Research Question:
Can one use basic language processing techniques to classify which broad group of subreddits a subreddit belongs to? If so, which of these methods draws the most promising results?

Word Count: 2420 words (including the results as text)

Introduction:

Natural Language Processing(NLP) techniques have been found to have many practical uses in the field of text analytics and acts as a bridge both between humans and machines as well as humans and other humans. The subfield of text classification has been one of the most utilized uses and as seen in the form of spam-filtering emails and sentiment analysis(Meurers, 2012). NLP has some core techniques which are relatively easy to implement and provide that can be used for such tasks and are relatively easy for humans to interpret. This research seeks to investigate the application of these techniques as a tool for deciphering the subject group of subreddits. The experiment conducted will check for patterns of language across the groupings and try to determine whether the methods chosen have potentially promising applications in allowing an analyst to make such classifications. The groupings used are intentionally chosen to test the limits of these methods, with some topics being sufficiently different so that there is an expected greater ease of telling them apart, (e.g sports themed subreddits and educational subreddits) and others which are expected to have stronger links (e.g movies and TV series themed subreddits). The eight groups chosen are classed as “Discussion”, “Gaming”, “TV Series”, “Movies”, “News and Politics”, “Humour”, “Sports” and “Learning”. The chosen analytical methodologies are lexical diversity comparison, trigram comparison and word frequency distribution comparison.

Research Method:

The basis of my methods for this experiment can be broken into three separate parts. Firstly the sourcing of data, secondly the processing steps I took with that data, and thirdly the evaluations methods I adopted.

Souring the Data:

Selecting my data was a somewhat imperfect task, given that the data I was acquiring was quite large and it would be a timely task to go through each subreddit individually and categorize them myself. The vast majority of the subreddits I used were sourced from a reddit post (**LINK CITE**) which included categorization. I did some brief verification of my own, removing ones I felt were too borderline to be clearly categorized so that I could reduce error in the experiment data. I then ran a loop using the python Praw library and downloaded the comment data from each of the selected subreddits into text files. Each of the groups were separated into their own arrays so that I could run the experiments on them individually. The groups are shown in the code extract below.

```
``DiscussionSubs = ["CrazyIdeas", "AskReddit", "fatpeoplestories", "DoesAnybodyElse",
"IAmA", "bestof", "TalesFromRetail"]
GamingSubs = ["gaming", "leagueoflegends", "pokemon", "Minecraft", "starcraft", "Games",
"DotA2", "skyrim", "tf2", "magicTCG", "wow", "KerbalSpaceProgram", "mindcrack"]
```

```
TVSubs = ["arresteddevelopment", "gameofthrones", "doctorwho", "mylittlepony", "community",
"breakingbad", "adventuretime"]
MoviesSubs = ["movies", "harrypotter", "StarWars", "anime", "batman", "moviecritic",
"MovieSuggestions", "Hungergames", "MarvelStudiosSpoilers"]
NewsSubs = ["politics", "worldnews", "news", "conspiracy", "Libertarian", "offbeat",
"TrueReddit", "Conservative"]
HumourSubs = ["funny", "fffffffuuuuuuuuuuu", "4chan", "lmGoingToHellForThis",
"firstworldanarchists", "circlejerk", "okbuddyretard", "facepalm", "Jokes"]
SportsSubs = ["nba", "soccer", "hockey", "nfl", "formula1", "baseball", "MMA",
"rugbyunion", "PremierLeague"]
LearningSubs = ["todayilearned", "science", "askscience", "space", "AskHistorians",
"YouShouldKnow", "explainlikeimfive"]``
```

Processing the Data:

Because my experiment was designed to incorporate different evaluations of the data, the data processing steps for each measurement could not be uniform and I believed there would be value in analyzing the results of different variations of the data. Firstly, the text data which was tokenized into words using *NLTK.word_tokenize*, setting all of the letters to lowercase using python's *.lower()* function, and the removal of punctuation using python's *.isalnum()* function. Secondly for the alternative dataset I removed all of the stop words NLTK.corpus.stopwords.

Although I believed having these two datasets would give a bit more context to the data, I did however, not use the data which did not omit stopwords from the frequency distribution analysis as the use of them would clutter each of the data sets with low context word tokens.

Evaluation Methods:

The methods for evaluation can be best understood using the following bullet points.

- Lexical Diversity Score:
    - The number of unique words in a text divided by the total number of words.
    - Given for each group as well as each subreddit within a group.
- Frequency Distribution of top 75 words:
    - The number of times a particular word appeared in the text data for each subreddit.
    - A list was generated for each of the words that occurred in the top 75 of each subreddit.
    - Noting of words which occurred in all but one and all but two of the subreddits in the group.
- Top 100 Trigrams:
    - The 100 most common instances of a grouping of three words.
    - Done with and without stopwords.
    - List generated bigrams which were most common across a grouping.

