Project 3: NLP

By Ben Peck

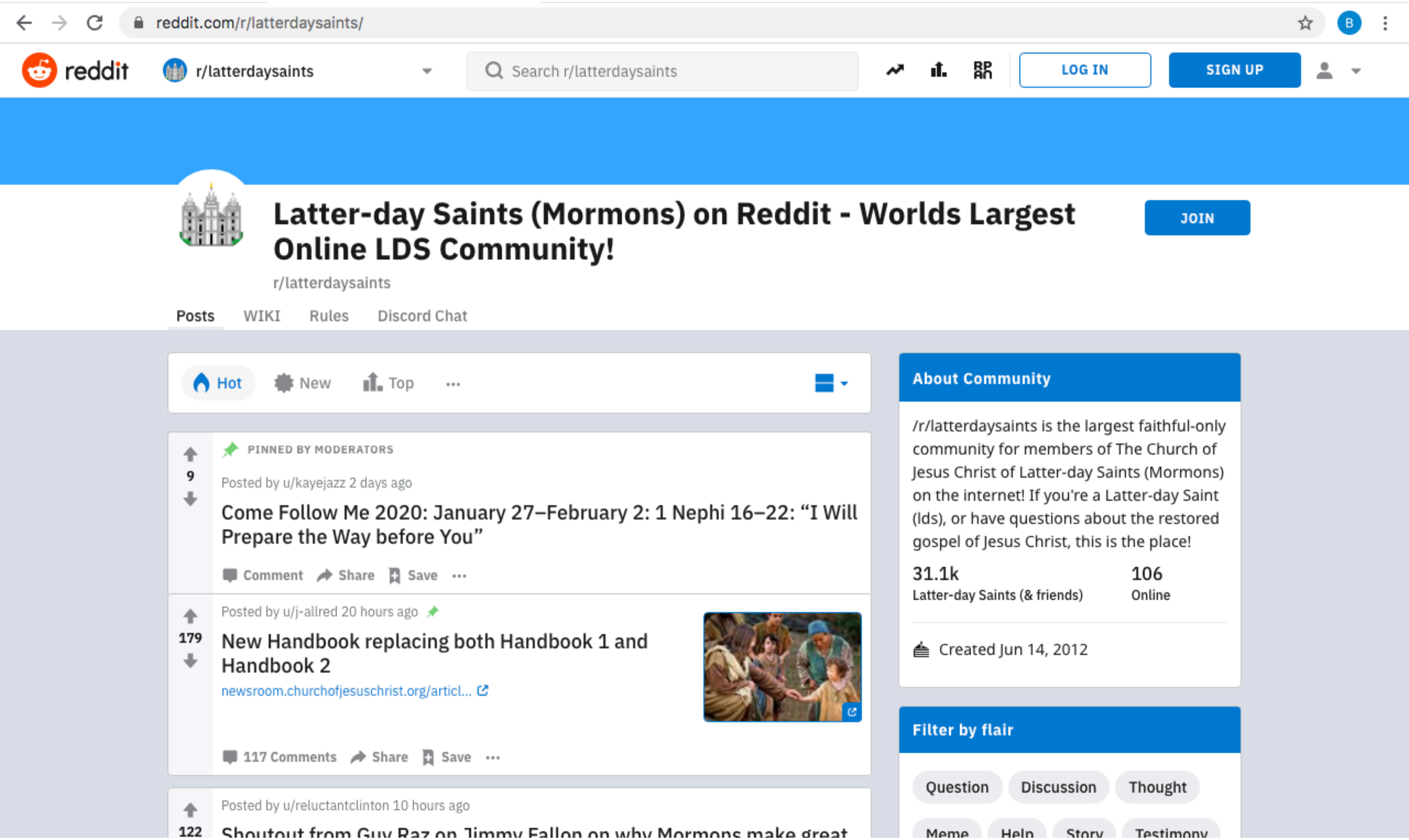
Objective

- Accurately predict which of two subreddits a post originated in based on their text, even when those subreddits have similar topics.

