stopTryingToMakeFetchHappen

age2, ddecastr, dromano, dsmits

INTRODUCTION

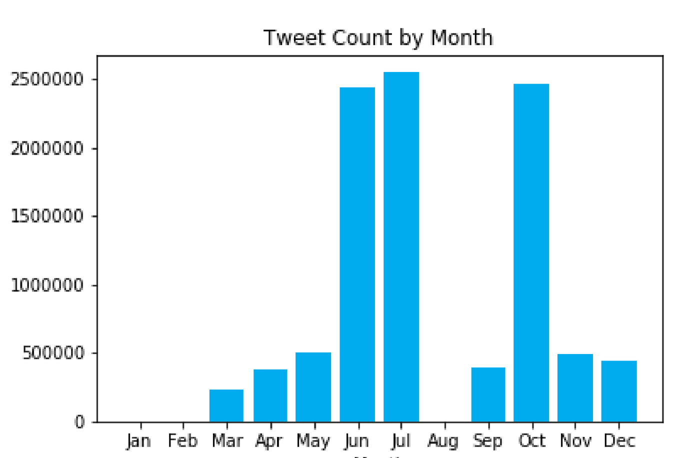
Slang has always been an integral part of online communication, with new slang terms arising every year. Slang provides insights into how language spreads online as it is popularized through social media. We focus on two questions: do different slang terms exhibit different lifecycles (catching on, falling off) on social media? Do some social networks, like Reddit, facilitate faster spreading of neologisms than others, like Twitter? Our hypotheses are first that different slang terms will have different lifecycles, and second that slang peaks earlier on Reddit than on Twitter.

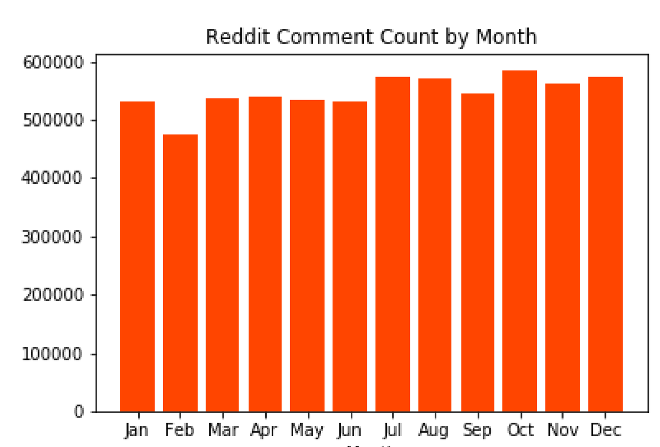
DATA

Our data comes from online archives of Twitter and Reddit. The amount and quality of data caused us to alter our analyses: we initially planned to analyze slang over four years on four social media platforms. Given the thousands of gigabytes of data on each platform each year, we focused on just tweets and Reddit comments from Mar-Dec 2015. Our slang terms were selected by looking at the intersection of online lists of slang popularized in 2015.

To preprocess our data, we took random samples of tweets and comments from each month, filtering non-English posts and irrelevant metadata, leaving 9,905,505 tweets and 6,558,630 Reddit comments. We then encoded each datapoint as a multi-hot vector for instances of each slang term.

The primary shortcomings of our Twitter data are that we could not obtain data from Jan-Feb 2015, and that our August data is sparse (less than 0.2% of other months). The primary shortcoming of our Reddit data is that it is limited to only comments. We keep these limitations in mind throughout our analyses. We relied on heuristics of how slang is used on social media, while avoiding confounding substring matches ("I love Africa" for "af") or homographs. Without external human validation, we cannot guarantee our data catches all of each slang's use cases and thus cannot compare between terms.





