Analyzing & Classifying Subreddits

Anel Akiyanova

January, 2021

Agenda

- Goal and potential use cases
- Data collection and subreddit title text decomposition
- Polarity analysis
- Model comparison
- Conclusion and recommendations

Goal and Potential Use Cases

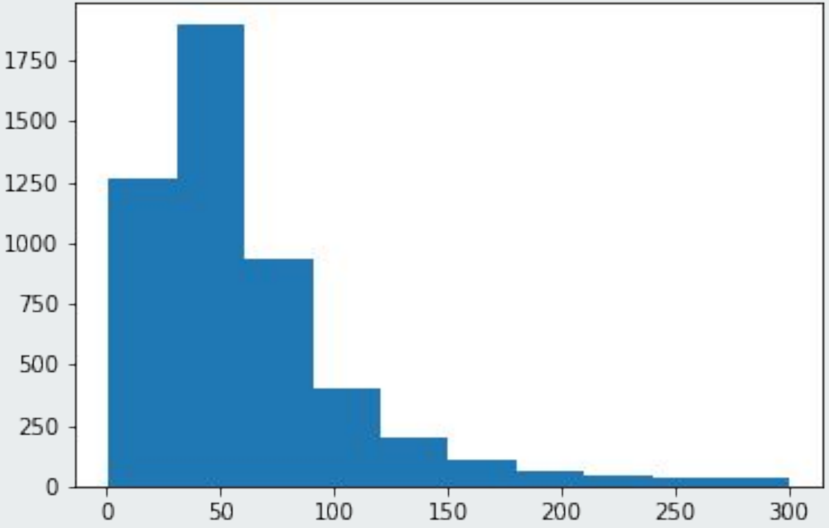
Goal:

- Perform exploratory analysis and classification of **chess**, **poker** subreddits
- Identify words and phrases that could be used in targeted advertisement

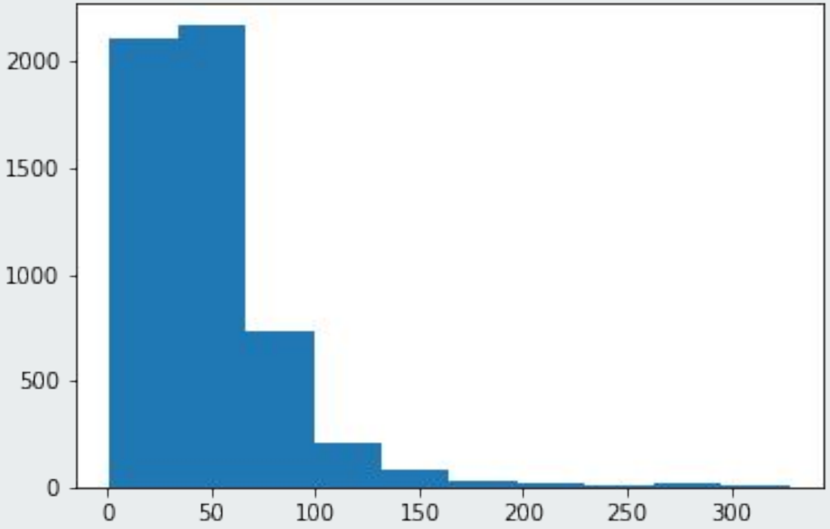
Use case:

- Use results of the analysis in targeted ads by online chess servers, chess classes, poker cardrooms (e.g. chess.com, pokerstars)

