Structure, Actor or Culture? A Predictive Model on Online News Popularity

Weiwei Zheng

Abstract

Social media has become an important platform for people to read news. Since people get much more autonomy online than in the analog sphere, it is hard to tell what kind of information delivered by media organizations will get popular in the long run on the Internet. In order to find out ways to better capture public opinion in the new media era, the study utilizes computational technique and tries to build a predictive model on online news popularity using the well-known social media website reddit.com. Analyzing the data scrapped by reddit's official API praw, the study made different models to predict popularity of news posted in 2014 and 2015 in r/worldnews community. The study categorizes predictors into three dimensions: structure, culture and actor, and looks at how powerful are online users' discussion pattern, news article content, and influence of users in predicting the upvotes news article got on the website. The result shows adding structural predictors in the model contributes the most in increasing prediction accuracy while the other two seem relatively trivial.

Key words: news popularity, network structure, social media, reddit

When communication scholars first raised the concept agenda setting, they believed media influences what people think about. Salience of content in newspapers was once found to accord with those important issues in people's mind (Lippman, 2017). However, in the new media age, no longer staying passive towards whatever told by media institutions, audience gain much more autonomy and become active participants alongside the process of information diffusion. People not only receive information delivered by media outlets, but they are also developing their own agenda of importance. People also

reproduce and transfer what they think is important to other people. This reciprocity between media institutions and the audience is called “agenda building” (Dearing & Rogers, 1996).

But how can we predict the public agenda? The development of public opinion is a dynamic process. The public form their own perceptions over different issues partly out of their own rationality, which cannot be simply generalized by looking at the media agenda. If institutes can effectively predict what kind of information can get the public attention, it must be helpful for us to analyze the public mindset and respond to public opinion. Nowadays, social media has already been a major platform for people to access different news sources. Thanks to Internet, the audience can easily start crowd discussion and engage in different types of collective activities distinct from offline life. They not only develop unique pattern of communication but can vote importance of information through collective activities (Horne et al. 2017). It must be illuminating to find out how we can effectively tell what kind of information is or has the potential to get popular and generalize the pattern of human communication in the virtual sphere.

There have been many articles trying to build predictive models on news popularity, but most of the research only focused on the content of news articles and did not frame their research question by integrating multiple theoretical construct. Admittedly, by categorizing articles based on their content, we can see how cultural difference affect the popularity of content. All texts are meaningful discourse, as said by cultural theorist Stuart Hall (2001). But “culture” shouldn’t be the only aspect to look at if we want to obtain a complete perception of certain social phenomenon. As Anthony Giddens proposed, a “duality of structure" is embedded in social practice (1984). There existed a social cycle in social life: People’s behavior and knowledge is filtered and changed by the dominant social structure, while people also keep reflexive and adapt their own actions to updated understanding within the social structure. Social structure can shape human’s perception. However, not all people stay passive in face with the dominant structure, some actors are found to have more influence than others to direct the structure. Based on the diffusion of innovation theory proposed by Lazarsfeld, opinion leaders, or the people with relatively large influence to divert the public opinion, take important roles in spreading and reshaping the information. Those actors possess more credibility and have many followers who are willing to mimic their behavior and easily get

convinced by what those important actors say. Those important actors, or opinion leaders, are usually the targets to look at when scholars study the flow of information among diverse subgroups of audience (Lazarsfeld et al. 1966).

Therefore, the article tries to build a predictive model on online news popularity which aggregates independent variables from the three dimensions structure, actor and culture using the well-known social media site reddit.com. In the r/worldnews community of reddit.com, people can post news articles about issues happening outside the United States and comment and upvote (support) each post. By looking at the how many upvotes an article gets, we can know how many positive feedbacks it wins. The study tries to answer the following research questions: 1) Do variables got from the three dimensions effectively predict the news popularity on reddit.com? 2) Which dimension seem to be the most important in getting predicting accuracy? After finding the best model, I will introspect on specific features to gain further insight.

