Classifying Reddit Posts into their Subreddits

Bala Krishnamoorthy

Contents

- Facts about “the front page of the internet”
- Problem Statement
- Methodology
- Modelling: Process
- Modelling: Performance
- Conclusion
- Next Steps

Facts about “the front page of the internet”

- Widespread popularity
  - 330 million monthly active users
    - ~62% of users visit Reddit for news
  - 160,000 pages are viewed per minute
  - More than 1 million subreddits (June 2017)
- Founded by Alexis Ohanian and Steve Huffman
  - Initially popularized the site by creating many fake profiles and comments
- A place of wonder and chaos
  - r/secretsanta (world’s largest)
  - r/showerthoughts
  - r/ roastme

Problem Statement

Given a post from one of two subreddits, can we build a model that predicts which subreddit the post came from?

Post: "Watching a graduation ceremony is like sitting through an entire movie that's end credits"

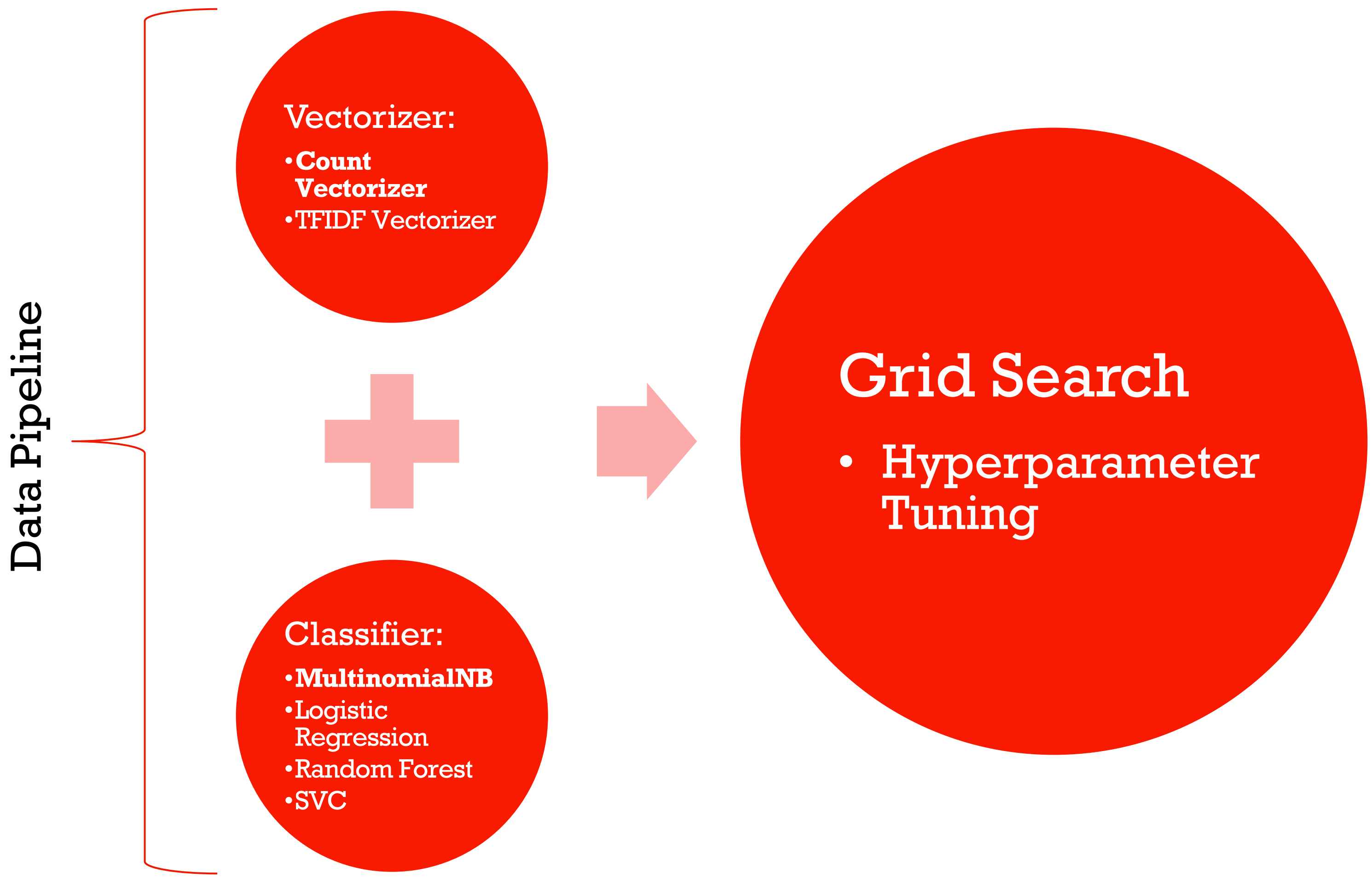
Subreddits: r/showerthoughts r/DeepPhilosophy

