Computer Aided Analysis of the Evolution of the Modern English Lexicon Milestone

Hans Alafriz

Abstract

The purpose of this study was to study the use of idioms and slang in everyday language, particularly when used in online social media such as Twitter. This was done by browsing public social media posts for relevant data points between 2010 and 2019, and using a supervised learning assisted k-nearest neighbors algorithm to determine the likely time period that the post was made, and to determine when new idioms are created and begin to join the modern lexicon. It was found that for the most part, new slang tends to join modern lexicon and steadily increase in usage after their initial usage but start to lose prevalence within 5 years, except for certain words that have other meanings, such as “shade” or “salty” which maintain their usage.

1 Technical Details

The project makes use of a vector that represents the dictionary of possible slang to look out for in each post. It uses a list of known slang words by year as training cases to use a k-nearest neighbor algorithm to predict the posting date of the test cases. The program asks for test cases from text files given from the user, which the model then predicts the year that it was published according to the slang used, and then the program asks the user what year it was actually published to calculate the error rate, adding the test data to the training data for future comparison, continually looping through until given an invalid filename. The model then uses the test data to graph the individual word usage by year, along with the error rate.

2 Data Analysis

From the current machine learning model, it was found that the use of individual idioms and slang seem to be used at least once by the year that they first appear, as seen in the figures. Occasionally slang is used at least once more within 2 years beyond their initial use. The model’s error rate however seems to have an increased error rate as it attempts to predict the publication date of Twitter posts later than 2012.

3 Interpretation

From the machine learning model, it was found that the use of individual idioms and slang seem to increase in frequency for about 2-3 years, after which their usage starts to wane for 1-2 years after that. It was also found that there was an increase in the variety of slang used towards the later end of the decade. There are also various pieces of slang that maintain steady use throughout the decade, complicating the model, requiring that the model be adjusted to account for essential slang used for the specific social media platform (Twitter).

4 Figures

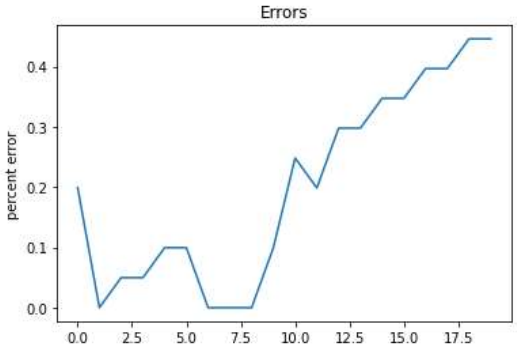

Figure 1-Error Rate Graph

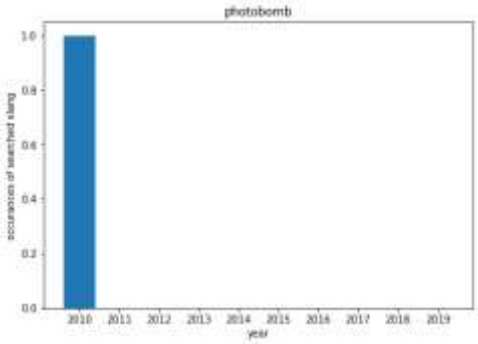

Figure 2-usage of "photobomb" by year

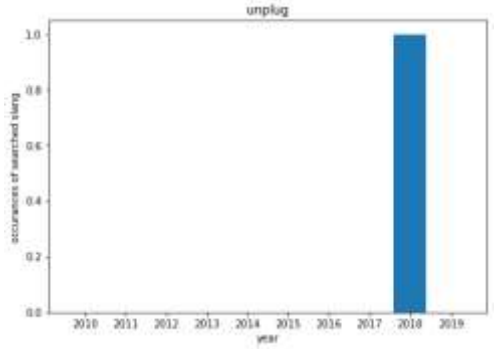

Figure 4-usage of "unplug" by year

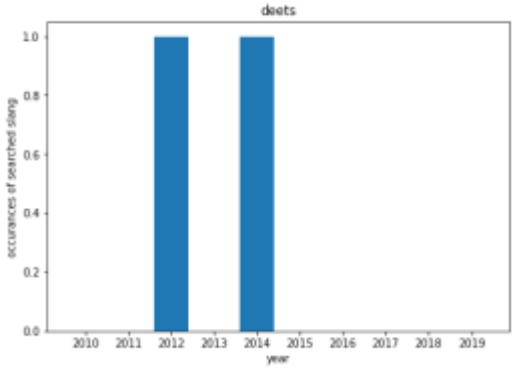

Figure 3-usage of "deets" by year

References

#slang words research
#https://bestlifeonline.com/2010s-slang/