Subreddit Text Decomposition



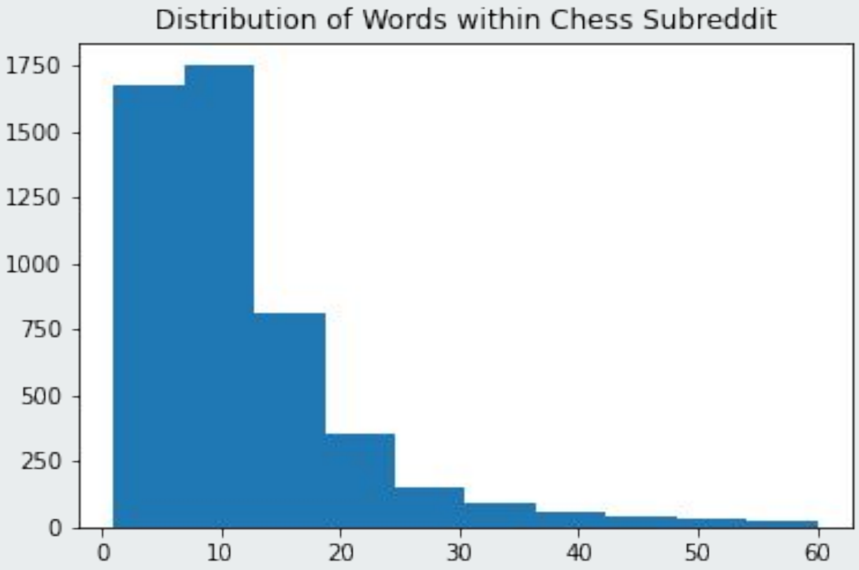
Title Length Distribution within Chess Subreddit

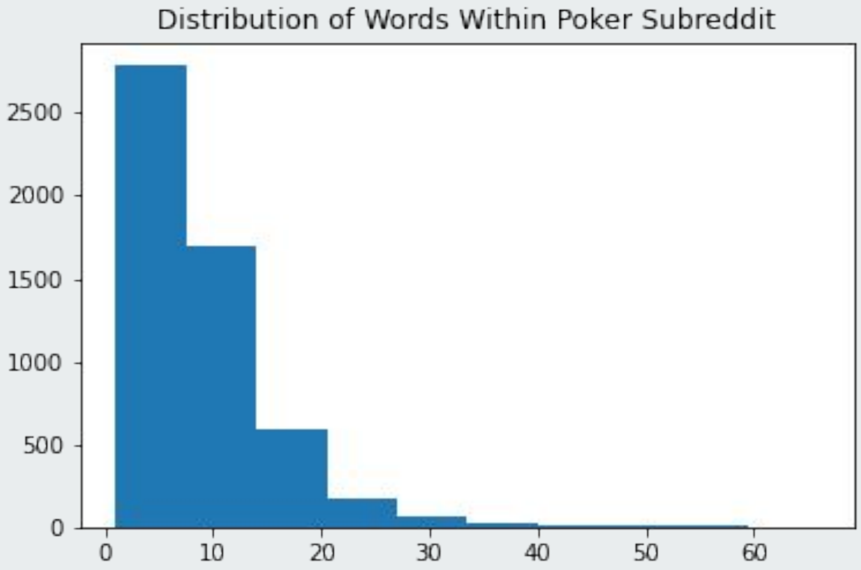


Title Length Distribution within Poker Subreddit

- Titles in the **chess subreddit** are 62 characters long on average
- Compared to 49 characters long in the **poker subreddit**

