Executive Summary

If you spend alot of time on online forums, there is a chance you are on Reddit. Being one of the most popular sites in the world, Reddit bills itself as the “front page of the internet” containing a massive collection of forums where people share news and content while also commenting on each other's posts.

Reddit is broken up into more than a million communities known as “subreddits,” each of which covers a different topic managed by moderators ("mods" for short) and Reddit users can post content. This especially is very useful for users with queries as they can expect specialised comments or replies from other users when they post their queries in specific communities. At the same time, users posting content of benefit is also sure that his community is benefiting from his content. However given a vast number of communities, users are limited in knowledge as to which community is the most relevant to their content or query and thus their reach is limited.

If this can be improved, content and queries can reach a greater audience which can achieve not only more narrowed down user engagement, it can also bring about multiple relevant perspectives to a particular post, be it a query or a content of benefit. As such, is the purpose of this project to classify textual posts to their relevant subreddits. The aim of the algorithm is to classify posts by any user to relevant subreddits based on the content of the post or hashtags mentioned by the author.

To begin, two subreddits have been identified to initiate a classification algorithm. They will be classified based on text alone to their individual subreddits. This will generate a model of high accuray which can then be worked upon to train more subreddits into the model to classify any text to relevant subreddits.

Objective of the project

1. Using Reddit’s API to collect posts from two subreddits
2. Use NLP to train a classifier on which subreddit a given post came from

Problem Statement

To create a classification algorithm that would automatically classify posts by anyone to relevant sub-reddit based on content of posts and hashtags mentioned by author

Workflow

Here lies the steps undertaken in this project

1. Data extraction

Data was scraped from two subreddits and converted into pandas data frame and saved as csv files

* Relationship advice:
* Legal advice:

1. EDA

Duplicate data were removed. Words were cleaned and transformed using lemmatising and word cloud was created to visualise data. Data was also split into training and testing set

1. Modelling and Evaluation

Modelling pipeline was used to generate several models

* Logistic regression with Count Vectorisation
* Logistic regression with TFIDF Vectorisation
* Multinomial Naïve Bayes with Count Vectorisation
* Binomial Naïve Bayes with TFIDF Vectorisation
* Random Forest and Extra Trees

After which, confusion matrix was used to evaluate the models based on accuracy score.

Conclusion and further recommendation

* Result: The logistic regression with count vectorisation performed the best with the highest accuracy
* Further recommendation: Currently, only the content of the post was used in classifying. As an improvement, other features such as the time the post was created or the number of comments or the type of comments could be used in classifying which might give other insights to post classification.