Judging a Book by Its Cover: Predicting the Marginal Impact of Title on Reddit Post Popularity

Evan Weissburg, Arya Kumar, Paramveer S. Dhillon
University of Michigan, Ann Arbor, MI 48104
{evancw|arkumar|dhillonp}@umich.edu

Abstract

Several factors influence the popularity of content on social media, including the *what*, *when*, and *who* of a post. Of these factors, the *what* and *when* of content are easiest to customize in order to maximize viewership and reach. Further, the title of a post (part of the *what*) is the easiest to tailor, compared to the post's body, which is often fixed. So, in this paper, we assess the impact of a post's title on its popularity while controlling for the time of posting (the *when*) by proposing an interpretable attention-based model. Our approach achieves state-of-the-art performance in predicting the popularity of posts on multiple online communities of various sizes, topics, and formats while still being parsimonious. Interpretation of our model's attention weights sheds light on the heterogeneous patterns in how the specific words in a post's title shape its popularity across different communities. Our results highlight the power of sentiment alignment, personal storytelling, and even personality politics in propelling content to virality.

Introduction

Online communities have an essential role in shaping public perception regarding critical societal issues. They serve as discussion forums and aid the dissemination of information. A crucial feature of most such communities is user feedback (both positive and negative) on posts, which determines content visibility and popularity. However, despite a deluge of social media posts, few reach a broad audience and become viral. Hence, it is crucial for social scientists to understand the determiners of content popularity, which could further shed light on the mechanisms by which online communities shape opinions. Social media platforms must also unpack the determiners of content popularity to improve the content they serve to users, as must content publishers that aim to produce compelling content.[1]

Although the problem is simple to formalize, modeling content popularity on social media is challenging because many factors determine popularity, including audience visibility, context, and time (Mazloom, Hendriks, and Worring, 2017). These factors can be distilled down to three main elements of content, the *what* (what content was posted, e.g., the words, pictures, gifs, etc.), the *when* (when was the content posted, e.g., time-of-day, month, etc.), and the *who* (who posted the content, i.e., the identity of the author).

Of these three factors, the *what* and *when* can be easily customized to maximize viewership or reach, unlike the author (*who*), which is typically fixed. For instance, some content publishers post content on social media only at certain times, e.g., on Mondays or in the mornings, to garner as much widespread attention as possible. Similarly, it is common to tailor the title of a post to attract attention to the post's actual content body, which is often predetermined in the form of text, gif, image, or video to be shared. The post's title is also its most salient "attention-grabbing" attribute and is a crucial driver of a user's decision to click-through and consume the actual content. Further, titles are also a required component of posts across all social media websites and are, therefore, a generalizable property of such platforms. Hence, we focus on the title of a post in this paper.

This paper seeks to unbox the impact of a social media post's title on its popularity and further study the heterogeneous patterns across multiple online communities of various sizes, topics, and formats. We study the above research question on the `Reddit.com` social media community. We chose Reddit for our study since its size makes it highly representative of the overall internet. With many different topic-driven "subreddit" sub-communities, Reddit offers a rich landscape to study community dynamics in closer detail. These subreddit communities are centered around a single topic, so we can easily control for the general topic and audience by only comparing posts within the same subreddit. Figure 1 shows a snapshot of a subreddit.

In order to isolate the impact of the title of a post, we condition on the body of the post and assume it is fixed. We further consider only posts made within a 30-minute timeframe of each other to control for time (*when*). The author's identity is less salient in determining content popularity on Reddit due to the lack of a significant "follower" mechanism. Hence, user identity effects are minimized, and controlling for the *who* is not warranted. Further, since we only compare posts from the same topic-driven subreddit community within a set time window, *context* effects are also minimized.

Though we chose Reddit as the focus of our study, our model applies to any text or caption-oriented social media platforms, such as Twitter or Facebook. Our model could

[1]Readers may recognize the task of content publishing when creating a picture caption for an Instagram post or crafting a humorous reply to a family member on Facebook.

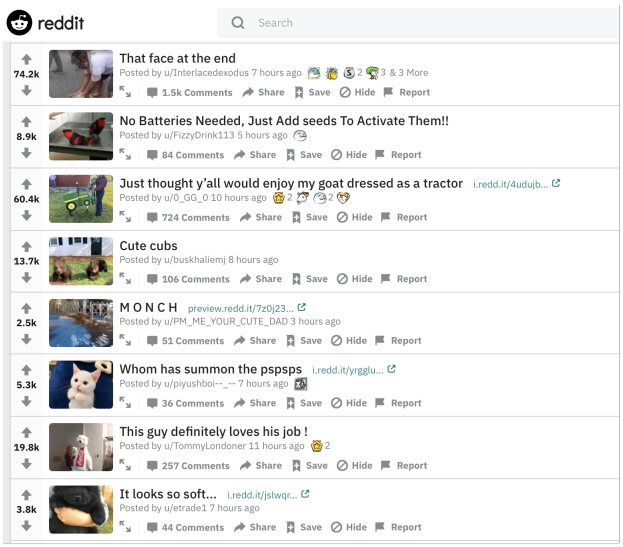

Figure 1: An example from subreddit /r/pics, where posts are ordered by user voting patterns. Older posts drop in position over time for freshness.

be trained at a platform level to recognize viral content or trained to learn the nuances of popular content on a smaller scale using specific community groups, as is the focus of our study.

Methodologically, we propose a simple novel neural network model that links the text of a post's title to its popularity. Our model extracts textual features by leveraging an attention mechanism (Vaswani et al., 2017). Unlike deeper frameworks that use multiple attention layers totaling millions of parameters, we use only one layer to extract contextual clues from the post title (Devlin et al., 2018). This makes our model flexible, parsimonious, and interpretable.

Next, we interpret our results and model weights within each subreddit's context, which allows us to understand how important the title is to different communities on Reddit. Finally, we interpret the model's attention weights to determine which title words are most influential in propelling content to virality, illuminating the power of sentiment alignment, personal storytelling, and even personality politics. These interpretability analyses also reveal how a content publisher can adapt titles to a specific subreddit to boost its popularity potentially.