Results:

Lexical Diversity Scores

Topic	***Sports***	***Movies***	***TV***	***News & Politics***	***Gaming***	***Learn***	***Humour***	***Discuss***
Avg. no stopwords	0.2834	0.30509	0.31553	0.30053	0.25608	0.26943	0.39611	0.28597
Avg. with stopwords	0.16818	0.17754	0.17663	0.1638	0.14916	0.14355	0.24189	0.15408
Lowest no stopwords	0.2359	0.18851	0.24621	0.23071	0.17725	0.18561	0.25246	0.23988
Highest no stopwords	0.39006	0.35866	0.35952	0.33919	0.33046	0.36193	0.679012	0.33182

Frequency Distribution of top 75 words:

Sports
9/9: *[see, go, got, going, would, think, fucking, know, love, get, really, good, time, like, one, man]*
8/9: *[fuck, could, never, shit, back]*
7/9: *[win, well, still, last, right, people, much, game, https, even]*

Movies
9/9: *[know, get, really, got, way, even, would, love, time, like, one]*
8/9: *[people, also, good, still, think, first, see, much, well, made, great]*
7/9: *[fucking, ever, man, never]*

TV
7/7 *[going, see, well, never, would, back, made, know, good, people, love, could, best, really, make, think, show, one, get, time, like]*
6/7: *[got, right, episode, great, https, watch, much, even, still, first]*
5/7: *[want, amazing, looks, ever]*

News and Politics
8/8: *[trump, one, going, good, know, way, even, really, https, time, right, think, would, like, much, get, want, people, well, could, see]*

7/8: *[make, say, also, go, deleted, still, got, years, never, us, fuck]*
6/8: *[shit, fucking, every, take, need]*

Gaming
13/13: *[like, one, really, get, time, know, would, good, make, going, got]*
12/13: *[think, love, much, never, game, people, see, even]*
11/13: *[made, could, fucking, way, well, still, first]*
Learning
7/7: *[would, really, know, way, years, see, even, much, people, good, make, think, back, get, could, well, like, one, time, us]*
6/7: *[still, say, many, go, want, work, also, something]*
5/7: *[every, removed, https, never, first, going]*

Humour
9/9: *[like, one, get, would, see, time]*
8/9: *[actually, know, think, post, even, deleted, shit, https, people, right, got, make, good]*
7/9: *[reddit, fuck, really, well, want, much]*

Discussion
7/7: *[go, thing, time, much, could, way, would, even, think, know, really, see, people, one, like, get]*
6/7: *[going, work, never, also, something, good, need, make]*
5/7: *[day, someone, back, feel, got, right, well, first, reddit, want, still]*

Frequency Distribution of top 100 trigrams:

Sports
No stop words: ***(3/9)****=[(top, post, time), (gon, na, win)],* ***(2/9 SRs)****=(indians, blew, lead), (13, year, old), (rip, kobe, bryant), (wan, na, see), (got, ta, love), (gon, na, end)*
With stop words: ***(9/9)****=[(one, of, the), (this, is, the)],* ***(8/9 SRs)****=[(this, is, a)],* ***(7/9 SRs)****=[(a, lot, of), (to, be, a), (going, to, be)]*

Movies
No stop words: ***(2/9)****=[(best, movie, ever), (max, fury, road), (horror, movie, ever), (yes, yes, yes), (every, time, watch), (one, favorite, films), (may, rest, peace), (amazing, well, done), (ca, wait, see), (mad, max, fury), (gon, na, say), (one, favorite, movies), (one, best, movies), (best, movies, ever), (top, post, time), (sure, welcome, one), (watched, last, night), (movie, ever, seen)]*
With stop words: ***(8/9)****=[(i, don, t), (one, of, my), (one, of, the)],* ***(7/9 SRs)****=[(i, feel, like), (this, is, the)],* ***(6/9)****=[(a, lot, of), (to, be, a), (of, all, time)]*

News and Politics
No stop words: ***(4/8)****=[(contact, moderators, subreddit), (action, performed, automatically), (subreddit, questions, concerns), (moderators, subreddit, questions), (please, contact, moderators), (automatically, please, contact), (performed, automatically, please), (bot, action, performed)],*

***3/8**=[(president, united, states)], **(2/8)**=(would, love, see), (queen, elizabeth, ii), (https, https, https), (na, na, na)]*
With stop words: *(8/8)=[(a, lot, of), (this, is, the), (one, of, the), (this, is, a)], **(7/8)**=[(is, going, to), (the, rest, of)], **(6/8)=[**(the, fact, that), (i, don, t), (to, be, a)]*

Learning
No stop words: *(4/7)=[(https, https, https), (removed, removed, removed)] , **(2/7 SRs)**=[(free, speech, democracy), (5, years, ago), (one, favorite, subs), (keep, good, work), (picture, black, hole), (supermassive, black, hole), (may, rest, peace), (innovation, free, speech), (net, neutrality, cornerstone), (neutrality, cornerstone, innovation), (cornerstone, innovation, free)]*
With stop words: *(7/7)=[(i, don, t), (one, of, my), (one, of, the)], **(6/7)**=(free, speech, democracy), (5, years, ago), (one, favorite, subs), (keep, good, work), (picture, black, hole), (supermassive, black, hole), (may, rest, peace), (innovation, free, speech), (net, neutrality, cornerstone), (neutrality, cornerstone, innovation), (cornerstone, innovation, free)*

