Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

RAFFMAN: Measuring and Analyzing Sentiment in Online Political Forum Discussions with an Application to the Trump Impeachment

Jakapun Tachaiya, Joobin Gharibshah, Kevin M. Esterling, Michalis Faloutsos
University of California, Riverside
{jtach001, jghar002, kevin.esterling}@ucr.edu, michalis@cs.ucr.edu

Abstract

Given an online forum, how can we quantify changes in user affect towards a person or an idea over time? We argue that online political forums constitute an untapped opportunity for understanding sentiment toward aspects under discussion. However, the analysis of such forums has received little attention from the research community. In this paper, we develop RAFFMAN, a systematic approach to quantify the impact of external events on the affect of forum users towards a concept, such as a person or an entity. First, we develop an approach to capture and quantify the observed activity: we identify related keywords, filter threads, and establish correlations between events and spikes in the activity. Second, we modify and evaluate state-of-the-art NLP techniques to achieve high accuracy (74%) in a three-class sentiment classification problem. As a case study, we deploy our method to quantify the effect of President Trump's impeachment on several concepts including: President Trump, Speaker Pelosi, and QAnon. Our data consists of 32M posts from Reddit and 4chan over a span of 6 months from September 2019 to February 2020. This initial analysis hints at an increase in political negativity, especially for people's affect towards the President. Overall, our work is a building block towards mining the affect of online forum user towards a concept, which constitutes a untapped, massive, and publicly-available source of information.

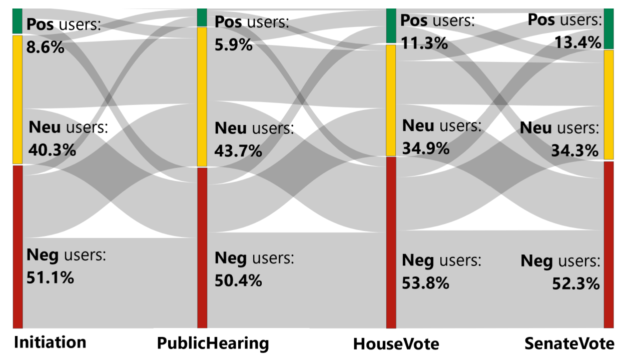


Figure 1: Our systematic approach in action: The evolution of users' affect on Reddit towards concept "Trump" on the significant dates (which coincide with changepoints) during the impeachment process. We observe that: (a) more than 50% of the users express negative affect towards Trump, and (b) the impeachment seems to have increased polarization toward concept "Trump" as neutral users decreased from 40.3% to 34.3%. Upon further investigation, we find that 12.1% of users flip-flop from negative to positive or vice versa.

Introduction

How can we assess the emotional affect toward the impeachment of Donald Trump among the users of an online discussion forums?

The more general question is how we can detect the evolution of the affect or **sentiment** of users in an online forum towards a **concept** (a person or an idea) in response to external real-world **events**. A major political event such as an impeachment is a concept that can evoke emotional affect in users that can manifest in discussions within a forum as the official proceedings unfold. The challenge is that a complex political event such as Trump's impeachment has many different concepts or aspects that can occur within a discussion, and the discrete political stages that unfold over time can change individuals' emotional affect toward those aspects.

Emotional "affect" is central to the field of political and social psychology. Human cognition connects emotional affect labels to objects such as concepts or ideas (Lodge and Tabor 2005). The affect label itself is *not necessarily a preference or opinion* regarding the object; even a supporter of an idea could express frustration or disappointment in series of posts. In the social sciences, however, the study of public opinion is not limited only to stance or preferences. Instead, these affect labels, in the form of positive and negative emotions, impact individuals' cognition and beliefs that are relevant to that object (Marcus, MacKuen, and Neuman 2011). For example, negative affect toward an object can lead individuals to more closely attend to threatening information (Brader 2005).

When users engage in online discussion forums, their messages contain latent affect or sentiment that we are able to connect to specific concepts or ideas. Because of the relationship between affect and information processing, having a flexible method to assess affect will be important to researchers who wish to understand how users learn about political information on social media and in online discussion forums (Lazer et al. 2018). We should clarify that the term "affect" can include the full range of human emotions, such as fear, anxiety, happiness, etc. Sentiment maps these emotions into **three categories**: positive, neutral and negative. Because of this mapping, we use "sentiment" and "affect" interchangeably with this relationship understood.

We examine the specific case of user engagement relevant to Trump's impeachment. While one can think of the impeachment as a singular event, in practice impeachment is a political process that is composed of different aspects, such as institutional settings, people and activities, as well as discrete stages, such as congressional committee hearings, votes, announcements, and the like. These discrete stages drive user engagement and the substance of the discussion regarding the different aspects. Because these events unfold over time, we are able to observe which specific stages drive user engagement, and the affect that users express toward the aspects that are under discussion during these events.

Given the context above, the problem we address here is inherently complex and difficult. The input to the problem is a concept, an event or group of events, and online forum data, and the desired output is the nature of the users' emotional affect toward that concept and its change over time. We want to model and quantify: (a) the intensity of user engagement, which is caused by the event, and (b) the change of the users' affect polarity, that is, their sentiment, towards concepts.

The problem introduces several challenges. First, we need a keyword expansion method to capture the complex set of aspects related to the political process of interest. Second, we need a time series statistical method that will allow us to make inferences about the discrete events that drive user engagement. Third, we need to accurately recover sentiment or emotional affect that is connected to aspects under discussion during the discrete events. Fourth, we need to handle the unstructured nature of forum data, in that the posts can vary in lengths, threads become discussions and it is difficult to follow the discourse. Finally, we need to consider that user participation varies over time, and that some forums allow for anonymous users.

There has been limited work on the problem as framed here. We can group related work in two main streams. First, quantifying and measuring studies in forum are appeared in (Hine et al. 2017) which analyzes general properties and characteristics of 4chan. (Schild et al. 2020) detect changes related to real-world events. Second, sentiment analysis on the web is mostly used on social media data like Twitter (Wang et al. 2017). Aspect-based sentiment analysis, a finer task of sentiment analysis, is usually implemented in the domain of user product reviews (Wang et al. 2017). We revisit previous work in the related section at the end.

Contribution: We propose, RAFFMAN[1], a systematic approach to measure the change in users' affect towards a complex concept in response to real-world events in online discussion forums. Our approach consists of the following key components: (a) filtering, (b) detection of change, and (c) sentiment analysis. We adapt, customize and synthesize state-of-the art methods to optimize aspect-based sentiment analysis using the unstructured data of political discussion forums, including (a) developing a keyword expansion method that can identify and filter for different aspects of a complex event, (b) adapt a time series statistical method to identify important discrete stages that compose the complex event, and (c) optimizing aspect-based sentiment analysis to political discussion. We show that our approach yields up to 74% of accuracy in our three-class classification (positive-neutral-negative). The classification accuracy increases to 81.1% if we only examine short posts with less than 23 words.