Why solve this problem? To evaluate the strengths and weaknesses of NLP classification tools.

Methodology

1. Data gathering
   - Access reddit's API, and collect reddit posts
2. Data cleaning
   - Extract text content from each reddit post
3. Modelling
   - Build helper functions
   - NLP: text vectorization
   - Run and evaluate several classifiers
4. Visualize Results

Modelling: Process



Data Pipeline

Vectorizer:

- **Count Vectorizer**
- TFIDF Vectorizer

Classifier:

- **MultinomialNB**
- Logistic Regression
- Random Forest
- SVC

Grid Search

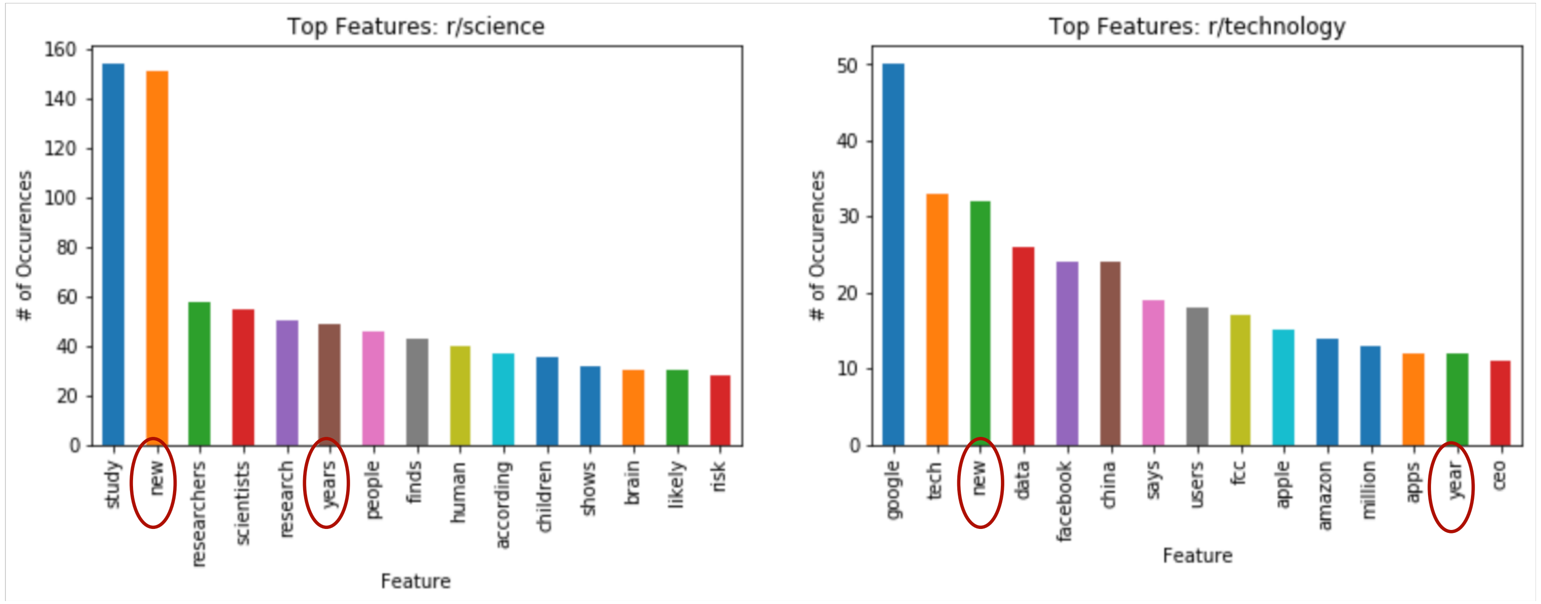
- Hyperparameter Tuning

Modelling: Performance (part 1)

r/science vs r/technology

Baseline Accuracy: 62%

Score Type	CV + LogReg	CV + RandomForest	CV + MultinomialNB	TFIDF + SVC
Cross-Val	92%	88%	**93%**	91%
Test	91%	88%	**92%**	92%

