Classification Model to distinguish posts from r/jobs and r/forhire

By: Lin ZheQin

Table of Contents

Problem Statement:

Build a classification model

that is able to identify whether a post belongs to **r/forhire** and not **r/jobs** with

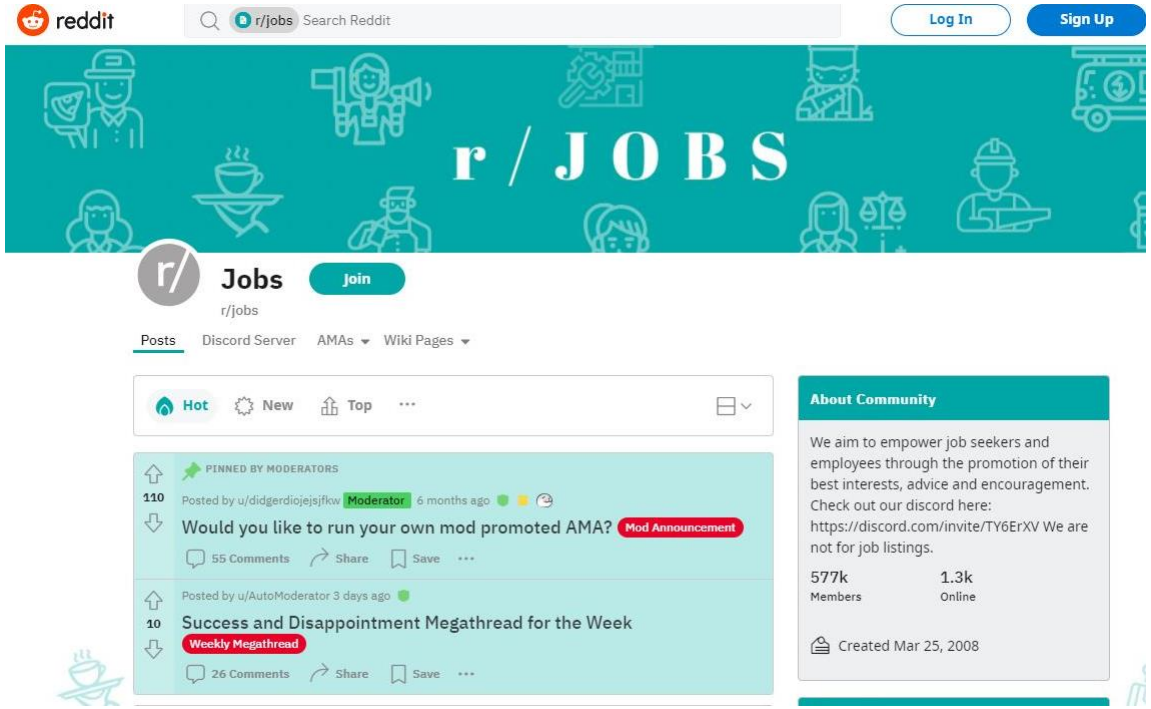
at least 90% accuracy

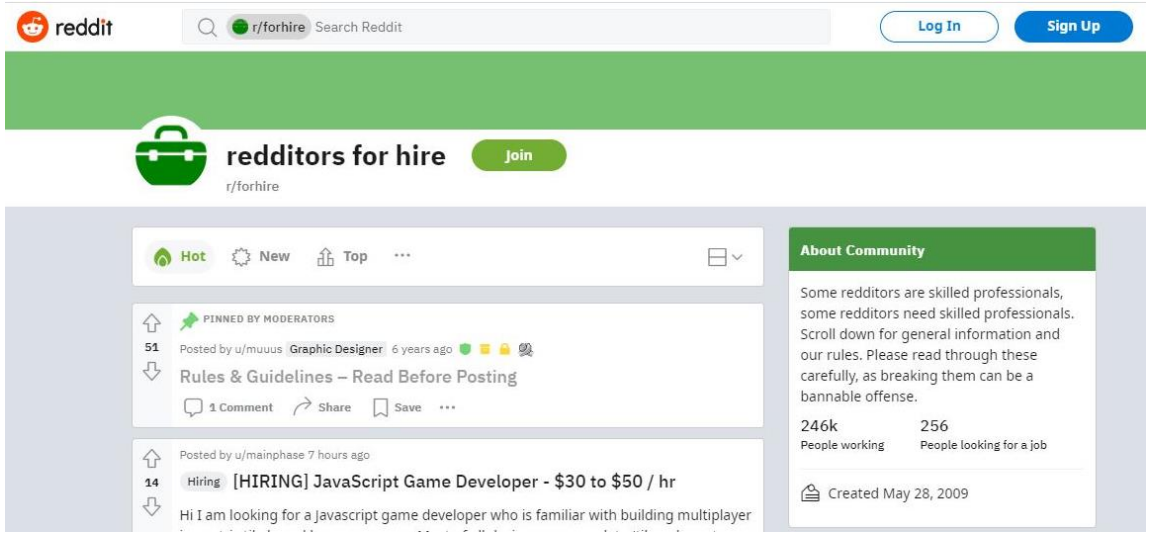
Stakeholders:

