An Analysis of the Popularity of Facebook News Posts

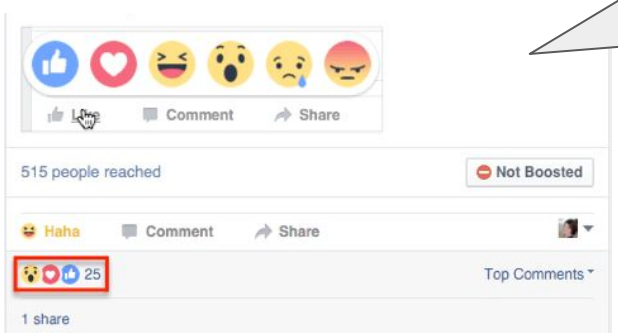
By: Thanika Reddy

Introduction

Some posts on social media become more popular than others

What factors lead to an increase in popularity?

- Does the **sentiment** of a post affect how quickly it gets comments?
- Do posts on weekends get more activity than posts on weekdays? Does the **time** of post creation matter?



Data and Preprocessing

Facebook News Dataset

~20K posts from 83 various news organizations & personalities , last 250 page posts.

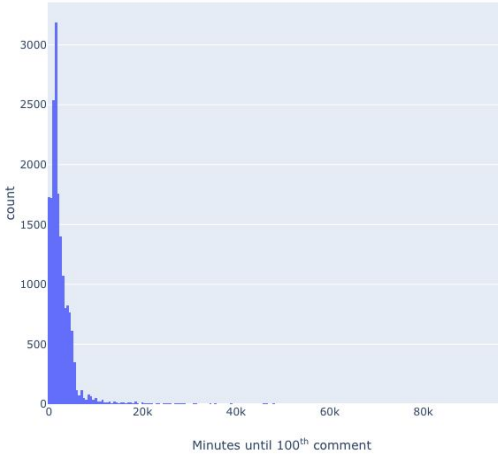
Each post has up to 100 comments for a total of ~1,000,000 comments.

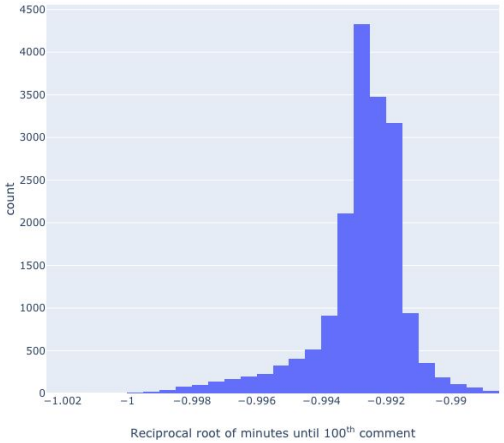
Features representing each post	Features representing each comment
Post creation time Post scrape time Description Link Contents of the post, Page ID Post ID, Number of "angry", "haha", "like", "love", "sad" and "wow" reactions Number of shares the post has	Parent post ID Comment creation time Name and ID of the user who created the comment Contents of the comment.

VADER (for sentiment)

Derived features
Day of post creation (7 binary features)
Time (on a certain day) of post creation (4 binary feature)
Average number of shares per second
Average number of "angry", "haha", "like", "love", "sad" and "wow" reactions per second
Sentiment, Positivity, Negativity and Neutrality
Minutes until a post gets its first comment
★ **Minutes until a post gets its 100th comment**

Heavily skewed data

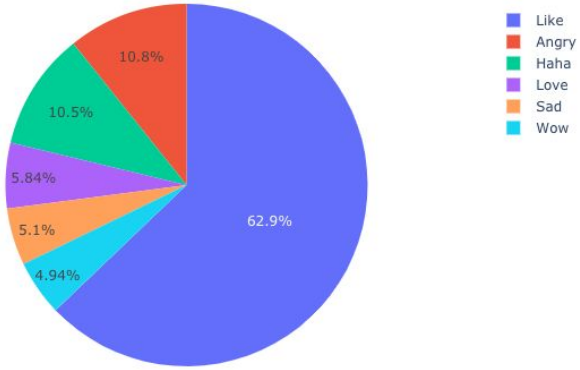


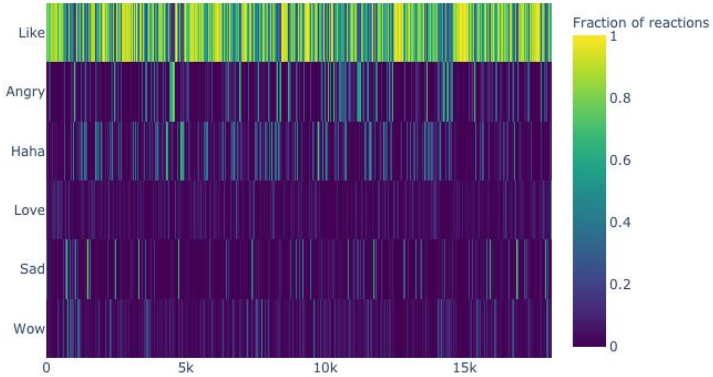


80.9% of the posts get their 100 comments within 67 hours.

