Effective Messaging on Social Media: What Makes Online Content Go Viral?

Maryam Mousavi
Hasan Davulcu
Mohsen Ahmadi
Arizona State University
maryammousavi@asu.edu
hdavulcu@asu.edu
pwnslinger@asu.edu

Robert Axelrod
Gerald Ford School of Public Policy
University of Michigan
axe@umich.edu

Richard Davis
Scott Atran
Changing Character of War Centre
University of Oxford; and
Artis International
rdavis@artisinternational.org
satran@artisinternational.org

ABSTRACT

In this paper, we propose and test three content-based hypotheses that significantly increase message virality. We measure virality as the retweet counts of messages in a pair of real-world Twitter datasets: A large dataset - UK Brexit with 51 million tweets from 2.8 million users between June 1, 2015 and May 12, 2019 and a smaller dataset - Nord Stream 2 with 516,000 tweets from 250,000 users between October 1, 2019 and October 15, 2019. We hypothesize, test and conclude that messages incorporating "negativity bias", "causal arguments" and "threats to personal or societal core values of target audiences" singularly and jointly increase message virality on social media.

CCS CONCEPTS

• **Human-centered computing** → **Social networks**; **Social media**; **Social content sharing**.

KEYWORDS

Message Effectiveness, Negativity Bias, Causal Arguments, Core Values, Message Virality

ACM Reference Format:
Maryam Mousavi, Hasan Davulcu, Mohsen Ahmadi, Robert Axelrod, Richard Davis, and Scott Atran. 2022. Effective Messaging on Social Media: What Makes Online Content Go Viral?. In *Proceedings of the Web Conference 2022 (WWW '22), April 25–29, 2022, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512016

1 INTRODUCTION

Building a foundation that allows machines to predict viral messages and campaigns requires sensitivity to the psycho-cultural factors that affect the political, economic and social shifts that their sponsors seek. Thus, knowledge of psychology and cultural belief systems [4] are needed to develop such algorithms. Overcoming viral campaigns ultimately relies on human appraisal of strategic aspects, such as importance of "core values" and the stakes at play (political, social, economic), and relative strengths of players in those stakes [6]. The critical role of social science goes beyond the expertise of engineers, analysts, and data scientists that social media platforms employ to moderate propaganda, disinformation and hateful content [56].

An acute problem concerns overwhelming evidence from cognitive and social psychology and anthropology, that truth and evidence – no matter how logically consistent or factually correct – do not sway public opinion or popular allegiance as much as appeals to basic cognitive biases that confirm deep beliefs and core cultural values [54]. Indeed, many so-called "biases" used in an argument do not reflect sub-optimal or deficient reasoning but rather suggest their efficient (even optimal) use for persuasion – an evolutionarily privileged form of reasoning to socially recruit others to one's circle of beliefs for cooperation and mutual defense. Thus, to combat false or faulty reasoning – as in noxious messaging – it's not enough to target an argument's empirical and logical deficiencies versus a counterargument's logical and empirical coherence. Evidence is mounting that value-driven, morally focused information in general [25] and social media in particular [20, 28], not only drives readiness to believe, but also concerted actions in support of those beliefs [14].

These findings suggest that one way to get the attention of a target audience and open it up to a shift in whom they share and follow is to provide strong and persistent evidence of the morally corrosive effects of disinformation on their own values and beliefs. Social psychologists have studied myriad cognitive biases and cultural preferences (e.g., confirmation bias [16, 52], in-group vs. out-group framing [1]) that influence decision making; however, we find that some influential and universal biases apparent in viral campaigns are better suited than others, in an initial stage, to machine measurements using existing word lists (e.g., negative vs positive sentiment [29], word-emotion association lexicon [30, 39], The Penn Discourse TreeBank [46]). These include several relation types that are relevant to causation, valence, arousal and dominance [38] for attributing information to authoritative voices, avoiding loss vs. seeking gain [60], and confirmation of one's beliefs [24, 53, 54]. Moreover, each country and component identity groups have core values (e.g., devotion to family, ethnic group, nation, religion, and political ideal) that underpin what is important at individual and group levels ("who I am," "who we are") [5, 48].