- Data science peer audience
- Job seekers
- Jobs posters

Data Gathering:

- 1,000 posts from each sub-reddit were gathered using Pushshift API
- Extracted and merged [‘title’] and [‘selftext’]
- 500 posts (Balanced classes) from each sub-reddit was used for the training set
  - ❑ r/jobs – 885 usable posts after data cleaning
    - ❖ 112 null posts (removed by reddit moderators & users)
    - ❖ 3 duplicate posts
  - ❑ r/forhire – 666 usable posts after data cleaning
    - ❖ 305 null posts (removed by reddit moderators & users)
    - ❖ 29 duplicate posts



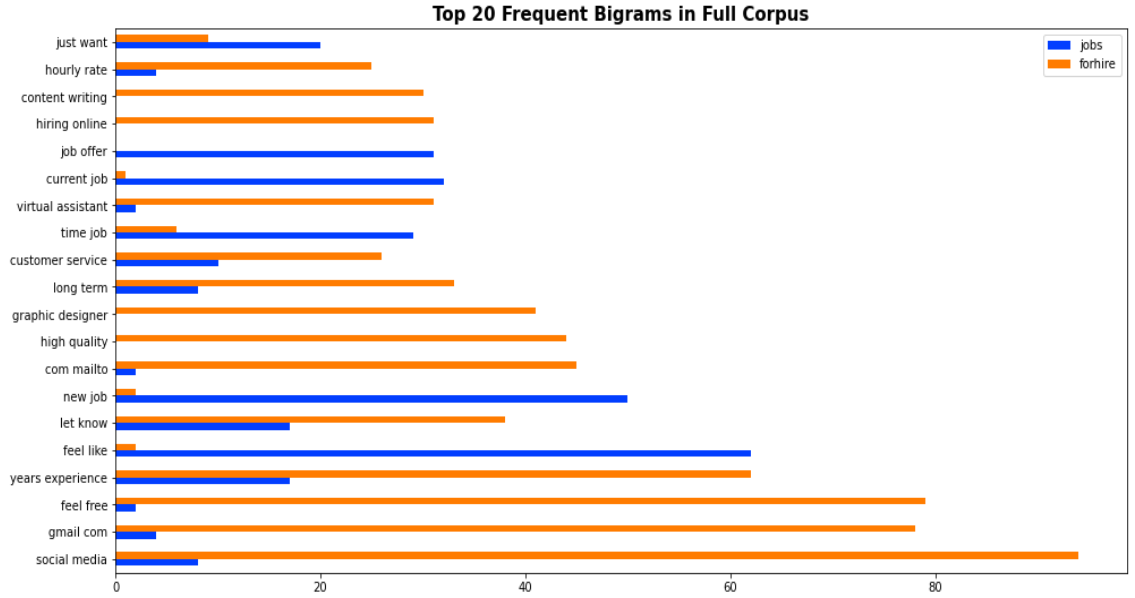


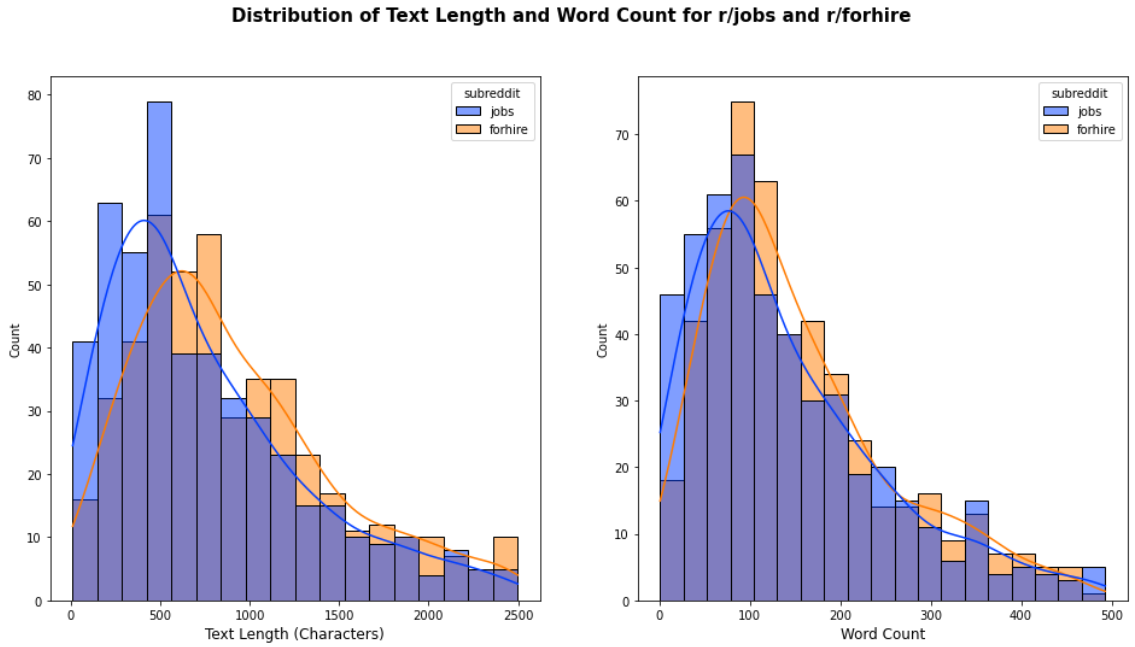
Preprocessing:

- Remove line breaks and URLs
- Tokenize & Lemmatize
- Stop Word removal

EDA:

- Bigrams (word pairs) using CountVectorizer
- Sentiment analysis with VADER
- Text length/Word count & Upvote Ratio were not very useful in distinguishing posts