Subreddit Text Decomposition



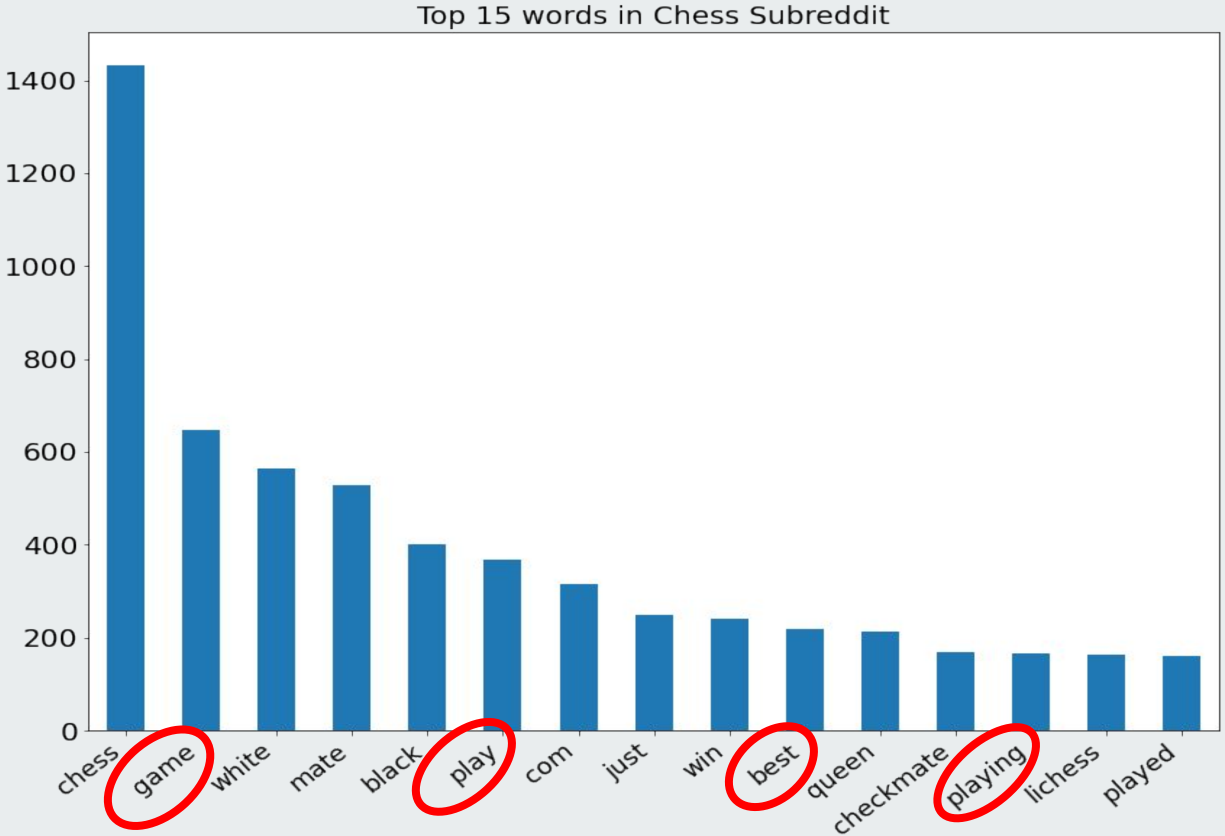


- In the **chess subreddit**, most titles contain 1-11 words
- Majority of the titles in the **poker subreddit** are 1-6 words long

Chess Subreddit: Most Common Words

- **‘chess’** most used word

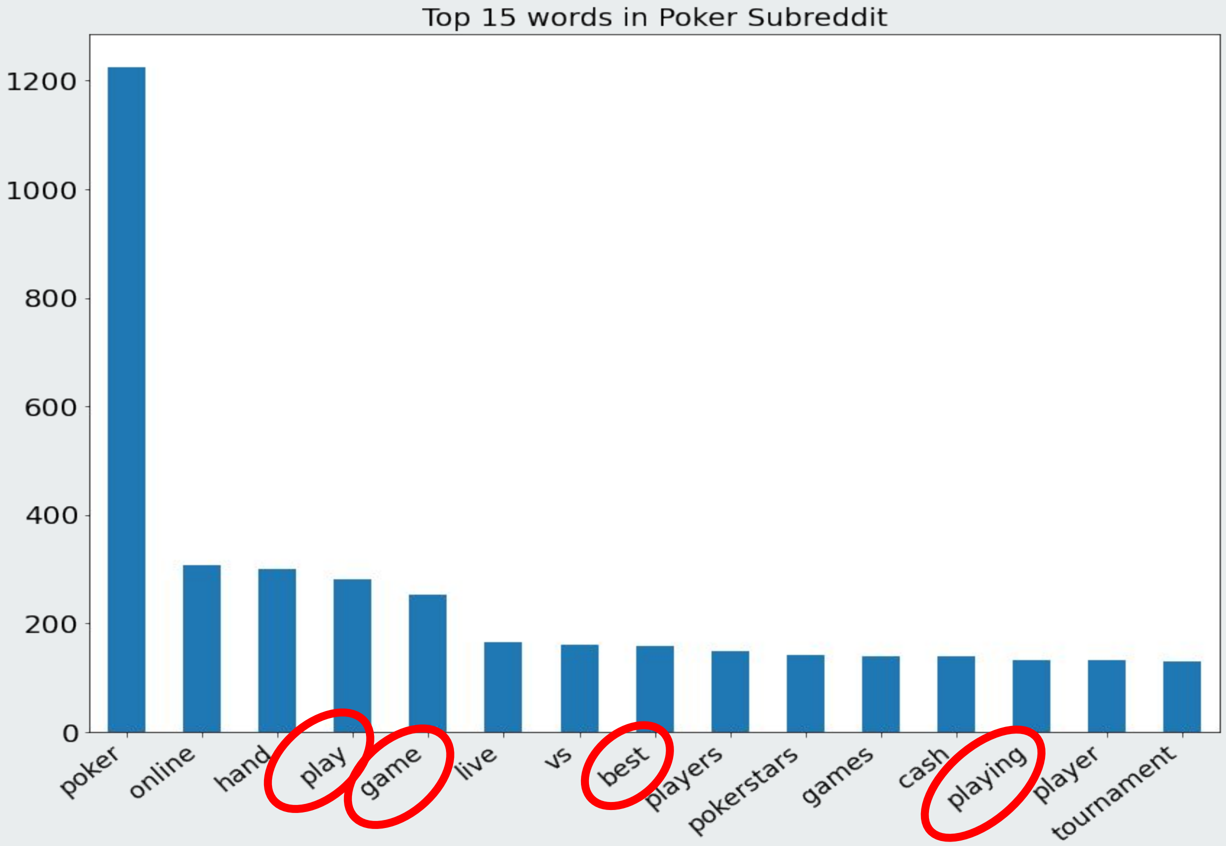
- Circled words repeat in the **poker subreddit:**
  - ❑ game
  - ❑ play
  - ❑ best
  - ❑ playing



