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

Kate Barnes[1], Péter Juhász[2], Marcell Nagy[1], Roland Molontay[1,3*]

[1*]Department of Stochastics, Budapest University of Technology and Economics, Müegyetem rkp. 3, Budapest, H-1111, Hungary.
[2]Department of Mathematics, Aarhus University, Aarhus, Denmark.
[3]ELKH-BME Stochastics Research Group, Budapest, H-1111, Hungary.

*Corresponding author(s). E-mail(s): molontay@math.bme.hu;
Contributing authors: kbarnes@edu.bme.hu; peter.juhasz@math.au.dk;
marcessz@math.bme.hu;

Acknowledgments

Thank you József Pintér for collecting the Reddit data, Donát Köller and Levente Murgás for their insights in our research meetings. Thanks to Fulbright Hungary for funding Kate Barnes at BME's Human & Social Data Science Lab. **The authors declare they have no competing interests.**

Abstract

This study sheds light on interconnected topic dynamics across traditional news sources and social media platforms, emphasizing the influential role of topicality in shaping content popularity in social media. Using the Latent Dirichlet Allocation and BERTopic models, we define sets of 120 New York Times (NYT) topics to compare with 899,766 image-with-text memes from Reddit, showing that social media content aligns with many of the same topical patterns observed in news outlets. Topicality is formalized based on the temporal distributions of the topics over the past 5 years. Using these topicality features, the investigation reveals significant correlations between the rising popularity of NYT topics and increased average upvotes on Reddit, particularly evident in "innovator" memes posted during the early stages of a topic's prevalence in the NYT. Furthermore, topicality features show significant predictive power over other content-based control features in a CatBoost classifier's prediction of viral Reddit memes.

Keywords: memes, popularity prediction, machine learning, New York Times, Latent Dirichlet Allocation, BERTopic

1 Introduction

Popular content from social media sites such as Reddit provides a window into the opinions and interests of internet users. As people are spending more time on the internet, the question of what they are engaging with has come to the forefront, but the forces underlying the emergence of popular internet content are still little understood. Particularly, the influence of memes' topics on their popularity has not yet been analyzed. Do social media sites discuss the same issues as verified news sources? Are topical subjects that are currently prevailing in the news more likely to go viral on social media too? Is it worth "riding the wave" of topicality, or does interest in a topical subject decline rapidly on social media sites?

The widespread prevalence of image-with-text memes online shows how well-suited the internet is to multi-modal reasoning and communication. Although video, sound, and image data are more complicated to analyze than pure text data, making it less studied in the literature, our understanding of online communication would be incomplete without these analyses [21]. To address this shortage of multi-modal analyses, here we study image-with-text memes. These are usually images with overlaid text that are copied and shared extensively across the internet. Due to higher information density, images are more succinct than pure text content. They grab our attention and are able to rapidly communicate complicated messages [32]. Like many forms of online media, the topic of a meme is not only communicated through the text caption but also through the underlain image.

In the digital age, many are concerned about losing depth of conversation in the short-form discourse of the internet. However, social media content may simply achieve nuance in different ways than longer-form written communication. Multi-modality is one way of conveying nuance online. The level of nuance conveyed in online communication is further enhanced by referencing relatable and popular topics in society. The mechanism of reference is especially important to memes, which are elsewhere defined exactly as a reference to reality followed by a punchline [41]. By identifying meme topics, we connect the individual pieces of media to the broader discourse they are a part of. Using big-data, we analyse overall trends in the discourse across 5 years of data.

Memes often reference cultural and political themes [7, 14]. They can be used to consolidate group identities [15] and influence opinions [31]. Zannettou *et al.* labeled memes from four social media sites with annotations from the Know Your Memes (KYM) website, indicating whether a meme template has political content. They found that memes on many social media sites contain political content, and on Reddit political memes are more popular than non-political memes [51]. Today, over half of U.S. adults get news from social media [25]. Furthermore, political memes have the potential to influence social actions and perceptions. For example, one study found that images shared online to promote the Black Lives Matter movement led to greater political action [28]. More recently, TikTok has been used to increase public engagement in the Israeli-Palestinian conflict [50]. Importantly, online information is not always trustworthy and memes containing quotes and factual information have been shown to change as they replicate across the internet [40]. Often memes do not make unbiased references but are imbued with opinions. In memes, the intertwining of informational and normative influence is abundantly clear [10].

Although many studies highlight the fact that online content references current societal events, topicality has yet to be studied more broadly. We introduce formal definitions of topicality in section 3. Informally, when a particular topic is widely prevailing, it is said to be topical.

Topic analysis can be used to connect internet memes to the broader societal themes that they reference. Memes are not made in a tabula rasa environment but can be grouped by their topics, which in turn makes them relevant. Liking and re-posting are implicit endorsements of the memes' topics and opinions [41], implying that topics may influence meme diffusion. Indeed, previous research shows that individuals are more likely to spread information about topics of interest to them. Twitter users are more likely to adopt hashtags that align with their own topical interests [19] and retweet hashtags that they have already tweeted about in the past [49]. Typically, the diffusion of memes through the internet is modeled as a simple contagion process, similar to the spread of disease [47], however, topical content clearly exhibits more complex diffusion mechanisms. Poux-Médard *et al.* studied the interactions between topics in the diffusion of information on Reddit news pages, showing that the topics of previous posts have minor explanatory power for predicting the topics of subsequent posts [35]. These analyses suggest that topics are interrelated and exhibit popularity trends.

The majority of previous research using topic analysis models such as Latent Dirichlet Allocation (LDA) [5] and BERTopic [20] on social media data focuses on the topical interests of internet users. To the best of our knowledge, this is the first study to address topicality in a comprehensive manner. Weng and Menczer show that if a particular hashtag on Twitter is adopted by people with diverse topical interests, the hashtag is more likely to go viral, due to exposure to a diverse, and presumably broadly connected, community [48]. In other cases, homogeneous communities with strong central figures promote content virality when the topic of the shared content matches the interest of the community members, implying the importance of shared beliefs and attitudes [10]. Preferential attachment can be described via topical analysis as well. For example, Instagram users tend to connect with other users with the same interests as them [16]. All of these topic analyses only used one document corpus to define and analyze topics, whereas our analysis uses a benchmark dataset (NYT) to identify the topics of Reddit memes, rather than simply extracting topics from the memes themselves. This comparative aspect leads to better identified topics and can provide insight for other cross-corpus topic analyses.

In our previous work, we showed that Reddit memes with COVID-19-related content were more likely to go viral in March 2020 at the beginning of the pandemic [2] than memes without COVID-19-related content. It is likely that other events exhibit similar properties. Our assumption is that, in general, topical posts about subjects that are currently widely discussed in society are more likely to go viral online. In the present work, we investigate this hypothesis using the New York Times (NYT) as a reference document for topicality. Specific contributions are listed below.