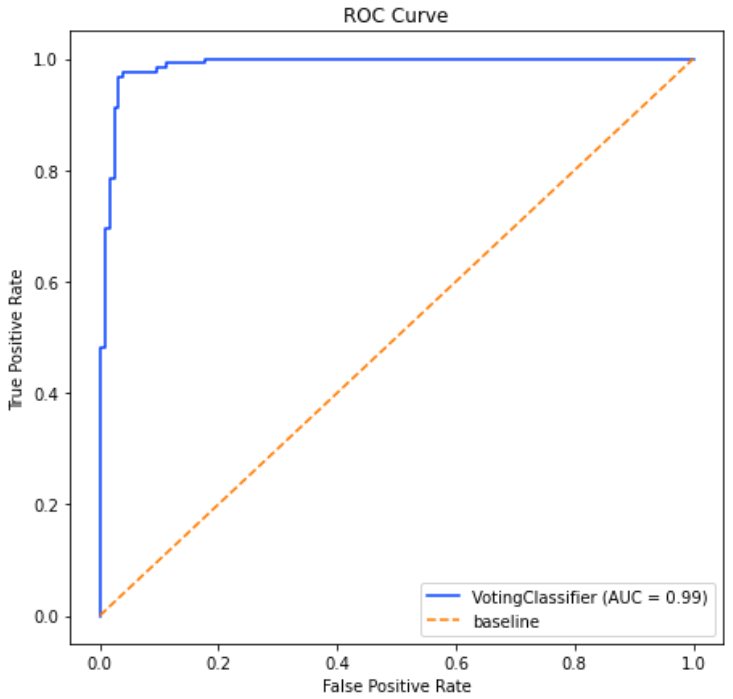


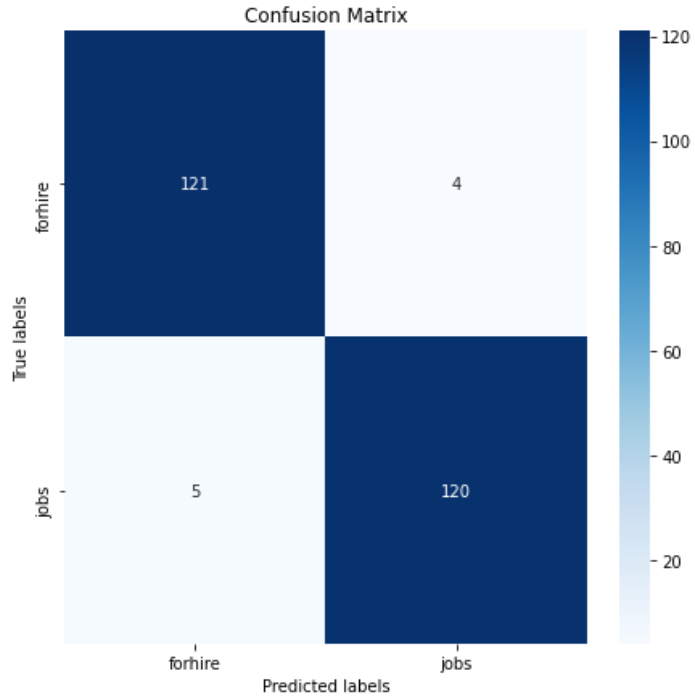


Modeling & Evaluation:

- **Models used** – Random Forest vs Voting Classifier (soft voting)
- **r/forhire** set as **True**
- **r/jobs** set as **False**
- Focus on minimizing **False Negative** *(Posts incorrectly predicted to be from r/jobs but are from r/forhire)*
- **Metrics** – F1 score, Recall & ROC AUC

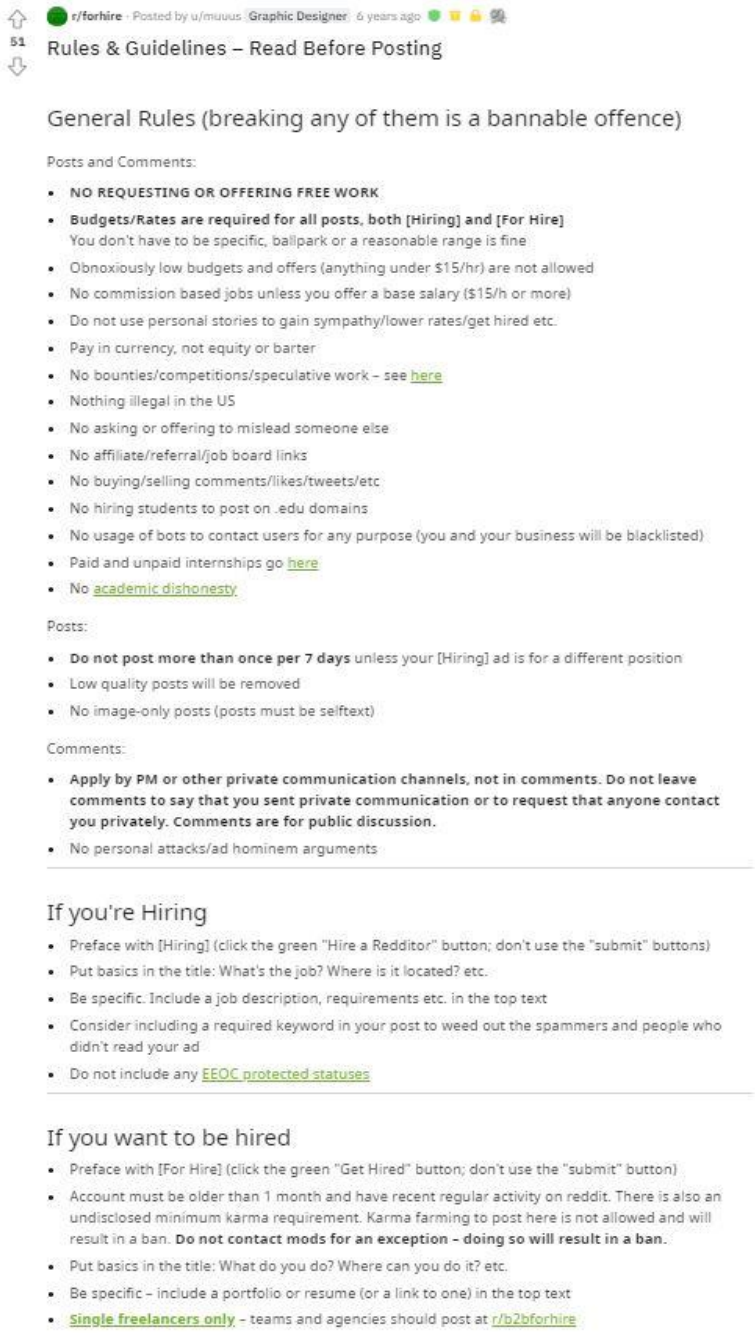
	F1 Score	Recall	ROC AUC	True Positives	False Positives	True Negatives	False Negatives
Baseline	0.66667	1.000	0.50000	125	125	0	0
Random Forest (default)	0.96356	0.952	0.98672	119	3	122	6
Random Forest (tuned)	0.95510	0.936	0.98726	117	3	122	8
MultinomialNB (default)	0.94071	0.952	0.97670	119	9	116	6
MultinomialNB (tuned)	0.94488	0.960	0.97952	120	9	116	5
SVC	0.94400	0.944	0.98618	118	7	118	7
Voting Classifier	0.96414	0.968	0.98861	121	5	120	4





