Science Communication on Twitter
- Investigate Dissemination Pattern in the New Media Environment

Yongle Zhang, Leyi Sun
College of Science & Engineering, University of Minnesota

Introduction

As a prominent online social-networking platform, Twitter may provide unique opportunities to enhance public trust in science. Some scientists now engage the public through posting “tweets” and sharing their scientific viewpoints, but it remains unknown how different modes of communication affect the dissemination of scientific knowledge.

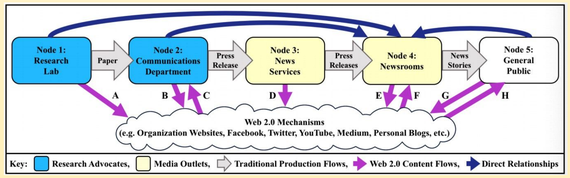


Figure 1: Media Production Pipeline for Scientific News

Theorists have divided **public communication pattern** into two modes:

- Within one-way communication, a continuum of information flow is transmitted from press agency at one end to the public at the other;
- In two-way communication, a continuum runs from persuasion at one end to conflict management with the conversation purpose.

Figure 2: one-way communication on Twitter is referred as using hyperlinks or posting plain tweets.

Figure 3: two-way communication is referred as using mentions “@”, retweets “RT” and hashtags “#”.

Objectives

RQ1: What account features are helpful to **identify researchers/ journalists** on Twitter? How can we develop a more effective way to **classify Twitter users** into roles such as researchers, organizations, news service, newsrooms, the general public and bots?

RQ2: How does one- and two-way communication affect **audience engagement** on Twitter?

Hypothesis: Two way communication predicts more audience engagement (i.e number of likes and retweets).

Methodology

Step 1: We collect data from several scientific topics such as ***#ClimateChange***, *#AntiVaccine, #FlatEarth and #Evolution* using Twitter API tool “Tweepy”. We decided the topic on climate change considering a enough amount of tweets (n=801,900) and less noisy data.

Step 2: A **codebook** was developed based on 300 user accounts to help role classification (Appendix 1). Human coders manually classified 1,000 accounts to build the **gold standard dataset** for further training as continued to refine the definitions in the codebook. The final **inter-rater reliability** (a degree of agreement among raters) is 0.78.

Step 3: Based on our ground truth data, we brainstormed several features that may helpful for training (Appendix 2). Next, Ground truth data that marked with identified features are passed to several machine learning models for training. After that, we cross-validated the models, and calculated the performance of each model (accuracy, precision and recall). Last, we applied the test model to a new data cluster for testing.

Step 4: For RQ2, we classify tweets of journalists and scientists by communication type (one/two -way) by looking at keywords “#”, “RT” and “@”. Since our data is not normally distributed, we conducted **Mann-Whitney U test** (in R script) to see the relationship between audience engagement and communication types.

Results

RQ1: Machine Learning Classification

First, we build our machine learning models for classifying all categories.

Table 1: Classification on All Roles (9 Categories)

Classifier	Accuracy
Logistic Regression	0.472076
Decision Tree	0.389953
Random Forest	**0.472157**
Multinomial Naive Bayes	0.245510
DummyClassifier (the Control Group)	0.178381

Our best model is the **Random Forest** with a 47.2% accuracy, which showed a great improvement on a **nine-category classification**, when compared with our Dummy Classifier.

Then, the binary classification on scientists and journalists:

Table 2: The Results of Binary Classification on Scientists

Classifier	Accuracy	Precision	Recall
Logistic Regression	0.6762	0.0349	0.0044
Decision Tree	0.7625	0.6399	0.6567
Random Forest	**0.8349**	**0.8654**	**0.5935**
Multinomial NB	0.4574	0.3508	0.7810
DummyClassifier	0.3241	0.3241	1.0

Figure 4 : Receiver Operating Characteristic (ROC) Curve on the Classification of Scientists