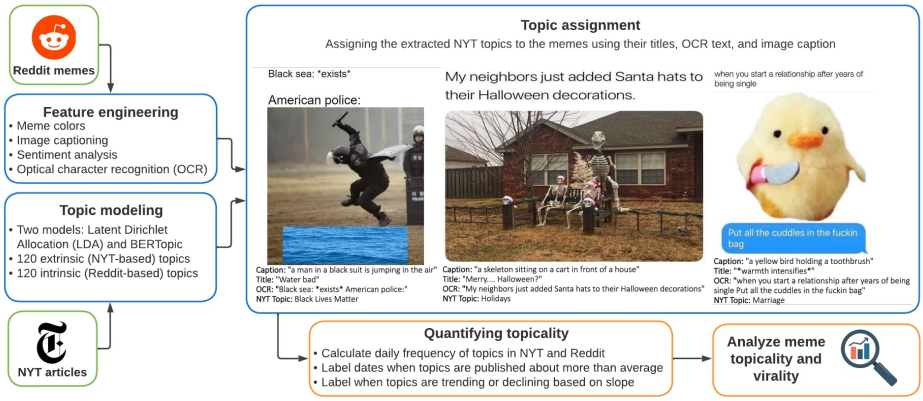

Fig. 1: Research workflow including examples of OCR text extraction, BLIP image captioning, and meme topic assignments.

- Comparing the temporal distributions of 120 topics on Reddit and NYT.
- Showing that Reddit posts about topics that are gaining prevalence in the NYT also receive more upvotes on Reddit, including an innovators' advantage for being among the first to post about a topic.
- Demonstrating topic-based features have significant predictive power in meme virality prediction.

In order to establish these findings, we performed a topical analysis of NYT and Reddit data sets, engineered topic-based features, and used these to predict viral memes with a CatBoost classifier [13]. The research workflow is summarized in Figure 1. Section 2 discusses the data sets and engineering of control features describing the Reddit memes. Topicality features, describing topics extrinsic and intrinsic to the Reddit data, are discussed in Section 3. Section 4 discusses the training of a Catboost classification model to predict viral memes using these topic-based features. We analyze the incremental predictive power of topicality features over other content-based control features, as well as the importance of intrinsic vs. extrinsic topicality features in the same section.

2 Data Description and Preparation

Data was collected from two sources: image-with-text memes from a popular social media site called Reddit, discussed in Section 2.1, and archived article metadata from the New York Times (NYT), see Section 2.2. Sections 2.3 and 2.4 discuss the engineering of content-based control features such as the color content of meme images and the sentiment of meme texts.

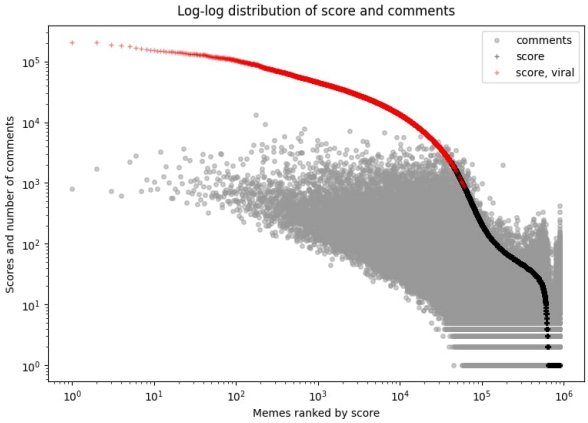

Fig. 2: Distribution of upvotes in Reddit data and corresponding number of comments each meme received.

2.1 Reddit

Reddit, nicknamed "the front page of the internet", is the source of a lot of viral internet data, making it likely that the viral content analysed here also circulates elsewhere on the internet [38]. Data from r/Memes, the largest community dedicated to sharing memes on Reddit, were collected using the Pushift API [3, 6, 34]. We limited collection to a maximum of 1000 randomly sampled posts per day. After removing items with broken image links and gifs, the Reddit data set contained 899,766 memes from between January 1, 2018 and November 14, 2022. The r/Memes subreddit is the 12th largest community on Reddit, with more than 26 million subscribers. Based on the above characteristics, we believe that the collected data is representative and our conclusions can be applied to memes appearing elsewhere.

In addition to the meme images, the API provides a number of meta data features including the title and caption posted with the meme, publication date, an over-18 content indicator variable, the number of comments and the score. On Reddit, the score of a post is calculated as the number of upvotes minus the number of downvotes it received. These attributes provide opportunities to engineer various other features.

As visualized in the memes' score distribution in Figure 2, few memes receive a lot of attention, while the vast majority go unnoticed. In general, memes with high scores also receive more comments, reflected by the Pearson's correlation coefficient of $r = 0.51$ ($p < 0.001$). In many places on the internet including Reddit, popular content whether measured by likes, followers or views, is sorted to the top of users' news feeds [17]. Reddit's default content sorting algorithm, used to curate the "hot" tab, prioritizes content based on a mixture of how recently it was posted and the logarithm of the score [1]. Content ranking methods such as this create a feedback loop between what is popular and what is visible online. While unpopular memes are visible briefly when they are first posted, popular memes are widely viewed. One study manipulated the score of Reddit posts, finding that by providing 10 upvotes soon after

Feature	Type	Description
Score	Int	Number of upvotes - downvotes
Comments	Int	Number of comments
Viral	Binary	1 if score in top 5% of memes
Date	String	Date on which meme was posted
day-of-week	Category	Day of week when meme was posted
all-text	String	Title, BLIP caption and OCR text
Title length	Int	Number of characters in title
OCR length	Int	Number of characters in OCR text
Over 18	Binary	Reddit metadata, content warning
Emoji	Binary	Indicates emoji in memes' all text
Sentiment	Float	Text valence score
Image	String	Link to image thumbnail
Height	Float	Thumbnail height
Width	Float	Thumbnail width
HSV	Float	Average HSV image components
RGB	Float	Average RGB image components
10 colors	Floats	Percentage of color in meme image
Face	Binary	1 if the meme image contains a face
Topic num.	Category	Topic assigned to meme
Topic prob.	Floats	Probabilities associated with topic
Topic entropy	Floats	Shannon's entropy of memes' topics
Monthly	Float	Monthly average topicality
Daily	Float	Daily average topicality
Slope	Float	Slope of topicality distribution

Table 1: Features describing the Reddit memes. Shaded rows indicate features used for the Catboost identification of viral memes including topicality-related features (red) and control features (gray).

the content was posted they could greatly increase the posts chance of going viral [8]. This can explain the heavy-tailed distribution of popularity measures on many sites, including in our Reddit data.