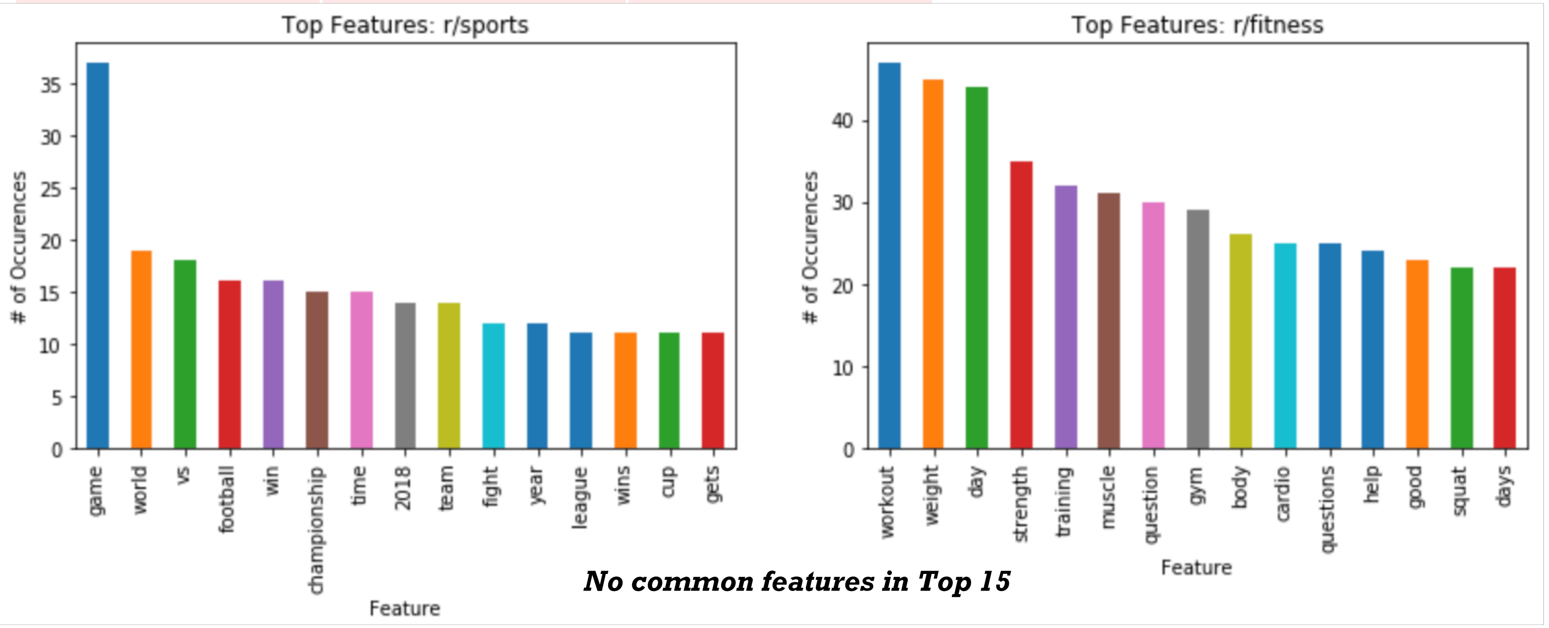


Modelling: Performance (part 2)

r/sports vs r/fitness

Baseline Accuracy: 64%

Score Type	CV + MultinomialNB	TFIDF + SVC
Cross-Val	**96%**	95%
Test	**96%**	95%



No common features in Top 15

Conclusion

- NLP is powerful
  - Relatively simple pipelines can produce accurate, reliable results
- As one would expect, the model performs better when there are more distinct differences between subreddits

Potential Improvements

- Test the model on *very* similar subreddits
  - And compare against human performance! ☺
- In this project, NLP was conducted solely on the "title" block within the post
  - We can expand our NLP curtain to include any text not within the "title" block, comments, urls of links embedded within posts, etc.
  - Combine numerical and text features (e.g. number of comments per post and text within comments)
- Automate model selection and hyperparameter tuning