Subreddit Text Classification

Austin Hillyard
Brigham Young University
ajh283@byu.edu

Abstract

This research attempts to use text-classification to distinguish 22 different subreddits, some very similar and others completely different as text classification on highly specific social media groups is not very common. The resulting model over-fit the training data, and has only about a 7% accuracy on data outside of training. Classifying a large amount of subreddits will not likely be very successful without representative data to train off of, and using a starter checkpoint model trained off of social media in general. Limiting research down to a select few groups that are highly related will also more likely achieve interesting and useful results.

1 Introduction

The social media side *reddit* is divided into various subreddits. These subreddits are topic specific groups that often can have a lot of overlap in their interest. A user might have an interest in a particular genre but not know which of several similar subreddits might be the best fit for them. Or, a moderator might want to have a tool that can automatically detect the relevance of a text to the subreddit they are a part of. Using text classification, This research has fine tuned the distilBERT model in an attempt to automatically categorize a sample of text into one of 22 pre-selected subreddits based off of real data pulled from reddit using the *praw* API.

2 Related Work

Social media has been incredibly useful to researchers to gather a large of data fairly easily and use that data for machine learning purposes. Oftentimes researchers will fine tune models as done in this paper. Buyukoz trained several different kinds of machine learning models, and found that distilBERT can transfer "generic semantic knowledge to other domains" quite well relative to other models, the same model used in this reddit research (Büyüköz et al., 2020).

Transformer models seem to perform better overall in these kinds of social media analysis. Guo found that models fine tuned to social media analysis actually outperform non social media models that are fine tuned on a specific topic Guo et al., 2020. The model used in this reddit research is only the generic distilBERT model. If this research was to be repeated, then a more specific model would likely be the first step in attempting to improve the results.

Fiallos has done very similar research to this paper. Fiallos uses reddit data to train a multi label text classification model. Funnily enough, the reddit trained model is then used to predict the topic interests of a Twitter user based on their own tweets with 75.62% accuracy (Fiallos and Jimenes, 2019). Twitter is an open forum platform, it is not divided into topic specific groups like reddit. However, the model made for this paper is trained on specific subreddit labels, instead of more generic topics, so it would be difficult to extrapolate the model to uses outside of reddit.

Finally, Sjaji has done research specifically on reddit flairs, rather than specific subreddits (Shaji, 2021). Usually subreddits can have their own specific flairs, but their use remains the same across subreddits; to categorize a post according to a subtopic or format within the subreddit. This research does not use transformers, but very similar research could be done using a transformer to classify text according to flairs within a subreddit.

3 Methods

The data was collected from reddit using the python API wrapper *praw*. The subreddits were manually selected. Most of the subreddits chosen reflect the interests of the writer, with some generic subreddits that reflect different topics.

3.1 Collection and Storing of Data

As mentioned, the data was gathered using the *praw* API. From each of the 22 subreddits, 50 posts were gathered from the "hot" category on reddit, meaning the posts are relatively new and gaining popularity within the subreddit. From each of those 50 posts, 100 top level comments were also gathered. Posts and comments are then stored in a CSV file, with the columns *subreddit*, *content*, and *type*. Type refers to whether the text is from a post or a comment, meaning posts and comments are separated on their own lines.

It should also be mentioned that posts have two components, title and content. Title is a mandatory field whereas content is not. Depending on the subreddit, content can often be left completely blank in place of a picture. When storing posts, the title and content fields were concatenated together, prefaced with distilBERT <CLS> token, and then separated with the <SEP> tokekn to try and tell the model these are different fields within the same item.

The data was gathered from reddit on April 13th. This is relevant as the "hot" category is dependent on time, so some of the topics will be local to relevant conversation, and may not translate to newer data very well. This is admittedly a shortcoming of the data, but it pulls from data that is not limited to most popular of all time, as "hot" will have both posts with thousands of up-votes, and ones with only ten up-votes but are starting to get more popular.

3.2 Machine Learning Pipeline

After gathering all of the data, the csv was then converted into a hugging face dataset and split 70% for training and 30% for testing. The dataset was then used to fine tune the distilBERT checkpoint model from hugging face over 20 epochs.

3.3 subreddits

3.3.1 Explanation of subreddits chosen

As mentioned, the subreddits do reflect mostly the interests and hobbies of the writer, though they are supposed to be varied in some aspects but similar in others. Warhammer is a tabletop game with several different systems. The two main systems are Warhammer40k and Warhammer Age of Sigmar. The first is a science fiction setting, and the second is a fantasy setting. Thus, they will have different factions in the game, characters, and references to events and places. However, they will both talk about models, painting, and the parent company GamesWorkshop. The attempt was to train the model on very similar groups with a lot of overlap, but with a few key topic differences in hopes that the model could distinguish them despite their similarities.

