Motivations for Language Learning Final Report

Madelyn Scandlen and **Shivali Pandya**
mscandlen3@gatech.edu spandya32@gatech.edu
College of Computing
Georgia Institute of Technology

Abstract

This project utilizes natural language processing methods to analyze why Reddit users are learning a new language (L2). Users are classified into motivation profiles based on their self-disclosed reason for learning Spanish (i.e. learning Spanish for a job, for a partner, moving to a new country, etc.). The model builds a motivation profile for each user and then classifies users into one of six groups based on their motivation: 0) native speaker, 1) culture, 2) interpersonal relationships, 3) school, 4) career, 5) traveling/moving. The models uses a text corpus of Reddit posts.

The code repository for this project can be accessed at github.com/mscandlen3/cs4650.

1 Introduction

The goal of the project is to identify why users in two subreddits are learning Spanish, identify which users are native versus non-Native Spanish speakers, and lastly analyze how motivations to learn Spanish may change over time. A model was created to classify L2 learners into proficiency levels by a written essay. This model was intended to be combined with the motivation classification to evaluate if a user's disclosed motivation could predict proficiency in their writing over time.

However, this goal has been adapted from its initial conception of classifying users' language motivations and understanding which language motivations lead to higher Spanish writing proficiency over time. In the data set, many users had only one or two posts and most posts were less than 50 words and most posts were in English. Even though some posts were in Spanish, users may have used online translation tools such as Google Translate or DeepL to write their posts, so the posts may not be a valid measure of Spanish language proficiency. Given the constraints of the data, the original goal shifted slightly to identify motivation from the post rather than proficiency level of the user.

1.1 Related Work

Learning can take place in a formal context such as in school or informally through conversations or reading books. Recently the COVID-19 pandemic has further underscored the growing trend of using online platforms for people to learn and connect. With this trend continuing, there is a growing body of natural language processing studies looking at language learning online. One study, "Learning in the Wild: Coding Reddit for Learning and Practice" focuses on creating a way to analyze informal learning content on Reddit by using data from four different Ask- subreddits (Kumar et al., 2018). Platforms like reddit are particularly unique because "participation engages self-motivated learners, occurs outside traditional professional settings (e.g., academic research, university lecture halls, workplaces), combines perspectives from experts and non-experts alike" (Kumar et al., 2018). This study created a way to understand how to set up a database using data scraped from Reddit.

Reddit as a platform is useful for its metadata about users. In another study, researchers scraped data from five subreddits including r/Europe, r/AskWomen, r/AskMen, r/relationship_advice, r/r4r and tried to train models to classify users' age, nationality, and gender (Pril, 2019). This study used three baselines of a random forest classifier, multilayer perceptron model, and K nearest neighbors. A graph convolutional network was created to maintain relationships between posts. All of the models were trained on the td-idf scores from the users' posts. The gcn model performed worse than the baseline models with a combination of a multilayer perceptron and K-nearest neighbors performing the best.

There is also a large language learning community on Reddit that would be interesting to study. Psychological researchers are interested in the importance of motivation in language learning, finding that urgent motivations such as communicating at a job or with a close relative will achieve a higher proficiency level, whereas motivations like being interested in culture will not achieve as high proficiency (Anjomshoa and Sadighi, 2015). Other psychologists have been interested in NLP analysis of motivations for participating and completing activies, such as Fukuoka et al.'s interest in evaluating women's motivation for exercise for health (Fukuoka et al., 2018). This work largely influenced our analysis and we hope to continue to expand this intersection of computing, language and psychology, so there is much to explore.

2 Data

For the task of classifying L2 learners by proficiency, a corpus from UC Davis Spanish courses was used, available for public use here. The data included for our model the essay, the essay id, the Spanish course level, and the calculated proficiency score. The calculated proficiency score was the rounded mean of the four self-reported proficiency scores in listening, reading, speaking, and writing. This calculation for the score seemed appropriate as the distribution resembled a normal distribution. Figure 1 separates the scores by the course level.