METHODOLOGY

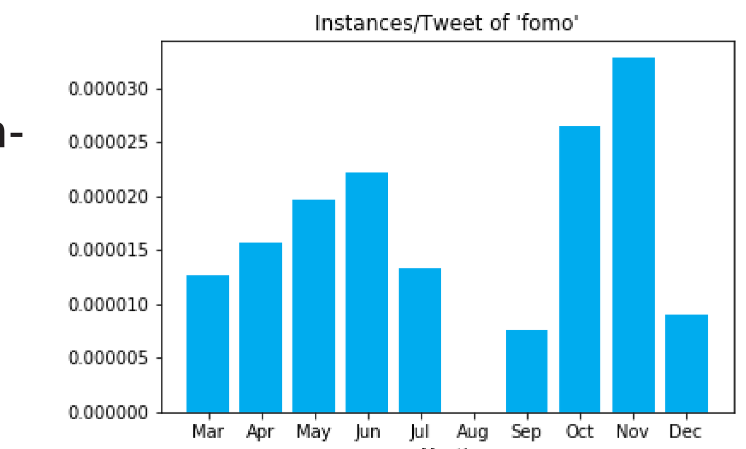
Experiment 1: Our first investigation is whether slang terms exhibit different lifecycles of growth and decay. We use the statsmodels package in Python for all statistical testing. To analyze the monthly differences in slang usage, we focus on the Twitter data and use the proportions_ztest function to compare the slang co unt per month, normalized for total tweets per month. To avoid an increased false positive rate, we define statistical significance as either multiple consecutive months of statistical significance at the .05 level (p<.05*.05=.0025) or Bonferroni-correction month-to-month analysis (p<.05/9 =.0056). We use the default two-sided test and would accept our original hypothesis if we find that at least two words exhibit different patterns in statistical significance. We would reject it otherwise. Note: due to potential confounds in preprocessing (discussed above), we cannot directly compare proportions between slang terms.

Experiment 2: To answer if slang spreads faster on Reddit versus Twitter, we first define "slang-month" —analogous to person-years, as the total amount of months all slang terms contribute to our study, in order to quantify the possible amount of time it takes for slang terms to reach their peak. We investigate our slang terms from Mar-Dec, giving us 81 total slang-months (9 slang terms x 9 months). We then use the proportions_chisquare test to compare the ratio of the total amount of months it takes each slang term to reach its peak over total slang-months on Reddit versus Twitter. If the null hypothesis were true and slang does not peak at different times on Reddit vs Twitter, we would not expect a significant χ2 value. If slang does peak earlier on Reddit or Twitter, we would expect a significant χ2 value (χ2 > 3.84 and p<.05).

RESULTS

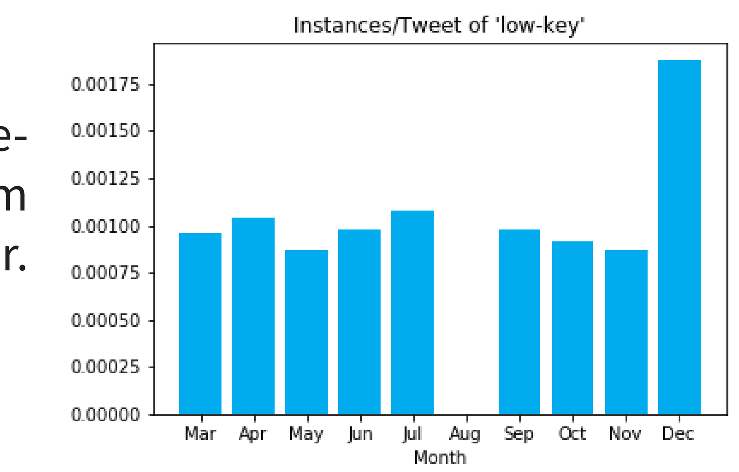
Question: Do different slang terms exhibit different behavior in terms of catching on and falling off?

We claim that different slang terms exhibit different usage behavior in terms of catching on and falling off. Some slang terms, such as "fomo", by our operationalization, show no statistically significant periods of change in usage.



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

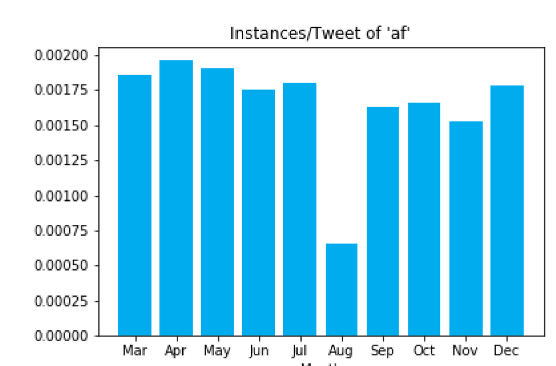
Other slang terms, such as "low-key", show distinct and significant patterns of change. Below, we see a significant decrease from April to May, followed by significant increases from May to June and June to July, and finally a significant increase from November to December.