Historical evidence and recent behavioral and neuroimaging experiments [43] suggest that people make their greatest exertions and sacrifices for so-called "sacred values." These are non-negotiable,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
TheWebConf 2022, April 25–29, 2022, Lyon, France
© 2022 Association for Computing Machinery.
https://doi.org/10.1145/3485447.3512016

immune from material tradeoffs and discounting, and which accelerate and amplify the flow of information in social media space.
In this paper, we develop and test three content-based hypotheses that *singularly* and *jointly* increase a message's virality on social media . We measure virality as the share or retweet counts of messages in a pair of real-world Twitter datasets: A large dataset - UK Brexit with 51 million tweets from 2.8 million users between June 1, 2015 and May 12, 2019 and a smaller dataset - Nord Stream 2 with 516,000 tweets from 250,000 users between October 1, 2019 and October 15, 2019. Given above, this article formulates and tests following three hypotheses:
H1 - Negativity Bias: *Negative framing increases the virality of a message on social media.*
We investigate the effects of negative framing by detecting sentiment of the messages as negative or non-negative using a state of the art sentiment analysis tool designed for social media (VADER[29]) augmented with an enhanced sentiment lexicon [30]. Next, the null hypothesis is formulated so that it can be tested for possible rejection and accept the alternative hypothesis.
H2 - Causal Arguments: *Causal arguments increase the virality of a message on social media.*
In this investigation, we are not concerned with identifying relationships mentioned in messages that are causal in some "true" metaphysical sense; what characterizes true causation is a highly contentious topic within philosophy [8]. Instead, we are concerned here only with what the text asserts – causal language and what is meant by it. If and only if the text explicitly appeals to some notion of promoting or hindering, then the relationship it asserts is one that we want to represent, whether or not it is accurate.
H3 - Threats to Core Values *Threats to core societal and individual values of a target audience increase the virality of a message on social media.*
We establish the presence of a threat to a core societal and individual value by tapping into the presence of either (i) negativity bias or (ii) causal arguments alongside a core value of the target audience in a message vs. the others.
The rest of the paper is organized as follows. In Section 2, we survey the related research through a discussion of existing literature. In Section 3, we present the study datasets, their political significance and key preprocessing steps. Research hypotheses, results of the hypotheses testing and findings are presented in Section 4. Section 5, concludes the paper with implications towards new theory and algorithm development for predicting message virality using content-based features, as well as some limitations and future work.

2 RELATED WORK

Identifying when a piece of information goes "viral" in social media is an important problem in social network analysis. This is often referred to as "cascade prediction". Recently, researchers attempted to predict the final size of information cascades by developing network and community structure based approaches and measures. Pei et al. [40] measured the influence of a root node by its k-shell number and related heuristics. Weng et al. [57] and Guo et al. [26] utilized features describing both structural and temporal properties of early-stage cascades. The works described in [61] and [51] modelled cascades as one-dimensional point processes. In [27], authors survey the performance of a variety of network-based cascade prediction methods and conclude that feature based methods generally provide better prediction accuracy. However, they suffer from heavy overhead such as community detection and computation of expensive features. Random point process based methods enable researchers to achieve the prediction with little preprocessing but they are shown to be less accurate compared to feature based methods. Another study [34] largely complements previous research, suggesting that actors with larger number of followers receive the bulk of attention – with some internal variation.
Content based cascade prediction works include [42] which employs Berger's STEPPS (social currency, triggers, emotion, practical value, public, and stories) framework to explore the relationships between the six principles and the level of online engagement, revealing interesting results about what people post initially versus what they pass along from others in their network. Authors of [7] observe that virality is driven, in part, by activation and arousal. Content that evokes either high-arousal positive emotions (awe) or negative emotions (anger or anxiety) tends to be more viral. Content that evokes low arousal or deactivating emotions (e.g., sadness) tends to be less viral. Authors of [33] measure and conclude with descriptive statistics that, since the beginning of the coronavirus disease 2019 (COVID-19) epidemic, misinformation has been spreading uninhibited over traditional and social media at a rapid pace. They provide an early quantification of the magnitude of misinformation spread and highlight the importance of early interventions in order to curb this phenomenon that endangers public safety at a time when awareness and appropriate preventive actions are paramount.