We validate and showcase our approach using data from the online discussion forums 4chan and Reddit, where users provide an untapped wealth of information on people's thoughts and sentiments in response to current events: there are 1,000 and 5,000 new posts per minute respectively to 4chan and Reddit. Since we investigate the effect of President Trump's impeachment, we focus on politically-oriented sub-forums. In the remainder of this paper, we will use the terms Reddit and 4chan to refer exclusively to Reddit's r/politics, and 4chan's /pol sub-forums. We collect 32M posts that occurred during the impeachment process between September 2019 to February 2020. We provide an overview of the key results below:

- **The user engagement doubles at significant stages.** We found that, during significant events such as the "House vote" for 4chan and the "committee public hearing" for Reddit, the number of posts related to the topic doubled as shown in Fig. 2. This indirectly increases our confidence for our keyword selection and thread filtering methods.
- **Reddit users are more engaged with impeachment compared to those of 4chan.** The percentage of impeachment related posts on Reddit (51.8%) is much higher than that on 4chan (13.3%). This suggests that Reddit users were more concerned about the impeachment as an event. For example, during the "House vote" event, 95% of posts in Reddit engaged to the topic compared to only 38% in 4chan.
- **More than half of all posts have negative affect toward the concept "Trump" throughout the impeachment.** We find that more than 50% of all posts exhibit negative affect toward aspect "Trump" in both Reddit and 4chan. This is true for the majority of the the 6-month impeachment period, namely for 83% of the days for Reddit and 98% of the days for 4chan.
- **The impeachment events increased the divergence of user affect for concepts "Trump," "Impeachment" and "Pelosi."** We find that around 6% of neutral users change their affect to either negative or positive on the key aspects "Trump" and "Impeachment" in Reddit. We also observe a similar divergence for "Pelosi," the Speaker of the House and a key figure in the impeachment saga.
- **The impeachment increased the negative affect towards the concept "Pelosi."** We find a 7.9% increase in users with negative affect and 6.6% decrease in users with positive affect between the two events "House vote" and

[1]Acronym not explained here for anonymity purposes.

"Senate vote" in Reddit.

There is a possibility that individuals or groups can engage in social platforms with an agenda to promote an idea or even to spread confusion and animosity. Although we did not address this issue here, we discuss it in our Discussion section.

Our work can be seen as a building block to harness the untapped potential of online discussion forums. We argue that effective ways to analyze such forums can provide valuable information on: (a) what resonates with people, as can be seen by increases in the engagement, and (b) how people feel about events and prominent people. Detecting deliberate misuse and social engineering is an important next step in this line of work.

Background and Datasets

Our work focuses on online discussion forums. We have collected data from two forums, Reddit and 4chan, over a span of six months during the impeachment period between September 2019 and February 2020. We discuss and present our datasets with their basic statistics in Table 1.

1. Reddit. We use Reddit a well-known text-based discussion forum with eponymous users. We select the "politics" subreddit (/politics/) because it is directly related to our main focus. The "politics" subreddit contains a large pool of 6.5M registered users with roughly 100k daily posts. The users that select into these forums, along with their posts, can function as a convenience sample for social science research (Coppock 2019) and serve as an interesting population of direct interest regarding online engagement. To collect this subreddit data, we use the archiver service, Pushshift (pushshift.io), that collects every post made in the main Reddit site and makes that data publicly available for academic purposes.

2. 4chan. We use 4chan, which is considered a fringe alt-right forum, as an interesting contrast to Reddit. On 4chan, users do not need to create an account to use the platform. As a result, most users remain anonymous while posting comments in the forum. We focus on the "politically incorrect" subforum (/pol/). This is the most active subforum in 4chan with an average of 150k daily posts as reported by 4stats.io. 4chan does not make their data publicly available and it routinely deletes data in the forum. Here, we collect data from a community-run archiver 4plebs (4plebs.org), which crawls and archives all the activity from 4chan and makes it publicly available.

3. Ground-truth for aspect-based sentiment analysis. We have access to a gold standard benchmark data set for aspect-based sentiment analysis (ASBA) obtained from the NLP workshop SemEval (Semantic Evaluation); however this benchmark data is mostly in the domain of restaurant reviews or laptop reviews which is not matched to our task in political discussion. To remedy that drawback, we create our own benchmark dataset using the existing posts in both Reddit and 4chan. We use two groups of annotators (a) five general annotators from Amazon's Mechanical Turk platform and (b) three political unbiased experts in the scientific field. The annotators labeled each post with sentiment toward a given aspect. The final label is produced by using a two-round majority vote approach from (a) and (b) to get a balanced and unbiased training set shown in Table 2. We assess our annotated data by using the Fleiss-Kappa coefficient on the two groups of annotators in Table 3. We observe the highest agreement in all aspects from experts. These results showcase the benefit of using politically unbiased experts in the ABSA annotation tasks.

Forum	Posts	Threads	Users
Reddit - politics (total)	10.5M	149K	509K
Reddit - politics (filtered)	5.4M	62K	392K
4chan - pol (total)	16.9M	411K	-
4chan - pol (filtered)	2.2M	38K	-

Table 1: Our datasets from Reddit and 4chan over a span of six months from September 2019 to February 2020. The term filtered refers to posts and threads that we identify as related to the event the impeachment of President Trump in Step 1 of our approach. Anonymity in 4chan prevents us from having the number of unique users.

Aspect	Negative		Neutral		Positive	
	Train	Test	Train	Test	Train	Test
Trump	412	102	406	102	413	103
Impeachment	184	46	183	46	184	46

Table 2: Our ground-truth dataset with more than 2K posts for concepts "Trump" and "Impeachment" using: (a) Mturkers, and (b) experts.

Aspect	Mturk	Experts
Trump	0.453	0.583
Impeachment	0.372	0.691
All	0.433	0.601

Table 3: Assessing the annotator agreement using the Fleiss-Kappa coefficient on ground-truth for aspect-based sentiment analysis.

4. Concepts, events, aspects, and keywords. Our goal is to study the effect of an event on user sentiment toward a concept. We use the term concept or aspect to refer to a person or an idea, and we can use a set of keywords to describe that concept. For example, "Trump" as an aspect can be referred to with keywords such as "Trump," "Potus," "Donald," . . . etc. We explore the following aspects in this paper:

- "Trump" : Donald Trump is the 45th president of the United States from the Republican party.
- "Impeachment" : Impeachment is a U.S. constitutional process to remove government official from the office.
- "Pelosi" : Nancy Peslosi is the Speaker of the House, a leading figure of the opposition Democrat party.
- "QAnon" : QAnon or Q is a far-right conspiracy theory.
- "Goodell" : Roger Goodell is the current American football league Commissioner (an aspect that should not be related to impeachment that we use below for a placebo test of our methods).