Figure 1: List of subreddits pulled from

Warhammer40k	StarWars
Warhammer	lotr
ageofsigmar	retrogaming
PrequelMemes	ProgrammerHumor
lotrmemes	CloneWarsMemes
politics	Funnymemes
Conservative	minipainting
democrats	SteamDeck
Republican	Steam
gaming	linux
StarWars	windows
lotr	Grimdank

There are also meme subreddits, but of different topics. Usually meme subreddits have very similar formats, with a short post title and no post content. Although the format can be similar, the topics and inside jokes discussed by users should be intrinsic to the subreddit itself and distinguishable. Of course, the meme subreddits are being trained alongside the other subreddits as well, meaning the model can also use format to distinguish different subreddits.

There are also some political subreddits included. These are of interest as they might often discuss the same topics in current news cycles in a similar format, but have completely different ideological interpretations of the same information. The goal was to have the model attempt to pick out these differences as well.

Finally, there are some more generic games reflecting elements of pop culture and technology. Some are more similar, like Steam and SteamDeck, but others are more distinct like StarWars and lotr (Lord of the Rings). Most of these should be clear from their topics. There are also the meme subreddits that mirror the same topics as these, so the model will also have to distinguish more humorous discussion with general fan discussion.

4 Evaluation

As mentioned, the model was trained over 20 epochs. In Table 1 are the data showing the progression of accuracy, training loss validation loss over each epoch. The accuracy is also not exceptional in training, but not horrible as the model is classifying 22 separate labels. However, this is undermined by how the model seems to be over-fitting based on the increase of validation loss over each epoch.

The precision, recall, and F1 are also reported in Table 2. All values are just below 50%, so there is a decent balance between the statistics.

Extra data outside the dataset was also gathered two days later from the original dataset, on April 15th. This data was much smaller, five posts per subreddit and 10 comments for each post. This data was then evaluated on the model, and the accuracy was only 7.66%. The likelihood of being randomly correct is about 4.5%. The model is unfortunately not that much better than random chance due to the over-fitting. Table 3 shows the incorrect classifications of each subreddit pair by count in this evaluation.

Epoch	Training Loss	Validation Loss	Accuracy
1	2.4323	1.8241	0.4602
2	1.5790	1.6159	0.5166
3	1.1723	1.5967	0.5341
4	0.8619	1.6179	0.5480
5	0.6133	1.7719	0.5412
6	0.4274	1.9060	0.5443
7	0.2831	2.0892	0.5365
8	0.1955	2.2329	0.5357
9	0.1340	2.4017	0.5448
10	0.0962	2.6054	0.5467
11	0.0652	2.7636	0.5419
12	0.0545	2.9018	0.5415
13	0.0504	3.0117	0.5448
14	0.0434	3.0829	0.5496
15	0.0398	3.1304	0.5538
16	0.0287	3.1799	0.5487
17	0.0288	3.1907	0.5537
18	0.0246	3.2181	0.5538
19	0.0259	3.2504	0.5542
20	0.0240	3.2479	0.5534

Table 1: Training and Validation Metrics

Precision	Recall	F1
0.49624	0.48252	0.48778

Table 2: Precision, Recall, and F1 at final epoch

5 Conclusions

The model was not very successful outside of training at predicting the subreddit of a sample of text. The main reason for this is the over-fitting of which there are several likely causes.

Firstly, the data was gathered from the hot category on a single day. As mentioned, the hot category of a subreddit is time dependent. It may be an average representation of the type of content on a subreddit, but that representation may not be accurate to other time frames. Topics on many of these subreddits change frequently, even over a short period of time. New releases in content, either movies or games or announcements will often dominate discussion for a few days in the entertainment subreddits. Having trained the model on a very specific time frame means that trying to predict text from later or earlier time frames will be more difficult for the model as many of the key words it encounters may be different.

Secondly, many of the classification labels might be far too similar. The point of the research was to try and disassociate several similar subreddits a part from each other, and be able to tell them apart. But the model may not have been able to learn the intricacies of two similar subreddits while also contrasting with very different subreddits, and instead resorted to basic topics and over-fitting. In the incorrect subreddits table 3, we can see a few instances of the model confusing similar subreddits. Steam and linux are conflated likely because the company Steam has a linux based operating system. SteamDeck and gaming are very closely related, as the SteamDeck is a gaming device. Conservative and democrats is an understandable conflation as they are both political, and likely talk about the same political events, albeit from different lenses. Other conflations might not be easily explained by similar interests, such as politics and Warhammer40k. Though, this particular error might be explained by the fact that in the Warhammer40k community, a political controversy occurred the same weekend as the data was collected. There also may be a lot of overlap of users between similar subreddits, and certain patterns of dialogue may also be common across subreddits.