Related Work

Reddit, the sixth-largest website on the internet by unique traffic, is home to over a million subreddit communities. Each of these user-created themed communities is a microcosm of the broader social news ecosystem, aggregating thousands of user-generated posts sorted by realtime user voting feedback (Medvedev, Lambiotte, and Delvenne, 2017). Since Reddit provides such a user-centric slice of social media, it is a natural site to study online community interaction and content virality at scale (Wang, Hamilton, and Leskovec, 2016; Zayats and Ostendorf, 2018). For example, many studies have attempted to predict *comment popularity* using linear regression, feed-forward neural networks, and recurrent neural networks (RNNs) (Fang, Cheng, and Ostendorf, 2016; Horne, Adali, and Sikdar, 2017; Wang, Ye, and Huberman, 2012). The most successful ones have used deep reinforcement learning, for instance, for predicting the popularity of future comment replies to existing threads (He, Ostendorf, and He, 2017; He et al., 2016). All these studies reveal that, on Reddit, context is key to predicting popularity.

A different thread of research focuses squarely on predicting *post popularity*, a more difficult and input-rich task; rather than a simple sentence or paragraph, a user-generated post can include a link, image, gif, or video (Berger and Milkman, 2012). Additionally, many variables are involved in post popularity including title, content, creation time, subreddit of origin, and author (Bakshy et al., 2011; He et al., 2016; Hessel, Lee, and Mimno, 2017; Hong, Dan, and Davison, 2011; Lakkaraju, Mcauley, and Leskovec, 2013; Mazloom, Hendriks, and Worring, 2017; Suh et al., 2010). Disentangling the related effects of these variables to assess the weight of a single one is a difficult task, yet it often yields fruitful community insights.

Content popularity prediction is challenging in general because it is the result of a stochastic user-influenced positive feedback loop. Salganik, Dodds, and Watts (2006) explore this phenomenon, concluding that although the best content tends not to be rated poorly by validators and the least popular content tends not to be rated well, for most other posts, validation results were inconclusive. Stoddard (2015) investigates this relationship between inherent post quality and post popularity further, introducing a viewership-corrected metric of inherent quality on dual datasets of Reddit and Hacker News. Contrary to Salganik, Dodds, and Watts (2006), Stoddard (2015) finds that his inherent quality metric has a strong correlation with popularity. However, Stoddard (2015) omits a large quantity of data in his analysis, preferring to focus on high-ranking posts.

Similar content "reposting" is another phenomenon of interest, first studied in the context of post popularity prediction by Lakkaraju, Mcauley, and Leskovec (2013). They developed a model based on title, image, and repost count to predict the relative popularity of reposted content generally on Reddit. Interestingly, they note that a repost following a previously popular high-visibility post is unlikely to be popular.

Studies in other online communities have also found that title plays a critical role in content popularity. On a dataset of online news articles, Piotrkowicz et al. (2017) improve on popularity prediction baselines using a title-only model and propose that the content title is often "the most prominent" component of online content. Undoubtedly, content titles serve as useful hooks to draw initial attention.

Prior to our work, the most generalizable model for predicting post popularity on Reddit was a multimodal study on the role of captions versus images by Hessel, Lee, and Mimno (2017). They use existing machine learning architectures to contrast the importance of post titles and related images on subreddits `r/aww`, `r/pics`, and `r/cats`. They

show that treating popularity prediction as a pairwise task can be effectively used to control for time, so we also adopt this relative popularity prediction task in our model. Their results suggest that both text and image features are essential components in a popularity prediction model. Our results show conversely that popularity prediction is tractable without using image features and with a much smaller model. Therefore, our work differs in scope since we exclusively focus on the impact of a post's title in driving popularity and assume the post's content is fixed ahead of time, as is often the case when posting an image or link online. Further, while Hessel, Lee, and Mimno (2017) analyze the role of caption versus image; our work focuses on presenting our state-of-the-art model for title-oriented popularity prediction, as well as developing a model-driven qualitative analysis of the community-level textual factors that propel content to virality.

Problem Formulation

In this section, we provide a brief overview of the Reddit platform and formalize the problem of post popularity prediction.

Reddit Overview

The community-driven discussion platform Reddit consists of multiple topic-focused subreddits. Within each subreddit community, users generate posts consisting of a title, content, and other associated metadata. Reddit allows registered users to upvote or downvote on posts, generating a per-post "karma" score, which determines a post's visibility (Medvedev, Lambiotte, and Delvenne, 2017). Reddit allows users to post a variety of multimodal content, though many subreddit communities place controls over post content; for example, `r/AskReddit` allows no content (title only), `r/politics` allows only URL link-based submissions, `r/pics` allows only images and `r/depression` allows only text-based posts. We view a user-generated post $p \in \mathcal{P}$ as a collection of associated data: $p =$ (score, title, content, subreddit, timestamp). While the user has no control over a post's score, they dictate its title, content, subreddit, and post creation time.

Problem Formalization

As described earlier, we focus our popularity study on post titles rather than content. Titles feature most prominently in post previews where many users vote. Additionally, users have greater control over their post's title than their post's content — we focus on the common task of "captioning" where a news link or image content is already fixed (Tan, Lee, and Pang, 2014). This title-centric approach also allows us to contrast the role of post title across different subreddits by comparing our model's accuracy and attentional output between these communities. We are aware of no prior work that makes this type of deep cross-subreddit comparison.

Following the lead of several previous papers on predicting post popularity, we formulate the problem as a pairwise ranking task (He et al., 2016; Hessel, Lee, and Mimno, 2017; Mazloom, Hendriks, and Worring, 2017). Given distinct posts $p_1, p_2 \in \mathcal{P}$, we aim to predict whether p_1 or p_2 has a higher associated popularity score using only the associated titles.

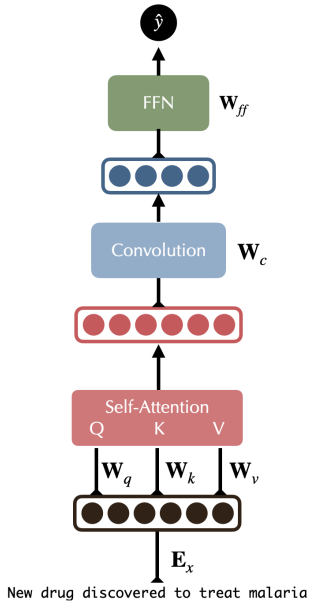

Figure 2: Model component illustration.