Simply put, an event is also a concept that can be "defined" by a set a keywords. As always, the scenarios can be more complex in practice. For example, the impeachment of Trump is a complex process that is composed of discrete political stages, where each stage can can span multiple days. The *New York Times* lists the following as major stages for Trump's impeachment:

- "Initiation" : Sep 24 2019, the Speaker of the House announced a formal impeachment inquiry.
- "Articles of Impeachment" : Dec 11-13 2019, Committee voted to approve two articles of impeachment.
- "House vote" : Dec 18 2019, House passed the two articles of impeachment.
- "Senate trial" : Jan 29-31 2020, Senators questioned and rejected for any new witnesses or documents
- "Senate vote" : Feb 5 2020, Senate rejected both articles of impeachment against Trump.

Word	Similarity	Word	Similarity
impeachment	1.000	dismiss	0.405
trump	1.000	contempt	0.403
censure	0.568	inquiry	0.403
bush	0.482	prosecute	0.387
trial	0.475	speaker	0.386
judiciary	0.454	resort	0.385
acquit	0.443	remove	0.381
perjury	0.437	cloture	0.380
resolution	0.430	evidence	0.373
witness	0.428	constitution	0.364

Table 4: Top 20 similar words acquired from initial event-keywords, "Trump" & "Impeachment" with keyword expansion techniques trained with data from Wikipedia. The higher the score the greater the similarities between that word to an initial keyword set.

Overview of Our Approach

Our approach provides a method to systematically quantify sentiment in online forums consisting of three major steps, which we outline below.

Step 1: Identifying Related Activity

Given a small set of keywords that are known to be relevant to an event of interest, we want to capture related activities in a forum without requiring specific domain knowledge. This step consists of (a) expanding a set of initial keywords, and (b) identifying related posts and threads in the forum.

a. Keyword expansion. We utilize an iterative embedding-based approach to expand a set of initial keywords. The key design elements of this approach are as follows: a) We use two similarity expansions, one in the word-word space and one in the post-post space, (b) we use an iterative approach in each of these expansions, and (c) we provide a flexible ranking of the identified words to meet the user needs. Specifically, in order to implement the keyword expansion step, we take following phases:

Phase 1: Domain representation. We represent words and posts of forums in an m-dimensional embedding space with the Word2Vec method (Mikolov et al. 2013).

Phase 2: Word-space expansion. We expand the initial set of keywords by adding relevant words iteratively.

Phase 3: Post-space expansion. We identify posts that are similar to the set of posts that contain the relevant words from the previous step.

Phase 4: Result Processing. We extract and rank the keywords from the posts of the previous step, based on several metrics like word-word similarity, post-post similarity and TF-IDF which is based on importance and relevancy. *Similarity score* is calculated by the average of cosine similarity in a Word2Vec embedding space (Mikolov et al. 2013) between the initial set and the expended set. Then, a subset of ranked keywords that passes a threshold of similarity will represent an expansion set. This threshold varies depending on the task of interest.

We implement our keyword expansion techniques on the initial event-keywords known to be related to Trump's impeachment, namely "Trump" and "Impeachment," on Wikipedia pages that contain those words. We selected Wikipedia to expand the event-keywords set because it is external to our forums and so prevents bias in the event-keywords expanded set that could occur if we used our forums' specific posts. The results of our keyword expansion technique are shown in Table 4.

b. Identifying related threads and posts. A key step in our approach is to identify the threads that relate to the event and concepts of interest, which we achieve as follows.

i. Identifying related posts. We label a post as related if it contains keywords in any part-of-speech obtained from the previous step. In our experiment, we select only keywords with *similarity score* more than 0.4, which yields 13 unique keywords. We discuss the selection of this value below.

ii. Identifying related threads. We label a thread as related if the title of the thread contains selected keywords or the percentage of related posts are more than the *post-relevance threshold*. In our case, we use 30% as a threshold which we justify below.

Threshold selection. We set the value of our two thresholds, 0.4 similarity and 30% post percentage, using the elbow method (Ketchen and Shook 1996) by comparing the quantity of related posts and threads obtained with different parameter settings. We identify overall related posts and threads shown in Table1, which are (51.8%, 41.6%) and (13.3%, 9.4%) of the total posts and threads in Reddit and 4chan, respectively. The parameters for similarity score and post percentage can be varied depending on the goal of the experiment task and one's preference in the trade-off between too much or too little inclusivity. Higher threshold values yield more posts and threads in exchange for possibly including more posts and threads that are unrelated to the concept and event of interest.

Step 2: Detecting Engagement Change

To identify real-world stages of Trump's impeachment event that impact a forum's engagement activity with respect to

Reddit		4chan	
Changepoint	Stage	Changepoint	Stage
1. 11/18/2019	11/18/2019-11/21/2019: Committee public hearings.	1. 12/19/2019	12/18/2019: House voted to pass the two articles of impeachment.
2. 12/19/2019	12/18/2019: House voted to pass the two articles of impeachment.	2. 01/03/2020	01/03/2020: Trump announced the death of Iranian general (unrelated to impeachment).
3. 02/05/2020	02/05/2020: Senate vote (acquitted).	3. 09/24/2019	09/24/2019: The initiation of impeachment.
4. 09/24/2019	09/24/2019: The initiation of impeachment.	4. 12/09/2019	12/04/2019-12/09/2019: Judiciary committee hearings.

Table 5: Correlated real-world stages ranked by changepoint algorithm (PELT) on a significant increase of #posts on the impeachment of Trump.

our concepts, we turn to changepoint algorithms that can detect significant changes in time series data. Specifically, we choose Pruned Exact Linear Time (PELT) (Killick et al. 2012), a parametric algorithm that can (a) detect changes and (b) rank them by maximizing its log-likelihood of mean and variance of the time series. In our case, we use a daily number of related posts containing our expanded set of keywords acquired from Step 1 as our time series data. We choose to model the number of posts rather than the number of threads or users because posting reflects the base activity of engagement in forums where users post in response to a topic of interest.

We apply the PELT algorithm to the daily number of related posts to get a list of dates ranked by significance. We then compare changepoints with the real-world events in our domain with a window of 1-2 days to accommodate asynchronous activity that occurs just after the event itself. Given (a) a daily number of related posts in forums and, (b) a list of real-world stages of the impeachment of Donald Trump obtained from the *New York Times*[2], we identify the most impactful real-world events, listed in Table 5. On Reddit, the "Committee public hearings" is the most statistically significant changepoint compared to the "House vote" on 4chan. Interestingly, we also find a non-related event to the "impeachment" in 4chan on January 03, 2020. This changepoint emerges from the increase in 4chan activities of the keyword "Trump," in response to the announcement by Trump himself of the assassination of Iranian general Qasem Soleimani.