Humour
No stop words: *(3/9)=[(gold, kind, stranger)], **(2/9)**=[(links, please, respect), (threads, info, contact), (bloop, someone, linked), (follow, links, please), (bleep, bloop, someone), (please, respect, rules), (respect, rules, reddit), (rules, reddit, vote), (reddit, vote, threads), (vote, threads, info), (bot, bleep, bloop), (thing, ever, seen), (thanks, gold, kind), (edit, thanks, gold)]*
With stop words: *(8/9)=[(this, is, the)], **(6/9)**=[(this, is, a)]*

TV
No stop word: *(3/7)=[(may, rest, peace)], **(2/7)**=[(https, https, https), (one, best, actors), (gon, na, get)]*
With stop words: *(7/7)=[(this, is, the), (one, of, the)], **(6/7)**=[(a, lot, of), (this, is, a), (i, feel, like), (one, of, my)]*

Gaming
No stop word: *(5/13)=[(rest, peace, john)], **(3/13)**=[(holy, fucking, shit), (thing, ever, seen), (rest, peace, tb), (may, rest, peace), (https, https, https), (peace, rest, peace)]*
With stop words: *(13/13)=[(this, is, the], **(12/13)**=[(one, of, the), (a, lot, of), (this, is, a)], **(9/13)=[**(going, to, be), (to, be, a)], **(8/13)**=[(i, do, know), (i, want, to)]*

Discussion
No stop word: *(3/7)=[(https, https, https)], **(2/7)**=[(every, single, time), (net, neutrality, rules), (save, net, neutrality), (net, neutrality, https), (yes, yes, yes)]*
With stop words: *(7/7)=[(a, lot, of)], **(6/7)**=[(this, is, a), (one, of, the), (this, is, the)]*

Discussion of Results:

In review of the results, the task of deciphering the correct class a subreddit belongs to appears difficult for a human observer. There appears to be more promise in the word frequency distribution lists than the other two methods although the value is more apparent when classifying when the groups are more broadly defined.

For words which occurred most in the top 75 of each group member there was a glut of low value data where many common words occurred in all categories. The group which was easily identifiable by a single key word was the “News and Politics” group which all had “trump” in their top results. When more broadly grouping the subjects into entertainment (Sports, Movies, TV, Gaming) and non-entertainment the word “love” was common in all but one of the subreddits in the four groups showing a high correlation. The “Learning” and “Discussion” groups could also be identified by a universal presence of the word “think”. The lack of a presence of profanity within the top word frequency distribution was also a notable metric which had unexpected results. The two academic based groups of “Learning” and “Discussion” had no instances of this which hinted at its usefulness, yet things were not so predictable in other groups such, with there being no presence in “TV” while “Movies” did have instances of these in the top words. It is my belief that this could be due to a poor categorization of the groups. Having better data that was categorized by someone with more domain knowledge would likely improve the results, at least minorly.

Lexical diversity across each of the groups appears relatively unpredictable in all instances apart from the “Humour" group which scored a notably high average of 0.39611, also containing the lowest low score and highest high score. Even this example shows the unreliability of the lexical diversity as a means of estimating the class of a text though as the range between the highest and lowest scores have an extremely large spread of 0.25246 to 0.679012. Each of the other groups were in a relatively small spread between 0.25608 for “Gaming” and 0.31553 for “Learning”. The prospect of grouping the subreddits more broadly here has potential as similar groups do show some similarities as shown with the “TV” and “Movies” groups as well as the “Learning” and “Discussion” groups which have very similar scores in each of the lexical diversity metrics noted.

Trigrams showed an extremely low level of replication across the subreddits of each group, although they appear to give more context about the individual topic of a subreddit. The top 100 with stopwords included gave back extremely common English language which gave very little context. Most of the entertainment categories had very little similarities in trigram distribution with each subreddit appearing to have their own niche topics. Death remembrance seems to be a common link between many of the topics although it still was not to the extent that it would be of much use in this context. The more academic subreddits did show some commonality in the apparent topics of some of the trigrams with the inclusion of the word “neutrality” appearing a few times in the “Discussion”, “Learning” and “News and Politics” groups. An issue with these results is that they are somewhat distorted by the inclusion of text data which came from the app and moderator bots so cleaning the data better might bear more useful insights.

In conclusion, no single one of the three techniques appears to give much definitive use in isolation when estimating which category a subreddit belongs to. Some of these techniques are more suitable for identifying certain groups over others, leading to the conclusion that one should take into account each of the features when determining the probable category of a subreddit.

Citations:

1. Meurers, D., 2012. Natural language processing and language learning. *Encyclopedia of applied linguistics*, pp.4193-4205.
2. u/douglasmacarthur " The 200 most active subreddits, categorized by content" *Reddit*. Accessed March 14, 2023.<https://www.reddit.com/r/TheoryOfReddit/comments/1f7hqc/the_200_most_active_subreddits_categorized_by/>.