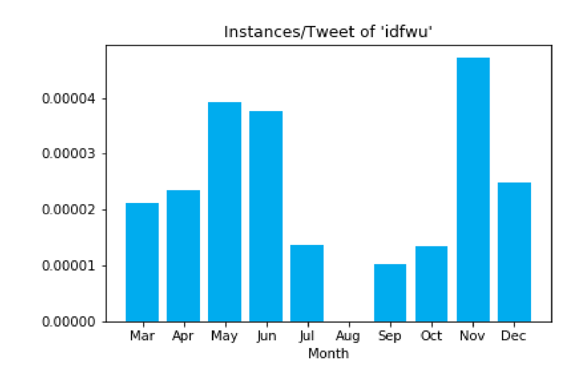
Slang	Month	Proportion	P-value
low-key	3	0.000960415	None
low-key	4	0.00104541	0.30757
low-key	5	0.000873777	0.00903
low-key	6	0.000982763	0.02249
low-key	7	0.00107924	0.00079
low-key	8	0	0.19855
low-key	9	0.000980901	0.22033
low-key	10	0.000912458	0.18996
low-key	11	0.000868066	0.34646
low-key	12	0.0018759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

The results from the rest of the slang terms:

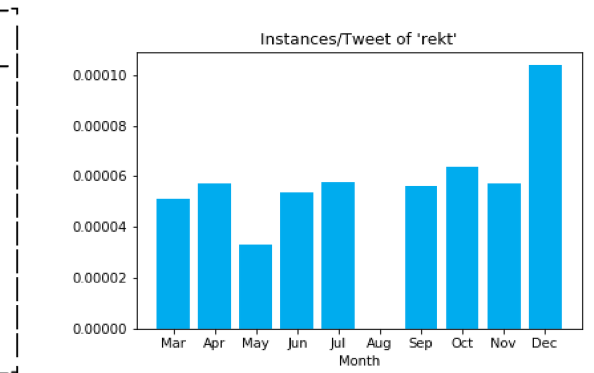
Month	P-value
3	None
4	0.35548
5	0.51694
6	0.02519
7	0.21618
8	0.2891
9	0.34247
10	0.75421
11	0.03723
12	0.00234



Month	P-value
3	None
4	0.85884
5	0.19647
6	0.87474
7	0.0
8	0.8848
9	0.9005
10	0.60682
11	0.0
12	0.07526

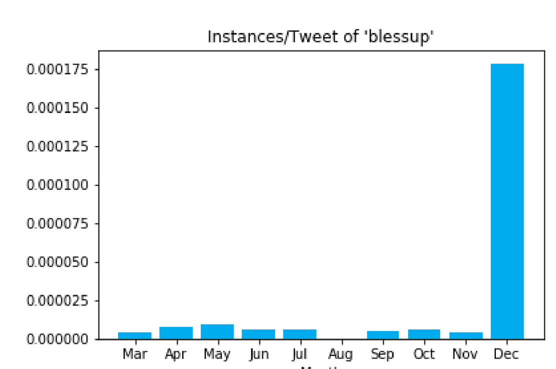


Month	P-value
3	None
4	0.74318
5	0.08838
6	0.06175
7	0.55313
8	0.76651
9	0.76934
10	0.57783
11	0.6107
12	0.01206

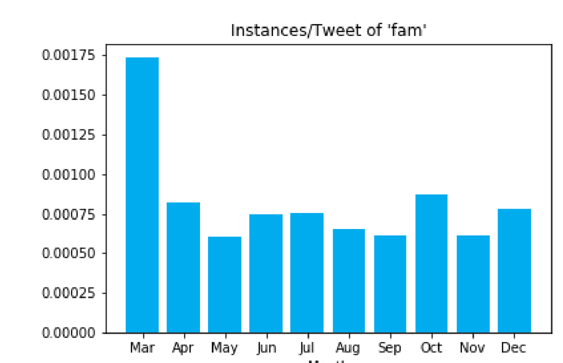


Month	P-value
3	None
4	0.92739
5	0.1486
6	0.0
7	0.0
8	0.57832
9	0.41082
10	0.02301
11	3e-05
12	0.78508