It is a common choice to model the heavy-tailed popularity distributions on social media as binary variables [48, 9]. Here, we define virality as a percentile. If a given meme is in the top 5 percent of memes posted within a +/-7-day period around it, it is labeled viral (1) and otherwise it is labeled non-viral (0). By measuring virality locally, in a two-week period, we avoid replicating seasonal popularity trends on Reddit. Additionally, as Reddit usership is constantly growing [44], this assures we do not overstate the influence of memes from recent years. The viral/non-viral target variable, also visualized in Figure 2, is used in the binary classification task discussed in section 4.

Table 1 summarizes all features describing the Reddit meme data. Shaded rows indicate features used for training the predictive model, including topicality-related features (red) and control features (gray).

2.2 New York Times

Data from the same timeframe as the Reddit memes, January 1 2018 to November 14 2022, was collected from the New York Times Archives API [33]. The API returns all article metadata for a given month, including the abstract, snippit, lead paragraph, headline, publication date, keywords and type of material. Most (72%) of the collected data was from the news section, but other types of material such as crosswords, obituaries, sports and book reviews were also included. The NYT published roughly 200 articles per day on weekdays and 100 articles per day on weekends, and publishing rates showed a slight decline over the 5-year period we examined. On a few dates, the archives contained more than 500 items on a single day due to the NYT updating the archived podcast episodes, or, during the first year of the Covid-19 pandemic and during US elections, due to state-by-state statistics reports. These statistics reports and podcasts were removed to prevent the topic models from forming topics specific to this type of post and to ensure an even daily distribution of articles. No other statistics that should be handled similarly were identified. After removing these entries, our data set contained 255,783 NYT articles in total.

2.3 Text features

We extracted text from the Reddit meme images using Optical Character Recognition [18], and generated image captions using Bootstrapping Language-Image Pre-training (BLIP) [29]. Examples of OCR and image caption results can be seen in Figure 1. The OCR text and image captions were combined with the title and caption posted with the meme image on Reddit into one features containing all text associated with the meme. This all-text feature was subsequently used for topic extraction and assignment in the Reddit data. In the NYT data set, the abstract, snippit, lead paragraph, headline and keywords were all combined into one all-text feature for topic extraction.

The all-text features for the Reddit and NYT data were cleaned identically to assure alignment when modeling the topics in both document corpora. We removed punctuation, made the text lower case, removed stop words and stemmed the words to their root forms using NLTK [4]. Additionally, we made custom edits to the NYT text data based on observed differences between the NYT and Reddit vernaculars. For example the NYT referred to the Covid-19 pandemic with the word "coronavirus" whereas posts on Reddit tended to use "covid", "rona", "corona" and "pandemic". This is important as the LDA algorithm identifies and assigns topics based on shared words. Therefore, if a post on Reddit referred to "covid", it would not necessarily be associated with a NYT topic that only included the word "coronavirus". To solve this issue, we added the words "covid", "rona", "corona" and "pandemic" to tthe all-text feature of every NYT article that was tagged with the NYT keyword "Coronavirus". Other examples include adding “RGB” to articles tagged with Ruth Bater Ginsberg

and "BLM" to articles tagged with Black Lives Matter. Additionally, acronyms of the form "G.O.P." and "N.F.L." were edited to "GOP" and "NFL" in order to not lose this information in when punctuation was removed in the text cleaning process.

In addition to text cleaning, we extracted numerical features from the Reddit text data. These features were used as control features describing the Reddit memes for classification. As seen in Table 1, we recorded the number of characters in the title posted with the meme and the OCR text extracted from the meme image. Other studies show that on Twitter, the length of the post is strongly correlated with popularity [46]. A binary variable indicating whether or not the all-text feature contained emojis was engineered using the Emoji 2.8.0 identification Python library [43]. The valence of the Reddit title and OCR-extracted text was analysed with the NLTK sentiment model [39]. Sentiment scores close to 1 are more positive while sentiment scores close to 0 represent more negative sentiment.

2.4 Image features

In addition to text features, we extracted numerical features from the meme images following a similar procedure to our previous work [2]. In total, we used 19 image-related features: average values of the HSV and RGB components of the meme images, the thumbnail width and height, 10 colors, and a binary face detection variable.

The pre-trained Multi-Task Cascading Convolutional Neural Network (MTCNN) was used to detect faces in the meme images [24]. The model returned the probability that a face was present and the number of faces present in the image. For simplicity, we elected to use a binary variable indicating whether or not the meme image contained a face. Using the OpenCV image segmentation technique to mask the meme images, we calculated what percentage of the image area contained each of 10 color [42]. The average hue, saturation and value components of the HSV representation of the images, and the average red, green and blue components of the RGB image were also used as control features.

Previous work has shown that low-level image features such as those mentioned above can have a large effect on popularity [37, 27]. However, while an earlier study on the aesthetics of images shows that high definition, bright-colored images are more appealing [12], in the case of memes it is the opposite. Popular memes generally contain dull colors [2]. For memes, these low-level features could encode the template image used to create meme, indicating that meme templates rather than the low-level features themselves, have an impact on popularity [11].

3 Topic Modeling

Using two topic models, LDA and BERTopic, we analysed the topics of Reddit memes and NYT articles. Each topic is represented by a list of its most common words, as shown in the example topics in Tables 2 and 3. To identify "extrinsic topics", we trained the topic models on the NYT data and performed inference on the Reddit data. To determine "intrinsic topics" the topic models were trained and inference was performed on the Reddit text data. From the distributions of each topic over time, seen in Figures 3, 5 and 6, we define topicality features for the Reddit data. These

%	Corpus	Type	Top words in topic's word distribution
6.3	NYT	NYT-based	US Politics & Government: "state", "unit", "govern", "polit", "predid", "trump", "nation", "american", "washington"
6.3	NYT	NYT-based	Coronavirus: "coronaviru", "pandem", "viru", "covid", "corona", "rona", "ncov", "quarantin", "reopen", "mask"
4.4	Reddit	NYT-based	Technology: "compani", "social", "compute", "internet", "industri", "medium", "facebook", "online", "technolog"
3.7	Reddit	NYT-based	Cuisine: "food", "restaur", "cook", "recip", "cookbook", "chef", "farm", "wine", "drink", "tabl"
0.8	Reddit	Reddit-based	Video Games: "game", "play", "video", "drive", "minecraft", "car", "skeleton", "mustach", "control", "player"
0.8	Reddit	Reddit-based	Cuisine: "milk", "thano", "slow", "soup", "mushroom", "popcorn", "spoil", "hunger", "drank", "waiter"

Table 2: Most common LDA-identified topics in NYT and Reddit corpora.