Model

Our model is shown graphically in Figure 2 and it consists of four layers: *embedding*, *attention*, *convolution*, and *feed-forward*. These layers are explained in more detail below.

Embedding Layer: The embedding layer encodes the title information for a given post into fixed-length vectors. We use pretrained GloVe vectors for our embeddings, with embedding dimension $d_m = 300$ (Pennington, Socher, and Manning, 2014).

Attention Mechanism: Though initially developed for sequence-to-sequence translation, the attention mechanism is now widely used for deep text-based natural language processing tasks, including studies with Reddit data (Devlin et al., 2018; Vaswani et al., 2017; Yu and Jiang, 2019; Zhang et al., 2018). We simplify the fully stacked attention mechanism as used by (Devlin et al., 2018; Vaswani et al., 2017) to a single self-attention module for our problem to make our model parsimonious. Through self-attention to word embeddings, this layer can add context to each word in the title, acting as a secondary embedding function serving input to the convolutional and dense layers described later.

The attention mechanism used in this paper is a version of scaled dot-product self-attention. The input consists of the fixed-length embeddings produced from the title, each of dimension d_m. These word embeddings are formulated into a matrix M, and the following attention product is computed.

$$\text{SelfAttention}(M) = \text{softmax}\left(\frac{MM^T}{\sqrt{d_m}}\right)M \qquad (1)$$

Convolutional Layer: Next, our model has a 1-dimensional convolutional layer, inspired by previous applications of convolution for text analysis (Kim, 2014). In this novel application, we convolve on attention outputs positionally with a fixed kernel size. Since we treat self-attention as a popularity contextualization function, inserting a convolutional layer between the attention output and the feed-forward module helps preserve positional information between word attention outputs. For each window of length k, this mechanism processes input of dimension (k, d_m) and produces a scalar. Overall, for an original input title with n words and kernel size k, this mechanism produces vector output with length $d_n - k + 1$. The convolutional layer also helps minimize per-token parameter counts before the final dense layer, thereby preventing overfitting.

Feed-Forward Layer: The final module in our model is a feed-forward concatenation layer. This dense layer squashes the convolution layer's positional output into a scalar popularity prediction.

Loss Function: Based on previous work in post popularity prediction, we use a pairwise ranking function to predict relative post popularity (Burges et al., 2005; Herbrich, Graepel, and Obermayer, 2000; Hessel, Lee, and Mimno, 2017; Joachims, 2002) which lets us control for time effects since we compare posts that were posted within a 30-minute timeframe. To incentivize predictive confidence, we would like to maximize the pairwise difference in model outputs, so we implement a contrastive max-margin loss function. Given two posts p_1, p_2 such that $\text{score}(p_1) > \text{score}(p_2)$, with model outputs x_1, x_2, we compute,

$$\text{Loss}(x_1, x_2) = \max(0, x_2 - x_1) \quad (2)$$

For inference purposes, it is also possible to interpret the model's output x, for an unpaired post p.

Experimental Setup

This section describes the training and evaluation details of our model and the Reddit post selection/pairing algorithm for our pairwise-ranking loss function. Our model and baselines were trained on an Nvidia 2080Ti with 512 GB RAM. Our model processed approximately 100 pairs per second without batching and around 3000 pairs per second with a batch size of 64. Our source code is accessible online[2].

Dataset Source

Our Reddit data is sourced from `pushshift.io`, an online dataset generated using Reddit's public API. This dataset was originally scraped in 2015, extended in 2016, and made public in 2020 (Baumgartner et al., 2020; Hessel, Tan, and Lee, 2016; Tan and Lee, 2015).

We focus our attention on Reddit submissions from 2017. For each subreddit of interest, we filter posts according to the following criteria. We remove posts that received few (less than 2) upvotes and the posts that were "stickied" on the subreddit — a mechanism that allows subreddit moderators to artificially boost the visibility of posts circumventing the normal voting-based process. Table 1 shows the details of the sixteen subreddits that we used for our study.[3]

[2]https://github.com/evanweissburg/judging-a-book/

Subreddit	Post Pairs
/r/Showerthoughts	44,863
/r/AskReddit	47,313
/r/news	27,692
/r/worldnews	38,576
/r/relationships	21,962
/r/depression	8,787
/r/aww	103,604
/r/pics	52,628
/r/politics	142,218
/r/The_Donald	103,774
/r/sports	3,019
/r/soccer	40,807
/r/science	4,957
/r/NoStupidQuestions	6,828
/r/funny	92,699
/r/jokes	29,971

Table 1: Dataset Size by Subreddit

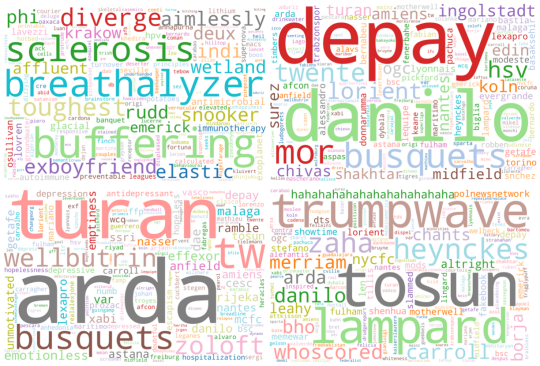

Figure 3: Word cloud visualization of (un)popular content on Reddit stratified by quartiles; Top Left = 75-100% (most popular), Top Right = 50-75%, Bottom Left = 25-50%, Bottom Right = 0-25%

Preliminary Analysis

To provide a preliminary view into the types of content that is popular on the sixteen subreddits that we study, we display word cloud visualizations, split into quartiles based on popularity. The score of any word w is computed as the average scores of the posts $p \in \mathcal{P}$ in which it appears. To prevent larger subreddits from dominating the results, each post's score is normalized by the mean of the top 100 posts in the subreddit. Additionally, to avoid misspelled words in low scoring posts, we only show words in the graphic that appear at least 10 times in the corpus. Once each word has an associated score, we generate a word cloud for the 150 ran-

[3]For computational reasons, subreddit /r/The_Donald was reduced in size by 1/6 by random sampling.

dom words whose score falls into the corresponding quartile, as shown in Figure 3.

