Topicality Boosts Popularity Online: A Comparative Analysis of NYT Articles and Reddit Memes

Kate Barnes[1], Péter Juhász[2], József Pintér[13], Marcell Nagy[1], and Roland Molontay[13]

[1] Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
[2] Department of Mathematics, Aarhus University, Ny Munkegade 118, DK-8000 Aarhus, Denmark
[3] HUN-REN–BME Stochastics Research Group, H-1111 Budapest, Hungary
kate.barnes@coloradocollege.edu, peter.juhasz@math.au.dk, {pinterj, marcessz, molontay}@math.bme.hu

1 Introduction

Popular content on social media sites can be seen as a window into the interests of society. As more people are turning to social media as a news source, many are concerned about these sites replacing conventional news. Do social media sites report on the same issues as verified news sources? Are topical subjects, which are currently popular in the news more likely to go viral on social media? If so, is it worth "riding the wave" of topicality, or does interest in topical subjects decline rapidly on social media?

Memes often discuss cultural and political themes. Research highlighting particular political moments such as the presidency of Donald Trump [5], the Black Lives Matter Movement [4], and the COVID-19 pandemic [1] indicates that topical content may be more likely to go viral online. To the best of our knowledge, the relationship between content topicality and popularity online has not yet been analyzed in general.

In this work, we define topicality using the New York Times (NYT) as a source document. Topics that are currently being published about more than usual in the New York Times are considered topical. We compare the temporal distributions of topics in the NYT and image-with-text memes from Reddit. Despite the vast differences between these data sets (community-run vs. institutional, comedic vs. informational), we observe significant topical alignments. Furthermore, memes that are about topics that are currently popular in the NYT are also more likely to go viral on Reddit.

2 Data and Methods

899,766 memes posted on r/Memes between January 1, 2018 and November 14, 2022, were collected from Reddit using the Pushift API [2]. In addition to meta-data features returned by the API such as the meme title and the number of upvotes it received, we engineered numerical features describing the meme images. Importantly, we extracted text from the meme images using Optical Character Recognition (OCR) and generated image captions using Bootstrapping Language-Image Pre-training (BLIP). Examples of the captions generated with this technique can be seen in Fig. 1.

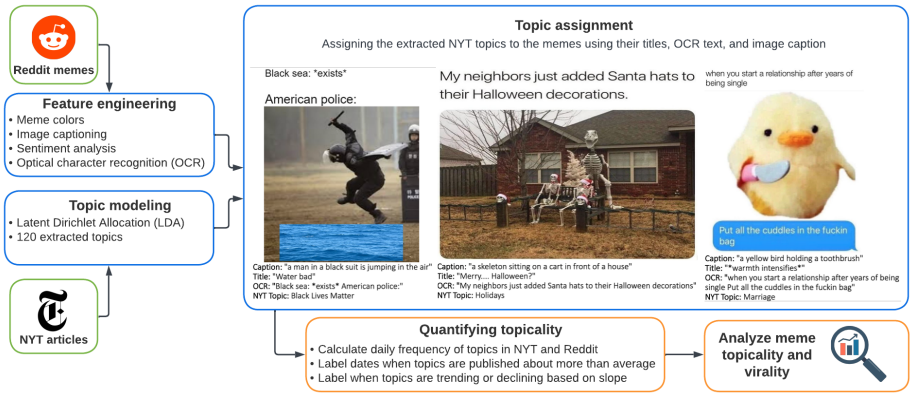


Fig. 1: Workflow and examples of memes, their titles, OCR text, caption, and the assigned NYT topics.

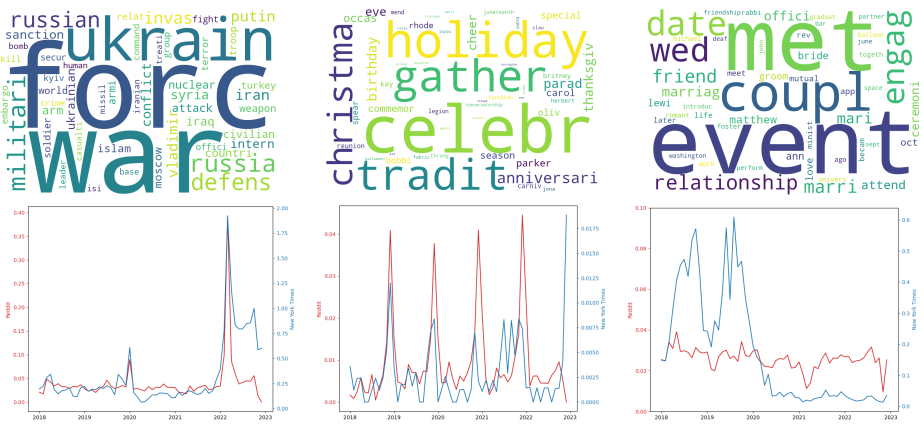
We used the New York Times (NYT) Archives API to fetch data about everything published by the NYT in the same time frame. 72% of the data was from the news section, but other types of material such as obituaries, sports, and op-ed pieces were also included, meaning the NYT topics we analyzed were not only political. In total, 255,783 NYT articles were analyzed.