3 STUDY DATASETS

The Brexit related keywords and #hashtags used to match and procure the UK Brexit dataset from Twitter dated between June 1, 2015 and May 12, 2019 are shown in Table 1. The Nord Stream 2 (NS2) gas export pipeline related keywords and #hashtags used to match and collect the second dataset from Twitter's Streaming Real-Time API[1] dated between October 1, 2019 and October 15, 2019 period are also shown in Table 1.

3.1 UK Brexit Dataset

The Key Events. The Brexit dataset compromises nearly 51 million tweets with a subset of 42 million re-tweets in English about the EU referendum (or the Brexit referendum), that took place on 23 June 2016 in the United Kingdom (UK) to ask the electorate whether the country should remain a member of, or leave, the European Union (EU). Remain Camp became the official group campaigning for the UK to remain in the EU, and was endorsed by the Prime Minister David Cameron and Chancellor George Osborne. Vote Leave was the official group campaigning for the UK to leave the EU, and was fronted by Conservative MPs Boris Johnson and Michael Gove, along with Labour MP Gisela Stuart. Other campaign groups, political parties, businesses, trade unions, newspapers and prominent individuals were also involved, with both sides having supporters from across the political spectrum. Parties in favour

[1]https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time

Table 1: List of #Hashtags and Key Phrases used for UK Brexit and NS2 Data Collection

#Hashtags/Key Phrases	#Hashtags/German Translations
Brexit: #brexit, #theresamay, #theresa, #time, #vote, #mps, #labour, #news, #country, #stopbrexit, #hardbrexit, #postbrexit, #nodeal, #remain, #leave, #parliament, #brexitdeal, #britain, #referendum, #government, #ireland, #tory, #brexitshambles, #yes2eu, #yestoeu, #betteroffin, #votein, #ukineu, #bremain, #strongerin, #leadnotleave, #voteremain, #no2eu, #notoeu, #betteroffout, #voteout, #eureform, #britainout, #leaveeu, #voteleave, #beleave, #loveeuropeleaveeu, #leaveeu, #FBPE, #AshamedToBeBritish, #NotMyVote, #NotInMyName, #NotInOurName, #PostRefRacism, #PostBrexitRacism, Cityoflondon, Boris Johnson, Nigel Farage	#europeancommission, #EconomicMigrants, #cancelbrexit, #Euro, #cyprus, #brexitlogic, #controversy, #youcouldntmakeitup, #DonaldTusk, #pensionslaw, #eeas, #eear, #Deutschland, #UNITY, #rightwingnigel, #unjust, #employmentnews, #UKgovernment, #poorharry, #ShameOnParliament, #dogwhistle, #NoDealBrexitOct31, #britswithgrit, #Nigel, #CryptoNews, #BrexitDarkMoney, #brexitdebate, #kremlin, #brexitLies, #hatebrexit, #Vauxhall, #QAnon2019, #QAnon2018, #questiontime, #labourmembers, #realsocialist, #hungparliament, #centralbanks, #tyranny
NS2: Nord Stream 2, Северный поток 2, NS2, natural gas, undersea pipeline, gas exports, US sanctions, Uniper BASF, Wintershall, gas as a weapon, energy choice, inexpensive gas, inexpensive energy, competitive gas, competitive energy, climate change, sustainable gas, reliable partner, stealing gas, Russian gas, energy transition	Nord Stream 2, Северный поток 2, NS2, Erdgas, Unterseeische Pipeline, Gasexporte, US Sanktionen, Uniper BASF, Wintershall, Gas als Waffe, Energie Wahl, Preiswertes Gas, Preiswerte Energie, Wettbewerbsfähiges Gas, Wettbewerbsfähige Energie, Klimawandel, Nachhaltiges Gas, Zuverlässiger Partner, Gas stehlen, Russisches Gas, energiewende