The result shows structural predictors seem to far outweigh cultural and actor predictors, which refutes the findings discovered by most of the articles related to the field. The study leverages a bunch of existing computational technique and innovatively incorporate measurements in network analysis to analyze human communication in social media. It paves way for research afterwards to look at interweaving effects between information features in different dimensions and information flow in time series perspective.

Related work and background knowledge

1. Reddit

Reddit.com is a user-generated news-aggregating platform in United States. Users can post links, original messages and comments in different communities called “subreddits” over a wide variety of topics on the website (Kilgo et al. 2016). In the r/worldnews subreddit, users post links to online news articles on worldwide events released by different news organizations. They can also upvote the articles which they find important or interesting (Horne et al. 2017). The ranking of the news articles in the front page “top” section in this subreddit is decided by the number of upvote (support) the specific posts got from the users.

Under each post, users can post comments. They can both initiate a discussion and reply to other people's comments or sub-comments. And the total number of comments and upvotes is attached to each post on the website. You can choose the range of time period to see news articles ranked by popularity over different period, including past year, past 24 hours, past week, past month and all time. The already popular articles are more likely than others to be read by users who simply read through the rank (Horne et al. 2017, Gilbert 2013).

2. News Popularity

There are many reasons why we need to study online news popularity. For one thing, what individuals consume on social media platforms depends not only one their free will but also on how the news feed ranking algorithm sorts these articles (Bakshy et al. 2015, Horne et al. 2016). Knowledge about the mechanism behind this crowd-generated news ranking in social media guides us to implement better communication strategy for a well-informed democracy. For another, in cases without such ranking criterion, it is hard for us to figure out whether something has already or has the potential to get hot online. There have been many studies already looking into the diffusion of online contents which tried to predict the short-term or long-term popularity of different contents in social media platform such as twitter and digg.com. They used a variety of measurements to measure popularity, such as page views, number of likes or shares, and number of searches (Zaman et al. 2014, Keneshloo et al. 2016, Lakkaraju et al. 2016). Number of page views, searches and comments are believed to reflect total amount of engagement while rating/upvote-downvote are believed to reflect the true motivation of users to share the content (Lakkaraju et al. 2013, Horne et al. 2017).

It is challenging to predict popularity of online contents with a mixture of miscellaneous endogenous and exogenous factors interweaving the process. Scholars have leveraged different attributes of online information in multiple dimensions including media users' network structure, crowd engagement, roles of important users, contents of information and temporal changes of users' attention (Horne et al. 2017). Though previous studies have delved into these fields respectively, most of them only focus on the

sentiment of articles and few studies have integrated variables from multiple perspective to build a cohesive model (Naveed et al. 2011). Those features of the online information can be generalized into contextual and content layers.

2.1 Context Features

Users structure attributes and activities of users of special roles (e.g. opinion leaders) can be both categorized into the contextual level. Though scholars have studied relationship between information diffusion and network structure or user activities, but not many studies have looked at the combined effect from both structural and user level on online news diffusion.

Regarding structure, social media users connect with each other through a variety of ways and network can be drawn by different types of connections such as mentioning and commenting posts (Yang et al. 2010). Network attributes are found out to have some influences on the process of information diffusion. For instance, whether a post will be shared depends on its locational and temporal position in the user network. It has also been found that whether a particular tweet actually is retweeted depends heavily on posters position in the social graph and the time of day the tweet is posted (Naveed, 2011). Network created by mentioning in twitter also has an un-negligible impact on the range, scale and speed of spread of posts (Yang et al. 2010). Besides, network structure is believed to influence people's cognitive ideas on the same contents. Linked (bilateral) and unlinked friends were found out to have different possibility to upvote the same content shared by the same person (Hogg et al. 2012). What's more, Posts shared by a well-connected user with active followers are more likely to be retweeted (Naveed, 2011).

Network structures have been considered by many scholars who looked in to the process of information diffusion and they conducted research on different social network like Twitter. But since network structure varies among different online communities, even in the same website, discovered patterns are hard generalized to some other social media platform like reddit (Naveed et al. 2011). Most the measurements of network structure proposed by other studies don't apply to our context. I will first start from some basic features of networks to build the model.