Reddit

The Subreddits

- r/latterdaysaints
  - Billed as “the largest faithful-only community for members of The Church of Jesus Christ of Latter-day Saints (Mormons) on the internet!”
  - 31.1k members
  - Created in June 2012



Latter-day Saints (Mormons) on Reddit - Worlds Largest Online LDS Community!

r/latterdaysaints

PINNED BY MODERATORS

9

Posted by u/kayejazz 2 days ago

Come Follow Me 2020: January 27–February 2: 1 Nephi 16–22: "I Will Prepare the Way before You"

179

Posted by u/j-allred 20 hours ago

New Handbook replacing both Handbook 1 and Handbook 2

newsroom.churchofjesuschrist.org/articl...

122

Posted by u/reluctantclinton 10 hours ago

Shoutout from Guy Raz on Jimmy Fallon on why Mormons make great

About Community

/r/latterdaysaints is the largest faithful-only community for members of The Church of Jesus Christ of Latter-day Saints (Mormons) on the internet! If you're a Latter-day Saint (lds), or have questions about the restored gospel of Jesus Christ, this is the place!

31.1k
Latter-day Saints (& friends)

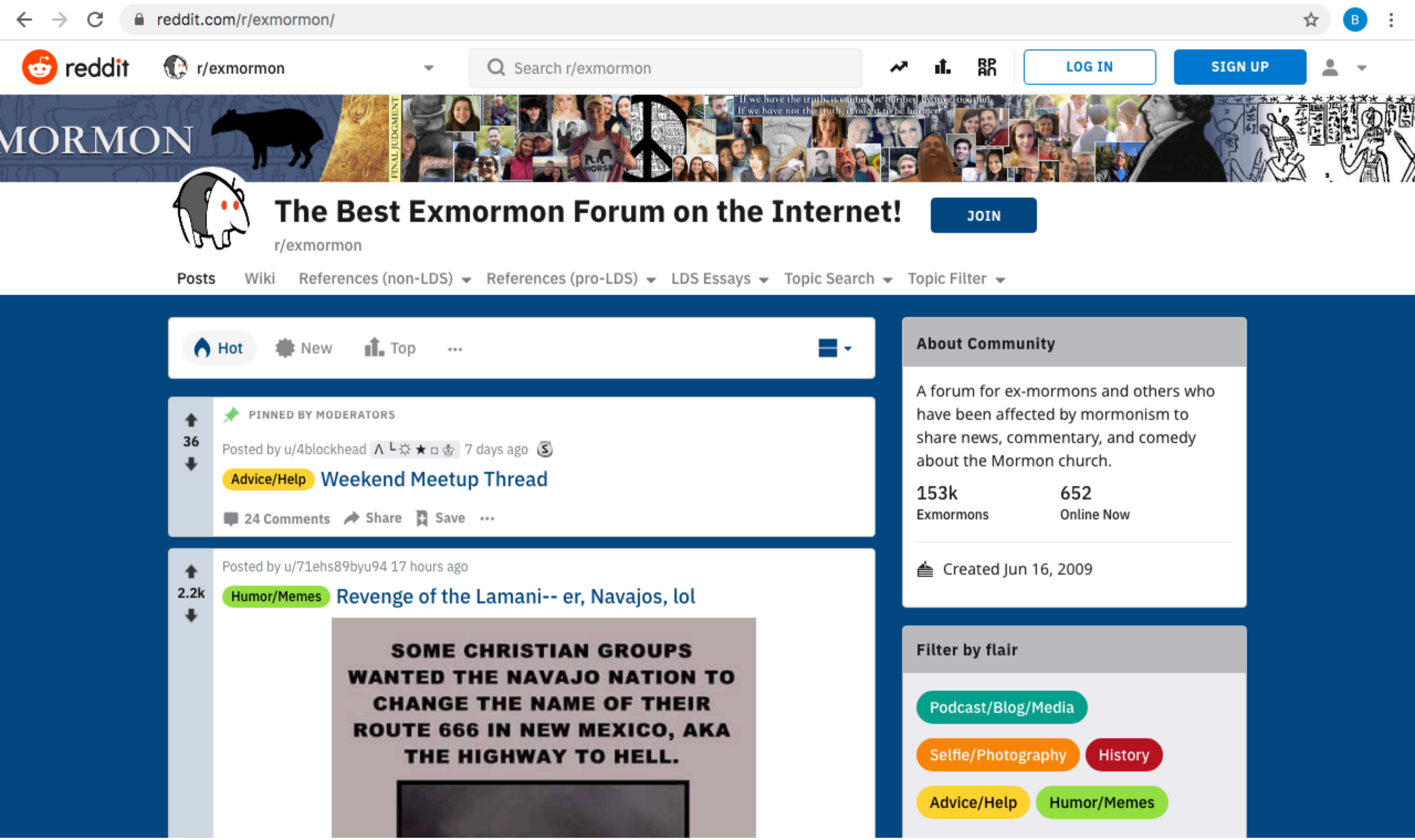
106
Online

Created Jun 14, 2012

Filter by flair

The Subreddits

- r/exmormon
  - “A forum for ex-mormons and others who have been affected by mormonism to share news, commentary, and comedy about the Mormon church.”
  - 153k members
  - Created in June 2009



The Best Exmormon Forum on the Internet!

r/exmormon

PINNED BY MODERATORS

36

Posted by u/4blockhead 7 days ago

Advice/Help Weekend Meetup Thread

24 Comments

2.2k

Posted by u/71ehs89byu94 17 hours ago

Humor/Memes Revenge of the Lamani-- er, Navajos, lol

SOME CHRISTIAN GROUPS WANTED THE NAVAJO NATION TO CHANGE THE NAME OF THEIR ROUTE 666 IN NEW MEXICO, AKA THE HIGHWAY TO HELL.

About Community

A forum for ex-mormons and others who have been affected by mormonism to share news, commentary, and comedy about the Mormon church.