Most other features have similarly skewed distributions

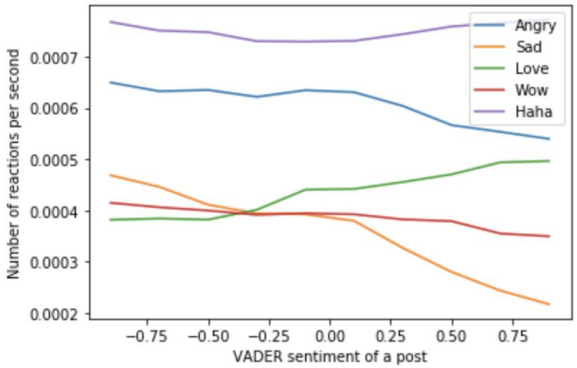
A log transform was not powerful enough for some features. The reciprocal root transform was then used, with a different fractional power for each feature



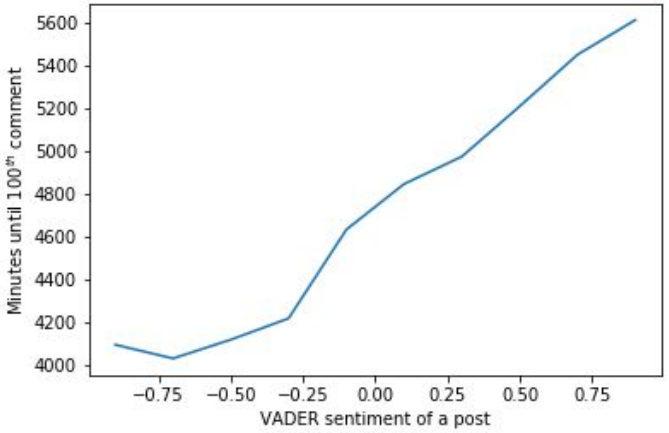


“Likes” are the most frequently occurring type of reaction

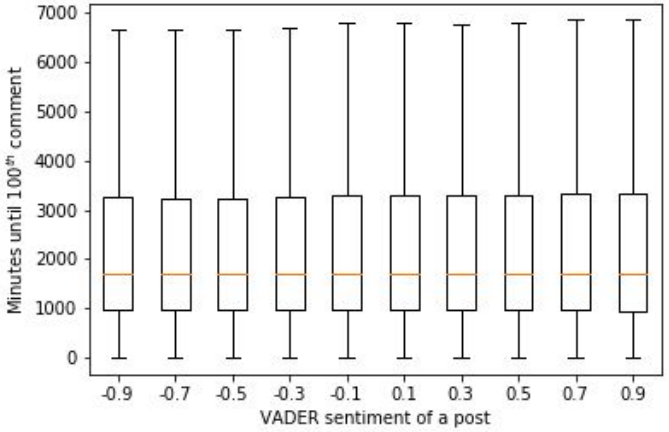
Potential relationships between features



The number of reactions (of a certain type) received by a post per second indicate the sentiment of a post.

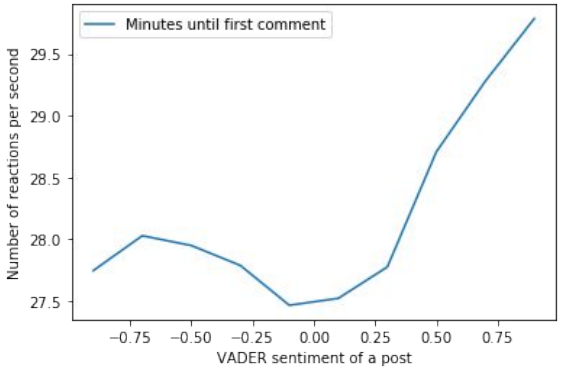


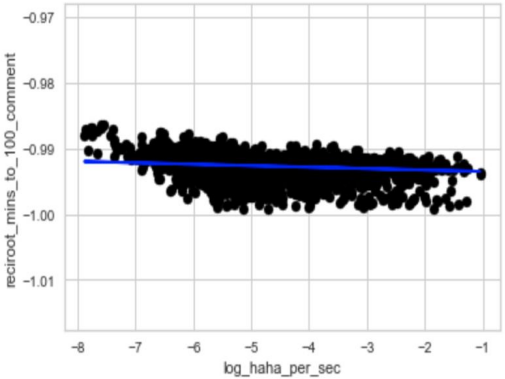
Posts with positive sentiment take longer to get 100 comments, i.e. they seem to be less popular