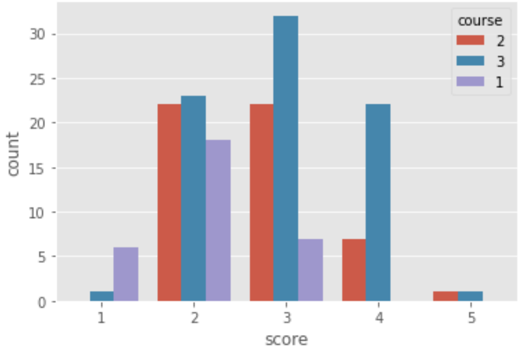

Figure 1: Countplot of Scores by Course

We pre-processed the data by first cleaning text by removing punctuation and special characters. Then we tokenized the essay and removed all the stop words. We padded the essays to a length of 200 words, which was the 75th quantile in length of words. We one-hot encoded the calculated proficiency scores to be used as ground truth labels in training and validating our model. We had 164 observations of 200 words to input to our model.

The Reddit data is a corpus of posts in 2019 to the subreddits *r/Spanish* and *r/LearnSpanish*, retrieved at https://files.pushshift.io/reddit/. The post were organized JSON objects and then loaded into a Pandas DataFrame containing information about the text, the timestamp of posting, the parent post, the original post, the author, and metadata about the author.

To clean the data, most metadata fields were discarded, so we only included information about the author, the author flair, the body of the post, and the timestamp. Any posts that did not have author information were removed.

For preprocessing, we first performed regex search on the author flair text to extract self-identified learners. We matched {*r"learner", r"heritage", r"native", r"[a-z]{1}[0-2]{1}", r"student", r"beginner", r"intermediate", r"advanced"*} to assemble a corpus of 73371 posts from identified learners.

We then manually annotated 800 posts from learners into 7 motivation classes: {0: "no motivation", 1:"culture", 2:"relationship", 3:"school", 4:"career", 5:"travel", 6:"heritage"}. We removed all posts labeled 0, and ended with 222 samples classified into 6 profiles. We subtracted one from each class to include 0, as shown in the categorical distribution in Figure 2.

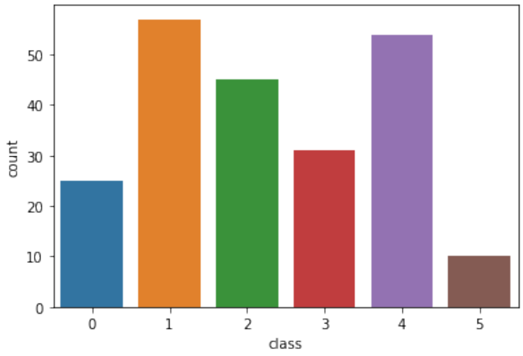

Figure 2: Countplot of Reddit Post Classes

A summary of the Reddit data is shown in Table 1.

Unique users	12092
Total number of posts	73371
Number of posts in *r/Spanish*	52601
Number of posts in *r/learninspanish*	20770
Number of posts with a flair	42699
Unique user flairs	806

Table 1: Summary of Reddit Data

3 Methods

We first performed *tf-idf* vectorization on the posts to create a Bag-of-Words classification task. We considered only 2000 features that occurred in a minimum of 5 posts and a maximum of 80% of the posts. This resulted in a final data frame of 222 samples with 535 features.

We performed Principal Component Analysis on the vectorized posts and found that only the first feature had high variance. We graphed only the first two features from the PCA scatter plot visualization for 6 classes, decided on by the qualitative annotation process. Figure 3 shows the scatter plot colored by the actual labels annotated by hand, whereas Figure 4 shows the scatter plot colored by the predicted clusters from the K-means analysis.