%	Corpus	Type	Top words in topic's word distribution
0.9	NYT	NYT-based	Arts: "book", "music", "art", "theater", "museum", "literatur", "literatur", "artist", "play", "novel"
0.5	NYT	NYT-based	Sports: "team", "soccer", "leagu", "game", "player", "basebal", "footbal", "tenni", "basketbal", "sport"
0.7	Reddit	NYT-based	Police Violence: "polic", "murder", "shoot", "homicid", "attempt", "murder", "homicid", "offic", "shot", "kill"
0.6	Reddit	NYT-based	Cuisine: "restaur", "cook", "recip", "cookbook", "cook", "chef", "food", "wine", "dish", "chicken"
0.6	Reddit	Reddit-based	Animals: "dog", "cat", "sit", "cat sit", "banana", "dog sit", "bear", "pet", "cartoon cat", "pictur dog"
0.6	Reddit	Reddit-based	Technology: "phone", "cell phone", "cell", "share", "vote", "reddit", "meme", "youtub", "comment", "post"

Table 3: Most common BERTopic-identified topics in NYT and Reddit corpora.

features are highlighted red in Table 1. In total, 12 topicality features were added for each model, BERTopic and LDA, twice the number listed in the table because the features are calculated using both the extrinsic and intrinsic topic models.

In this section, we discuss the LDA topic model (Section 3.1), the BERTopic model (Section 3.2) and the formalization of topicality into the numerical features (Section 3.3). At the end of Section 3.3 we report statistics describing the relation between topicality features and the popularity of the Reddit memes. The interested reader can see the complete sets of 120 NYT-based topics identified by LDA and BERTopic on our Github page [26]. The choice of 120 topics will also be discussed in the following section.

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a statistical language model used to distill topics from a corpus of documents using a bag-of-words approach in which word order is not considered. Words that commonly co-occur in documents are grouped into topics. Topics are then represented as a probability distribution over words in the corpus, with the words most highly associated with the given topic receiving the highest probabilities. The word distributions of the topics are compared with the word distributions of each document to assess the probabilities with which each document discusses each single topic. Documents are then represented as probability distributions over the set of topics [5].

One limitation of the LDA algorithm is the requirement to specify the number of topics in advance. We used the Keywords metadata provided by the NYT to estimate an appropriate number of topics in the NYT data set. The NYT indexed 1,583,868 unique keywords covering 92% of the articles, the most applied keywords being "US Politics and Government", "Trump, Donald J.", and "Coronavirus (2020-nCov)". We filtered the keywords list based on three criteria. The keyword should be used, 1. at least 100 times total, 2. on at least 50 distinct dates, and 3. at least 5 times on one

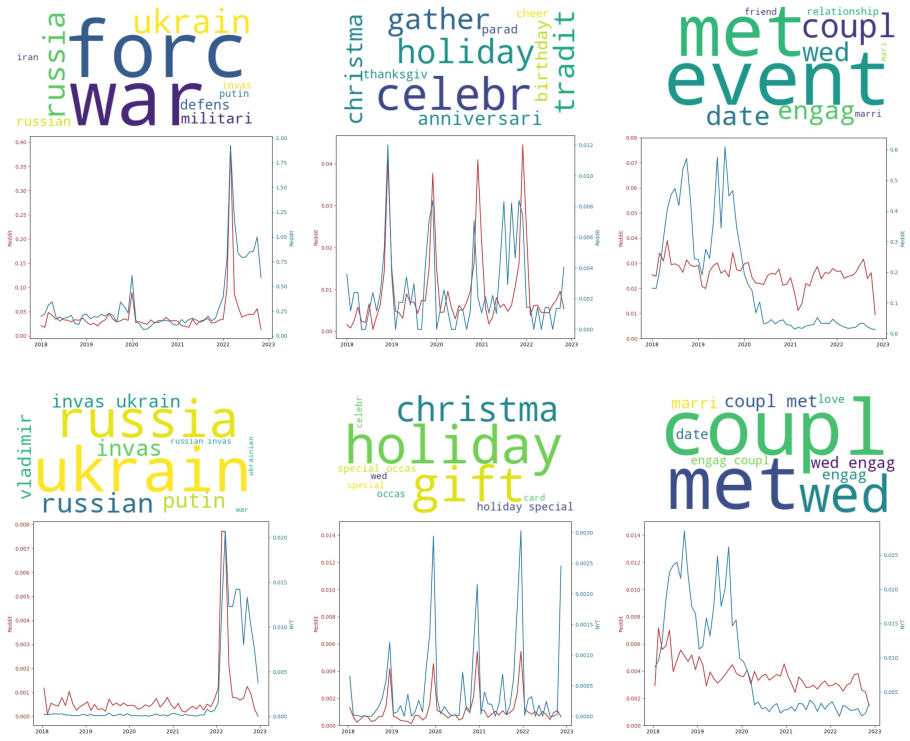

Fig. 3: Monthly average topicality in the NYT (blue line) and Reddit (red line) as identified by the LDA (top row) and BERTopic (bottom row) topic models. Word clouds above the distributions show the top ten words associated with the topic.

date. Only, 127 unique keywords fit these criteria, providing an estimated number of topics for the NYT corpus, which we rounded down to 120 for simplicity.

The NYT all-text feature was used to train the LDA model parameterized to find 120 topics in 1500 iterations over the corpus. We labeled articles in the NYT corpus according to their highest probability topic. Table 2 shows that the most prevalent topics identified in the NYT data are about US politics & government and Coronavirus, aligning with the most common keywords assigned by NYT. Using the trained LDA model we then performed inference on the Reddit all-text feature, resulting in a topic distribution assigned to each meme in the Reddit data, which described the probability that the meme is about each of the 120 NYT-based topics.

We labeled the Reddit memes according to their highest probability topic, saving that topic's probability of being assigned to the meme. We also calculated the Shannon's entropy of the memes' topic distributions. Memes uniformly assigned to all topics received high entropy values and memes assigned with high probability to one or two topics and low probability to the rest received low entropy values. These are the "topic", "prob" and "topic entropy" features described in Table 1.

In the same manner, a set of 120 intrinsic topics (500 iterations over the Reddit corpus) were extracted by training the LDA algorithm on the Reddit all-text feature. Again, the highest probability intrinsic topic, its associated probability and the Shannon's entropy of the topic distribution were saved for the Reddit memes.

Both the intrinsic and extrinsic topic models created one miscellaneous topic in which the most important words in the topic were common words like "already", "life", "way", "time" and "person". This topic was the most prevalent in both the NYT and Reddit corpora, but it was always assigned to less than 10% of the documents. It could be that the memes assigned to this miscellaneous topic are outliers which have no reasonable topic. Table 2 shows the top 2 most frequent topics in the Reddit and NYT data excluding this miscellaneous topic.

3.2 BERTopic

BERTopic is a newer topic modeling technique that uses a transformer, rather than bag-of-words based approach [20]. The model generates high-dimensional document embeddings, reduces the dimensionality of these embeddings with UMAP and then clusters the vectors into topic groups using HDBSCAN. Word order is important for the transformer-based model to understand the context of texts, so it is worth noting that word order was not changed during the text cleaning steps described in section 2.3. To ensure comparability, we used the same text data for both the BERTopic and LDA models.