At first glance, words that appeal to high-stakes events and emotions perform favorably, from medical terms ("autoimmune", "sclerosis") to descriptions of relationships and trauma ("exboyfriend", "homicides", "unhappiness"). We see references to famous individuals, like Arsenal striker Pierre-Emerick Aubameyang, and to shared frustrations, like slowly-buffering content. In the bottom quartile, we see that certain controversial sentiment ("trumpwave", "fakebook", "whiteness") appears to score poorly overall. Similarly, simple rules such as "soccer players score well" do not appear to be true — names of soccer players appear in every quartile of the results, emphasizing that constructing a viral post on Reddit requires nuance.

Post Pairing

Our post pairing algorithm controls for audience and time to minimize confounding in the dataset. First, since Reddit subreddits vary in viewership, we only pair posts from the same subreddit. To control for time, we greedily select posts ordered by time so that posts in a pair are from a similar time window. Since gaps can exist in a subreddit's posting history due to temporary closures or site unavailability, we also guarantee posts must be sourced within a 30-minute time frame. Following Hessel, Lee, and Mimno (2017), we minimize noise by pairing posts if their voting score difference is at least 20 and the more popular post's score is at least double the score of the less popular post.

Baseline Models

We implement four strong baselines against which we compare our model, as illustrated in Table 2. First, we implement a simple logistic regression using one-hot sentence encoding and a vocabulary of 20,000 words. With its low parameter count, this model provides a simple and interpretable baseline result. Next, we implement a two-layer multilayer perceptron model (MLP) that takes a GloVe sentence vector average as input. Our next baseline model is a more sophisticated deep learning method for textual data—a bidirectional LSTM model (BiLSTM) that takes GloVe word vectors as input. Finally, as is standard practice these days, we provide a fine-tuned BERT baseline for the task (Devlin et al., 2018). We used BERT-Base from the Huggingface[4] library and added a 2-layer dense module to the output. To make a fair comparison, we freeze all but the final dense layers for training purposes. We also benchmarked against a lightweight BERT variant (DistilBERT), but the results were very similar to BERT-Base, so we only report the BERT-Base results. Although BiLSTM and BERT-Base generally outperform weaker baselines, their outputs are difficult to interpret due to their vast number of parameters. In comparison, our model has approximately 1/4 the number of parameters as BiLSTM.

Finally, we perform 5-fold cross-validation for all our experiments. We randomly shuffle the data (subreddit post pairs) and train our model on 4/5 of the data and test on the remaining 1/5. This also allows us to measure variability in our results and hence also assess statistical significance.

[4]https://huggingface.co/transformers/model_doc/bert.html

1Hot Logistic	GloVe MLP	BiLSTM	BERT-Base	Ours
20k	2M	600k	110M	500k

Table 2: Parameter Count by Model

Results

Accuracy

Table 3 presents our results on eight popular subreddits divided into four categories by *subreddit submission type*. We note that our model performs best compared to baselines on subreddits where the content submission type is title-only and link, as expected. Overall, our model is competitive across all content types but is not statistically distinguishable in terms of its accuracy on subreddits with image content, matching BiLSTM in both cases. This suggests that the post's title is a less important determiner of its popularity when the content that it points to is image-based.

Next, we present another eight subreddits in Table 4, divided into four categories by topic. Here, we note that our model is effective on a wide range of subreddit discussion topics and vocabularies.

Ablation Study

Though our model is relatively simple, we performed an ablation study on the subreddit `/r/travel` to determine the efficacy of the convolutional layer. It turns out that the predictive accuracy of our model drops significantly from 0.823 to 0.808 if we remove the convolutional layer, which shows its importance in terms of contributing to sparsity and minimizing overfitting on smaller subreddits.

Interpreting Attention Weights

Next, we interpret our model's output attention weights. Since this is not a prediction task, we train our model on the entire dataset. For an input title of length k, the corresponding model self-attention output weights have shape (k, k).

Researchers have used attention mechanisms to understand model behavior in tasks such as recommender systems, neural machine translation, and text labelling (Wu et al., 2019b,a; An et al., 2019; Ding, Xu, and Koehn, 2019; Mullenbach et al., 2018; Xie et al., 2017). Theoretically, we frame our qualitative analysis as a view into the words and phrases that the downstream model is most interested in. We perform three types of qualitative studies interpreting relative attention weightings, as outlined below.

Top Subreddit Attention Weights Looking at the top few attention weights provides a simple but broad insight into popular content on a given subreddit. For each subreddit, the 15 words with highest attention weight are reported in Table 5.

Though the performance of the 1Hot Logistic regression baseline is inferior to our model, it is similarly easy to interpret. Thus, Table 6 reports the 20 top (absolute) feature weights from the 1Hot Logistic model as well as

Submission Type	Subreddit	1Hot Logistic	GloVe MLP	BiLSTM	BERT-Base	Our model
Title-only	/r/Showerthoughts	0.577 (0.004)	0.583 (0.008)	0.593 (0.006)	0.596 (0.004)	**0.603 (0.001)**
	/r/AskReddit	0.607 (0.004)	0.628 (0.003)	0.632 (0.003)	0.630 (0.004)	**0.644 (0.004)**
Link	/r/news	0.595 (0.006)	0.605 (0.004)	0.613 (0.005)	0.602 (0.006)	**0.637 (0.005)**
	/r/worldnews	0.633 (0.008)	0.638 (0.006)	0.650 (0.001)	0.647 (0.004)	**0.663 (0.002)**
Text	/r/relationships	0.695 (0.005)	0.732 (0.004)	0.748 (0.006)	0.747 (0.003)	**0.756 (0.004)**
	/r/depression	0.649 (0.010)	0.706 (0.010)	0.733 (0.005)	0.704 (0.006)	**0.741 (0.006)**
Image	/r/aww	0.597 (0.003)	0.594 (0.004)	**0.609 (0.003)**	0.596 (0.003)	**0.609 (0.001)**
	/r/pics	0.596 (0.007)	0.605 (0.004)	**0.618 (0.005)**	0.594 (0.003)	**0.618 (0.003)**

Table 3: Model Results by Subreddit Submission Type. *Note:* 5-fold cross validation accuracy is reported with standard deviation in parenthesis; results significant at the $\alpha = 0.05$ level in a pairwise t-test are bolded.

