Vocabulary Models of Informal Language Production in Reddit

Alexander Brooks, Andrew Turner, and Matthew Berland
albrooks@cs.wisc.edu, ajturner3@wisc.edu, mberland@wisc.edu
UW–Madison

Abstract: In this work, we present an analysis of word choice data within several sub-communities of the social media website reddit.com. We examine how the notion of a community member's "veterancy" within the community affects their language production.

Introduction

Social media makes it easier for learners to find and interact with experts in a wide variety of fields, and social networks enable us to acquire and analyze discourse data on a vast scale. Participation in informal online communities has been examined through a variety of lenses, but there remain open questions around how participation affects the language learners use in these spaces. Word choice is a convenient level of analysis, as vocabulary is easily encodable and parseable. In this poster, we apply machine learning (ML) and natural language processing (NLP) tools in a novel domain, exploring the impact of veterancy on word choice in three informal language communities within the social network Reddit (reddit.com). This work is significant because it demonstrates that relatively rudimentary ML and NLP methods can be used to characterize the developing expertise of users. We used the following research question to guide our work: in communities with multiple "types" of participation, does (and how does) veteran vocabulary differ from the other types' vocabularies?

Prior work

Computational sociolinguistics includes a significant body of work which blurs the lines between online personas and offline identities. In their survey of the field, Nguyen et al describe this subset of work as analysis of online language production to "automatically infer social variables from text" (2016).

We extend this body of literature by considering language production through the lens of *veterancy* within the community. We define a veteran in an online community loosely as someone whom a member of the community might consider by virtue of experience to be an authority on the community. Quantity of participation is thus a symptom of veterancy. Veterancy itself is not a totally novel demographic in this field: Fields et al (2017) consider length and frequency of membership in their analysis of programming patterns in the Scratch online community. Veterancy is similar to expertise, examined in the Scratch online community by Huang and Peppler (2019), but whereas an expert is presumed to be objectively better at some relevant task, veterancy does not include this connotation.

Methods

We addressed our research question through vocabulary classification via a recurrent neural network, trained unsupervised on a randomly selected half of two or more given post corpi. It considers each post using a "bag of words" model that entirely removes any order information, and removes the repetition of words. We generate our bag of words using the Natural Language ToolKit (NLTK) default word tokenizer, which conducts a number of small optimizations such as separating common contractions into their component words (Bird et al, 2009).

In binary classification settings, we calculate the "receiver operating characteristic area under the curve" (ROC-AUC) for our input sets. This provides a measure of true positives (correct guesses by the classifier) in comparison to false positives (incorrect guesses). We calculate this score for both the test set and the training set. Higher ROC-AUC values above 0.5 indicate more success separating the languages of the input sets. A ROC-AUC value on the test set comparable to that on the training set indicates that we avoided overfitting to features present specifically in the randomly selected training set posts.

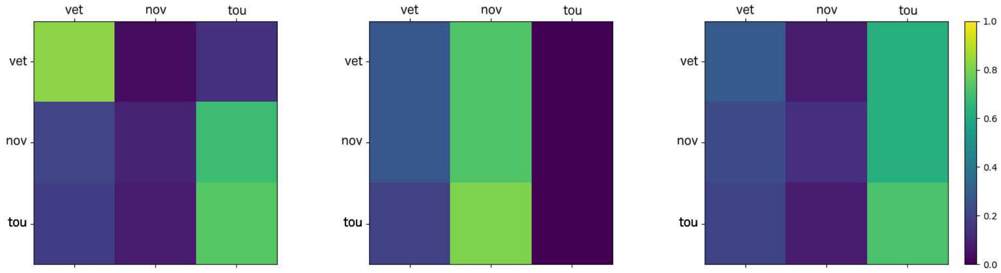
In classification settings with more than two input sets, we use a subset of the test set to plot a "confusion matrix" indicating the true source of each post and the provenance that our classifier guessed for that post. Heavy concentrations along the main diagonal (where the post really was from source A and the classifier guessed source A) indicate successful classification. Heavy concentrations along a vertical line indicate that the classifier mistook posts from many different sources as being part of the same language, and therefore from the same source. As in the binary classification case, we compare a test set confusion matrix against a training set confusion matrix, with significant differences between the two indicating overfitting.

With a default veterancy threshold of 10 posts (a parameter we adjusted throughout our experiments without noticeable effect), we divide a given subreddit's posts into a veteran set, a novice set, and a tourist set.

We measured confusion between these three sets for each of our chosen subreddits. We then repeated the classification process with a binary classification focusing on the pair of groups which the confusion measure suggested were most different to get an ROC-AUC approximation of how different the vocabulary distributions were.

Results and discussion

In r/ArtFundamentals, tourists and novices were not distinguishable from each-other, but veterans were with a true positive rate >0.8 and low false positive rates. In Figure 1 below, we report the confusion of our classifier for each subreddit. Each row indicates the true source of a post, and the column indicates the guess that the classifier made. Thus, we can see in r/productivity that the majority of all posts were guessed to be novice posts, indicating a failure of the classifier to learn the difference between categories. A similar pattern is displayed in r/vim, where most posts were guessed to be tourist posts. Rerunning the categories in pairs produced ROC-AUCs of 0.59 for novices vs tourists, 0.53 for novices vs veterans, and 0.47 for tourists vs veterans in r/productivity; in r/vim we found ROC-AUC scores of 0.55 for novices vs tourists, 0.52 for novices vs veterans, and 0.55 for tourists vs veterans.



(a) (b) (c)

Figure 1. Confusion matrices for veteran posts, novice posts, and tourist posts in r/ArtFundamentals (a), r/productivity (b), and r/vim (c).

The significantly higher distinguishability measurements for veterans in r/ArtFundamentals suggest a meaningful distinction between veteran and non-veteran members of that community. This distinction, and the lack of detectable difference in r/productivity and r/vim, could be attributable to the teaching and learning focus of the r/ArtFundamentals community, with community members learning to produce vocabulary in a distinct distribution as they gained veterancy. Our results highlight how structurally similar communities (in our case three sub-communities within the same social media website) may exhibit distinct patterns of language change as members participate in these communities. Our work provides an example of how those patterns in language change can be measured quantitatively, paving the way for future research on the sources of variation across these communities.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Every publicly available Reddit comment. (2015). Retrieved from https://www.reddit.com/r/datasets/comments/3bxlg7/ i_have _every_ publicly_available_reddit_comment/

Fields, D. A., Kafai, Y. B., & Giang, M. T. (2017). Youth computational participation in the wild: Understanding experience and equity in participating and programming in the online scratch community. *ACM Transactions on Computing Education (TOCE)*, 17(3), 15.

Huang, J., & Peppler, K. (2019). Studying Computational Thinking Practices Through Collaborative Design Activities with Scratch.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational linguistics*, 42(3), 537-593.