As a robustness check, we specifically look into the impeachment concept where we only expand one initial keyword, "Impeachment," and identify the expanded impeachment keywords with Step 1. With this filtering, we find that "House vote" becomes the most statistically significant changepoint on both Reddit and 4chan. Also, the Soleimani changepoint now becomes non-significant on 4chan because it is not directly related to the impeachment event itself.

Validation of an expanded set of keywords. With the changepoint detection algorithm, Fig. 2 plots the number of related posts acquired from Step 1 with the most significant changepoints from Table 5. We see the correlation of an increase in an activity of engagement on the topic of interest with the significant changepoints on both forums. This verifies our filtering techniques and expanded set of keywords from Step 1.

[2]https://www.nytimes.com/interactive/2019/us/politics/what-is-impeachment-process.html

Step 3: Aspect-based Sentiment Analysis

To determine users' affect toward a concept, we use aspect-based sentiment analysis (ABSA), a subtask of sentiment analysis in the natural language processing field. While traditional sentiment analysis captures the overall sentiment in text, ABSA aims to detect the corresponding sentiment towards a specific aspect, which in our application are keywords. That is, ABSA can associate specific (negative, neutral and positive) sentiment with different aspects in the same post.

BERT (Devlin et al. 2018), a recent language model from Google, outperforms other traditional techniques like neural networks (Huang, Ou, and Carley 2018) in many NLP tasks including sentiment analysis because it has the ability to capture the context around words. While there are many variations of ABSA with BERT, we choose (Xu et al. 2019) as our implementation due to a simplicity while yielding reasonable accuracy when compared to a very complex model like (Rietzler et al. 2019).

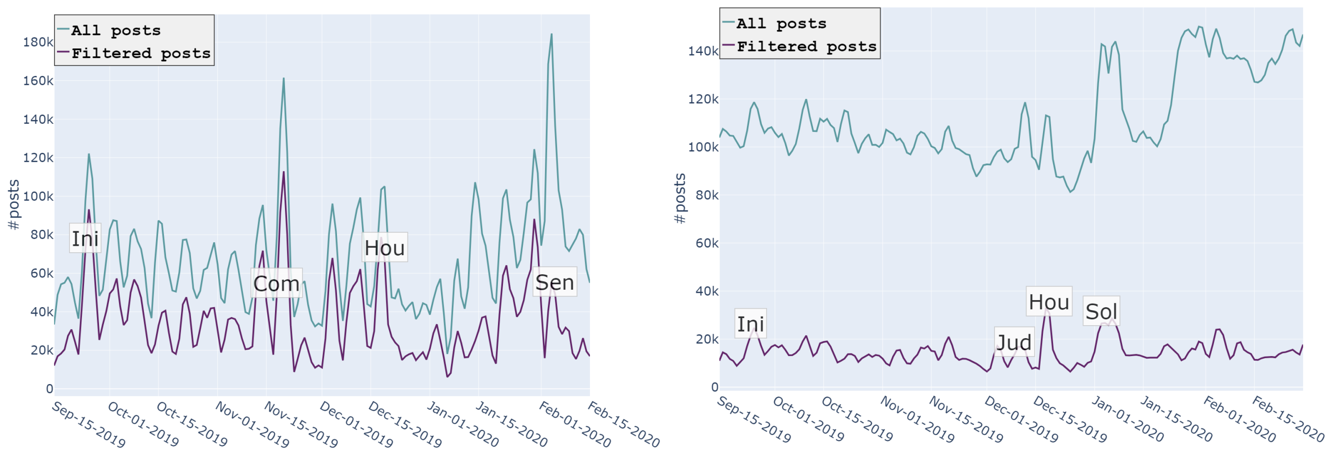
ABSA consists of two main subtasks: (a) detecting aspects in a sentence, and (b) determining a sentiment associated with an aspect.

a. Detecting aspects. To determine which word is an aspect in the sentence, ABSA employs IOB (Inside-outside-beginning), a common tagging techniques for an NLP task such as POS (part-of-speech) tagging (Toutanova et al. 2003) and NER (name-entities-recognition) tagging (Finke et al. 2005). However, since we focus on specific aspects such as "Trump" and "Impeachment," we use our expanded set of keywords, the 13 unique keywords from Step 1, as aspects in ABSA.

b. Determining sentiment associated with aspects. ABSA aims to classify a text with respect to a given aspect into the three different classes of polarity (negative, neutral and positive). BERT implements ABSA using a sequence-pair classification task. First, we transform our posts into tokens with a corresponding format. Let x represent BERT embedding sequences:

$$x = [CLS]a_1, ..., a_m[SEP]t_1, ..., t_n[SEP] \qquad (1)$$

where $a_1, ..., a_m$ are tokens of an aspect, $t_1, ..., t_n$ are tokens of words in a post, $[SEP]$ is separation token, and $[CLS]$



(a) Reddit

(b) 4chan

Figure 2: The temporal view of the posts in forums over a span of 6 months. These figures are labeled with significant change-points that correlated to the real-world stages in Table 5. The purple line represent the amount of related posts filtered with our approach from Step 1. The green line is the total number of posts made daily in each forum. We observe that: (a) user engagement doubles at significant stages (b) Reddit users are more engaged with the impeachment than 4chan users.

is a special token that can represent the whole embedding sequence. Second, we feed these embedding sequences into the BERT model $h = BERT(x)$. Third, $h[CLS]$ that represents the last hidden representation of embedding sequence is an input to a softmax layer for a sequence-pair classification task which generate the probability of each sentiment's polarities which we show with p.

$$p = softmax(W \cdot h[CLS] + b) \tag{2}$$

where $W \in R^{3\times768}$ (weights of our embedding sequence for each polarity on BERT), $b \in R^3$, $p \in [0,1]^3$, 3 is the number of polarities (negative, neutral and positive), 768 is a default length of embedding sequence on BERT. Finally, $argmax(p)$ returns the classification result.

To maximize our ABSA task, we experiment with different language models using the same testbed:

- **NLTK+VADER**: traditional rule-based sentiment analysis that captures the overall sentiment of a post. Stopword removal is performed during the prepossessing of a post.
- **BERT-baseline**: Original pre-trained model and fine-tuned with our ABSA dataset.
- **BERT-custom**: Post-trained model with review data from Yelp and Amazon reviews and fine-tuned with our ABSA dataset (Devlin et al. 2018).
- **XLNet**: A larger language model that claims a better performance over state-of-the-art BERT (Yang et al. 2019).