Topic	Subreddit	1Hot Logistic	GloVe MLP	BiLSTM	BERT-Base	Our model
Politics	/r/politics	0.612 (0.005)	0.631 (0.002)	0.638 (0.004)	0.616 (0.003)	**0.642 (0.004)**
	/r/The_Donald	0.637 (0.003)	0.631 (0.004)	0.640 (0.003)	0.634 (0.001)	**0.652 (0.001)**
Sports	/r/sports	0.511 (0.012)	0.589 (0.012)	0.617 (0.024)	0.577 (0.013)	**0.640 (0.006)**
	/r/soccer	0.635 (0.003)	0.642 (0.007)	0.651 (0.010)	0.625 (0.005)	**0.661 (0.003)**
Science	/r/science	0.620 (0.009)	0.641 (0.010)	0.674 (0.011)	0.673 (0.013)	**0.695 (0.015)**
	/r/NoStupidQuestions	0.532 (0.013)	0.594 (0.014)	0.611 (0.003)	0.599 (0.009)	**0.634 (0.005)**
Humor	/r/funny	0.560 (0.004)	0.565 (0.004)	0.574 (0.003)	0.566 (0.005)	**0.580 (0.002)**
	/r/jokes	0.586 (0.004)	**0.591 (0.009)**	**0.596 (0.006)**	0.590 (0.003)	**0.600 (0.004)**

Table 4: Model Results by Subreddit Topic. *Note:* 5-fold cross validation accuracy is reported with standard deviation in parenthesis; results significant at the $\alpha = 0.05$ level in a pairwise t-test are bolded.

the top-20 attention weights output by our model for the `/r/politics` dataset. As can be seen, there is minimal overlap in the two sets of word distributions, which highlights the ability of our model to learn highly discriminative words to predict post popularity.

Title Weights by Subreddit Model To understand subreddit popularity more deeply, it is helpful to visualize the same title with models trained on different subreddits. In Figure 4 we choose a viral title from some *origin* subreddit and examine differences in word-level attention weighting using a non-origin subreddit model. By analyzing the same title using model variants trained on different subreddits, we gain insight into how our model interprets community-level variation in popular content. For example, we might expect our model to pay more attention to "Trump" in `/r/The_Donald` than `/r/politics`, and we observe this behavior in the right-panel of Figure 4.

Subreddit Attention Directed Graphs A more powerful way to compare subreddit communities at scale using our model is by generating an attention directed graph, as in Figures 5-6. These graphs are generated from the trained subreddit model by collecting average attention weights for each input word-output word pair. To create each visualization, we systematically graph the strongest average input-output attention weights as directed edges, with stopwords removed for visual clarity. Therefore, a directed edge from word A to word B implies that our model believes A is an influential context for B with respect to the task of popularity prediction.

/r/AskReddit	/r/relationships	/r/depression	/r/pics
underwear	surgery	isolate	married
9/11	update	hug	weirdest
yelp	injury	convince	upvotes
woods	disability	dog	diet
nsfw	son	killed	marry
panties	option	cat	lbs
bridges	stalking	cried	filmed
sfw	therapy	kill	redditors
flash	twins	pet	questions
sleeper	property	picture	answers
pg-13	terrified	hardest	lb
wwe	creepy	terrified	filming
tortured	devastated	crush	diagnosed
erection	autistic	coward	answer
briefs	bathroom	knew	skyrim

Table 5: Top-15 words according to our model's attention weights for different subreddits.

Discussion

As shown in Tables 2, 3, and 4, our approach consistently beats strong baselines by 1-3% for the pairwise prediction task with a comparably small number of parameters. Hence, our model achieves both state-of-the-art performance and parsimony.

Previously top-performing baselines like BiLSTM and BERT-Base preclude model interpretability and are highly overparameterized. On the other hand, our approach combines high accuracy with weight interpretability akin to logistic regression. But unlike a logistic regression, our mod-

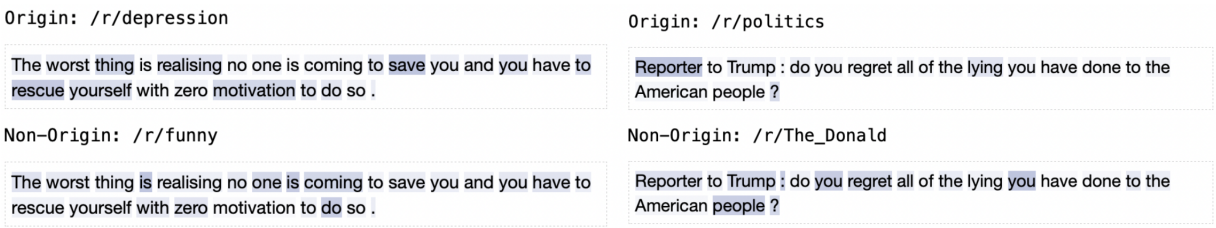

Figure 4: Comparing attention weights generated by two different subreddit models for the same title: "Save" and "rescue" contribute to popularity on /r/depression, but not on /r/funny (*left panel*). Similarly, references to "Trump," "regret," and use of the second person boost popularity more on /r/The_Donald than on /r/politics (*right panel*).

1Hot Logistic	Weight	Our model	Weight
Korea	-0.22	Las	4.49
North	-0.21	Venezuela	3.99
poll	0.14	Vietnam	3.90
report	0.14	Olbermann	3.63
Trump	0.13	WikiLeaks	3.60
Donald	0.13	transcript	3.56
Fox	0.13	Warren	3.50
neutrality	0.13	MSNBC	3.48
dem	0.12	neutrality	3.40
China	-0.12	WSJ	3.40
US	-0.12	marijuana	3.19
Russia	0.11	Fox	3.11
Syria	-0.11	look	3.04
net	0.11	Assange	3.03
immigration	-0.11	Maddow	3.01
court	-0.11	impeached	3.00
Supreme	-0.11	EU	3.00
travel	-0.11	explains	2.96
americans	0.10	tells	2.92
Jeff	0.10	trial	2.91

Table 6: Comparing the top-20 words based on their weights for the baseline 1Hot Logistic and our model on /r/politics

els' attention weights can be analyzed more deeply at the community level to uncover contextual trends in online popularity, as we explore below.

It is worth emphasizing that our study's primary objective is not to produce a more accurate popularity prediction model for Reddit. If that were the desiderata, then we would augment our model with more complex features rather than just the title. Instead, our focus is to capture the differential impact of post title across various communities and uncover interesting popularity patterns, as we discuss next.

