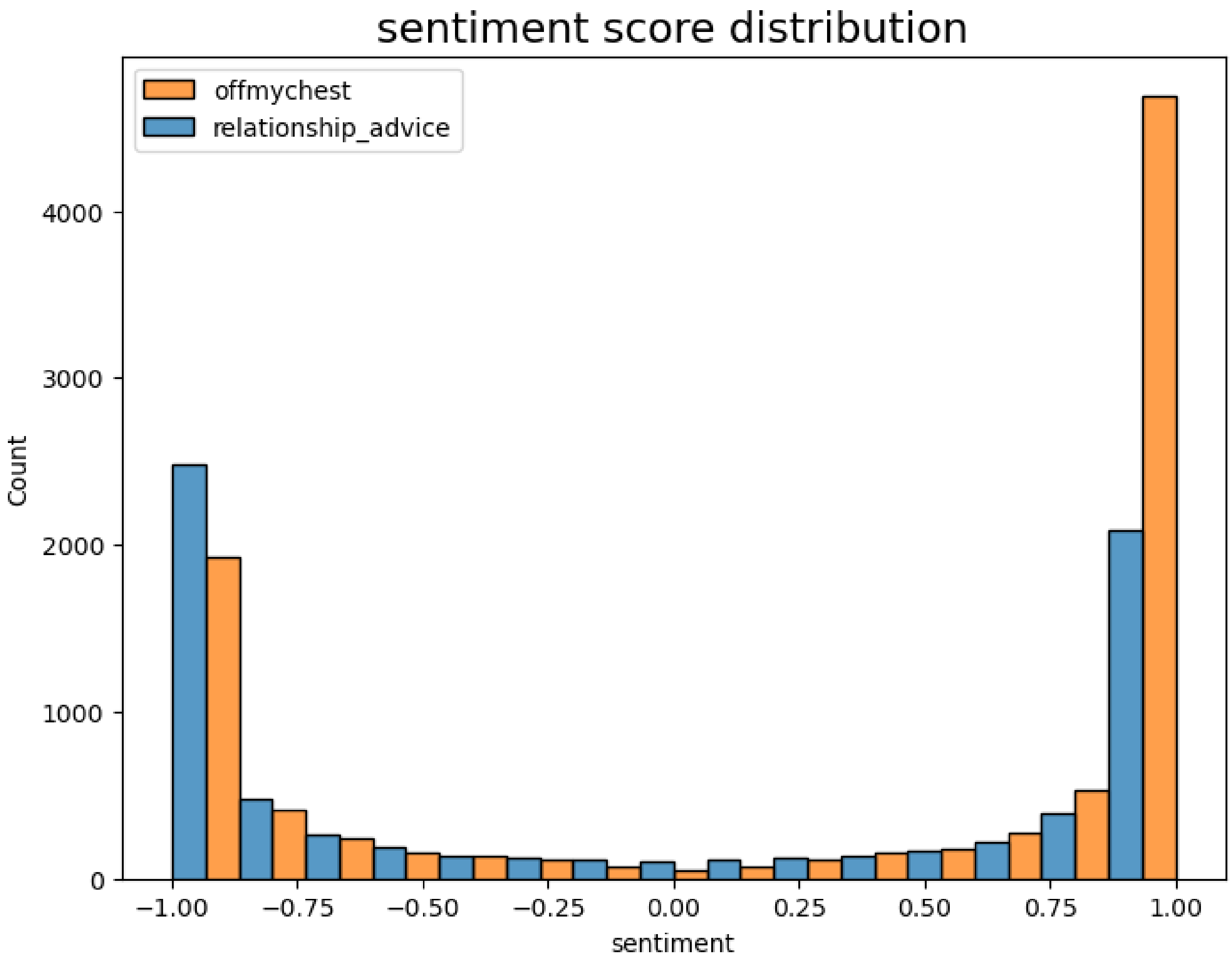
- Needs data collection in longer time spans