Using Latent Dirichlet Allocation (LDA) [3], we extracted a set of 120 NYT topics. LDA uses a probabilistic approach to identify word groups that commonly co-occur in documents. Fig. 2 shows word clouds describing three topics found using this method. Each NYT article was assigned to a topic based on the similarity of the articles' and topics' word distributions. Text about each Reddit meme, including the title posted with the meme, OCR-extracted text from the image, and the image caption, was combined. Then, this all-text feature was used to assign the NYT-based topics to our Reddit memes.

3 Results and Discussion

Temporal distributions of the NYT topics were calculated daily by summing the probability with which the memes or articles were assigned to the given topic and dividing by the total number of memes or articles posted on that day. Fig. 2 illustrates the temporal evolution of the weighted proportion of NYT articles (blue line) and Reddit memes (red line) posted about the given topic. Some topics exhibit spikes as in Fig. 2a while others show oscillating patterns such as in Fig. 2b. Reddit memes and NYT articles exhibit topical alignment in many issues such as the invasion of Ukraine (Fig. 2a), holidays (Fig. 2b), Covid-19, the Olympic games and the death of Queen Elizabeth (not pictured). Other topics, such as the romance and marriages topic in Fig. 2c, do not align between the two sources.

We developed variables to analyze whether trending, "topical", topics were more popular than non-trending topics. On dates when the topic is posted about more than average in NYT, we consider it topical and the binary topicality variable takes the value



(a) *UkraineInvasion* (b) *Holidays* (c) *Marriage*

Fig. 2: Topical alignment between the NYT and Reddit.

1. Conversely, when a topic is discussed less frequently than the NYT average, the topicality variable takes the value 0. By this metric, we compared topical and non-topical memes using a pooled T-test, finding that topical memes recieve more upvotes on average on Reddit than non-topical memes (mean scores: topical $= 598$, non-topical $= 573$, $p < 0.001$, Cohen's $d = 0.068$).

We also measured the importance of topicality to identifying viral memes. Many social media sites, including Reddit, exhibit heavy-tailed popularity distributions [1]. Few posts receive a lot of attention while the majority go unnoticed. We define virality based on a percentile. Memes with scores in the top 5% of posts +/-7 days around it are labeled with 1, otherwise 0. A CatBoost classifier was trained to identify viral memes based on three sets of features. The classifier trained with only topicality related features performs significantly better than a random guess (AUC=0.65). Topicality-related features also have incremental predictive power over other content-based variables such as color content and meme text sentiment: they increase the AUC from 0.70 to 0.73. Furthermore, the NYT-topicality features are important to the model's prediction, second only to title length and image dimensions based on the CatBoost's feature importance.

Summary. These results lead us to conclude that topicality does have an impact on the popularity of memes on Reddit. Topical memes are more popular overall, and topicality-based features play a major role in predicting viral memes. While NYT and Reddit are qualitatively very different sources of information, they do show significant alignment. NYT-based topics are good descriptors and predictors of popular Reddit memes.

References

1. Barnes, K., Riesenmy, T., Trinh, M.D., Lleshi, E., Balough, N., Molontay, R.: Dank or not? Analyzing and predicting the popularity of memes on Reddit. Applied Network Science 21(6) (2021)

2. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J.: The pushshift Reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 14, pp. 830–839 (2020)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Leach, C.W., Allen, A.M.: The social psychology of the Black Lives Matter meme and movement. Current Directions in Psychological Science 26(6), 543–547 (2017)
5. Zannettou, S., Caulfield, T., Blackburn, J., Cristofaro, E.D., Sirivianos, M., Stringhini, G., GuillermoSuarez-Tangil: On the origins of memes by means of fringe web communities (2018)