In terms of the actor level, it is worthwhile to also look at the influence of different users to look at how their impact on information popularity varies. Different type of users has different influence online. It is reasonable to categorize different users in the network and look at their specific roles and general patterns of the mass during the process. Information diffusion is influenced by social influence. Choices or opinions of others affects a user's behavior (Lerman et al. 2010). Out of all types of people studied in the diffusion of public opinion, opinion leaders are believed to be the most important. According to Lazarsfeld and his colleagues, opinion leaders who have more insights and influence than the majority, are inclined to receive news quite early and take the responsibility to spread the information to the general public (1965). Opinion leaders have been identified from reddit by other scholars who looked at users' profile characteristics, such as commenting, longevity, karma scores, posting frequency, and posting scores (Kilgo et al. 2016). These opinion leaders are more likely to post stories from professional news source and generate more top comments. Hence, their reputation must have an impact on the spread of the content they repost or comment (Leavitt et al. 2014).

Besides opinion leaders, there are also other different types of users which have been identified in reddit network which might be helpful for building our own model (Buntain et al. 2014). A type of "answer-role" users has been found in different communities of reddit who tend to answer a lot of questions while engaged in limited discussions. Though it is impossible for this study to observe this type of users at the current stage, categorizing users into different type must do us some good for building a well-rounded model.

2.3 Content Features

Content features of text data can be divided into different categories, including sentiment of a text, emotions within the text, subjectivity of its language, named entities, readability, part of speech, freshness of a content and the time it's posted (Keneshloo et al. 2016, Lakkaraju and et al. 2013 Tsagkias et al. 2009), which are all considered to be useful to represent specific text and predict content popularity in social network websites such as digg.com and twitter (Horne et al. 2016).

Apart from the basic attributes text data such as word count, sentiment features are always the focus of text analysis. In studies on reddit, popular news turns out to usually have a negative or positive title (Reis et al. 2015) but negative content (Horne et al. 2017). Articles with these kinds of features are more likely to get more comments (Reis et al. 2015). Sentiment should be one of our main foci we are studying text data.

Except for the semantic attributes, other content attributes like distribution of different topics among the corpus also influences how popular one piece of news can be (Keneshloo et al. 2016). News articles can not only be categorized based on their sentiment score but also by their topics (Reis et al. 2015). It would be interesting to look at separate effects from topic, sentiment and some other basic textual features certain article on its communication effect.

Method

1. Data Sets

I utilized the official API praw released by reddit.com to extract information of posts in r/worldnews community. However, praw only allows users to get 1000 observations from one single request, which means if I extract data from the top ranking of r/worldnews I can only get information of the 1000 top posts. Considering the limited size might lead to unobserved bias in the foregoing analysis, I refer to a dataset released by a Kaggle competition to supplement the analysis. The dataset includes basic information of posts on r/worldnews released from 2008 till November 2016. However, the datasets only records news title of each post but not their unique post id, so I used the title of news articles to get detailed post information with praw. In order to look at the long-term effect of posts' attributes on their popularity, I only keep the posts which were created in 2014 and 2015. I also filtered the entries which were recorded in the dataset to have 50 upvotes or more to shrinkage our scale of analysis. The focus of this study is about how to identify popular articles, so it is reasonable we get rid of unimportant samples rated by the users. The exact sample size can be seen in Table 1.

Here is the how I collect the data. I first used the titles of news articles extracted from the Kaggle dataset to get post id of each post. Though it is possible there might be multiple or no post with the exact title of a given news article, most of the titles only have one submission result returned by praw. For some unknown algorithm confounding issue, a handful of different news titles turn out to have the same post id. I drop all the duplicate post ids and the final sample size can be seen in Table 1. After I got their post id, I scraped post information including original article URL, upvotes, author id (the id of user who post the article on r/worldnews), date of created and number of comments. I also crawled all the comments of each post, including comment body and comment author id. Though a number of comments of each post have already been deleted the author id of each post is still accessible. After that, I also used the author id to get the user information. A bunch of author information is missing again, but the average percentage of missing user information of those who comment on a post is no more than 5 percent.