Although it is not necessary to specify the number of topics in advance of fitting BERTopic, the number of topics can be reduced after the training. We trained BERTopic to cluster topics with at least 100 documents, resulting in 194 NYT topics. Then, we reduced to a set of 120 topics using manual and automatic topic reduction techniques provided by BERTopic. This was done to make results from the two topic models as comparable as possible.

The set of 120 extrinsic, NYT-based topics were extracted using BERTopic from the NYT all-text document corpus, and inference was run on the Reddit all-text corpus to assign these extrinsic topics to the Reddit memes. As in the LDA topic modeling process, we saved the top topic, its probability, and topic entropy for each meme.

As with LDA, BERTopic initially generated miscellaneous topics with common words such as "like" and "also". These topics were combined into one miscellaneous topic which, as in the case of LDA, was the most prevalent topic assigned to the NYT and Reddit documents, but again this topic accounted for less than 10% of both data sets. We trained another model to find 120 intrinsic topics using only the Reddit all-text features, following the same procedure.

A disadvantage of BERTopic is the number of outliers produced, documents which are not assigned to any topic. To address this, we automatically assigned every document to its highest probability topic, resulting in no outlier documents. This was done to match the LDA results, in which every document is represented by a probability distribution over the topics. We used the BERTopic parameter to calculate probabilities to attain the same result. In addition to the length 120 topic vectors assigned to each meme, again we supplemented the meme data with the top topic, its probability

and the Shannon's entropy of the topic distribution to use as features for classification in section 4.

Table 3 shows the top 2 most common topics identified by BERTopic in the NYT and Reddit data sets. Although, the top topics in the NYT corpus differ between NYT and BERTopic, topics about US Politics and Coronavirus were among the top 10 most common topics found by BERTopic in the NYT corpus as well. Furthermore, topics about police brutality and sexual assault were among the top 10 most prevalent NYT-based topics in the memes data as identified by both the BERTopic and LDA models. Figure 3 further shows similarity between the 5-year temporal distributions of topics identified by the LDA and BERTopic models.

3.3 Topicality Features

In addition to saving the topic, probability and topic entropy associated with each Reddit meme, we developed variables to assess whether trending, "topical", topics were more popular than non-trending topics. These variables were calculated for both the LDA and BERTopic models and for both the NYT-based, extrinsic topics and the Reddit-based, intrinsic topics.

Topicality, $\tau_t(d)$, was calculated as the normalized sum of the probability with which documents were assigned to a given topic t on a given date d. Formally,

$$\tau_t(d) = \frac{\sum_{a \in A_d} p_a(t)}{|A_d|}, \tag{1}$$

where A_d is the set of articles published on date d and $p_a(t)$ is the probability that document a is about topic t. Note that topicality can be calculated for either daily or monthly granularity. In Figure 3, d in equation (1) was the month and year in which the document was posted, resulting in monthly average topicality, whereas in Figures 5 and 6, d was the date on which the document was posted. Furthermore, topicality was calculated for both NYT articles and Reddit memes. We use "document" to refer to *either* an article or a meme. The blue lines in the above mentioned figures show NYT topicality distributions in which a in equation (1) represents NYT articles, whereas for the red lines, a represents Reddit memes. The plots serve to compare topical trends in the NYT and Reddit and thus are only based on the sets of extrinsic, NYT-based topics.

Figure 3 shows monthly topicality distributions for three topics found by LDA and BERTopic in the NYT. There was a considerable amount of overlap between the topics found by the LDA and BERTopic models. Topics about current events typically exhibited a rapid spike followed by a slower decline. Oscillating topics appeared for sport events, holidays and and other regular celebrations such as Pride, as these occur periodically. In many topics, such as the invasion of Ukraine and Christmas holidays shown in Figure 3, the distribution of Reddit memes posted about the given topics aligns very well with the NYT topicality distributions. In other topics, such as one about marriage and romance in Figure 3, the two sources differ. We can also see that while the topic models identify the documents associated with each topic well, they

(a) binary topicality (b) positive/negative slope (c) combined

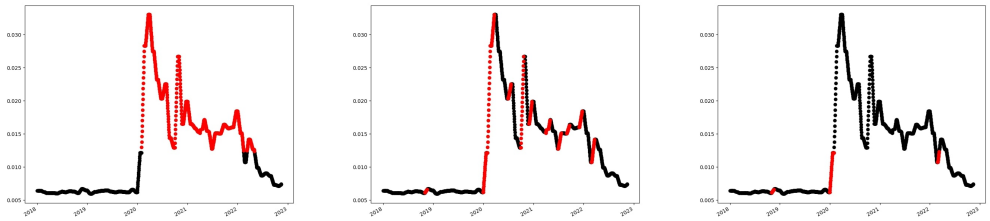

Fig. 4: Reddit extrinsic Covid-19 topicality distribution colored according to three binary topicality features (red=topical, black=non-topical).

are not perfect. For example, some NYT articles about Russian interference in the US 2020 presidential elections are associated with the topic about the invasion of Ukraine by the LDA topic model, as seen in the spike in 2020 in the LDA Ukraine invasion topic in Figure 3. Despite these disadvantages, the identified topics describe the majority of the memes well for both algorithms used.

Applying formula (1) for each topic on the NYT document corpus results in length 120 topicality vectors for each date d, $\underline{\tau}_d$, describing how frequently each topic was being discussed on a given month or day in the 5 years that our data span. This vector can be formalized as,

$$\underline{\tau}_d = \sum_{t \in T} \tau_t(d) \cdot \underline{i}_t, \tag{2}$$

where T is the set of topics, and $\underline{i}_t$ is the unit vector on $|T|$ coordinates with the coordinate corresponding to topic t being 1. Using the topicality vector $\underline{\tau}_d$ and the topic distribution $\underline{\tau}_m$ assigned to a meme m in the model inference step, we define the topicality of a Reddit meme m as,

$$\text{top}_m = \underline{\tau}_{d_m} \cdot \underline{\tau}_m \tag{3}$$

where $\underline{\tau}_{d_m}$ is the topicality distribution of the reference data (NYT for extrinsic topics, Reddit for intrinsic topics) on the given date d_m that the meme m was posted, and $\underline{\tau}_m$ is the topic distribution assigned to the particular meme instance. Note that again, this can be calculated for either a daily or monthly granularity, meaning d_m is either the day on which the meme m was posted or the month in which the meme was posted. These are the "daily topicality" and "monthly topicality" features listed in Table 1. We suspect the monthly topicality variable to have reduced noise. For example, the monthly average topicality for the holidays topic, seen in Figure 3, is higher for all memes posted in December due to the effect of Christmas, but the daily average topicality variable has higher values only on the few days around Christmas and the other holidays.