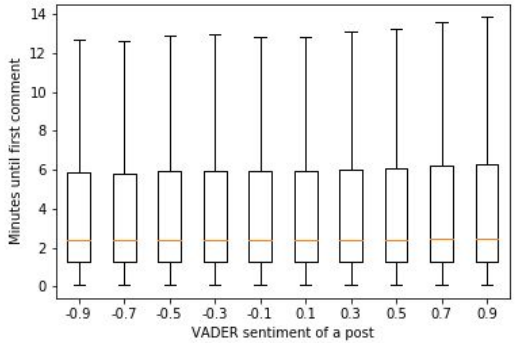


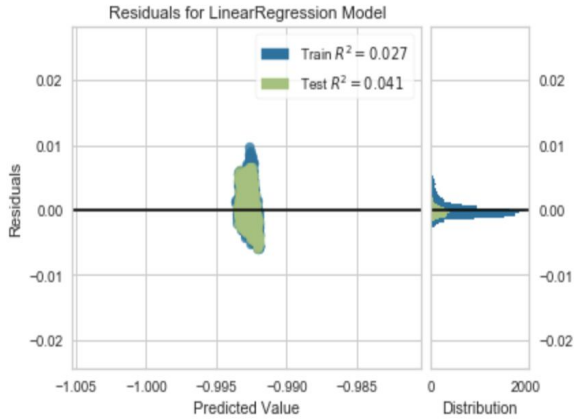
This pattern seems to be observed only for the 20% of posts that get their 100th comment after 67 hours

Potential relationships between features



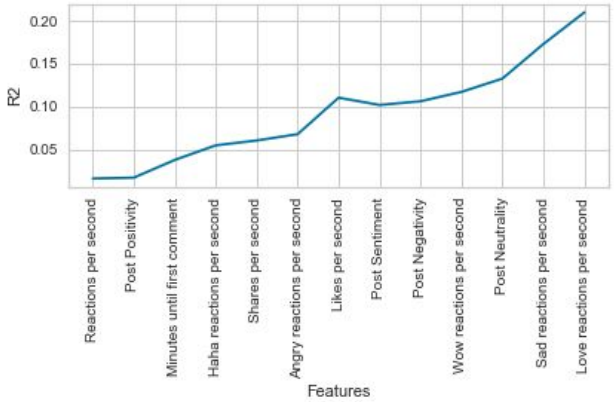


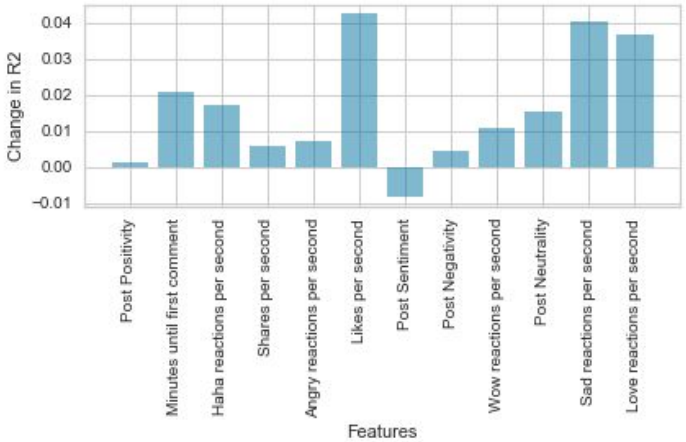




Fitting **univariate linear regression** between the time until 100th comment and each other feature