# EDA-Word Counts



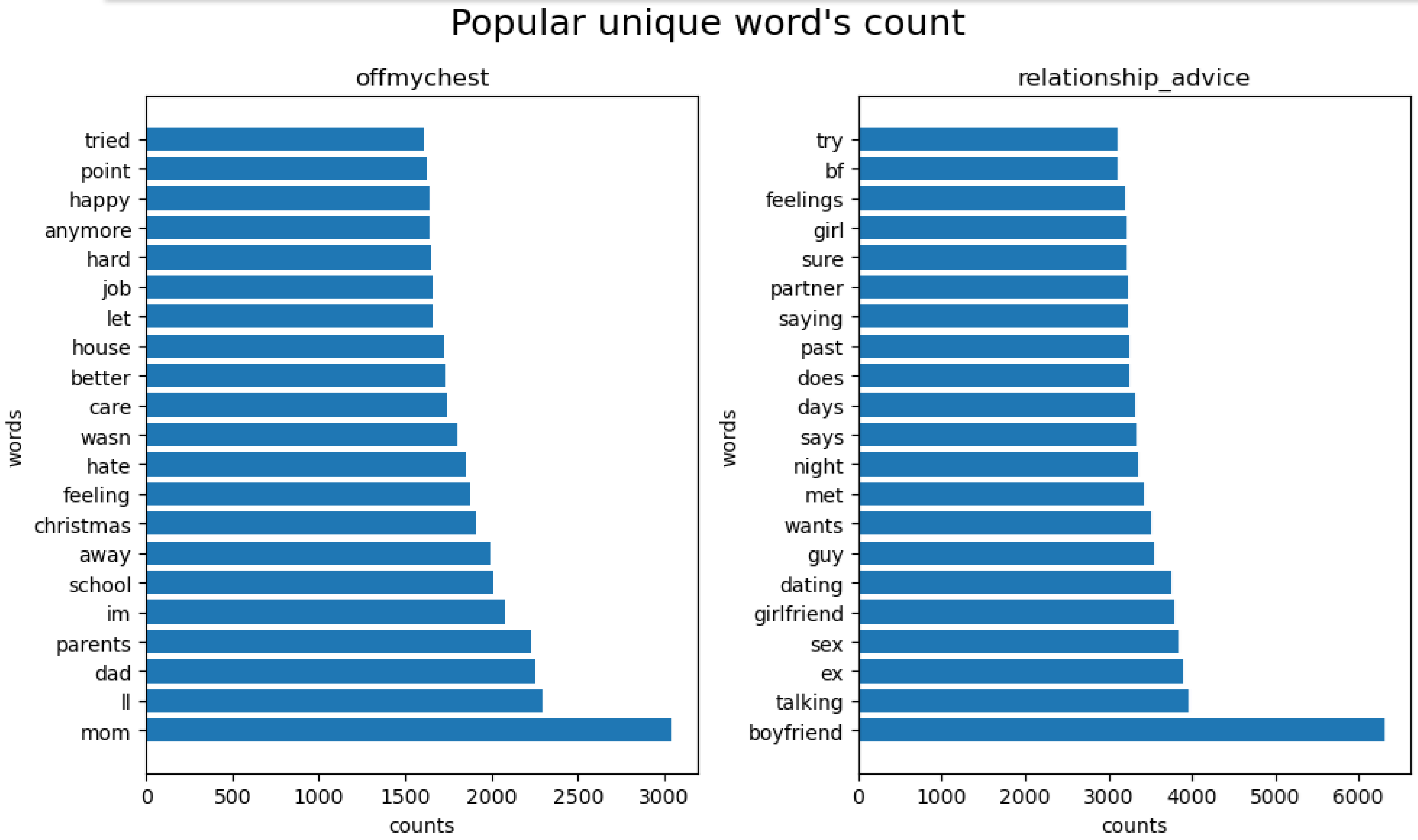
- Clear differences in words counts

# EDA-Sentiment Scores



- r/offmychest has a higher positive ratio

# EDA-Exclusive Popular Words



- 75% overlap in popular words

# Model Benchmarking

- Null model
- Model based on numeric features
  - Sentiment
  - Word counts
- Base NLP model
  - CountVectorizer
  - Logistic Regression (Regularized)

| Model | Accuracy score |
|---|---|
| Null model | 0.56 |
| Numeric model | 0.65 |
| Base NLP model | 0.88 |

- **Why accuracy score?**

# Model Comparison

- Logistic Regression
- KNN
- Naïve Bayes
- Random Forest

- **The odd case of high variance**

| Model | Accuracy score |
|---|---|
| Log Reg | 0.88 |
| Naïve Bayes | 0.85 |
| Random Forest | 0.84 |
| KNN | 0.76 |

# Balanced vs. Imbalanced Data

- LogReg (no class weighting)

| Score | Balanced data | Imbalanced data |
|---|---|---|
| Accuracy | 0.89 | 0.92 |
| Precision | 0.9 | **0.5** |
| Recall | 0.9 | **0.28** |
| f1-score | 0.9 | **0.36** |