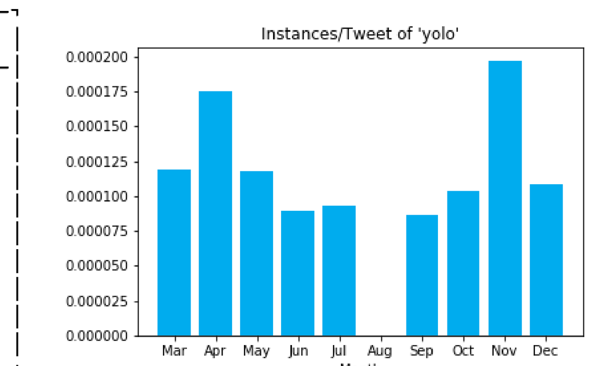
Month	P-value
3	None
4	0.59162
5	0.75738
6	0.36216
7	0.95441
8	0.92196
9	0.92955
10	0.81458
11	0.59717
12	0.0

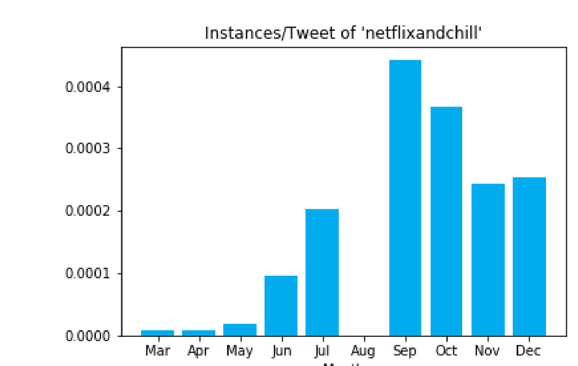


Month	P-value
3	None
4	0.0
5	8e-05
6	0.00036
7	0.69318
8	0.88107
9	0.94904
10	0.0
11	0.0
12	0.00235



Month	P-value
3	None
4	0.08608
5	0.0249
6	0.06286
7	0.64266
8	0.70496
9	0.71545
10	0.32279
11	0.0
12	0.00061





Question: Does slang spread faster on one social media site over another?

There is no evidence that slang spreads faster on one social media site over another. We found the slang "af", "fam", and "Netflix and chill" peaked on Twitter first, whereas "idfwu", "low-key", "rekt", and "yolo" peak on Reddit first. "Bless up" and "fomo" see peak usage in the same month across platforms (in Dec and Nov, respectively). We ran a chi-square test on the "Months to Peak" between the two sites -- 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.

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

The chi-square test yields no statistical significance between Twitter and Reddit Months-to-peak (χ2=0.72, p=0.40). To the right are charts of when each peaked on Twitter vs Reddit.

Reddit Peak Proportion	Month	Appearances	Total_Tweets	Proportion
af	11	179	562324	0.000318322
blessup	12	14	574524	2.4368e-05
fam	12	66	574524	0.000114878
fomo	11	11	562324	1.95617e-05
idfwu	3	2	536042	3.73105e-06
low-key	11	55	562324	9.78084e-05
netflixandchill	10	44	584791	7.52406e-05
rekt	3	252	536042	0.000470112
yolo	7	94	573984	0.000163768

Twitter Peak Proportion	Month	Appearances	Total_Tweets	Proportion
af	4	753	383581	0.00196308
blessup	12	79	442455	0.000178549
fam	3	409	235315	0.0017381
fomo	11	16	487290	3.28347e-05
idfwu	11	23	487290	4.71998e-05
low-key	12	830	442455	0.0018759
netflixandchill	9	173	391477	0.000441916
rekt	12	46	442455	0.000103965
yolo	11	96	487290	0.000197008

CONCLUSION

Our analysis of whether slang terms have different lifecycles allows us to reject the null hypothesis, providing evidence that terms rise and fall differently. Some terms experience drastic increases and decreases, while others remain relatively uniform. Comparing the rise-to-peak patterns across our slang terms on Reddit vs Twitter does not yield statistically significant results, and we fail to reject the null hypothesis.

NEXT STEPS

A caveat to our analysis is that data limitations may skew our results. Our analysis could yield more robust results if we could directly sample from the true distributions of Twitter, Reddit, Face-book, and Instagram activity over a several-year period. If we could overcome limitations in storage space and processing power, this methodology would give a more representative sample of word use and allow us to gain a clearer picture of the lifecycle of different slang terms. Additionally, we would analyze our data over time at a larger scope, e.g. two- and three-month iterations, as well as shorter iterations, e.g. weeks, day-to-day.