Table 3 shows, albeit a bit unsurprisingly, that our model demonstrates the most significant improvement over baselines on title-only and link-based subreddits. We expect it to perform best on title-only subreddits since, by definition, our model is title-oriented. These results also point to a secondary result: post title is also crucial on link-based subreddits. A plausible explanation for this is that Reddit provides an easy mechanism to vote on a link-based post after only viewing the post title. Therefore, many Redditors may fail to read externally-linked material before voting (Holmström et al., 2019). Next, we look at performance of our model on subreddits which contain topical content. As can be seen in Table 4, we find that title is a particularly important factor in popularity for science and sports-related content. Interestingly, our approach does not demonstrate considerable improvements on humor-related subreddits.

In Tables 5 and 6, we list the words to which the model attributes the highest attention weight; theoretically, per subreddit, these words are the most effective captioning tools to improve the popularity of content. Table 6 also provides a comparison between our model's top words and the baseline 1Hot Logistic regression, which yield slightly different results. On `/r/politics`, we find only two words of twenty ("neutrality" and "fox") are the same between both models, as shown in Table 6. While both models emphasize the importance of key names such as "Donald Trump," "Warren," or "Maddow," our model places additional significance on personal opinions, such as "explains" and "tells".

Exploring differences in language between subreddits through these word weights yields community insight. Previous work in online content virality found that popular content is mainly propelled by its positive valence and physiological arousal. For example, posts that inspire awe, rage, or anxiety tend to be more viral than posts that create a deactivating feeling of sadness (Rubenking, 2019; Tellis et al., 2019). Our qualitative findings align well with these hypotheses, though the relative roles of valence versus arousal differ between communities.

On `/r/AskReddit`, a subreddit based on asking opinion-based questions of the broader community, it is evident from Table 5 that the most viral posts reference emotional extremes such as fear, violence, and sexual content. `/r/relationships` indicates a similar trend: posts about creepy, stalking, or abusive relationships engender the most upvotes and community engagement. Based on these observations, we conclude that both of these communities tend to promote content that produces high arousal, whether through eroticism, anger, or a sense of injustice.

The subreddit `/r/depression` tells a slightly different story. Although high-arousal negative words like "isolate" and "killed" are common, lower-arousal positive valence words that indicate support systems are also prevalent. From "hug" to "cat," "dog," or "pet," this community clarifies that online post popularity allows for a full range of

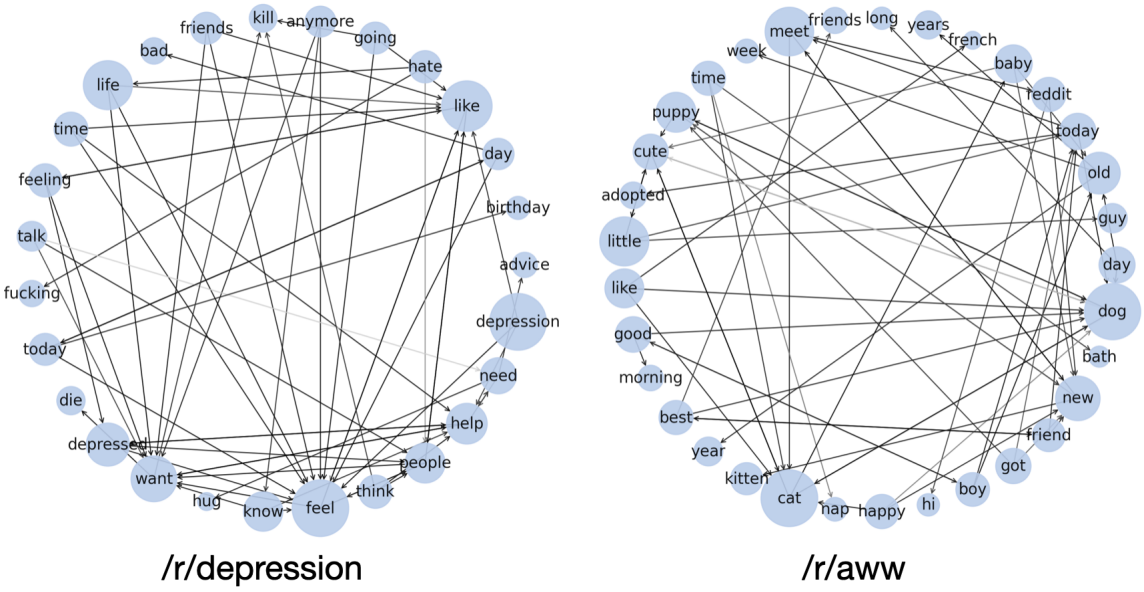

Figure 5: Subreddit graphs generated from attention weights showing which word nodes the popularity model links to others via directed edge; *Left Figure:* Word Graph for /r/depression; *Right Figure:* Word Graph for /r/aww.

emotion.

An unexpected trend comes from the picture-sharing subreddit `/r/pics`. Instead of words indicating interest, beauty, or landscape, this community primarily popularizes pictures regarding personal health and weight loss. Although this subreddit may have begun as "[a] place for pictures and photographs," per its current description, it appears to have evolved in focus through user-curation, a community migration phenomenon unique to Reddit. Rather than being drawn to exciting or beautiful images in isolation, users of `/r/pics` prefer pictures that come with emotional positive personal narratives, like overcoming cancer, losing weight, or getting married (Xu, 2019).

When the simple top-N word weights are insufficient insight, our model allows deeper analysis at the subreddit level using an attention-directed word graph. `/r/depression` shows a range of emotions in popular posts, as indicated in Figure 5. One can visually note certain themes from the graph vocabulary and connection network: loneliness ("want talk [to] people," "need friends/hug," "today [is my] birthday") and depression ("feel depressed," "want kill/die") are associated with the strongest attentional word connections, indicating popular subreddit sentiment. Compared to the word-level attention weights in Table 5 which show a significant number of viral positive words, we find fewer positive valence word chains in this figure. Since our model identifies fewer common positive valence word chains, we infer that within `/r/depression`, popular positive-valence posts tend to be more complex and varied in sentence structure than comparatively popular negative-valence posts, which appear to revolve around the same word chains and topics. This narrative is supported by the left panel of Figure 4, an example comparing attention outputs for a title originating on `/r/depression`. We find that "save you" and "rescue yourself" are strong popularity boosters on `/r/depression`, whereas the `/r/funny` model variant's attention output is distributed evenly, indicating that these negative-valence word chains would fail to boost post popularity on `/r/funny`.