All models (except the NLTK+VADER model) are pre-trained model which we fine tune them for our own specific task. We evaluated their performance with 5 fold cross validation. The results of our classification are shown in Table 6. BERT-custom is able to achieve a competitive result with a larger model like XLNet because training from the reviews data transfers to our political forum dataset. BERT takes less time in this classification task and yields higher accuracy on short posts than XLNet, so we choose BERT-custom to associate sentiment with aspects.

Model	All posts		Short posts	
	Accurcy	F1	Accurcy	F1
NLTK + Vader	51.1%	50.0%	56.7%	54.4%
XLNet	74.6%	74.3%	75.2%	75%
BERT baseline	70.7%	70.7%	75.9%	75.4%
BERT custom	74.3%	74.4%	**81.1%**	80.9%

Table 6: The summary of result of accuracy and F1 scores on aspect-based sentiment classification with our political dataset where short posts contain less than 23 words. For NLTK+VADER, we use a traditional rule-based sentiment analysis model to determine overall sentiment of a post.

Shorter post lead to higher (81.1%) classification accuracy. We investigate if the length of the post affects the accuracy of our ABSA model with BERT-custom performs. We compare between 234 short posts that contain less than 22 words, which is at 50 percentile of post length distribution, and 242 long posts that have more or equal than 23 words. Our BERT-custom model achieves 81.1% of classification accuracy on the set of short posts compared to 67.6% on the set of long posts. We conjecture that the longer posts may provide the user the ability to ramble and even mix discussion topics, which could affect the classification accuracy. Upon manual investigation, one post in Reddit not only uses many cursed words regarding "Trump," doubts the so called fair "Trial" procedure but it is also strongly in favor of "Impeachment." This shows longer posts introduce and mix several arguments and discussions and even appear self-contradicting at times.

Determining user's affect from her posts. We now want to identify the sentiment of a user towards a concept based on posts at given time. Intuitively, the process will aggregate the sentiment of the posts of that user during an appropriate interval of observation. Following our event-driven approach, we introduce the *temporal focus* parameter, windowInterval , to be the number of days around a given date, during which we collect the user posts for that concept. We then use a majority vote on the sentiment of these posts toward an aspect to determine the user's affect on that time period. Note that, if there is a tie between any two categories, we assign the user's affect as neutral. Finally, to increase our confidence in the outcome, we can require a minimum number of posts that a user has to have in that time period in order to be included in the report.

Case Study: The Impeachment

In this section, we study the dynamics in user affect related to the impeachment of President Trump, which consists of several related stages, as a case study on the political subforums of Reddit and 4chan. To recap, our case study involves concepts (a) "Trump," (b) "Impeachment," (c) "Pelosi" and (d) "Qanon" captured by 13 event-keywords shown in Table 4. We also investigate an unrelated aspect, "Goodell" representing Roger Goodell, the current NFL commissioner, as a placebo test of our methods that can help to show our analysis is indeed specific to the political impeachment process. We expect to see significant changepoints in user engagement in the first four impeachment aspects but not in the Goodell aspect.

Identifying Engagement Changes

a. User engagement correlates with impeachment related events. Our changepoint detection algorithm of user engagement identifies spikes that coincide with key stages of the impeachment process as identified by the *New York Times*. This observation acts as an indirect validation that our keyword selection and thread filtering follow the user activity adequately. Interestingly, in 4chan, we do not observe significant activity change on the stages "Articles of impeachment" and "Senate trial."

b. User engagement doubles in reaction to impeachment-related events. We find that there is a 218.7% increase in impeachment-related posts to "Committee public hearing" in Reddit, which is shown in Fig. 2a. We also see a 186.7% increase of such posts during the "House vote" in 4chan, as shown in Fig. 2b. The increase in the most significant changepoint in both forums is calculated by comparing the average of a two-day window before the change point and the peak of the changepoint itself.

c. Reddit users are more engaged regarding impeachment compared to those of 4chan. The trends in Fig. 2 show that posts made in Reddit are two times more likely to be a post about the impeachment and Trump compared posts made in 4chan. This observation shows that Reddit users were more engaged over the impeachment of Trump. Furthermore, the impeachment related posts dominate even more during significant external stages. This again is more pronounced for Reddit. For example, we observe the highest percentage of the related posts to the total of all daily posts at "House vote" for both Reddit and 4chan, equal to 95% and 38% of posts respectively.

This contrast in engagement coincides with the different lifespan of threads in Reddit and 4chan. 4chan has significantly shorter lifespan of threads than Reddit, since it is known to regularly delete their threads due to its infringed nature. Most threads live up to 3.9 minutes (median) and the fastest thread to expire was around 28 seconds (Bernstein et al. 2011). With those properties, users are less likely to follow conversations in specific threads and post which make a topic of interest diverse.

Identifying User Affect Changes

We assess how user affect towards a concept evolves over time in response to external events.

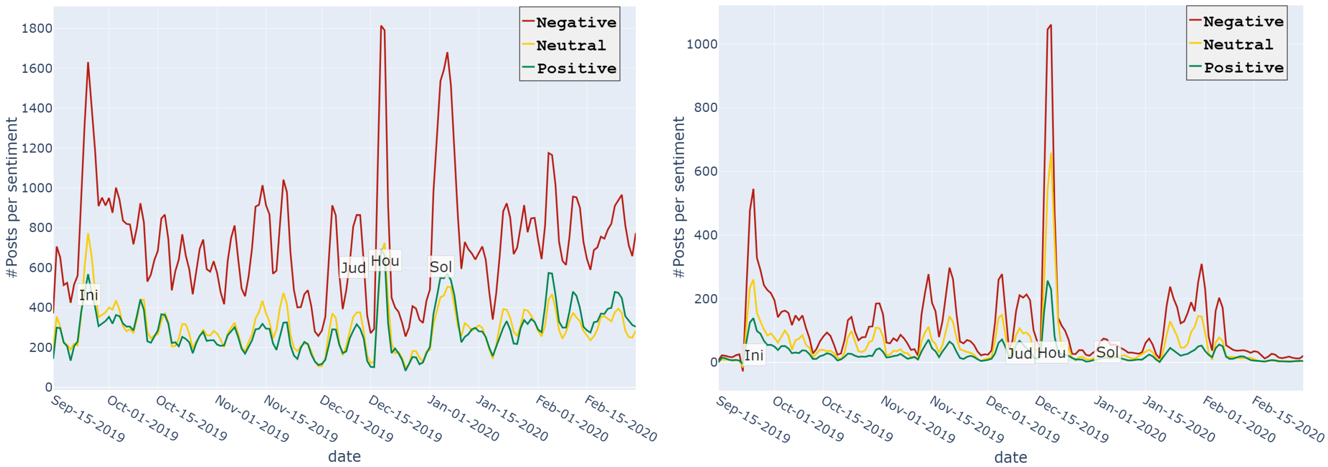
a. More than half of all posts are negative toward the aspect "Trump." We find that more than 50% of all related posts are negative in Reddit and 4chan on the aspect "Trump." We observe this trend at 83% and 98.8% of the days in the 6-month impeachment period, with respect to Reddit and 4chan. The peak of negative posts on "Trump" reaches 61% of all related posts on 4chan on January 03 where he announced the death of Soleimani, which is an event that we identify in our changepoint analysis, but is only tangentially related to impeachment (that is, a newsworthy event that possibly diverted public attention away from the impeachment proceedings). Given the lesser coherence of discourse found in 4chan, it is not surprising that our changepoint method appears to work better in Reddit.