(Correct, Predicted)	Count
(politics, Warhammer40k)	42
(StarWars, windows)	34
(gaming, Funnymemes)	32
(linux, Warhammer)	29
(Steam, linux)	29
(ProgrammerHumor, SteamDeck)	27
(Warhammer40k, Grimdank)	25
(SteamDeck, gaming)	23
(retrogaming, ProgrammerHumor)	22
(ageofsigmar, Republican)	21
(minipainting, StarWars)	19
(windows, politics)	17
(Conservative, democrats)	17
(Warhammer, PrequelMemes)	16
(Grimdank, ageofsigmar)	15
(democrats, Warhammer40k)	15
(SteamDeck, Funnymemes)	15
(Funnymemes, retrogaming)	15
(Conservative, Warhammer40k)	15
(democrats, 40kLore)	13
(PrequelMemes, windows)	13
(retrogaming, Funnymemes)	12
(Steam, Funnymemes)	11
(Warhammer, ageofsigmar)	11
(ProgrammerHumor, Warhammer)	11
(PrequelMemes, Conservative)	11
(ageofsigmar, PrequelMemes)	9
(minipainting, Grimdank)	9
(Republican, democrats)	9
(Warhammer40k, StarWars)	7

Table 3: Table of the pairs of subreddits the model confused. The first subreddit name is the correct subreddit. The second is the predicted subreddit from the model. The count refers to the number of times the model made this error.

Thirdly, there are too many subreddits. This is likely the overall explanation for the model's overfit. With 22 labels, some being very similar, and others being completely different, the model likely has to resort to over-fitting to be able to decrease the training loss in the training loop as it has a very low chance of being right by chance.

5.1 Future Work

After having done this research, there are several things that have been learned that can be done to improve the performance.

5.1.1 Data Collection

The biggest problem with this model is its limited data. The data should have been collected over a large period of time to get a better representation of a subreddit through many different "news cycles". As mentioned, different events will cause subtopic changes in a subreddit. If there were more diverse data to learn from, the model could have learned to focus on what was consistent between those subtopics, instead of perhaps learning one subtopic as the larger whole of a subreddit.

reddit also has up-votes as a quality system for users to rate the relevance of a post within a subreddit. Up-votes could have been used to increase the weight of importance of a specific comment or post so the model can know to learn more or less from that item.

5.1.2 Starting Checkpoint

As Guo mentioned, there are models that have been fine tuned for social media analysis from the get-go, and they outperform models that are topic specific (Guo et al., 2020). Starting from a social media checkpoint would likely improve the model by a large margin. Social media discourse, especially on reddit, can be very unique to other registers, or even other platforms. Having to learn that discourse and also 22 subreddits on top of that might be very challenging to a generic distilBERT model. Using a social media checkpoint, or even a reddit specific checkpoint would likely help the model learn more efficiently.

5.1.3 Using another classifier instead of subreddits

One primary issue of scaling up this kind of research is the amount of subreddits total on the platform. There are over 100,000 active subreddits on reddit, meaning training AI to distinguish subreddits as you add more subreddits quickly becomes infeasible. Many topic classifiers already exist, and reddit even has such a feature for new accounts. reddit will query a new user on what interests they have, and it will show several suggested subreddits as a result.

To advance research on finer topics, there are a few other options. Training an AI on a select few and highly related subreddits might be successful as you have a limited number of labels for the AI to learn. There is still a risk of over-fit as the subreddits are very similar, so having a lot of data on each subreddit over a large range of subtopics and time

would be necessary to overcome that. Or, creating a subtopic classifier for a specific topic might be more successful. You could gather data on a select topic that is not too broad, such as video games, movies, or a specific table top game system, and train an AI to recognize subtopics in each subreddit. Many subreddits are as narrow as a specific subtopic already, such as the specific game systems in the Warhammer franchise.

5.2 Summary

Social media provides a very large dataset to train AI, either to develop new features or for academic research. One such area is being able to distinguish similar yet distinct forums or groups automatically to filter unrelated posts, or help users find the right group for them. Training a text-classification model on such intricacies in social media presents several challenges, such as getting representative data, having a good base AI model, choosing appropriate labels for classification, and avoiding over-fit. More research and care will need to be taken in order to accurately train a model to distinguish both highly specific and highly related groups from each other.

`https://github.com/austinhillyard/LING581-FinalProject`

References

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, Marseille, France. European Language Resources Association (ELRA).

Angel Fiallos and Karina Jimenes. 2019. Using reddit data for multi-label text classification of twitter users interests. In *2019 Sixth International Conference on eDemocracy eGovernment (ICEDEG)*, pages 324–327.

Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Mollá Aliod. 2020. Benchmarking of transformer-based pre-trained models on social media text classification datasets. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 86–91, Virtual Workshop. Australasian Language Technology Association.

Reshma Shaji. 2021. Exploratory data analysis on reddit data: An efficient pipeline for classification of flairs. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)*, pages 65–68.