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

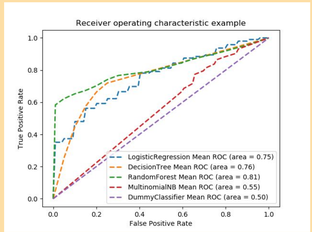
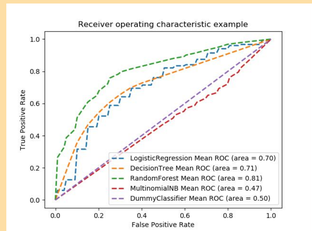


Table 3: The Results of Binary Classification on Journalists

Classifier	Accuracy	Precision	Recall
Logistic Regression	0.6119	0.6526	0.1986
Decision Tree	0.6944	0.6540	0.6361
Random Forest	**0.7484**	**0.7589**	**0.6220**
Multinomial NB	0.4770	0.4263	0.6115
DummyClassifier	0.4302	0.4302	1.0

Figure 5 : ROC Curve on the Classification of Journalists

Final Test Accuracy of Random Forest: 0.6846



RQ2: Statistical Analysis

Table 4: Wilcoxon Rank Sum Test with Continuity Correction

p-value	Engagement of “retweets”	Engagement of “likes”
Scientist	**p < 2.2e-16** (W = 48.248)	**p < 2.2e-16** (W = 13.544)
Journalist	**p < 2.2e-16** (W = 12.851)	**p < 2.2e-16** (W = 304.19)

The p values of four tests **are all blew 0.05,** which is the usually agreed alpha risk of 5 percent. We can assume that scientists and journalists who use two-way communication (mentions, retweets and hashtags) will get **significant more amount of audience engagement** (retweets and likes) on Twitter, which **support our hypothesis**.

Conclusion

Our study contributes a promising way to classify Twitter users who are involved in media production pipeline for scientific news, based on their account features. The finding also confirmed our hypo that journalists and scientists may get more audience engagement by using two-way communication.
We hope to help provide communication models that scientists can follow to effectively engage broader audiences on Twitter.
The study outcome aims at improving public engagement, as a crucial step in restoring public trust in science.

Appendix

Appendix 1: Twitter Role Codebook

Role	Definition
Researcher (C1)	A user who appears to participate in valid peer-reviewed scientific research. To be included in this category, a user’s profile includes both 1+ of the identified keywords, and at least one additional indicator that they disseminate or engage in scientific research. This includes features like: mentioning their research area in their description or recent tweets, retweeting scientific research or blogs from other profiles, or describing research-related activities. Default to 5 if a profile includes a keyword, but does not make any other indication of research-related activities. **Keywords:** professor, phD/phd,post doc/postdoc, faculty, scientist, principal investigator (PI), scholar, researcher, graduate/master student.
Public Information Officer/Scientific Research Organization (C2)	A person who appears to be the Public Information Officer employed by an industrial scientific research lab or academic institution; Or an institution, center, group, or community founded to carry out valid scientific research or facilitate fundamental research processes like data collection, tool implementation, collaboration forming, etc. for science. **Keywords for PIOs**: Communications Director, Media Relations Manager, etc. **Keywords for Scientific Research Orgs:** research agency/institute/lab, university, think tank, conference, journal, etc.
Non-Research or Non-Mainstream Media Organization (C3)	An entity of people advocating for a shared point of view on twitter, but not primarily news/research-oriented.
Journalist (C4)	A person who appears to write for newspapers, magazines, or news websites or prepares news to be broadcast. **Keywords:** journalist, reporter, photojournalist, editor, correspondent, columnist, etc.
General Public (C5)	People whose profiles do not provide any evidence that they participate in scientific research or media production including lay citizens, celebrities, writers, activists, politicians, etc.
News Service (C6)	The official Twitter account of an established News Service.
Newsrooms (C7)	The account of newspaper or broadcasting office where news is written and edited.
Bot (C8)	An account that appears to send out automated content on Twitter.
Unclassified (C9)	The account that does not fit into any other categories.

Appendix 2: Machine Learning Features

favourites_count, followers_count, friends_count, listed_count, statuses_count, bot, contributors_enabled, default_profile, default_profile_image, verified, lang, created_at

Acknowledgements

We’d like to thank the help of the PhD candidate **Estelle Smith**, and the support from our mentors, Dr. **Lana Yarosh** and Dr. **Haiyi Zhu** at the **Grouplens Research Lab.**