An Analysis of the Spread of Slang

age2, ddecastr, dsmits, dromano

Hypothesis

Language is a living, ever-changing phenomenon. Slang is at the leading edge of this dynamism, as young people constantly invent new words and phrases. With the onset of the digital age, slang has proliferated into an integral part of online communication. But how does slang rise and fall? Do certain social networks facilitate the spread and use of slang more than others? In particular, we investigate the behavior of how different slang terms gain and lose popularity over time and whether different slang terms exhibit different lifecycles on Twitter versus Reddit. Our hypothesis about the rise and fall of slang is that different slang terms demonstrate different patterns of increase and decrease in popularity. Comparing slang terms' lifecycles on Twitter versus Reddit, we hypothesize that Reddit will see the peak of slang terms come sooner than they do on Twitter.

Data

Due to the massive number of Tweets and Reddit comments, we focus on just 2015. First, we assembled a set of slang terms popularized around this year. We then compiled a database of nearly 10,000,000 tweets from March-December, 2015 using the Internet Archive's Twitter stream grab and over 6,500,000 Reddit comments from PushShift's archive. We then filtered for English language data and multi-hot encoded each point for instances of each slang term.

Methodology

To investigate whether slang terms exhibit different lifecycles of growth and decay, we calculate the proportion of (appearances/total tweets) for each month and run consecutive proportional z-tests using the statsmodels package in Python. To negate possible false positives, ie p-hacking, we are only interested in multiple consecutive months of statistically significant differences in usage at the $p < .05$ significance level, which has at most a $.05*.05 = .0025$ probability of happening by chance or a $p < .05/9 = .0056$ Bonferroni-corrected significance level from month to month.

To determine whether slang spreads faster on Reddit versus Twitter, we compare the number of months it takes each slang term to reach its peak on both platforms using the proportional chi-square test, as we are ultimately dealing with count data, also using the statsmodels package. By our definition, if Reddit were to see each of the slang terms peak in March, the first month of our data, it would have a "Months to Peak" of 0, and if we were to see each of the slang terms peak in April, it would have a "Months to Peak" of 1 month * 9 slang terms = 9. There are 81 total slang-months, which is the total amount of time all slang terms contribute to our study (9 months x 9 slang terms).

	Months to Peak	Total Slang-Months
Twitter	58	81
Reddit	53	81

Findings

Claim #1: Different slang terms on Twitter exhibit significantly different usage behavior.
Support for Claim #1: We support our claim with the behaviors of two 2015 slang terms: “low-key” and “fomo”.

Slang	Month	Proportion	P-value
low-key	3	0.000960415	None
low-key	4	0.00104541	0.30757
low-key	5	0.000873777	0.00903
low-key	6	0.000982763	0.02249
low-key	7	0.00107924	0.00079
low-key	8	0	0.19855
low-key	9	0.000980901	0.22033
low-key	10	0.000912458	0.18996
low-key	11	0.000868066	0.34646
low-key	12	0.0018759	0.0

Slang	Month	Proportion	P-value
fomo	3	1.27489e-05	None
fomo	4	1.56421e-05	0.77201
fomo	5	1.95914e-05	0.6622
fomo	6	2.21214e-05	0.72408
fomo	7	1.33288e-05	0.01934
fomo	8	0	0.88644
fomo	9	7.66329e-06	0.91377
fomo	10	2.64068e-05	0.02566
fomo	11	3.28347e-05	0.4341
fomo	12	9.04047e-06	0.01349

Here we see by our definition that “low-key” has a statistically significant decrease from April to May, followed by significant increases From May to June and June to July, and finally a significant increase from November to December. While “fomo” does see some significance at the .05 level, it doesn’t quite meet our definition of statistically significant at any point.

Claim #2: There is no evidence to suggest that Reddit sees slang develop before Twitter or vice versa.
Support for Claim #2: We support our claim with the previously described chi-square test, which yields a test statistic of 0.72 giving a p-value of 0.40, meaning there is not a statistically significant difference between Twitter and Reddit Months to Peak.