In addition to the daily and monthly topicality features, reflecting how frequently a given topic was being discussed in the NYT or Reddit, we designed a feature describing

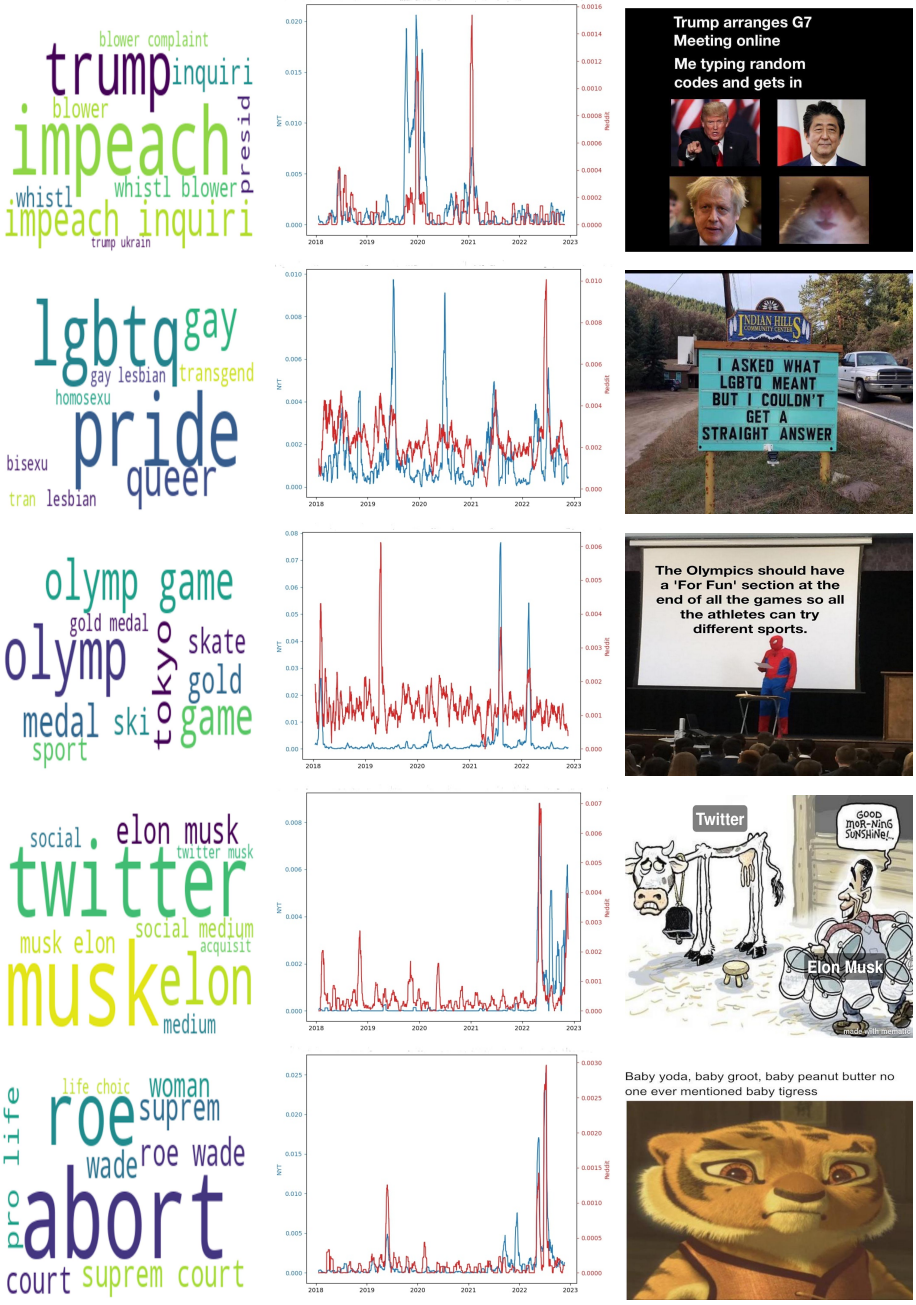

Fig. 5: BERTopic-identified topical allignments between the NYT and Reddit.

Fig. 6: LDA-identified topical allignments between the NYT and Reddit.

the slope of the daily topicality distributions (e.g. Figure 6). This was calculated as the average slope of the reference data distribution for the +/-2 days around when the meme was posted. Given that topics about current events tend to spike rapidly, we used a small time frame. This is the "slope topicality" feature in Table 1. We suspect that a topic will be more popular on Reddit when the topicality distribution has a positive slope, indicating increasing topicality trends.

Two binary topicality features were engineered based the slope and daily topicality features. These binary variables were used for statistical tests showing that topical memes are more popular than non-topical memes. The first binary topicality variable, visualized in Figure 4a, indicates whether the meme's topic is being posted about more (labeled 1) or less (labeled 0) than average in NYT on the date the meme was posted. The second binary variable, visualized in Figure 4b, was calculated based on the slope topicality. This variable takes the values 1 when the slope of the reference topicality distribution is positive, and 0 when the slope of the distribution is negative. On average, memes about topics that show increasing, positive slope, topicality trends receive a higher score than memes about topics with decreasing topicality trends (LDA mean scores: 632, 541, $p < 0.001$, BERTopic mean scores: 694, 672,$p < 0.01$).

The cross tabulation of these two binary topicality variables makes a categorical topicality feature with 4 categories. This categorical feature indicates both whether the slope of the reference distribution was positive or negative on the date when the meme was posted and whether the topic was being published about more or less than average on the date when the meme was posted.

Figure 4c shows category 1, when the slope of the reference distribution is positive but the topic is not yet being posted about more than average. A pooled T-test showed that memes posted in this period recieve higher average scores than than the other 3 groups (LDA mean scores: 644, 571, $p < 0.001$ BERTopic mean scores: 609, 575, $p < 0.001$). Differences between the other groups, for which T-test results were not reported, were not significant.

4 Identifying Viral Memes

In order to gain a better understanding of how topicality influences the virality of a meme, we trained a CatBoost classifier. CatBoost, an ensemble learning algorithm for gradient boosting on decision trees, has also been shown to outperform other algorithms on a variety of problems [23]. The classifier is able to parse categorical features, meaning the 120-category topic features did not have to be one-hot-encoded before training. We used the image, text and topicality features summarized in the shaded rows of Table 1 for the binary classification task of identifying viral memes in our Reddit data. Figure 7 shows the algorithms performance identifying viral memes.