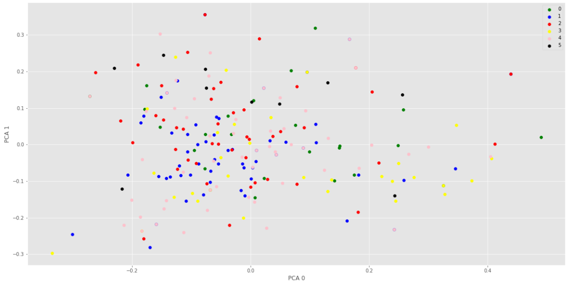

Figure 3: Scatter plot for Annotated Classes

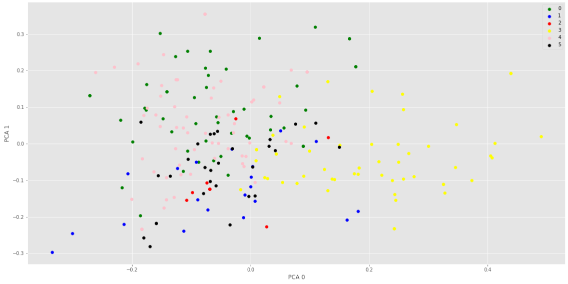

Figure 4: Scatter plot for K-Means Clusters

We chose to perform k-means clustering on the data before manually annotating it to see if there were any obvious clusters that emerged. We tested different numbers of clusters from 1 cluster to 6 clusters. We then used the elbow method to find the optimal number of clusters, shown in Figure 5. The elbow method identified 2 clusters as the optimal number, which is supported by the visualization, though we concluded that more clusters should be used after our annotation process. Using only two clusters would be the equivalent of creating a binary classifier and upon looking at the data qualitatively, having only two clusters would be a gross oversimplification of the data though it would lead to a cleaner split between the two classes.

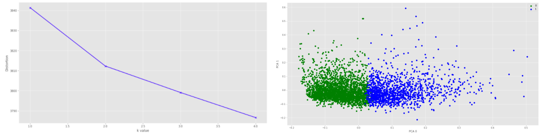

Figure 5: K-Means Elbow Criterion and Clusters

We ultimately decided that the K-means clustering was not capturing enough qualitative information about the data and manually annotated the posts instead. We also were unable to automate a process of throwing out posts that did not contain motivation through regular expression search, so we incorporated labeling posts to discard in the annotation process.

4 Models

To classify Spanish learners into proficiency levels, we created a bidirectional LSTM with four layers. The input to the LSTM was an array of all the essays padded or truncated to a length of 260 word tokens, the 75th quantile of essay length after removing stop words. This input was passed to an embedding layer with an embedding dimension=64 and an input size = 5000. This output was passed through a bidirectional LSTM layer with an output size of 64. There was a dropout layer that dropped 20% of the sample, and the output was passed to a dense layer that performed softmax activation function and outputted the relative probabilities for the 5 classes.

For the second task of classifying posts in motivation profiles, we built another bidirectional LSTM to compare against baseline probabilistic and regression models. The input to this LSTM was the body of the post, padded or truncated to a length of 57 word tokens, also the 75th quantile of post length after removing stop words. The model had four layers: an embedding layer with an input size of 5000 and embedding dimension of 64, a bidirectional LSTM layer with an input size of 64 and an output size of 128, a dropout layer that sampled 80% of the data, and a softmax activation layer that outputted a probability distribution for 6 classes. Because the labels were categorical and lacked a numeric relationship, we chose a sparse categorical cross entropy loss function.

4.1 Baseline Models

The baseline models we used were a probabilistic classifier and a regression classifier to compare against the sequential model. We had initially intended to only perform Bag-of-Words classifications that disregarded sequential modeling, but when the models did not perform as well as expected, we decided to use them as a baseline to compare to the later implemented LSTM.

We first started with a supervised multinomial Naive Bayes classifier with no-smoothing. The input was the vector space of tf-idf scores for all the annotated posts. We then hypertuned the parameters to use add-one smoothing, which performed a small amount better.

We created a multinomial Logistic Regression classifier to compare. The input was also the vector space of tf-idf scores for all the annotated posts. We hypertuned the parameters of this model by choosing liblinear optimization and added a random state shuffle.

5 Results

5.1 Experiment Setup

For the proficiency classification, we calculated the mean of the four score components (listening comprehension, reading comprehension, speaking ability, and writing ability) and rounded to have an integer 0, 1, 2, 3, 4 score for every essay. We performed one-hot encoding on these scores to use as ground truth labels to validate the model's predictions. We choose *categorical cross-entropy* as the loss function because of the one-hot encoding and *adam* as the optimizer.

For the motivation classification, we used manually annotated labels as the ground truth labels for the posts. We identified 22k posts as being made from learners, and two annotators "read the content of the post and identify the user's motivation to learn Spanish as one of the following categories: 0) no motivation mentioned, 1) culture, 2) interpersonal relationships, 3) school, 4) career, 5) traveling/moving " for 3.5% of the sample. For posts which mentioned multiple motivations, the data point was duplicated for each motivation mentioned, and each of the duplicate data points were marked as one of the different motivations. There was a 100% agreement score among the annotators for 800 posts.

The probabilistic and regression models trained and validated on the tf-idf vectorized posts and categorical labels, using precision, recall, and f1-scores as the metrics for comparison. The sequential model was trained on padded and cleaned posts with the corresponding categorical labels, and accuracy and loss were used as metrics to compare to the baseline models.