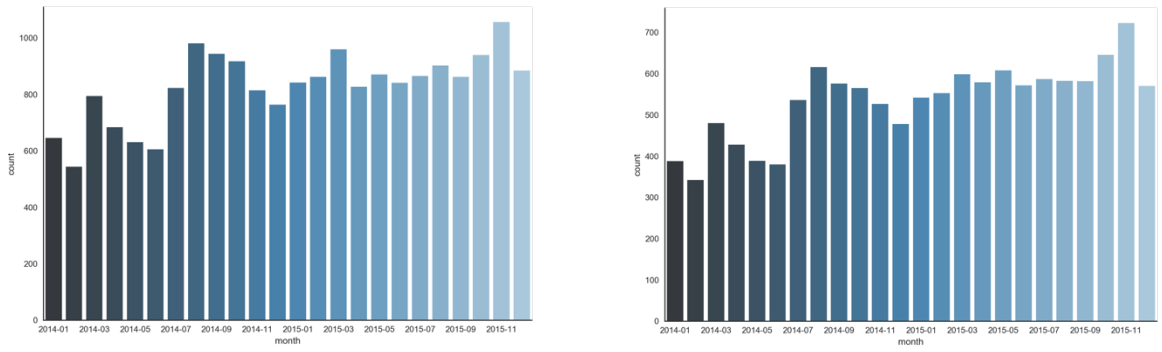
On the other hand, I tried to use the information I got from praw to link the text data of each article. I used another API newsarticle3k and got all the text of the news from the given URL of each post I scraped from reddit. After preprocessing the data match post id of both reddit and news article datasets and getting rid of posts with missing data, 19869 observations are left.

Table 1. Sample size in the two years

reddit	year	Kaggle dataset	reddit dataset	article dataset	preprocessing
r/worldnews	2014	92030	10743	(many articles	9154
r/worldnews	2015	94621	11947	are missing)	10720
total		186651	22690	19987	19869

The descriptive statistics of our post information can be seen in Figure 1 and Figure 2. Even though there is noticeable variance of posts in each month, the distribution of posts number and news article number are basically the same, so we can preclude the potential bias causing by missing dates and text information of news articles.

Figure 1. Monthly Distribution of number of posts (left) and number of articles with published date (right) in each month



2. Features

Comment structure

The first part of our predictor is extracted from the comment information. By using the unique id of each comment and their author under of post of news article, I can build a network structure of different users to represent their discussion pattern of each post. When a user comment on other people's comments or posts, I draw a directed edge from the sender to the recipient. For each post, I can get information of the basic attributes of the comment network such as number of nodes (unique participants) and number of edges. In order to measure how closed the nodes connect with each other in the network, I get more complex structural attributes from the comment network such as density, transitivity, number of strongly connected component and etc.

Since the total comment network of most posts are pretty disperse and scattered, which is a common feature of most social media network, I extract the strongly connected component of each post's total network to calculate other more intensive features such as average shortest path, distribution of eccentricity, diameter, radius and etc. The list of predictors can be seen in Table 2.

Roles of important actors

I also look at the performance of active users (opinion leaders) in each post's discussion. By active users, I mean those with high "karma". "Karma" is a metrics used by reddit to signal the influence of users. There are two kinds of "karma" information we can get from praw, "link karma" and "comment karma". They respectively represent how many upvotes each user got from the posts and comments the user released. In order to measure the participance of opinion leader in each post, I first located the top ten percent of nodes measured by different metrics of network analysis using API network, such as betweenness centrality, closeness centrality, hub score, authority score, and etc., then I separately calculated the average link and comment "karma" of those users. In this way, I can see whether the opinion leaders in the discussion are influential or not. On the other hand, I also calculated the two average "karma" of centers and periphery of each network measured by networkx, which can reflect the overall influence of users participating in the discussion of certain post. The measure measurement I used to find out important users in the network can also be seen in Table 2. The link and comment "karma" of each post's author are also put in the model.