The binary target variable was highly unbalanced because it was defined as a percentile: only the top 5 percent of Reddit memes were considered viral. We used the CatBoost weighting parameter to improve performance on the imbalanced class prediction. The CatBoost classifier was trained using data from 763,388 memes, and tested using 84,821 memes. The model was fine-tuned with 5-fold cross validation, and sklearn's GridSearchCV for parameter selection. Notably, title length feature is the

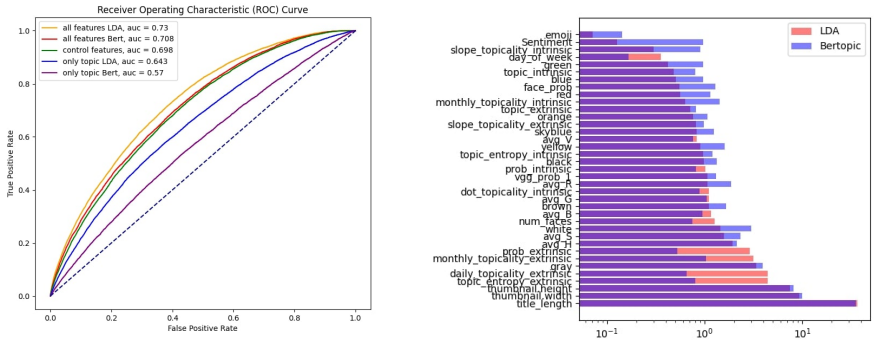

Fig. 7: ROC and features importances results identifying Reddit meme virality with CatBoost model

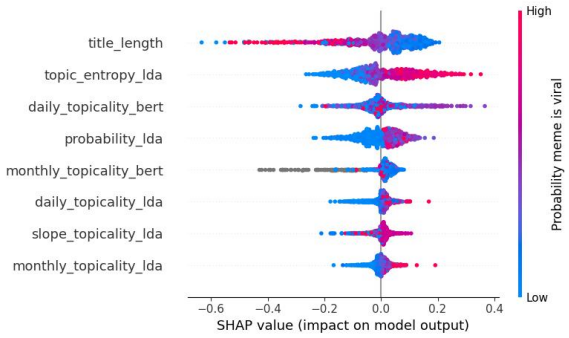

Fig. 8: SHAP values for extrinsic topicality features indicating that topicality positively affects the CatBoost probability memes are predicted viral.

most important for classifying viral memes. Meme's posted with short titles preform better than meme's accompanied by a lot of text. This mirrors earlier work showing brevity is the most important feature for viral tweets [46]. The reason could be that messages that are easier to convey is understood better. Given the importance of the text length variable, we used CatBoost's per-float-feature-quantization parameter to increase the number of decision boundaries allowed for this feature, taking advantage of its predictive power.

The classifier was trained separately for the LDA-based and BERTopic based topicality feature sets. In order to estimate the importance of the topic-based features for predicting meme virality, we trained the CatBoost model with different subsets of features. The first models were trained using only topicality features, highlighted red in Table 1. Alone, the topicality features (AUCs: LDA=0.64, BERTopic=0.57) are already able to predict meme virality better than a random guess (AUC=0.5).

It could be the case that the explanatory power of the topicality features is already contained by information in the control features set. To test this, we trained the model with only control features, described in sections 2.3 and 2.4. Then, we assessed the incremental predictive power of topicality features over control features. Models trained with the entire feature set, all shaded rows in Table 1, were better able to identify viral memes than the model trained with only the control features set. We tested the statistical significance of these AUC differences using a procedure from previous research [22], finding the engineered topicality features have significant predictive power over the control features for both BERTopic ($AUC = 0.71$, $p < 0.01$) and LDA ($AUC = 0.73$, $p < 0.001$). Results were more significant for the LDA topic model, and topicality features ranked higher in the LDA features importances than for BERTopic, as seen in Figure 7.

Figure 8 shows the SHapley Additive exPlanations (SHAP) values for given extrinsic topicality features [30]. The text length feature, which is negatively correlated with meme virality, is provided for reference. The topicality features positively impact the CatBoost's estimation of meme virality.

5 Conclusion

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. Here we showed:

- Many topics on Reddit and the NYT show similar topicality distributions.
- Memes that are about topics with increasing topicality trends in the NYT receive more upvotes on average.
- "Innovator" memes posted when a topic is just beginning to be prevalent in the news receive the most upvotes.

These findings can be interpreted from the perspective of agenda-setting theory, which states that media sources shape the interests of society by selecting which topics to publish about. Our finding that news topicality has a positive effect on social media virality also indicates that it may be harder to get attention for and spread non-topicality information. This study does not unveil a causal link between topics in NYT and on Reddit, however the great overlap between the sources indicates that topically they exert similar influence on their audiences. Future research could delve into differences between the opinions purported by news and social media about the same topics.

Articles from many disciplines explore the diffusion of internet memes through online sites. Our results showing that topicality plays a role in meme popularity, suggest that topics and topicality should be taken into account to model the more complex diffusion process of online media.

Most topic analyses have analysed the relationship between user topical interests and content. Our analyses of topics in general and use of a reference document to assess the topicality of the topics is a unique contribution. Defining topicality based on only the NYT can also be seen as a limitation of this work too. It would be an interesting

challenge to develop a more complete description of what subjects are topical on the internet by combining multiple reference documents. While NYT does report on global events and on issues other than news, using NYT as a source document skewed our set of topics to be more USA- and politics-centred. The USA-skewed results was appropriate for our purposes to match the large proportion of Reddit users based out of the USA, 48% of all users [44]. However, this means our results are less relevant to other parts of the globe.

Finally, the results presented here contribute to the growing body of research modeling multi-modal data. With the rise of internet communication, multi-modality is a paradigm shift in the forms discourse takes in general. Predicting the popularity of multimedia content is more difficult than text-only data. For example, impressive results have been achieved predicting viral tweets [49, 45], but we have not yet seen such success predicting the popularity of image-with-text memes. One article achieved an AUC of 0.86 predicting image-with-text memes, but they worked with a small data set labeled by hand by humans, and kept only the highest and lowest scoring memes, thereby eliminating all moderately-popular memes and reducing the difficulty of the problem [9]. Recent improvements in generative AI will soon make modeling multi-modal content more tractable. For now, predicting multi-modal content remains a challenge, and our AUC of 0.73 stands out in the literature.

With increasing availability of internet data, social media has become a popular source of data for academics. Reddit is an especially popular data source, and the ethical guidelines around using this data are not yet firmly established. In light of this, we heed recommendations from studies surveying social media users about research data use. Namely we collected data from a large, public community, assured no personal identifying information was associated with the data, and we plan to share our results back to Reddit upon completion [36].

References

[1] Amir Salihefendic. *How Reddit Ranking Algorithms Work*. Accessed June 1, 2023. https://www.yale.edu/about-yale/yale-facts. 2015.

[2] Kate Barnes et al. "Dank or not? Analyzing and predicting the popularity of memes on Reddit". In: *Applied Network Science* 6.1 (2021), pp. 1–24.

[3] Jason Baumgartner et al. "The pushshift Reddit dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 830–839.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[6] Bryce Boe. *PRAW: The Python Reddit API Wrapper*. https://github.com/praw-dev/praw. Accessed: 2022-12-15. 2016.