5.2 Model Performance

For the proficiency classification, we initially chose an 80-20 train-test split on the corpus, with the LSTM model training on 99 samples and validating on the other 65 samples, but we found that our model was overfitting to the training data. After 20 epochs, our training accuracy was 78% and training loss was 0.7071. The validation accuracy was 38% and validation loss was 1.4313. We reduced the number of epochs to 15 and increased the validation set to be 30% of the samples to find improved results. The new training accuracy and loss were 65% and 0.9779, and the validation accuracy and loss were 44% and 1.3287.

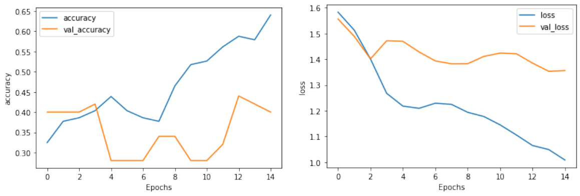

Figure 6: Proficiency LSTM

For the motivation classification, we first chose a 80-20 train-test split of the data. The multinomial Naive Bayes model without smoothing initially achieved 20% accuracy. We added add-one smoothing and increased the validation set to be 30% of the data and achieved 36% accuracy. The f1-score for each class is shown in 2.

class	f1-score
0	0.00
1	0.48
2	0.36
3	0.00
4	0.41
5	0.00

Table 2: Smoothed Naive Bayes

The confusion matrices for the initial Naive Bayes and the smoothed Naive Bayes models are pictured in Figure 9.

The model was unable to correctly identify posts that fell under motivation categories of { 0:native speaker, 3:interpersonal relationships, and 5:traveling/moving}. The multinomial Naive Bayes model may make a faulty assumption that all features are independent. For example, the content of the post and the users' self identification flairs about their fluency in the language may be related. In addition, posts may contain multiple reasons for why a user is learning Spanish which may confuse the classifier, as demonstrated in Figure 7.

"*I live in South Florida where it is densely populated with Hispanic people. It's to the point where sometimes you need to speak Spanish for someone to help you find what you're looking for in the grocery store. You also need to speak Spanish to apply for most jobs.*"

Figure 7: post by a Spanish learner on Reddit

The poster may be motivated to learn Spanish to be considered for a job or to participate actively in the community or to partake in daily activities.

The multinomial Logistic Regression achieved 38% accuracy on the 80-20 train-test split. When adding liblinear optimization and a random state shuffle, the model achieved 44% accuracy. The f1-score for each class is shown in table 3.

class	f1-score
0	0.00
1	0.42
2	0.40
3	0.00
4	0.69
5	0.00

Table 3: Optimized Logistic Regression

The sequential LSTM was split 70-30 into train-test sets. After 10 epochs, the model achieved a 88% training accuracy and 0.8258 training loss and 46% validation accuracy and 1.5772 validation loss.

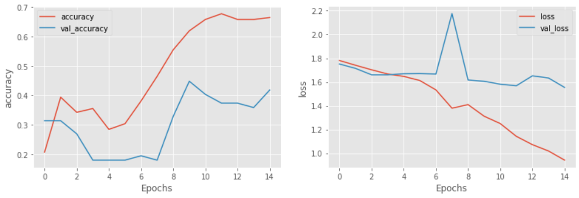

Figure 8: Motivation LSTM

5.3 Result Comparison

The Logistic Regression models performed much better than the Naive Bayes models, especially upon optimization. The Logistic Regression also performed better than the sequential LSTM, suggesting that our initial idea that a Bag-of-Words captures a better picture of motivation classification than preserving sequential dependencies. However a true comparison between the Logistic Regression and sequential LSTM cannot be drawn, as the Logistic Regression uses the tf-idf vector as its features to predict on and the LSTM uses the tokenized sequence.

