BACKGROUND AND RELATED WORK

1. Reddit

Reddit.com is a user-generated news aggregating platform in United States. Users post links, original messages and comments in different communities called "subreddits" which cover a wide variety of topics (Kilgo et al. 2016). In the r/worldnews and r/news communities, users post links to online news articles on worldwide/local events released by different news organizations, and vote on perceived importance or interest on the articles (Horne et al. 2017). The ranking of the news articles in the front page "top" section in this subreddit is determined by the number of total support or "upvotes" by the users. Under each post of news article, users can post comments to initiate a discussion or reply to other people's previous comments or sub-comments, and check the total number of comments and upvotes. You can choose the range of time period to see news articles' popularity ranking over different period, including past year, past 24 hours, past week, past month and all time. Crowds influence the ranking and diffusion of news and they are more likely to get exposed to information which is agreed by the majority others (Horne et al. 2017, Gilbert 2013).

2. News Popularity

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. There are a bunch of studies having looked into the diffusion of online contents and tried to predict the popularity of different contents over short term or long term in social media platform such as twitter and digg.com. They used a variety of measurements to capture difference in popularity including 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 reflects total amount of engagement while rating/upvote-downvote reflects the true motivation of users to share the content (Lakkaraju et al. 2013, Horne et al. 2017). However, it is challenging to predict popularity of online contents with a mixture of miscellaneous endogenous and exogenous factors altogether mediates this process.

Scholars have leveraged different attributes of online information of different formats to its popularity including videos, images and texts that circulated on different online platforms (Keneshloo et al. 2016). Many researches have already looked into online news articles' popularity and that turns out to be difficult. Predictors lie in multiple

dimensions and they are hard to measure, 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, few studies have integrated all these aspects together to build a predictive model (Naveed et al. 2011). Features of the online information commonly can be generalized into contextual and content layers.

2.1 Context Features

Users structure attributes and activities of users of special roles are both categorized in contextual level. Though scholars have studied relationship between information diffusion and network structure and user activities, but not many studies have looked at the combined effect from both structural and user level on online news diffusion.

Online users connect with each other through a variety of ways and network structures can be visualized from these different types of connections such as mentioning between different users, commenting on the same posts and answering questions of previous users (Yang et al. 2010). Network attributes are found out to have some influences 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 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 a negligible impact on the range, scale and speed of spread of posts (Yang et al. 2010). Network structure also influences people's cognitive ideas on the same contents. Linked (bilateral) and unlinked friends were fond out to have different possibility to upvote the same content shared by the same person (Hogg et al. 2012). Posts shared by a well-connected user with active followers are more likely to be retweeted (Naveed, 2011).

The network where information circulated determines its popularity. Network structures have been modeled by many scholars to predict information diffusion. Many previous studies have already been conducted on social network like Twitter, but since network structure varies among different online communities, even in the same website, discovered patterns cannot be generalized to other social media platform like reddit (Naveed et al. 2011). Structure is a complex and multifaceted concept to study.

Individuals get empowered in the process of information diffusion in social media. Interactivity and immediacy is the two key features of online news. People not only gain more autonomy in determining what kind of news they like to read, but they are also able to get latest news with no much delay (Reid et al. 2015). 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, which occurs when choices or opinions of others affects a user's behavior (Lerman et al. 2010). Scholars have built up various models to predict information diffusion by studying users' dynamic behavior, such as the study to predict daily variation of news popularity on digg.com (Hogg et al. 2012, 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 type. According to Lazarsfeld and his colleagues, opinion leaders who use to have more insights and influence, take a more important role to receive news in the first stage then pass it to the majority public (1965). Opinion leaders, who have more influence than normal people in public sphere have been identified from reddit by using user profile characteristics, such as commenting, longevity, karma scores, posting frequency, and posting scores (Kilgo et al. 2016). These opinion leaders were 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. Their roles in circulating information is still uncertain, so it would be interesting to identify and observe her influence on news articles' popularity.

The comments over online news articles also have some impacts on their popularity. Popular news always comes with more comments (Reis et al. 2015) which might also impose a "rich-get-richer" effect on the articles. A stark difference exists between attention paid to the new retweeted posts and currently most popular contents (Gilbert, 2013). So existing number of comments and upvotes also exert an influence on contents' popularity in the future.

2.3 Content Features

Content features of text data can be divided into various 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 related to specificity of text data and powerful predictor of content popularity in different social network such as

digg.com and twitter (Horne et al. 2016). Apart from the basic attributes text data such as word count, sentiment features are usually leveraged to predict online news popularity. In studies on reddit, popular news usually has a negative or positive title (Reis et al. 2015) and negative content (Horne et al. 2017), where the tone of title and content are not necessarily the same. These kinds of title and content body are more likely to get more comments in return (Reis et al. 2015). Sentiment should be the main focus of text attributes if we are going to build an overarching model on news articles' popularity.

Except for the semantic attributes, other basic facts about articles are also essential in modeling news popularity, such as the author name, news organization, category, and etc. (Keneshloo et al. 2016). News from reputational organizations are prone to attract more attention News articles of different categories of news articles were found to have different distribution of sentiment scores (Reis et al. 2015). It would be interesting to look at separate effects from categories and sentiments on the communication effect.

References

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.

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

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

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.

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

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.*

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.

*Kilgo, Danielle K., Joseph J. Yoo, Vinicio Sinta, Stephanie Geise, Melissa Suran, and Thomas J. Johnson. "Led it on R*eddit: An exploratory study examining opinion leadership on Reddit." *First Monday* 21, no. 9 (2016).

Lakkaraju, Himabindu, Julian J. McAuley, and Jure Leskovec. "What's in a Name? Understanding the Interplay *between* Titles, Content, and Communities in Social Media." *ICWSM* 1, no. 2 (2013): 3.

Lazarsfeld, Paul Felix, Bernard Berelson, and Hazel Gaudet. *The people's choice: how the voter makes up his mind in a presidential campaign, by Paul F. Lazarsfeld [et al.]*. Columbia Univ. Press, 1965.

Lerman, Kristina, and Tad Hogg. "Using a model of social dynamics to predict popularity of news." In *Proceedings* of *the 19th international conference on World wide web*, pp. 621-630. ACM, 2010.

Leavitt, Alex, and Joshua A. Clark. "Upvoting hurricane Sandy: event-based news production processes on a social news site." In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1495-1504. ACM, 2014.

Naveed, Nasir, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. "Bad news travel fast: A content-based analysis of interestingness on twitter." In *Proceedings of the 3rd International Web Science Conference*, p. 8. ACM, 2011.

Reis, Julio, Fabrıcio Benevenuto, P. Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. "Breaking the news: First impressions matter on online news." In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*. 2015.

Tsagkias, Manos, Wouter Weerkamp, and Maarten De Rijke. "Predicting the volume of comments on online news stories." In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1765-1768. ACM, 2009.

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. "Sentiment strength detection in short informal text." *Journal of the Association for Information Science and Technology* 61, no. 12 (2010): 2544-2558.

Wang, Feng, Haiyan Wang, Kuai Xu, Jianhong Wu, and Xiaohua Jia. "Characterizing information diffusion in online social networks with linear diffusive model." In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference On*, pp. 307-316. IEEE, 2013.

Yang, Jiang, and Scott Counts. "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter." *Icwsm* 10, no. 2010 (2010): 355-358.

Zaman, Tauhid, Emily B. Fox, and Eric T. Bradlow. "A Bayesian approach for predicting the popularity of tweets." *The Annals of Applied Statistics* 8, no. 3 (2014): 1583-1611.