Poker Subreddit: Most Common Words

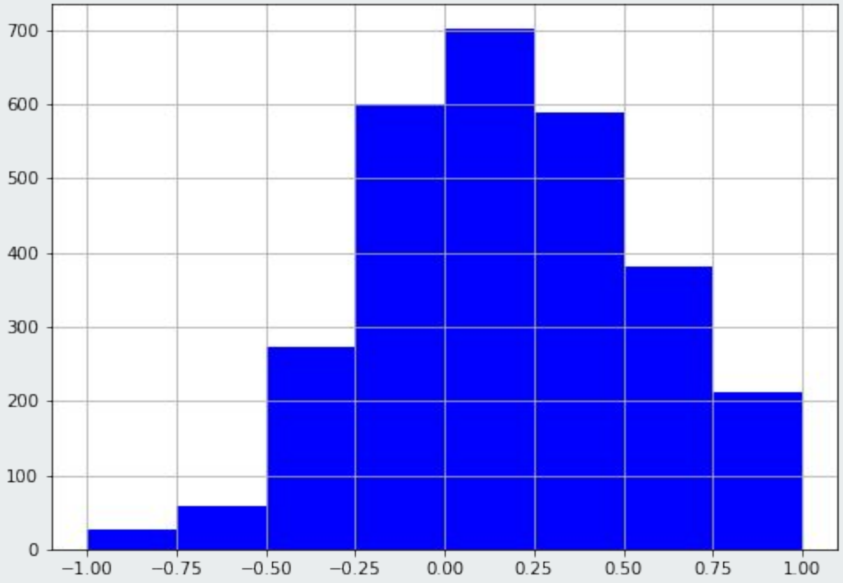
- ‘**poker**’ most used word

- Familiar words from **chess subreddit:**
  - ❏ game
  - ❏ play
  - ❏ best
  - ❏ playing