Interesting, during February 5, 2020 in 4chan, positive posts (26.4%) outpace neutral posts (19.9%) on "Trump," a results shown in Fig. 3a, a result we do not observe in Reddit. This change corresponds to the "Senate Vote" to acquit all articles of the impeachment on Donald Trump. This is no surprise for 4chan, the forum known for alt-right and supporting Trump.

b. The impeachment increased the polarization of user affect for the aspects "Trump," "Impeachment" and "Pelosi." Polarization (Levendusky 2013) of a user's affect occurs when users tend to change their sentiment from neutral to become either more positive or more negative. We find that neutral affect among users is decreased by 6% on "Trump" and 6.4% on "Impeachment" when compared to the start of impeachment process, "Initiation," and the end of the process "Senate vote." We also observe a similar polarization for "Pelosi," the Speaker of the House, and a vocal critic of President Trump as the neutral users are also decreased by 1.2% at "House vote" and "Senate vote," two events where we have enough users' affect on "Pelosi" to draw a conclusion.

Although most users are anonymous in 4chan, we try to gauge polarization by comparing the percentage of posts per sentiment expressed during the same interval above. We find that there is a 3.8% decrease in neutral posts on "Trump," an 8.9% decrease on "Impeachment" and a 10.5% decrease on "Pelosi."

c. The impeachment process increased negativity towards the aspect "Pelosi." In Reddit, we find a 7.9% increase in users with negative affect and a 6.6% decrease in



(a) “Trump” concept. (b) “Impeachment” concept.

Figure 3: The temporal view of the number of posts per sentiment acquired from our ABSA BERT-custom model on 4chan in a span of 6 months. These figure are labeled with significant changepoints that correspond with the real-world stages in Table 5. We observe that: (a) 50% posts about Trump are negative at 98.8% of the days (b) find a spike in posts only on “Trump” at (2) Jan 03 2020, the death of Soleimani.

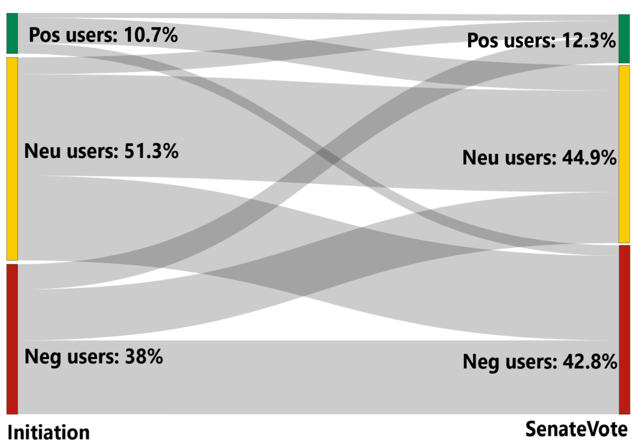


Figure 4: The evolution of users’ affect towards concept “Impeachment” between events: “Initiation” and “Senate vote” on Reddit. We observe that: (a) 12.8% of users flip-flop from negative to positive or vice versa, and (b) a increase in polarization as neutral decreased from 51.3% to 44.9%.

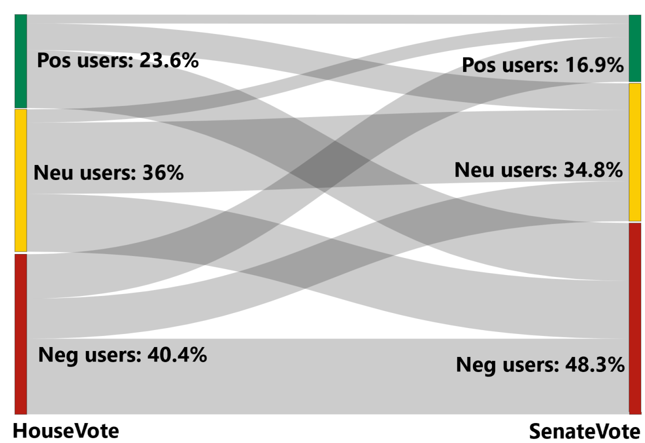


Figure 5: The evolution of users’ affect on Reddit toward an concept ”Pelosi” in the significant dates “House vote” and “Senate Vote” in the impeachment process. We observe that: (a) 25.8% of users flip-flop between negative to positive, and (b) the impeachment increased polarization slightly towards concept “Pelosi” as neutral decreased from 36% to 34.8% while the negative affect increased by roughly 8%.

users with positive affect toward Pelosi between the two significant events, “House vote” and “Senate vote” where we have enough number of users for conclusive analysis. This shows how user’s affect toward “Pelosi” develop in response to the impeachment.

d. Concepts “QAnon” and “Goodell” and placebo test. QAnon or Q is a far-right conspiracy theory that was created online by a user with the name Q. This theory claims that “a cabal of Satan-worshiping pedophiles running a global child sex-trafficking ring is plotting against President Donald Trump, who is battling them” according to Wikipedia. Reporting suggests that many Trump supporters and the president himself are sympathetic to this idea. We observe some “QAnon” engagement and changepoints during the impeachment period, but they are not aligned with the impeachment events on either Reddit or 4chan.

As a placebo test, we also consider the concept “Goodell,” which represents Roger Goodell, the NFL Commissioner. We include this aspect as a placebo test to ensure that our methods not only show relevant engagement, but they also do not pick up on irrelevant engagement. We observe that

Goodell's appearance on the forum is limited and also his name does not show any increased engagement aligned with the impeachment events.