Text features

The last group of the predictors I examine is related to the text of news articles. I basically study the sentiment, topics and other basic features of news articles. As for sentiment, I look at both the sentiment of headline and article body of each post. I used the algorithm VaderSentiment algorithm used by most scholars conducting sentiment analysis in text to get an index scaled from -1 to 1 to represent the sentiment level of the article's text body and title (Araújo et al. 2014). If a sentence's sentiment score is -1, it is identified by the algorithm as extremely negative, and 1 extremely positive.

And I also look at both and macro and micro level of the content feature of the text body. On one hand, in the macro level, I obtain ten topics of the news article corpus by adopting the Gensim's LDA (Latent Dirichlet Allocation) topic modeling algorithm. After getting the ten topics of the news corpus, I create ten variables to represent each article's likelihood of being categorized into each topic. The value of these ten variables of one news article should add up to one. On the other, in the micro level, I study the general summarizing features of each article's text body, including word count, average length of each

sentence, average word length and distribution of different part of speech, and etc. Considering most of study on predicting online news popularity so far mostly look at the effects from topics or/and sentiment, only when I control other important features of different text, I can see whether comment structure and roles of opinion leaders can exert a considerable effect in predicting news popularity. The detailed text variables I put in the model can also be seen in table two.

Table 2. Predictors used in this study (n = 50)

Structure (the largest strongly connected component is call the main network)				
number of nodes in the whole network	density of the whole network	number of weakly connected components	number of strongly connected components	number of nodes in the main network
density of the main network	transitivity of the main network	average shortest path of the main network	diameter of the main network	radius of the main network
mean of eccentricity of the main network	standard deviation of eccentricity of the main network			
Actor (except for the last three cell, only top 10% of nodes found by the measurements are used, the average of both the comment karma and link karma of each measurement is used)				
closeness centrality of the main network	betweenness centrality of the main network	in degree centrality of the main network	out degree centrality of the main network	hub score distribution of the main network
authority score distribution of the main network	center of the main network	periphery of the main network	author of the post	
Culture				
word count of the article	average word length of the article	lexical diversity of the article	sentence count of the article	average sentence length of the article
percentage of verb in the article	percentage of adjective in the article	topic 1: natural science and research	topic 2: Malaysian airline incident	topic 3: Islamic countries and terrorism
topic 4: women and children's rights	topic 5: national defense and nuclear weapons	topic 6: Information security	topic 7: health and drug	topic 8: climate change and mineral energy
topic 9: China, Korea peninsula and Europe	topic 10: Russia	sentiment score of the article body	sentiment score of the article title	length of the title

But one problem of using the network measurement is, since the comment network of the posts, especially those with low upvote and few participance in discussion, is pretty small, they don't even have any strongly connected component or more than 10 nodes in the total network and it is impossible for me

to calculate some network attributes or the average “karma” of the top 10 percent important nodes in the main network. Therefore, I drop all those observations with missing network and actor measurement and there 4584 observations left in the final analysis are only . Given that most of the dropped observations are posts with low upvotes, dropping them will not cause considerable damage to the study. Figure 3 shows that the distribution over different months are still similar as before and Figure 4 shows the different average of upvote score between datasets with and without missing measurements.

Figure 3. Counts of posts with and without missing measurements over different months

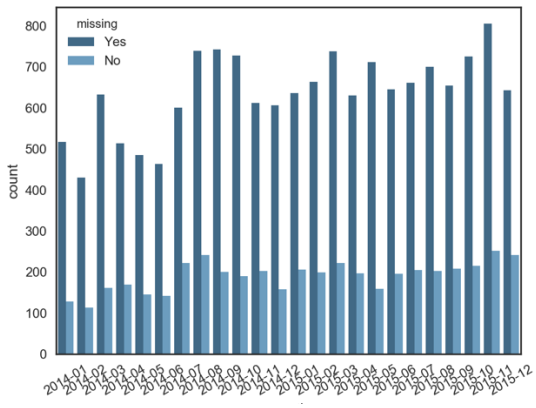
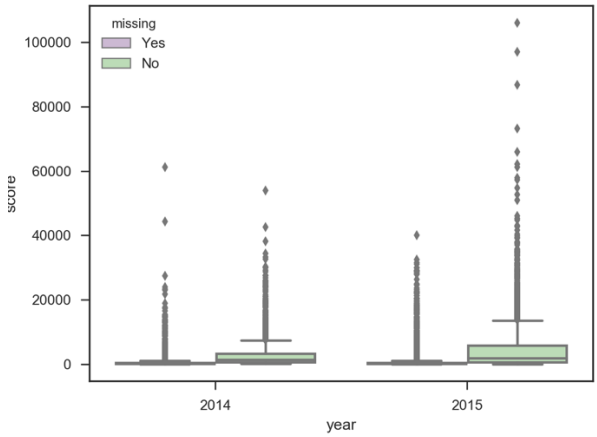