153k Exmormons

652 Online Now

Created Jun 16, 2009

Filter by flair

Data Collection

- Collected the 10,000 most recent posts from each of the two subreddits using Reddit's Pushshift API
- The only fields saved were subreddit, title, text content, date created, and score

EDA and Cleaning

- Dropped posts where the text of the post was blank, [removed], or [deleted]
- Proportion of posts from each sub roughly unchanged
- Left just over 10,000 posts, approximately 5,000 from each subreddit
- Title and main text of posts were combined and edited to remove formatting strings and streamline tokenizing

The Models

- First Pipeline: Count Vectorizer and K Nearest Neighbors
  - Grid Search CV over params:
    - KNN n_neighbors: 3, 5, 7
    - Cvec max_features: 1000, 2000, 5000
    - Cvec stop_words: english, None
    - Cvec ngram_range set to (1, 1)

The Models

- Best Params:
  - KNN n_neighbors: 3
  - Cvec max_features: 5000
  - Cvec stop_words: english
- Accuracy score with best params: 62.7%

This is barely better than the baseline score!
We can do better than that

The Models

- Second Pipeline: Count Vectorizer and Naïve Bayes Multinomial
  - Grid Search CV over params:
    - NBM alpha: 0.1, 0.5, 1, 5
    - Cvec max_features: 1000, 2000, 5000, None
    - Cvec ngram_range: (1, 1), (1, 2)
    - Cvec stop_words set to english

The Models

- Best Params:
  - NBM alpha: 0.5
  - Cvec max_features: None
  - Cvec ngram_range: (1, 2)
- Accuracy score with best params:
  - On Training Data: 99.7%
  - On Testing Data: 85.4%

Conclusions and Next Steps

- The subreddits could be distinguished with reasonably high accuracy using a multinomial naïve bayes model
- To further refine, other features (such as score) could be brought in and potentially correlated with sentiment
- Further exploration of the effects of individual features

Any Questions?