The attentional graph for `/r/aww` lies in sharp contrast with `/r/depression`, indicating substantial differences in the format and relative complexity of subreddit title construction. The largest nodes reflect universally cute imagery (cats and dogs, puppies, kittens, and babies). With `r/aww` in particular, strong word chains become clear ("Reddit, meet my cute new puppy," or "Good morning from my happy kittens"), indicating the simplicity and repetitive nature of successful titles. Here in particular, the value of self-attention is clear; with these visualizations, we gain insight into particular phrases and pairings that are successful on a given subreddit. In `/r/aww`, for example, cats and dogs are frequently used together, and "new" is usually used in conjunction with "Reddit" and "meet".

Similar patterns can be found for `/r/jokes` in Figure 6. Just as with `/r/aww/`, there are clear word chains ("A man walks into a bar" or "Wanna hear a joke"). However, the graph as a whole highlights an interesting issue. Many of the most common words in the titles are verbs, and these connect almost exclusively to subjects — the most salient features our model can extract here are the actions that the subject of the joke takes. However, often the most crucial element of a joke is a pun or clever use of word choice. These elements can be difficult to capture with the GloVe word embeddings that our model uses, since these vectors struggle to accurately represent multiple meanings of a given word at once.

Finally, we use our model outputs to provide insight into political discussion on both sides of the aisle. In our

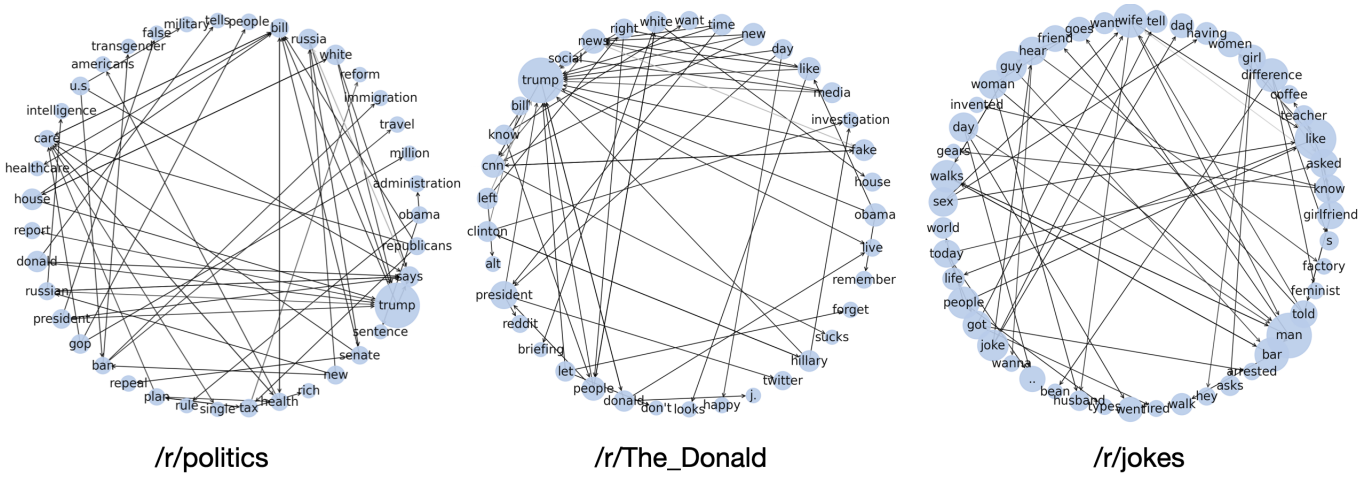

Figure 6: Subreddit graphs generated from attention weights showing which word nodes the popularity model links to others via directed edge; *Left Figure:* Word Graph for /r/politics; *Middle Figure:* Word Graph for /r/The_Donald; *Right Figure:* Word Graph for /r/jokes

analysis of `/r/politics` (a subreddit for US politics that typically leans liberal) and `/r/The_Donald` (a subreddit for supporters of Donald Trump), we find significant thematic differences in attentional word associations. Though both subreddits attribute what "Trump says" to content popularity, `/r/politics` focuses on Trump's standpoints on political issues whereas `/r/The_Donald` is more interested in Trump's opinions about figures and institutions. Note that the strongest connections surrounding Trump on `/r/politics` are "bill," "immigration," and "healthcare," compared to "CNN," "Clinton," "Obama," and the "media" on `/r/The_Donald`. Our model highlights that those who frequent the pro-Trump `/r/The_Donald` promote content that centers around Trump's personality rather than his political agenda, compared to the more agnostic `/r/politics`. In the right panel of Figure 4, we see a comparative attention visualization of an August 2020 news title exemplifying further differences. While `/r/politics` seems to be interested in the "reporter" and potential "lying," `/r/The_Donald` is intrigued that "Trump," "regret[s]," "you," and "people" are mentioned, indicating that the model trained on `/r/The_Donald` believes this populist sentiment will help propel this particular post to virality within the scope of its community.

Conclusion

In this paper, we unpacked the marginal impact of the title of a social media post in driving its popularity while controlling for the post body and the post's timing. We leveraged a self-attention based model to extract salient features from post titles. Our model showed strong performance on various subreddits while being parsimonious. Further, the model's attention weights permit an in-depth analysis of individual posts and community trends beyond simply identifying popular keywords. We compared online communities at scale through subreddit-level attention visualizations, highlighting differences in sentiment, word choice, and political ideology. Through our per-subreddit qualitative analysis, we find that viral content is extremely heterogenous on Reddit. This emphasizes the importance of considering the audience of a target community in order to tailor the text of a title; while an uplifting and positive narrative might be effective in generating popular content within one community, a caption arousing frustration might perform better in another. Our work provides an intuitive attention-based framework to study inter-community differences in detail.

Future work on online popularity prediction could consider the addition of new features to maximize performance, such as submission time, image or text content, or authorship in the context of an ensemble-type model. It might also be desirable to adapt and apply our model to other text-oriented platforms, such as Twitter or Facebook. For example, it could be enlightening to obtain a cross-platform view of content popularity to compare and contrast viral content at an internet scale on Reddit versus Instagram. Since our model is parsimonious and attention-based, new interpretation or visualization techniques could also be developed in future work to obtain novel insights into content virality.