Discussion

Here, we discuss the scope and limitations of our work. **a. Emotions, Sentiment and Stance.** RAFFMAN is designed to detect sentiment in discussions about specific aspects that compose discrete stages of an event. Sentiment is of significant emerging interest to social scientists. The use of automated sentiment analysis is only just beginning to emerge in social science; for example (Adams-Cohen 2020). Sentiment is closely related to emotion, and the study of emotion is a vast field in the social sciences. It has long been established that emotions are central to cognition and the information processing that informs individuals' preferences. For example, (Marcus 2000) is a highly cited review of the field of emotions and politics; (Lynggaard 2019) is a more recent overview of the methodological considerations in quantifying emotions and the impact of emotions; (Gross 2008) and (Brader 2005) are highly cited applications demonstrating the role of emotions in persuasion research; and (Brader, Valentino, and Suhay 2008) is a highly cited application investigating the role of emotions in information processing. We argue that RAFFMAN provides a set of tools that would help to advance the study of the role of sentiment in information processing and opinion formation on events discussed in social media settings, in parallel to emotions in the social sciences, and so will make a strong contribution.

Stance detection differs from sentiment analysis in that it is an NLP task to infer the preferences of individuals in favor, against or neither towards the aspect. Some (Mohammad, Sobhani, and Kiritchenko 2016; AlDayel and Magdy 2019) have studied the relation between stance and sentiment, showing sentiment by itself is not enough to detect a person's stance; however, sentiment can be used as one of the important features to detect a stance.

b. Dataset size and Google's BERT. BERT (Devlin et al. 2018) is a state of the art language model that we use here. BERT has transfer learning capability and has proven to be very effective in providing good accuracy with fewer labeled datasets in many classification tasks. A recent stufy (Romero et al. 2019) shows that BERT works well in an image classification task with around 1,000 labeled datasets. Another project[3] uses only 500 labeled datasets to do sentiment analysis on IMDB movie reviews with BERT and was able to yield 83% classification accuracy. These sutdies show that our 2,000 labeled dataset is ample enough to be used on the ABSA task with BERT transfer learning.

c. Who could use our tool in practice, and how? RAFFMAN is a powerful tool to gauge user affect towards any concept that users discuss online. Sentiment is of interest to itself to social scientists, and in addition our methods could be adapted to the study of individual engagement and information processing in online settings. The additional power lies in that: a) it uses organically derived responses, reflecting genuine engagement, b) it collects opinions *in vivo*, namely at the time that different events take place, and c) it can identify the sentiment evolution at the level of individuals, and not simply in the aggregate. The latter, of course, requires that the forum uses permanent user-names, like Reddit.

From a practical point of view, a user can specify: (a) the forum, (b) a time interval, (c) the concept as a group of keywords, and (c) the events of interest as a group of keywords. The outcome could be a plot as shown in Fig. 1: the identified engagement spikes, and a the evolution of the user sentiment between several time-points of interest.

d. The potential impact of our tool. The value and impact of our tool could be quite broad. The potential users could span a wide range: (a) politicians and political advisors, (b) policy makers, (c) marketing firms, and (d) social science researchers. We think that the last group could derive immediate and significant benefits (as enthusiastically argued by the political scientist in our team). In particular, RAFFMAN will enable social scientists to test hypotheses about the role of sentiment in opinion formation that occurs over time in response to social media engagement, which is a core interest in the fields of political psychology, communication and public opinion.

e. How can we detect and account for bots and manipulation? In this work we do not investigate social engineering or deliberate campaigns in our forums. Such misuse from individuals or foreign state actors has been observed in other social media and has sparked national debates. One could argue that this kind of behavior may be less prevalent in our two forums as they attract significantly fewer views compared to, say, Twitter. However, we have manually identified a few cases of such parasitic behavior. In Reddit, users named "GoldyTSA" and "OriginalWorldliness" exhibit a spamming behavior where they posted the exact posts, 43 and 47 times respectively, using a cursed word regarding Trump on the day of "House Vote."

In our future work, we intend to investigate such phenomena and: (a) develop techniques to identify misbehaving users and bots, and (b) quantify the effect of these behaviors on the forum discussions. In that effort, we will leverage the vast literature on detecting fake users and accounts and our own experience in identifying parasitic behavior in online commenting platforms, such as Disqus (disqus.com).

f. How representative is the data? This is the usual concern in every data-driven study. We argue that for the purposes of political discourse our data represents a reasonable case study to illustrate our methods. First, we use two different discussion forums with significantly different appeal and focus (Reddit and 4chan). Second, we consider a substantial amount of time (6 months) with a total of 32M posts during a significant event in U.S. political history. Naturally, for events of smaller magnitude, the intensity of the engagement will be lower, but our approach should provide accurate results.

Furthermore, although we focus on political events and discussions here, our method could apply more generally to other forums and other domains, such as discussions on

[3] https://blog.insightdatascience.com/using-transfer-learning-for-nlp-with-small-data-71e10baf99a6

sports, business, health, entertainment etc.

Related Work

We summarize related works in the following general areas.

Forum analysis. There are several studies trying to understand general activities in web-based discussion forums. (Hine et al. 2017), (Papasavva et al. 2020) and (Thukral et al. 2018) work on understanding properties, trends and characteristics of forums like ephemerality, heavy-tail and anonymity on posts, threads and users. Some focus on specific tasks in forums (Macdonald et al. 2015), (Munger et al. 2015) (Shrestha et al. 2019) and try to identify main actors like hacker users, depressed users and influential users using a variety of techniques including linguistics, behavioral modeling on user activities and graph-based approaches. The most relevant study to our paper (Schild et al. 2020) explores new emerging words and trends from the concept, "Covid-19," to see how the engagement and topics evolve over time, but they do not study the user sentiment towards these topics.

Sentiment analysis on the web. Most studies in sentiment analysis focus on the area of product reviews (Wang et al. 2017) or social media like tweets (Caetano, Almeida, and Marques 2018). Although (Park et al. 2016) and (Hsu, Hsu, and Tseng 2019) analyze forum data, they mostly use a base sentiment classification model to capture overall polarity on the text. Aspect-based sentiment analysis (ABSA) is rarely used on forums. Some of the relevant studies include an effort (Chakraborty, Goyal, and Mukherjee 2020), which applies ABSA using a neural network on reviews of scientific papers to quantify the reviewers' sentiment towards an aspect like originality, and a recent effort (Xu et al. 2019) that focuses on training the BERT approach for different domains using knowledge transfer.

Conclusion

The key contribution of our work is RAFFMAN, a systematic approach to quantify the change of forum user affect towards a concept in response to real events. Our approach consists of three phases: (a) identifying the related posts, (b) detecting changes in engagement, and (c) conducting sentiment analysis. These three components work synergistically to quantify user sentiment towards a concept is response to a complex event, which could consist of many sub-events. To quantify sentiment, we customize and synthesize state-of-the art methods to classify posts into three categories: positive, neutral and negative. We show that RAFFMAN achieves a classification accuracy of 81.1%, if we focus on posts with less than 23 words and up to 74% of accuracy with all posts.