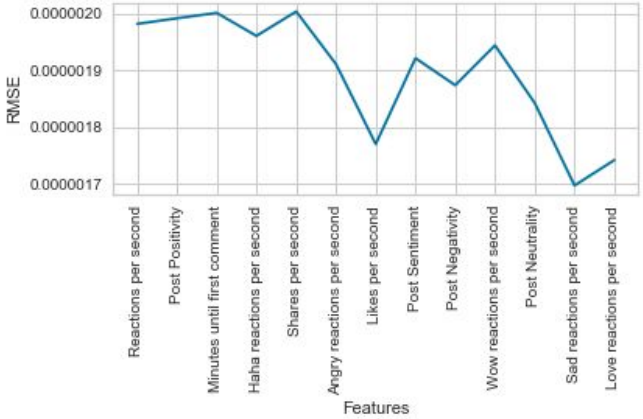
Multivariate Linear Regression

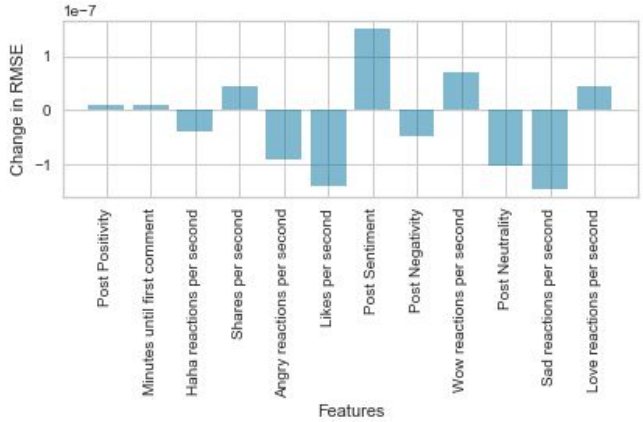




Add features one at a time, observe the change in the R^2 score, pick the six features with that increase the R^2 score the most over 100 iterations:

Log-scaled number of "love", "sad", "like" and "wow" reactions per second,
Reciprocal root transformed number of reactions per second,
Minutes until a post gets its first comment