[7] Richard Brodie. *Virus of the mind: The new science of the meme*. Hay House, Inc, 2009.

[8] Mark Carman et al. "Manipulating visibility of political and apolitical threads on reddit via score boosting". In: *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE. 2018, pp. 184–190.

[9] Chen Ling, Ihab Abuhilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, Gianluca Stringhini. "Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes". In: *Association for Computing Machinery* 5 (2021). DOI: https://doi.org/10.1145/3449155.

[10] Robert B Cialdini and Noah J Goldstein. "Social influence: Compliance and conformity". In: *Annu. Rev. Psychol.* 55 (2004), pp. 591–621.

[11] Michele Coscia. "Average is boring: How similarity kills a meme's success". In: *Scientific reports* 4.1 (2014), p. 6477.

[12] Ritendra Datta et al. "Studying aesthetics in photographic images using a computational approach". In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III 9*. Springer. 2006, pp. 288–301.

[13] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support". In: *arXiv preprint arXiv:1810.11363* (2018).

[14] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. "Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 153–164.

[15] Marta Dynel and Thomas C Messerli. "On a cross-cultural memescape: Switzerland through nation memes from within and from the outside". In: *Contrastive Pragmatics* 1.2 (2020), pp. 210–241.

[16] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. "Online popularity and topical interests through the lens of instagram". In: *Proceedings of the 25th ACM conference on Hypertext and social media*. 2014, pp. 24–34.

[17] Flavio Figueiredo et al. "On the dynamics of social media popularity: A YouTube case study". In: *ACM Transactions on Internet Technology (TOIT)* 14.4 (2014), pp. 1–23.

[18] a9t9 software GmbH. *OCR.space Free OCR API and Online OCR*. URL: https://ocr.space/ (visited on 03/01/2023).

[19] Przemyslaw Grabowicz, Niloy Ganguly, and Krishna Gummadi. "Distinguishing between topical and non-topical information diffusion mechanisms in social media". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 10. 1. 2016, pp. 151–160.

[20] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[21] Alon Halevy et al. "Preserving integrity in online social networks". In: *Communications of the ACM* 65.2 (2022), pp. 92–98.

[22] James A Hanley and Barbara J McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1 (1982), pp. 29–36.
[23] Abdullahi A Ibrahim et al. "Comparison of the CatBoost classifier with other machine learning methods". In: *International Journal of Advanced Computer Science and Applications* 11.11 (2020).
[24] Iván de Paz Centeno. *MTCNN 0.1.1 for Python*. Accessed Feb 1, 2022. https://pypi.org/project/mtcnn/. 2022.
[25] Jacob Liedke and Luxuan Wang. *Social Media and News Fact Sheet*. Accessed Dec 1, 2023. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/. 2023.
[26] Kate Barnes. *Topical allignments between the NYT and Reddit*. Accessed Dec 1, 2023. https://k-barnes.github.io/memes_topicality.html. 2023.
[27] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. "What makes an image popular?" In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 867–876.
[28] Colin Wayne Leach and Aerielle M Allen. "The social psychology of the Black Lives Matter meme and movement". In: *Current Directions in Psychological Science* 26.6 (2017), pp. 543–547.
[29] Junnan Li et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12888–12900.
[30] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
[31] Brian McClure. "Discovering the discourse of internet political memes". In: 2016.
[32] Douglas L Nelson, Valerie S Reed, and John R Walling. "Pictorial superiority effect." In: *Journal of experimental psychology: Human learning and memory* 2.5 (1976), p. 523.
[33] New York Times. *NYT Developers: Archives API*. Accessed Jan 15, 2022. https://developer.nytimes.com/docs/archive-product/1/overview. n.d.
[34] Matthew Podolak. *PMAW: Pushshift Multithread API Wrapper*. https://github.com/mattpodolak/pmaw. Accessed: 2022-12-15. 2021.
[35] Gaël Poux-Médard, Julien Velcin, and Sabine Loudcher. "Properties of reddit news topical interactions". In: (2022), pp. 16–28.
[36] Nicholas Proferes, Naiyan Jones, and Michael Zimmer. "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics". In: *Social Media and Society* 7 (2 2021). DOI: https://doi.org/10.1177/20563051211019004.
[37] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115 (2015), pp. 211–252.
[38] Beth Sanderson and Miriam Rigby. "We've Reddit, have you?: What librarians can learn from a site full of memes". In: *College & Research Libraries News* 74.10 (2013), pp. 518–521.

[39] Prajwal Shreyas. *Sentiment analysis for text with Deep Learning*. 2019. URL: https://towardsdatascience.com/sentiment-analysis-for-text-with-deep-learning-2f0a0c6472b5 (visited on 10/01/2020).

[40] Matthew Simmons, Lada Adamic, and Eytan Adar. "Memes online: Extracted, subtracted, injected, and recollected". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1. 2011, pp. 353–360.

[41] Brian H Spitzberg. "Toward a model of meme diffusion (M3D)". In: *Communication Theory* 24.3 (2014), pp. 311–339.

[42] R Stone. *Image Segmentation Using Color Spaces in OpenCv+Python*. 2018. URL: https://realpython.com/python-opencv-color-%20spaces/ (visited on 10/01/2020).

[43] Taehoon Kim, Kevin Wurster. *Emoji 2.8.0 for Python*. Accessed May 1, 2022. https://pypi.org/project/emoji/. 2023.

[44] Tiago Biachi. *Reddit - Statistics & Facts*. Accessed August 1, 2023. https://www.statista.com/topics/5672/reddit/#topicOverview. 2023.

[45] Oren Tsur and Ari Rappoport. "Don't let me be# misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes". In: *Ninth International AAAI Conference on Web and Social Media*. 2015.

[46] Oren Tsur and Ari Rappoport. "Don't let me be# misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 9. 1. 2015, pp. 426–435.

[47] Lin Wang and Brendan C Wood. "An epidemiological approach to model the viral propagation of memes". In: *Applied Mathematical Modelling* 35.11 (2011), pp. 5442–5447.

[48] Lilian Weng and Filippo Menczer. "Topicality and impact in social media: diverse messages, focused messengers". In: *PloS one* 10.2 (2015), e0118410.

[49] Lilian Weng et al. "Competition among memes in a world with limited attention". In: *Scientific Reports* 2 (2012), p. 335. DOI: https://doi.org/10.1038/srep00335.

[50] Moran Yarchi and Lillian Boxman-Shabtai. "The Image War Moves to TikTok Evidence from the May 2021 Round of the Israeli-Palestinian Conflict". In: *Digital Journalism* (2023), pp. 1–21.

[51] Savvas Zannettou et al. "On the origins of memes by means of fringe web communities". In: *Proceedings of the internet measurement conference 2018*. 2018, pp. 188–202.