In the future, our overarching goal is to make RAFFMAN an open-source platform that will catalyze and integrate the efforts of the research community. Our ambition is to see our approach as a step that will increase and enhance democracy by providing a direct way to hear people's voices especially in the political arena. To achieve this, we plan to expand our work in several different directions. First, we want to develop techniques to detect misuse, tampering, and social engineering, which may be present even in these forums. Second, we want collect more datasets and create more labelled benchmark datasets to use as ground-truth. Third, we want to develop a web-based platform with an easy to use interface for our approach to accelerate its adoption by end users.

References

Adams-Cohen, N. 2020. Policy Change and Public Opinion: Measuring Shifting Political Sentiment With Social Media Data. *American Politics Research* 48. doi:10.1177/1532673X20920263.

AlDayel, A.; and Magdy, W. 2019. Assessing Sentiment of the Expressed Stance on Social Media. *CoRR* abs/1908.03181. URL http://arxiv.org/abs/1908.03181.

Bernstein, M. S.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. G. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In Adamic, L. A.; Baeza-Yates, R.; and Counts, S., eds., *ICWSM, 2011*. The AAAI Press. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2873.

Brader, T. 2005. Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions. *American Journal of Political Science* 49(2): 388. doi:10.2307/3647684.

Brader, T.; Valentino, N.; and Suhay, E. 2008. What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat. *American Journal of Political Science* 52: 959 – 978. doi:10.1111/j.1540-5907.2008.00353.x.

Caetano; Almeida; and Marques. 2018. Characterizing Politically Engaged Users' Behavior During the 2016 US Presidential Campaign. In *2018 IEEE/ACM (ASONAM)*.

Chakraborty, S.; Goyal, P.; and Mukherjee, A. 2020. Aspect-based Sentiment Analysis of Scientific Reviews. *ACM/IEEE 2020* doi:10.1145/3383583.3398541.

Coppock, A. 2019. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. doi:10.1017/psrm.2018.10.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805. URL http://arxiv.org/abs/1810.04805.

Finke; Rose; Grenager; Trond; Manning; and Christopher. 2005. Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. ACL '05.

Gross, K. 2008. Framing Persuasive Appeals: Episodic and Thematic Framing, Emotional Response, and Policy Opinion. *Political Psychology* 29: 169 – 192. doi:10.1111/j.1467-9221.2008.00622.x.

Hine, G. E.; Onaolapo, J.; Cristofaro, E. D.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. URL https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670.

Hsu; Hsu; and Tseng. 2019. Discovering negative comments by sentiment analysis on web forum. *World Wide Web* 22(3): 1297–1311. doi:10.1007/s11280-018-0561-6. URL https://doi.org/10.1007/s11280-018-0561-6.

Huang; Ou; and Carley. 2018. Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks. *CoRR* abs/1804.06536. URL http://arxiv.org/abs/1804.06536.

Ketchen, D. J.; and Shook, C. L. 1996. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal* 17(6): 441–458. ISSN 01432095, 10970266. URL http://www.jstor.org/stable/2486927.

Killick; Rebecca; Fearnhead; Paul; and Eckley, I. 2012. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association* 107: 1590–1598. doi:10.1080/01621459.2012.737745.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The science of fake news. *Science* 359(6380): 1094–1096. ISSN 0036-8075. doi:10.1126/science.aao2998. URL https://science.sciencemag.org/content/359/6380/1094.

Levendusky, M. S. 2013. Why do partisan media polarize viewers? *American Journal of Political Science* 57(3): 611–623. ISSN 15405907. doi:10.1111/ajps.12008. URL https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12008.

Lodge, M.; and Tabor, C. 2005. The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis. *Political Psychology* 26(3): 455–482.

Lynggaard, K. 2019. Methodological Challenges in the Study of Emotions in Politics and How to Deal With Them. *Political Psychology* 40. doi:10.1111/pops.12636.

Macdonald, M.; Frank, R.; Mei, J.; and Monk, B. 2015. Identifying digital threats in a hacker web forum. In *2015 IEEE/ACM International Conference (ASONAM)*, 926–933.

Marcus, G. 2000. Emotions in Politics. *Annual Review of Political Science - ANNU REV POLIT SCI* 3: 221–250. doi:10.1146/annurev.polisci.3.1.221.

Marcus, G. E.; MacKuen, M.; and Neuman, W. R. 2011. Parsimony and Complexity: Developing and Testing Theories of Affective Intelligence. *Political Psychology* 32(2): 323–336. ISSN 0162895X. doi:10.1111/j.1467-9221.2010.00806.x. URL http://doi.wiley.com/10.1111/j.1467-9221.2010.00806.x.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781.

Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S. 2016. Stance and Sentiment in Tweets. *CoRR* abs/1605.01655. URL http://arxiv.org/abs/1605.01655.

Munger; Tyler; Zhao; and Jiabin. 2015. Identifying Influential Users in On-Line Support Forums Using Topical Expertise and Social Network Analysis. ISBN 9781450338547.

Papasavva, A.; Zannettou, S.; Cristofaro, E. D.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board.

Park; Beck; Fletche; Lam; and Tsang. 2016. Temporal analysis of radical dark web forum users. In *ASONAM, 2016*.

Rietzler, A.; Stabinger, S.; Opitz, P.; and Engl, S. 2019. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. *arXiv preprint arXiv:1908.11860* .

Romero, M.; Interian, Y.; Solberg, T. D.; and Valdes, G. 2019. Training Deep Learning models with small datasets. *CoRR* abs/1912.06761. URL http://arxiv.org/abs/1912.06761.

Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2020. "Go eat a bat, Chang!": An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. *CoRR* abs/2004.04046. URL https://arxiv.org/abs/2004.04046.

Shrestha; Anu; Spezzano; and Francesca. 2019. Detecting Depressed Users in Online Forums. ISBN 9781450368681.

Thukral, S.; Meisheri, H.; Kataria, T.; Agarwal, A.; Verma, I.; Chatterjee, A.; and Dey, L. 2018. Analyzing Behavioral Trends in Community Driven Discussion Platforms Like Reddit. In *(ASONAM) 2018*, 662–669.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. NAACL '03, 173–180. USA: Association for Computational Linguistics. doi:10.3115/1073445.1073478. URL https://doi.org/10.3115/1073445.1073478.

Wang, C.; Day, M.; Chen, C.; and Liou, J. 2017. Temporal and Sentimental Analysis of A Real Case of Fake Reviews in Taiwan. In *(ASONAM) 2017*, 729–736.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. S. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In *ACL 2019*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR* abs/1906.08237. URL http://arxiv.org/abs/1906.08237.