

Engagement

Messier

Reddit

Methodology

Models

Getting People to Engage

A Look At Reddit Posts

Christopher Messier

General Assembly

Washington, D.C.

November 3, 2017

Overview

Engagement

Messier

Reddit

Methodology

Models

1 Reddit

2 Methodology

3 Models

What is Reddit?

Engagement

Messier

Reddit

Methodology

Models

Reddit is a large link sharing website

- "The front page of the internet"
- Types of posts
 - Content - links, pictures, videos, articles
 - Self Posts - text posts by the user
- Users share links with communities known as subreddits

Graph of Subreddits by User Cross Posting

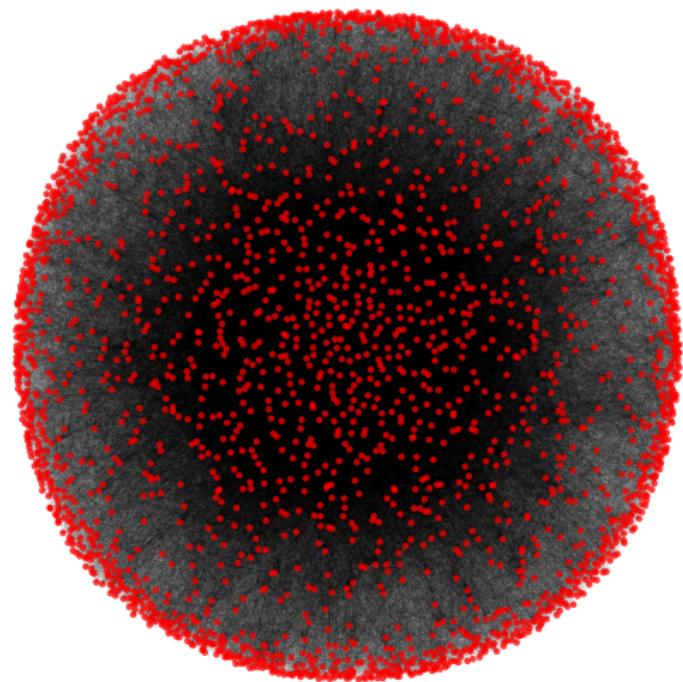
Engagement

Messier

Reddit

Methodology

Models



Measuring Engagement

Engagement

Messier

Reddit

Methodology

Models

High engagement posts will be defined as:

- Posts that have the top 10% most comments
- Binary Classification
 - High Engagement = 1
 - The rest = 0
- Determined for each subreddit

Why Comments?

Engagement

Messier

Reddit

Methodology

Models

Reddit has several intrinsic post metrics:

- Upvotes/Downvotes
 - Count of other users choice to like or dislike a post
- Rank
 - The page position of a post, determined by its popularity
- Comments capture active engagement with the site, interaction

Data Collection

Engagement

Messier

Reddit

Methodology

Models

Data was collected during the week of 10/30/2017

- 2,923 subreddits
- 1,049,005 users
- 2,590,769 unique posts
- 42,374,242 comments

Data Collection (cont.)

Engagement

Messier

Reddit

Methodology

Models

Collected using a proprietary python web scraper

- Attempted to get the most representative sample possible
- Used `r/random` to generate a random sample of subreddits
- Captured up to 2,000 most recent posts per subreddit

Omitted Data

Engagement

Messier

Reddit

Methodology

Models

The following data was removed/omitted:

- Duplicate posts
- Posts from bots, such as AutoModerator
- Non-english/non-ascii posts
- Adult Material

Omitted Data (cont.)

Engagement

Messier

Reddit

Methodology

Models

Analysis was performed on a subsample, $n= 363,660$:

- Computational Reasons
- Balanced Classes

Understanding the Data

Engagement

Messier

Reddit

Methodology

Models

- TFIDF Vectorization of the title

Modeling was performed using the python package Scikit-Learn

Models used include:

- Decision Trees
- k -Nearest Neighbors
- Random Forest

Scoring

Engagement

Messier

Reddit

Methodology

Models

Models are evaluated using the following metrics:

- Precision:

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

- Recall:

$$R = \frac{T_p}{T_p + F_n} \quad (2)$$

- F1:

$$F1 = \frac{2(P \times R)}{P + R} \quad (3)$$

Measuring the effects of the content of a post

- Used a decision tree model
 - Minutes since posting
 - Whether it's a question
 - Features outside content
 - Has an emoji in the title

Content Decision Tree Results

Engagement

Messier

Reddit

Methodology

Models

Content Decision Tree Results

	Predicted True	Predicted False
True	27398	26740
False	27166	27494

- Precision: 0.50
- Recall: 0.50
- F1: 0.50

Measuring the sentiment of a post's title, using nltk

- Used a decision tree model
 - Positive
 - Negative
 - Neutral scores

Sentiment Decision Tree Results

Engagement

Messier

Reddit

Methodology

Models

Confusion Matrix

	Predicted True	Predicted False
True	52569	1569
False	53061	1599

- Precision: 0.50
- Recall: 0.50
- F1: 0.36

k-Nearest Neighbors

Engagement

Messier

Reddit

Methodology

Models

Measuring the sentiment of the post title

- Sentiment Analysis performed using `nltk`
- The number of comments on a post,
- Weighted?

k-Nearest Neighbors Results

Engagement

Messier

Reddit

Methodology

Models

Confusion Matrix:

	Predicted True	Predicted False
True	32203	21935
False	26670	27990

- Precision: 0.55
- Recall: 0.55
- F1: 0.55

Measuring the sentiment of the post title

- Sentiment Analysis performed using `nltk`
- The number of comments on a post,
- Weighted?

Random Forest Results

Engagement

Messier

Reddit

Methodology

Models

Confusion Matrix:

	Predicted True	Predicted False
True	41136	13002
False	16322	38338

- Precision: 0.73
- Recall: 0.73
- F1: 0.73

Engagement

Messier

Reddit

Methodology

Models

Thank You