Conclusion/Next Steps:

- **Voting Classifier** – **96.8%** ability to recognize **r/forhire** posts
- Scores are similar to Random Forest
  - Ensemble Method (Voting Classifier)/ Hyperparameter optimization had modest gains
  - r/forhire moderators doing a good job in reviewing and removing posts due to stricter community guidelines
- Recommendations –
  - Expand to other subreddits that concern employment *(e.g. r/careeradvice)*
  - *Get more data from other employment resources (e.g. LinkedIn, JobStreet)*



r/forhire · Posted by u/muuus Graphic Designer 6 years ago

51

Rules & Guidelines – Read Before Posting

General Rules (breaking any of them is a bannable offence)

Posts and Comments:

- **NO REQUESTING OR OFFERING FREE WORK**
- **Budgets/Rates are required for all posts, both [Hiring] and [For Hire]**
  You don't have to be specific, ballpark or a reasonable range is fine
- Obnoxiously low budgets and offers (anything under $15/hr) are not allowed
- No commission based jobs unless you offer a base salary ($15/h or more)
- Do not use personal stories to gain sympathy/lower rates/get hired etc.
- Pay in currency, not equity or barter
- No bounties/competitions/speculative work – see here
- Nothing illegal in the US
- No asking or offering to mislead someone else
- No affiliate/referral/job board links
- No buying/selling comments/likes/tweets/etc
- No hiring students to post on .edu domains
- No usage of bots to contact users for any purpose (you and your business will be blacklisted)
- Paid and unpaid internships go here
- No academic dishonesty

Posts:

- **Do not post more than once per 7 days** unless your [Hiring] ad is for a different position
- Low quality posts will be removed
- No image-only posts (posts must be selftext)

Comments:

- **Apply by PM or other private communication channels, not in comments. Do not leave comments to say that you sent private communication or to request that anyone contact you privately. Comments are for public discussion.**
- No personal attacks/ad hominem arguments

If you're Hiring

- Preface with [Hiring] (click the green "Hire a Redditor" button; don't use the "submit" buttons)
- Put basics in the title: What's the job? Where is it located? etc.
- Be specific. Include a job description, requirements etc. in the top text
- Consider including a required keyword in your post to weed out the spammers and people who didn't read your ad
- Do not include any EEOC protected statuses

If you want to be hired

- Preface with [For Hire] (click the green "Get Hired" button; don't use the "submit" button)
- Account must be older than 1 month and have recent regular activity on reddit. There is also an undisclosed minimum karma requirement. Karma farming to post here is not allowed and will result in a ban. **Do not contact mods for an exception – doing so will result in a ban.**
- Put basics in the title: What do you do? Where can you do it? etc.
- Be specific – include a portfolio or resume (or a link to one) in the top text
- **Single freelancers only** – teams and agencies should post at r/b2bforhire

Questions?