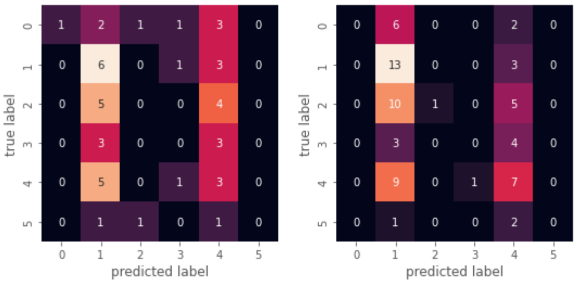

Figure 9: Confusion Matrices for non-smoothed and smoothed Naive Bayes models

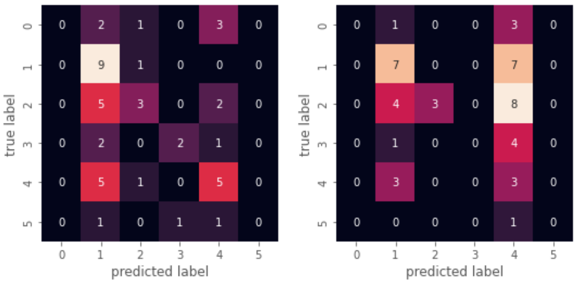

Figure 10: Confusion Matrices for non-optimized and optimized Logistic Regression models

The confusion matrices for the probabilistic and regression models are shown in Figures 9 and 10.

5.4 Hyper-parameter Tuning

In our initial exploration of data, we performed hyper-parameter tuning on our K-means clustering in order to find the most promising clustering. We tested was the initial number of clusters such as 6 (since this was the number of classes we have annotated), 10, 12, and 15. The model improved as we moved from 6 to 10 clusters. However, with 12 and 15 initial clusters, the model appeared to classify all of the posts into culture or career depending on which data points the centroids were initialized to.

A third hyper-parameter that we tested was how the initial centroids were initialized. Instead of using a random method for initialization, we followed a k-means++ model where centroids are initialized by picking points farthest from the previous centroid. We chose to explore the K-means clusters in depth as this topic is more exploratory for future expansions and k-means clustering was important in previous literature (Fukuoka et al., 2018).

5.5 Interpretability and Negative Results

The results of both the proficiency classification and the motivation classification were not enlightening to the overall idea. Almost every model overfit to the training data and did not accurately predict classes for the data. This is likely due to the limited data and lack of balance in the ground-truth labels.

For the proficiency data, we had 164 data points and a normal distribution of proficiency scores, seen in Figure 11. However, this caused the classes to be imbalanced and was likely a reason the model overfit so much; there were not enough points in the other classes. Potentially k-fold cross validation could alleviate this issue of imbalanced classes and would be implemented in future undertakings of this topic.

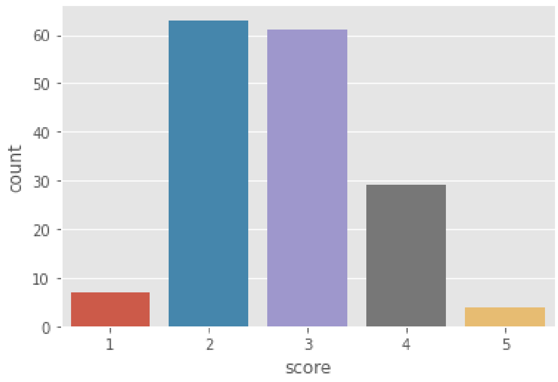

Figure 11: Proficiency Scores Countplot

The motivation class also lacked enough data points, due to the constraint of needing manual annotations. We ended up annotating only 3.5% of the learner data and we had to discard almost 75% of those annotations for not containing motivation. To fix this problem, having more annotations would be preferable to build a larger structure of the data. Semi-supervised classification using k-fold cross validation may also be a possible solution so that the data trains and validates with the ground-truth annotations and without them.

5.6 Synthesization

The overarching goal of the project was to perform NLP analysis on the motivations of language learners. First we wanted to see if we could build a model to predict proficiency level from a language learners' essays in the L2. Then we wanted to analyze the meta-information about a language learner shared on a social media platform such as Reddit. The ultimate task was to combine these earlier two tasks and see if motivation can predict proficiency, especially over time.

This lack of defined results prevented the final task from being attempted, as neither the proficiency nor motivation models were promisingly functional. The results were not necessarily negative however, as accuracy was greater than random guessing for all models. Ultimately more data is needed to see the patterns more clearly. The confusion matrices in Figures 9 and 10 show that the probabilistic models did not have enough data to learn some of the classes at all.

We did begin to explore users' engagement on the language learning subreddits over time. Figure 12 looks at the frequency of motivations over the first nine months of 2019. For example, in the months of July, August, and September when schools are starting, there are more posts that were identified as learning Spanish because of classes and school. The data, however, limited our ability to do a long term sequential analysis by user since many users did not have multiple posts that were far apart in time.