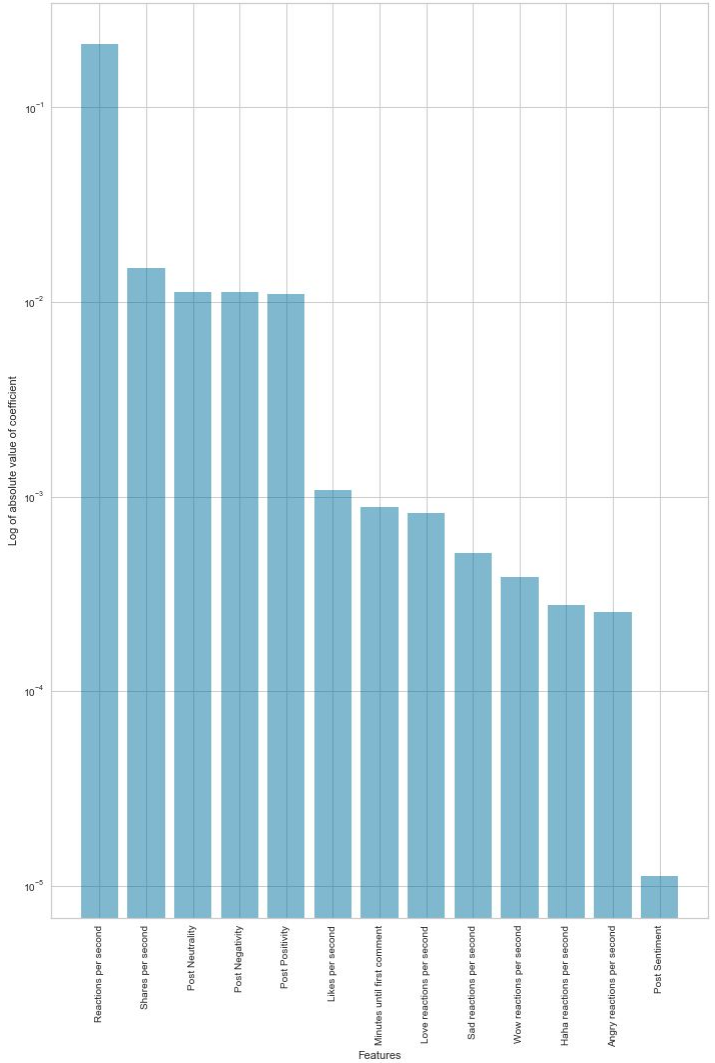




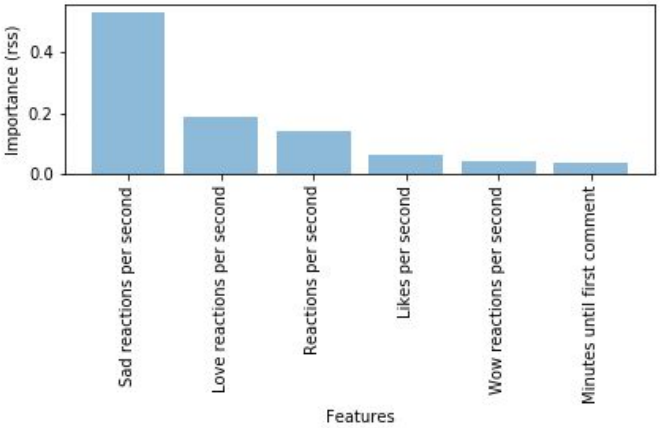
Multivariate Linear Regression

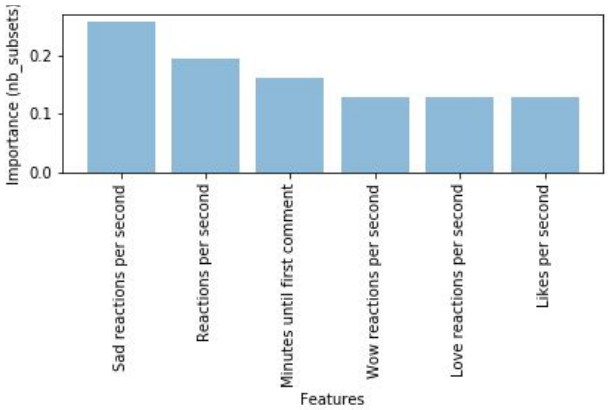
The number of reactions received by a post per second has the largest coefficient, followed by shares per second, positivity, negativity and neutrality.

Feature	Coefficient	P-value
Likes per second	-0.0010	$\ll 0.001$
Post negativity	-0.0176	0.701
Reactions per second	0.2091	$\ll 0.001$
Shares per second	0.0157	$\ll 0.001$
Minutes until first comment	-0.0009	$\ll 0.001$
"Wow" reactions per second	-0.0004	$\ll 0.001$
"Angry" reactions per second	-0.0002	$\ll 0.001$
"Love" reactions per second	-0.0008	$\ll 0.001$
"Haha" reactions per second	-0.0003	$\ll 0.001$
Post positivity	-0.0180	0.695
"Sad" reactions per second	-0.0005	$\ll 0.001$
Post neutrality	-0.0177	0.700
Post sentiment	4.58×10^{-5}	0.297



MARS and SVR





The MARS model chooses both negative sentiment and popularity of the news page/ its followers to be the most predictive. This is unlike multivariate linear regression, which chooses only the number of reactions per second to be the most predictive (with no sentiment).

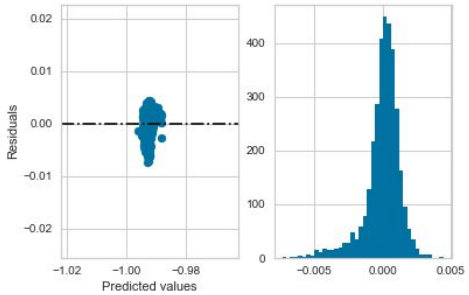
For SVR, a grid search was performed over three kernels (linear, RBF and polynomial), four values of C (0.1, 1, 100, 1000), eleven values of e (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10) and seven values of g (0.0001, 0.001, 0.005, 0.1, 1, 3, 5).

The RBF kernel with values C = 0.1, g = 0.1 and e = 0.0001 performed the best.

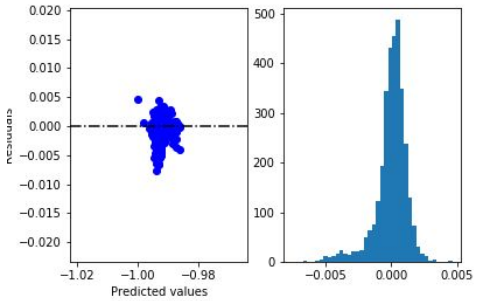
Comparison of model performance

Model	R^2 Score	RMSE
Multivariate Linear Regression	0.212	1.65×10^{-6}
MARS	0.239	1.56×10^{-6}
SVR	0.273	1.58×10^{-6}

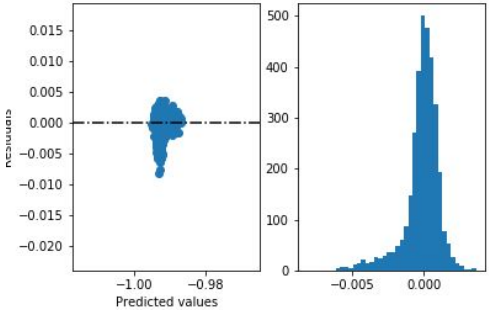
MARS improves the R^2 score by 12.7% over Multivariate Linear Regression
SVR further improves it by 14.2%
MARS reduces the RMSE by 5.4%



Residuals for multivariate linear regression



Residuals for MARS



Residuals for SVR

The residuals become progressively more heavy tailed.

Thank you!