Figure 4. Boxplot of upvote scores of posts with and without missing measurements



3. Model

The dependent variable I predict is the upvote score of each post. I try to build a predictive model using all variables from the three dimensions structure, actor and culture over news article's popularity. I use both ridge and lasso regressions in my methods. Considering I have so many variables (in total 50) in the model, I had better some method where I can use regularization approach to control complexity of the model to prevent overfitting. The formula of both methods is shown below.

Figure 5. Formula of Ridge regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Figure 6. Formula of Lasso regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\ =\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

Before I put all variables into the model, I normalized all dependent variables over the range 0 to 1, then I can compare the effect of each variable on the same scale. Ridge regression can magnify the effect of trivial predictors and performs better than lasso regression when there are many variables with small and medium size effect, while lasso regression can shrinkage the coefficient of trivial predictors to zero and performs better than ridge regression when there are only a few variables with medium and large size effect. Given that I am still not sure what variable would be important in the three dimensions I use both the methods and compare their performance. I split the dataset into training set (80 percent) and testing set (20 percent) and use 10-fold cross validation to find the best L2 regularization parameter of each model by comparing R square.

I also built separate models using only variables from each dimension to see which dimension has the largest predictive power. I compare the best model in the three dimensions with the best whole model

using all variables by MSE (mean square error). However, even though I can get the MSE and R square of each model, it is hard to tell whether this model is good or not. Therefore, I also build models with dummy regressor which predicts all the values of the dependent variable in the testing set as the mean of the dependent variable in the training set and then compare each model with the dummy regressor to evaluate the performance.

Results

The result of the modeling section is listed below in Table 3. It is easy to tell that the whole model performs the best in both methods and the r square of the whole models both exceed 0.6. The whole model also performs much better than the dummy regressor which uses the mean of dependent variable in training set as prediction. Features we extracted from posts seem to be effective predictors over their popularity.

By comparing the performance of separate models, we can see the whole model made by both methods are the best compared with the separate models. However, it is easy to tell the predictive power of the whole models mostly comes from the structure variables. The R squares of structural models are pretty closed to those of the whole models and they also exceed 0.6. This suggests users' discussion network seem to be an important indicator of whether the topic is popular or not. How people discuss a news incident tend to be highly correlated with whether it will get popular or not. On the other hand, actor and culture variables seem relatively trivial. Though actor model also performs slightly better than the dummy regressor, the MSEs are pretty slow and the R squares of both methods with actor variables are below 0.1. Even though scholars believe users' influence might have some impact or implication on the popularity of what they discuss, the effect is comparatively negligible with the impact from structure, as the results shows. And cultural models perform the worst among the three, even worse than the dummy regressor. Standing in remarkedly contrast to what most social media research found, the qualitative attributes of text like sentiment and content cannot imply people's attitude towards the content or they might also be totally unrelated with news articles' potential to get popular.

