**Executive summary**

**Developing classification models in distinguishing source of subreddit posts**

“Reddit is a network of communities where people can dive into their interests, hobbies and passions.” Two popular subreddit communities discussing about American footballs are NFL(National Football League) and NCAA ([National Collegiate Athletic Association](https://en.wikipedia.org/wiki/National_Collegiate_Athletic_Association)). However, these two treads are not mutually exclusive. Occasionally, there are discussions about NFL in NCAA threads both mistakenly or intentionally, and vice versa.

In order to automatically distinguish the source of subreddit thread and identify or suggest an appropriate thread suitable for the post, classification models are developed using texts and/or qualitative features extracting from texts, such as length of text, numbers of words used.

There are 1,001 posts extracting from the NFL subreddit thread and 976 posts from the NCAA subreddit thread. After duplicated posts are deleted, there are totally 1,611 posts using in classification models developments. 20 most common word used in each subreddit thread are deleted because. After data are cleaned and explored, various classification models such as logistic regression, Naïve Bayes, and Random Forest are developed using 1) stemmed and lemmatized texts and/or 2) qualitative features extracting from texts, such as length of text, numbers of words used. Then the performance of each model is evaluated and compared using accuracy, sensitivity and precision metrics.

The results show that the best classification model using only text feature from the posts is logistic regression with Tfidf vectorizer and the hyperparameters are set as max\_features of 900, unigram, and English stopwords. The model can predict the source of post thread with 89.47% accuracy, 93% precision and 83% sensitivity. While, the best classification model using qualitative features extracting from texts is also logistic regression with number of words, number of unique words, number of characters, number of stopwords, number of punctuations, and average word length as predictors. The model can predict the source of post thread with 70% accuracy, 79% precision and 70% sensitivity. The best feature to predict the source of post thread is number of words, thus the model usually performs poorly when the posts are short.

In conclusion, using logistic regression classifier model on texts from the posts from two subreddit threads, we can classify an unlabeled post into NFL and NCAA thread with accuracy of **82.04%.** In future study, models using combination features of texts and quantitative features extracting from texts and more advanced classification models such as BERT should be explored. Moreover, more data about inappropriate threads of posts should be collected to develop a better model so that we can automatically detect unsuitable threads for a post and suggest a more suitable thread with higher accuracy.