- LogReg (with class weighting)

| Score | Imbalanced data |
|---|---|
| Accuracy | 0.91 |
| Precision | **0.4** |
| Recall | **0.6** |
| f1-score | **0.48** |

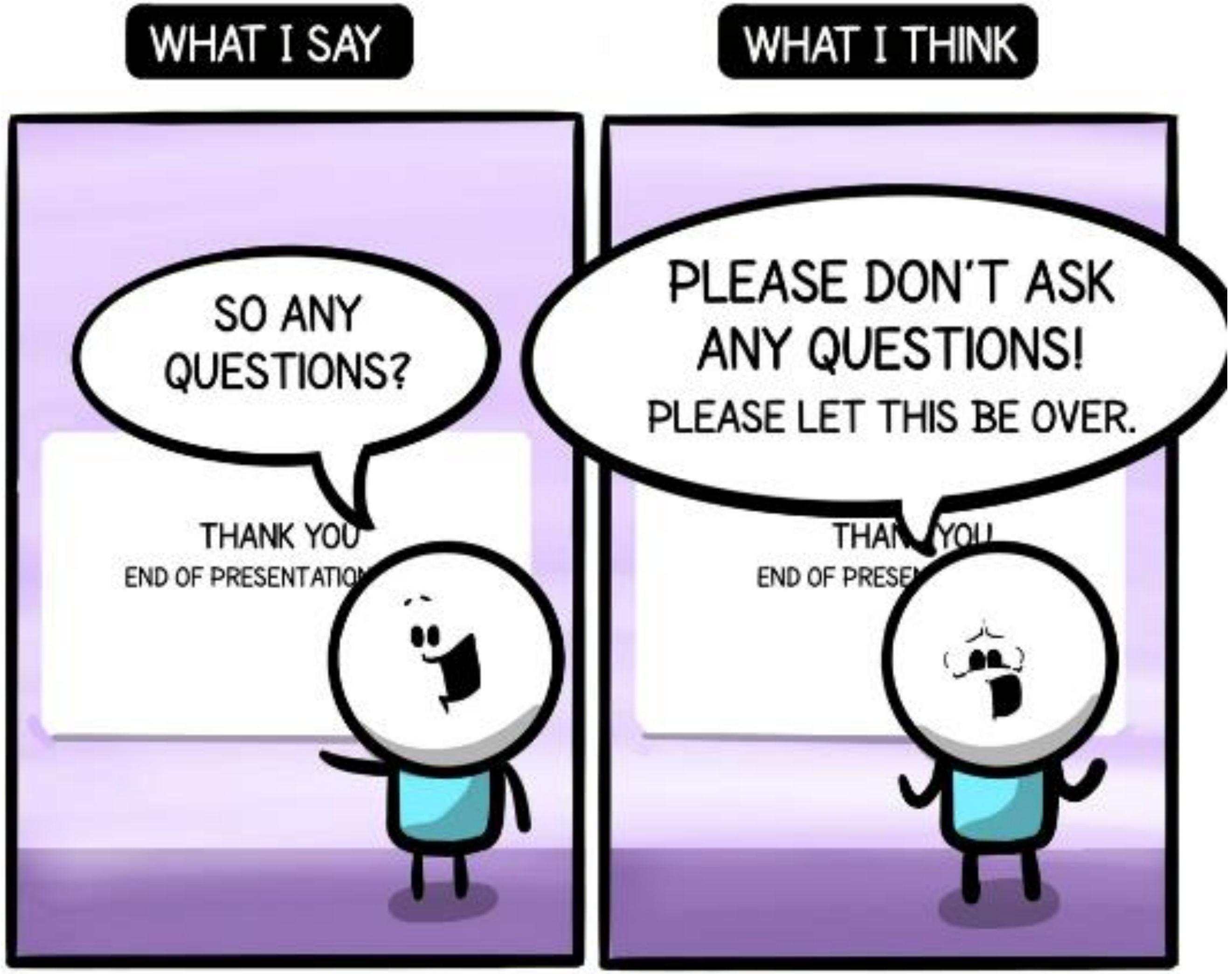
- Data imbalance ratio (94 to 6)

# Conclusions

- Best model showed 88% ‘accuracy’ in classification.
- Logistic regression outperformed other estimators.
- Classifiers could easily go into the overfitting territory.
- Imbalance classes pose challenges for our classifiers.

**Future work**

- Consider words in the context of sentences and relations (LM)



from reddit.com