Ethics Statement

This study was conducted on public data, in accordance with the Reddit Terms of Service. However, users may not expect their data to be used and disclosed for research purposes. We acknowledge and mitigate this consideration by focusing our analysis on broad community-level phenomena, with the goal of understanding trends in online interaction.

Another ethical consideration is use of this research for tailoring misinformation or propaganda. These issues make it more critical to analyze the factors involved in online content virality in order to understand and combat these dangers.

References

An, M.; Wu, F.; Wu, C.; Zhang, K.; Liu, Z.; and Xie, X. 2019. Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 336–345. Florence, Italy: Association for Computational Linguistics.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 830–839.

Berger, J., and Milkman, K. L. 2012. What makes online content viral? *Journal of Marketing Research* 49(2):192–205.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, 89–96. New York, NY, USA: Association for Computing Machinery.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

Dhillon, P. S., and Ungar, L. 2009. Transfer learning, feature selection and word sense disambiguation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 257–260.

Dhillon, P. S.; Foster, D.; and Ungar, L. H. 2008. Efficient feature selection in the presence of multiple feature classes. In *2008 Eighth IEEE International Conference on Data Mining*, 779–784. IEEE.

Dhillon, P. S.; Foster, D.; and Ungar, L. H. 2011. Minimum description length penalization for group and multi-task sparse learning. *The Journal of Machine Learning Research* 12:525–564.

Ding, S.; Xu, H.; and Koehn, P. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 1–12. Florence, Italy: Association for Computational Linguistics.

Fang, H.; Cheng, H.; and Ostendorf, M. 2016. Learning latent local conversation modes for predicting comment endorsement in online discussions. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, 55–64. Austin, TX, USA: Association for Computational Linguistics.

He, J.; Ostendorf, M.; He, X.; Chen, J.; Gao, J.; Li, L.; and Deng, L. 2016. Deep reinforcement learning with a combinatorial action space for predicting popular Reddit threads. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1838–1848. Austin, Texas: Association for Computational Linguistics.

He, J.; Ostendorf, M.; and He, X. 2017. Reinforcement learning with external knowledge and two-stage q-functions for predicting popular reddit threads. *arXiv preprint arXiv:1704.06217*.

Herbrich, R.; Graepel, T.; and Obermayer, K. 2000. Large margin rank boundaries for ordinal regression. In Smola, A.; Bartlett, P.; Schölkopf, B.; and Schuurmans, D., eds., *Advances in Large Margin Classifiers*, 115–132. Cambridge, MA: MIT Press.

Hessel, J.; Lee, L.; and Mimno, D. M. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. *CoRR* abs/1703.01725.

Hessel, J.; Tan, C.; and Lee, L. 2016. Science, askscience, and badscience: On the coexistence of highly related communities. *CoRR* abs/1612.07487.

Holmström, J.; Jonsson, D.; Polbratt, F.; Nilsson, O.; Lundström, L.; Ragnarsson, S.; Forsberg, A.; Andersson, K.; and Carlsson, N. 2019. Do we read what we share? analyzing the click dynamic of news articles shared on twitter. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, 420–425. New York, NY, USA: Association for Computing Machinery.

Hong, L.; Dan, O.; and Davison, B. D. 2011. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, 57–58. New York, NY, USA: Association for Computing Machinery.

Horne, B. D.; Adali, S.; and Sikdar, S. 2017. Identifying the social signals that drive online discussions: A case study of reddit communities. *CoRR* abs/1705.02673.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, 133–142. New York, NY, USA: Association for Computing Machinery.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Doha, Qatar: Association for Computational Linguistics.

Lakkaraju, H.; Mcauley, J.; and Leskovec, J. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013* 311–320.

Mazloom, M.; Hendriks, B.; and Worring, M. 2017. Multimodal context-aware recommender for post popularity prediction in social media. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Thematic Workshops '17, 236–244. New York, NY, USA: Association for Computing Machinery.

Medvedev, A. N.; Lambiotte, R.; and Delvenne, J.-C. 2017. The anatomy of reddit: An overview of academic research. *Dynamics on and of Complex Networks* 183–204.

Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; and Eisenstein, J. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1101–1111. New Orleans, Louisiana: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Piotrkowicz, A.; Dimitrova, V.; Otterbacher, J.; and Markert, K. 2017. Headlines matter: Using headlines to predict the popularity of news articles on twitter and facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 656–659.

Rubenking, B. 2019. Emotion, attitudes, norms and sources: Exploring sharing intent of disgusting online videos. *Computers in Human Behavior* 96:63–71.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.

Stoddard, G. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 416–425.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, 177–184.

Tan, C., and Lee, L. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. *CoRR* abs/1503.01180.

Tan, C.; Lee, L.; and Pang, B. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 175–185. Baltimore, Maryland: Association for Computational Linguistics.

Tellis, G.; Macinnis, D.; Tirunillai, S.; and Zhang, Y. 2019. What drives virality (sharing) of online digital content? the critical role of information, emotion, and brand prominence. *Journal of Marketing* 83:002224291984103.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 5998–6008.

Wang, A.; Hamilton, W. L.; and Leskovec, J. 2016. Learning linguistic descriptors of user roles in online communities. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 76–85. Austin, Texas: Association for Computational Linguistics.

Wang, C.; Ye, M.; and Huberman, B. A. 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, 244–252. New York, NY, USA: Association for Computing Machinery.

Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; and Xie, X. 2019a. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 3863–3869. International Joint Conferences on Artificial Intelligence Organization.

Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; and Xie, X. 2019b. Npa: Neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2576–2584. New York, NY, USA: Association for Computing Machinery.

Xie, Q.; Ma, X.; Dai, Z.; and Hovy, E. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 950–962. Vancouver, Canada: Association for Computational Linguistics.

Xu, Z. 2019. Personal stories matter: topic evolution and popularity among pro- and anti-vaccine online articles. *Journal of Computational Social Science* 2(2):207–220.

Yu, J., and Jiang, J. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Zayats, V., and Ostendorf, M. 2018. Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics* 6:121–132.

Zhang, W.; Wang, W.; Wang, J.; and Zha, H. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 1277–1286. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.