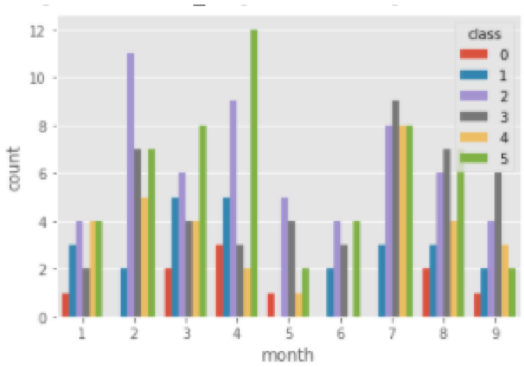

Figure 12: Distribution of motivation classes over time

5.7 Work Division

Madelyn pre-processed the data and created the probabilistic and sequential models. Shivali tuned the hyper-parameters and created data visualizations. Both partners annotated the data and wrote the final report.

6 Conclusion

The lack of data points for each class did not produce significant results for the analysis of language learning. However, with an increase in sample size, the patterns that are demonstrated in this initial experiment will only become more significant if the patterns exist.

The model to classify users by proficiency performed better than random assignment but the results were not exciting. The highest performing model to analyze motivation of Reddit users to learn Spanish was a supervised learning Logistic Regression model with six clusters on *tf-idf* vectorized posts. The largest division in motivation was between people motivated to learn Spanish because they are moving to or visiting a Spanish speaking area versus people learning Spanish for a job. There was overlap in the motivations for school/classes and visiting/moving. People learn languages for reasons which are often dependent and related. For example, a person may move to a Spanish speaking area (visit/move) for a job (career). An experiment with more independent features is needed to see a clearer stratification. The motivations may also change over time such as being motivated to learn Spanish because of school, forgetting some of it after graduating, and then being motivated to learn it again because of a job.

Potential future developments include creating a way for the model to identify posts that refer to multiple motivations. Another development could also be collecting time series data from users' posts and modeling how sentiment/motivation of learning a language from a user changes over time. Data could also be collected to analyze the frequency by which users post and how that might relate to their motivation level to learn Spanish. For example, are users who post on the thread weekly more motivated to learn Spanish compared to users who post monthly? Another future development would be to further annotate the data (at least 10,000 posts) and re-run the models. Lastly, the data could be reformatted so that there is one document containing all of a users' posts and then each user could be classified into a motivation group rather than the model classifying the posts themselves. The idea of building motivational profiles, inspired by Fukuoka, is largely undiscovered and it lays at an intersection of language, psychology, and computing (Fukuoka et al., 2018).

7 Ethical Impact

This data can be used in a positive context in order to provide better resources to L2 Spanish learners. For example, someone trying to learn Spanish because they are trying to move to a new place might want to be connected to native Spanish speakers in the place they are moving to so that they get used to that specific dialect of Spanish. Understanding reasons why people learn languages in general might help promote learning a new language and becoming bilingual. There have been documented neurological benefits to speaking multiple languages such as slowing cognitive decline as someone ages and executive control while switching tasks (Gold et al., 2013). There could, however, be some negative implications for personal data privacy if data on people's motivation to study language are shared with for-profit language learning companies such as Rosetta Stone or Duolingo who could then use the data to market their products. There is a rising concern in the ethics of using users' data on social media platforms and attempting to build a profile about the person.

8 Code Repository

The project can be accessed at the GitHub repository github.com/mscandlen3/cs4650.

References

Leila Anjomshoa and Firooz Sadighi. 2015. The importance of motivation in second language acquisition. *International Journal on Studies in English Language and Literature*, 3:126–137.

Yoshimi Fukuoka, Teri Lindgren, Yonatan Dov Mintz, Julie Hooper, and Anil Aswani. 2018. Applying natural language processing to understand motivational profiles for maintaining physical activity after a mobile app and accelerometer-based intervention: The mped randomized controlled trial. *JMIR mHealth and uHealth*, 6.

Brian T. Gold, Chobok Kim, Nathan F. Johnson, Richard J. Kryscio, and Charles D. Smith. 2013. Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *Journal of Neuroscience*, 33:387–396.

Priya Kumar, Anatoliy Gruzd, Caroline Haythornwaite, Sarah Gilbert, Marc Esteve del Valle, and Drew Paulin. 2018. Learning in the wild: Coding reddit for learning and practice. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, pages 1933–1942.

Robin De Pril. 2019. User classification based on public reddit data.