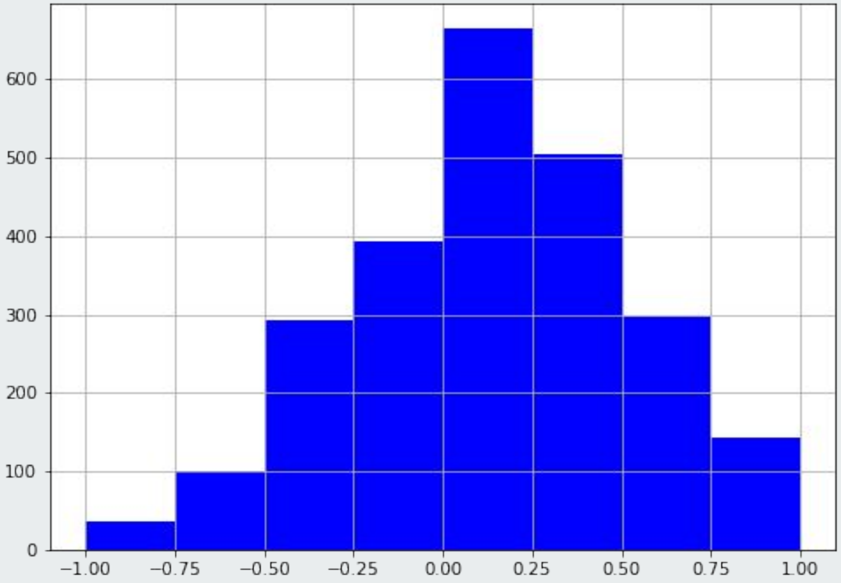


Polarity Comparison

Sentiments in the Chess Titles



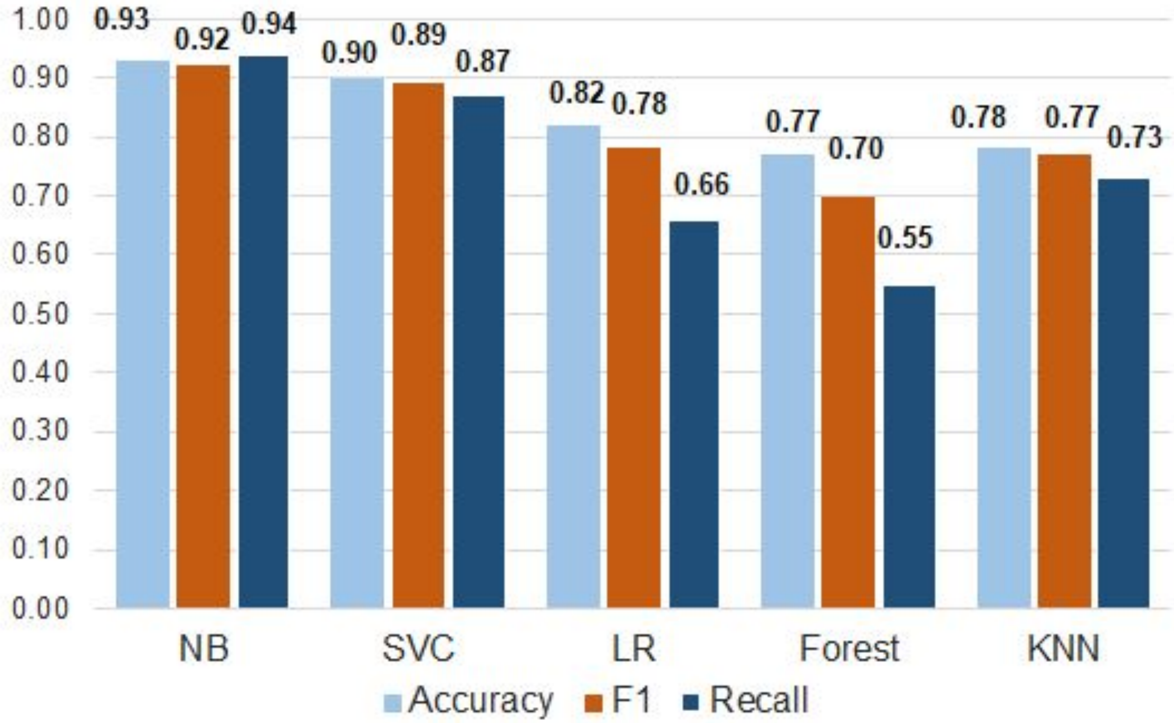
Sentiments in the Poker Titles



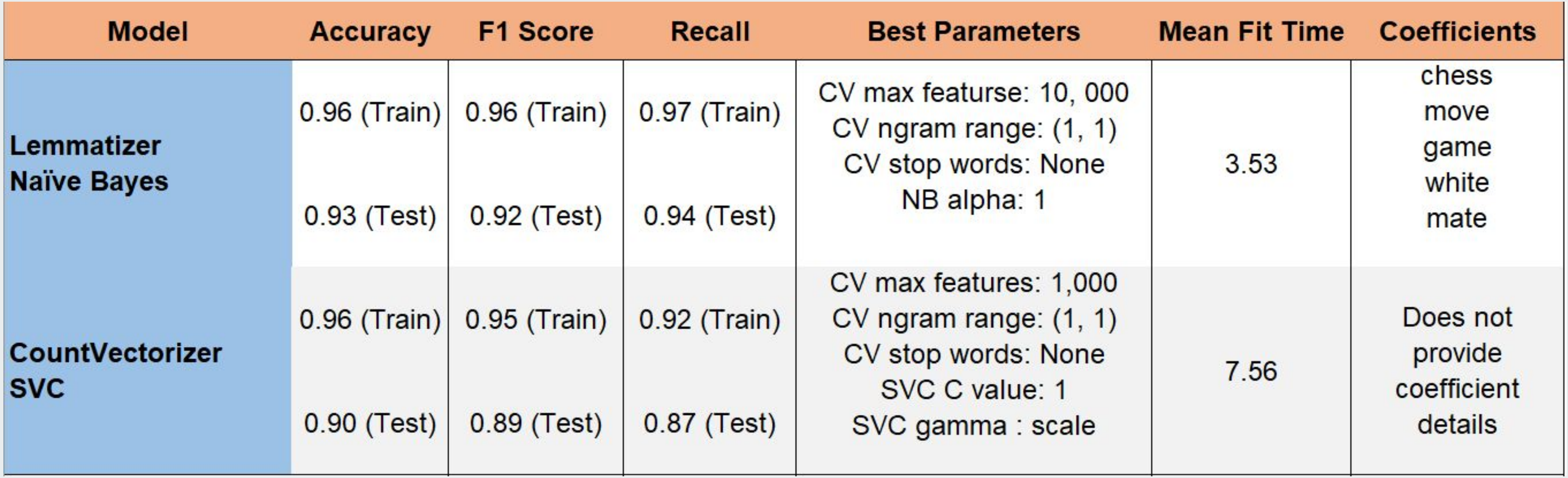
- In the **chess subreddit,** mean polarity score is 0.16
- On average, 0.12 polarity for **poker subreddit**

Model Comparison

- **Naive Bayes** with Lemmatizer worked best
- **Random Forest** overfit on Accuracy, F1, Recall
- **Logistic Regression** and **Random Forest** models scored lowest on Recall (tp / (tp + fn))



Naive Bayes vs SVC



Model	Accuracy	F1 Score	Recall	Best Parameters	Mean Fit Time	Coefficients
Lemmatizer Naïve Bayes	0.96 (Train) 0.93 (Test)	0.96 (Train) 0.92 (Test)	0.97 (Train) 0.94 (Test)	CV max featurse: 10, 000 CV ngram range: (1, 1) CV stop words: None NB alpha: 1	3.53	chess move game white mate
CountVectorizer SVC	0.96 (Train) 0.90 (Test)	0.95 (Train) 0.89 (Test)	0.92 (Train) 0.87 (Test)	CV max features: 1,000 CV ngram range: (1, 1) CV stop words: None SVC C value: 1 SVC gamma : scale	7.56	Does not provide coefficient details

Recommendations: Chess

Keywords search advertising

Bid on:

words:

- chess
- mate
- checkmate
- lichess
- moves

phrases:

- chess set
- new chess
- blitz game
- playing chess

Avoid:

words:

- game
- best
- playing
- play

Recommendations: Poker

Keywords search advertising

Bid on:

words:

- poker
- tournament
- pokerstars
- stakes
- hands

phrases:

- online poker
- live poker
- cash game
- bad bet

Avoid:

words:

- game
- best
- playing
- play

Future Research

- Build a neural network
- Include posts and comments as features
- Collect and add data from chess and poker blogs

Thank you!