Table 3. Results of modeling

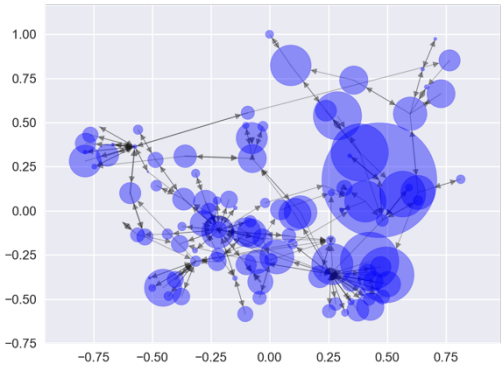
Ridge Regression	**Whole Model** (n = 50)	**Culture Model** (n = 20)	**Structure Model** (n = 12)	**Actor Model** (n = 18)
Alpha	0.2	> 100	0	4.2
R square (training)	0.634	0.003	0.631	0.014
R square (testing)	0.62	0	0.61	0.02
MSE (testing)	21181300.33	55331421.29	21493650.37	54137675.17
R square (dummy)	0	0	0	0
MSE (dummy)	55271237.27	55271237.27	55271237.27	55271237.27
Lasso Regression				
Alpha	0.5	> 73	0.3	2
R square (training)	0.637	0	0.631	0.015
R square (testing)	0.62	0	0.61	0.02
MSE (testing)	21247059.25	55271237.27	21454728.37	54042340.47
R square (dummy)	0	0	0	0
MSE (dummy)	55271237.27	55271237.27	55271237.27	55271237.27
#non-zero predictors	41	0	11	11

By using lasso regression, effect of trivial variables can be minimized to zero. We can see what specific variables are important in each dimension by looking at the scale of their coefficient. In the whole model, the variables with largest coefficients are all structure variables, followed by several variables in actor dimension. And the coefficients of most of the culture variables are pretty insignificant among all the non-zero predictors. The details are shown in Appendix.

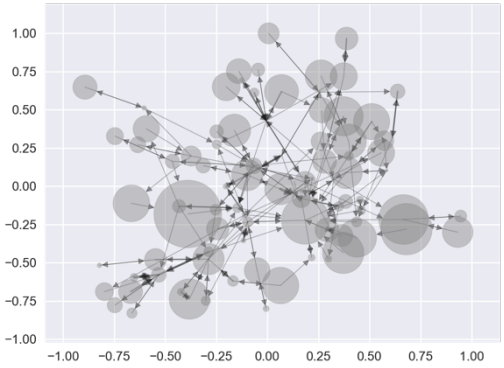
Figure 7 shows the structure of the main network (largest strongly connected component) of three posts. The width of links shows the weight of edges between the two nodes and the size of nodes shows the sum of the two kinds of "karma" the node has. The post with highest score not necessarily has the most number of nodes or highest density in its main network but is very likely to have many nodes with high "karma".

Figure 7. Sample main network structure of three posts

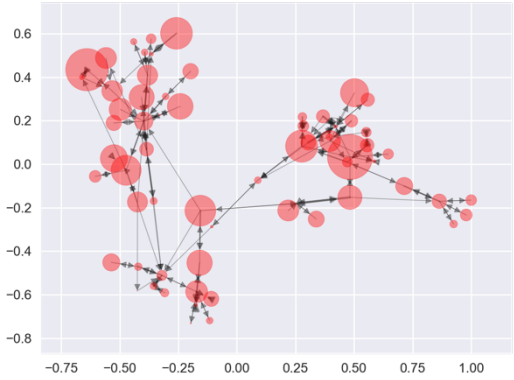
Upvote score = 86743



Upvote score = 4922



Upvote score = 2110



Conclusion

The study uses variables from structure, actor and culture dimension to build a predictive model on online news popularity using the social media site reddit. Structure, or users' discussion pattern seem to be the most important to tell whether a news issue is popular or not. Though the performance of opinion leaders is supplementary to structural predictors in the model, the size of is relatively trivial. Effect from the culture dimension seem totally unrelated, which challenges what most scholar found when they try to predict popularity of online news content. Structure seems to be one important aspect which most scholars neglect when they try to evaluate the diffusion of information. The roles of opinion leaders and cultural mechanism might not be as significant or worthy of being studied as what most social science researchers believe.

The study provides a ground-breaking approach to deduct public agenda or public opinion. In order to tell what actually is in the majority of people's mind, to look at the scale of discussion or the quantitative measurement of people's connection when talking about certain topics seems to be more useful than pondering over in-depth and qualitative attributes of what people are actually talked about. The study also inspires public relation events where organizers can analyze the effect of their communication strategy. Moreover, the study provides insight for following research to extend my perspective and approach over other social media site to generalize patterns of human communication in offline world.

However, the study still has some limitations. Without looking at the temporal dimension, I cannot observe the change in the network and how the change correlates with news popularity. It is hard to draw causal relation between news popularity and network structure and there might be other unobserved but important mechanism intermediating process where information gets attention. If the following studies keep track of how information popular in the long run, it might get more illuminating findings over the interaction between different dimensions. In addition, ridge or lasso regressions might not be the best approach to model this relationship, more machine learning approaches had better to be tried to validate the result obtained by this study.

References

Anthony Giddens. *The constitution of society: Outline of the theory of structuration*. Univ of California Press, 1984.

Araújo, Matheus, Pollyanna Gonçalves, Meeyoung Cha, and Fabrício Benevenuto. "iFeel: a system that compares and combines sentiment analysis methods." In *Proceedings of the 23rd International Conference on World Wide Web, pp. 75-78. ACM, 2014.*

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348, no. 6239 (2015): 1130-1132.

Buntain, Cody, and Jennifer Golbeck. "Identifying social roles in reddit using network structure." In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 615-620. ACM, 2014.

Dearing, James W., and Everett Rogers. *Agenda-setting*. Vol. 6. Sage publications, 1996.

Gilbert, Eric. "Widespread underprovision on Reddit." In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 803-808. ACM, 2013.

Hall, Stuart. "Encoding/decoding." *Media and cultural studies: Keyworks* 2 (2001).

Hogg, Tad, and Kristina Lerman. "Social dynamics of digg." *EPJ Data Science* 1, no. 1 (2012): 5.

Horne, Benjamin D., and Sibel Adali. "The impact of crowds on news engagement: A reddit case study." *arXiv preprint arXiv:1703.10570* (2017).

Horne, Benjamin D., Sibel Adalı, and Kevin Chan. "Impact of message sorting on access to novel information in *networks*." In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pp. 647-653. IEEE, 2016.

Karlsson, Michael, and Jesper Strömbäck. "Freezing the flow of online news: Exploring approaches to the study of *the liquidity of online news." Journalism Studies 11, no. 1 (2010): 2-19.*

Katz, Elihu, and Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers, 1966.

Keneshloo, Yaser, Shuguang Wang, Eui-Hong Han, and Naren Ramakrishnan. "Predicting the popularity of news articles." In Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 441-449. Society for Industrial and Applied Mathematics, 2016.

Lippmann, Walter. *Public opinion*. Routledge, 2017.

Appendix

Coefficients of non-zero variables in the lasso regression model (n = 41)

Variable	Coefficient
number of nodes in the whole network	110469.433
number of nodes in the main network	-41112.217
diameter of the main network	-7889.020
number of strongly connected component	7591.090
standard deviation of eccentricity of the main network	5597.957
density of the main network	-3369.253
average shortest path of the main network	-2963.904
comment karma of nodes with high hub score	2659.597
mean of eccentricity of the main network	2302.781
sentence count	-2225.927
average comment karma of the network center	2034.369
link karma of the author	1974.547
comment karma of the author	1707.347
average word length	1671.397
comment karma of nodes of closeness centrality	1370.051
percentage of adjective	1215.104
topic: Malaysian Airline incident	1197.487
word count	1184.894
average link karma of the network center	1184.349
title sentiment	-781.817
comment karma of nodes with high authority score	-768.091
density of the whole network	-752.920
lexical diversity	-726.918
percentage of verb	592.104
topic: Russia	-554.241
topic: natural science and research	506.588
link karma of nodes of closeness centrality	-480.903
number of weakly connected component	401.743
comment karma of nodes of betweenness centrality	-395.849
link karma of nodes of out degree centrality	-393.513
topic: climate change and mineral energy	382.520
topic: health and drug	-285.318
topic: nuclear weapons and national security	-195.877
topic: information security	137.292
radius of the main network	134.417