of 'remain' included Labour, the Liberal Democrats, the Scottish National Party (SNP), Plaid Cymru and the Green Party; while the UK Independence Party (UKIP) campaigned in favour of leaving the European Union;and the Conservative Party remained neutral. Immediately after the result, financial markets reacted negatively worldwide, and Cameron announced that he would resign as Prime Minister and Leader of the Conservative Party, having campaigned unsuccessfully to remain in the European Union.

This dataset also features 2017 UK General Election that was held on Thursday, 8 June 2017. The governing Conservative Party remained the largest single party in the House of Commons but lost its small overall majority, resulting in the formation of a Conservative-led minority government with a confidence-and-supply agreement with the Democratic Unionist Party (DUP) of Northern Ireland. Next spike in the volume following the General Election is resignation of Boris Johnson as Foreign Secretary on 9 July 2018 where in his resignation letter, Mr Johnson said the prime minister was leading the UK into a "semi-Brexit" with the "status of a colony". On 15 November 2018, Dominic Raab also announced his resignation as Brexit Secretary, citing his disapproval over the Cabinet position on the draft Brexit withdrawal agreement.The last volume spike on 12 June 2019 corresponds to the Member of Parliament (MPs) rejection of the Labour plan for a no-deal vote.

Detected Episodes. Initially, we developed a 20-Day moving average $\mu(20)$ signal on the daily volume chart, alongside an enveloping upper and lower band that is moving at $\mu(20) + 2 * \sigma(20)$ and $\mu(20) - 2 * \sigma(20)$. Anytime a daily volume spiked above the upper-band, we recorded it as the beginning of a new episode which lasted until the beginning of the next episode after falling below the $\mu(20) - 2 * \sigma(20)$. This simple episode detection measure adapted from a popular process control method [13] detected 13 episodes in the dataset, many of them corresponding to key political events outlined above.

Social Network Structure and Flows. For each of the 13 episodes, we performed community detection on the user-to-user re-tweet graph by running the popular Louvain [9] algorithm. After finding the communities, we created a Sankey diagram [49] where a column of boxes correspond to the relative sizes of the detected communities during an episode and the belts between the boxes representing the flows where the width of a belt is proportional to the shared numbers of users between current and the next episode's boxes (or user communities). The Sankey diagram that we use in Figure 1 exhibits similarities to "alluvial diagrams" which are a type of flow diagram originally developed to represent network changes; such as *fragmentation* of one cluster (or box) in a present episode into many clusters in the next episode, or *coalescence* of many flows (or belts) from current episode into a new community in the next episode.

The Sankey diagram in Figure 1 exhibits 61 unique communities (or boxes) with belts between episodes pointing at enduring community trends and potentially their drivers (i.e. wedge [47] and attractor [37] issues) and content types giving insights about the network dynamics.

Leave vs. Remain Camp Coloring in the Sankey Diagram. In order to detect the majority Leave or majority Remain types of each community in the Sankey diagram, first we identified top-400 most re-tweeted users and top-400 most re-tweeted tweets and then asked a team of English speaking coders who are intimately familiar with UK Brexit politics to mark each user and tweet as a follower of either the Leave Camp or the Remain Camp.

Next we customized and used a label propagation algorithm which incorporates ideas from [44, 62] to label unlabeled users based on the labels of already known top-400 re-tweeted users and tweets. By detecting the majority camp for each box (or community) as Leave or Remain, and applying the corresponding colors on the Sankey diagram, we obtain the final camp colored boxes in

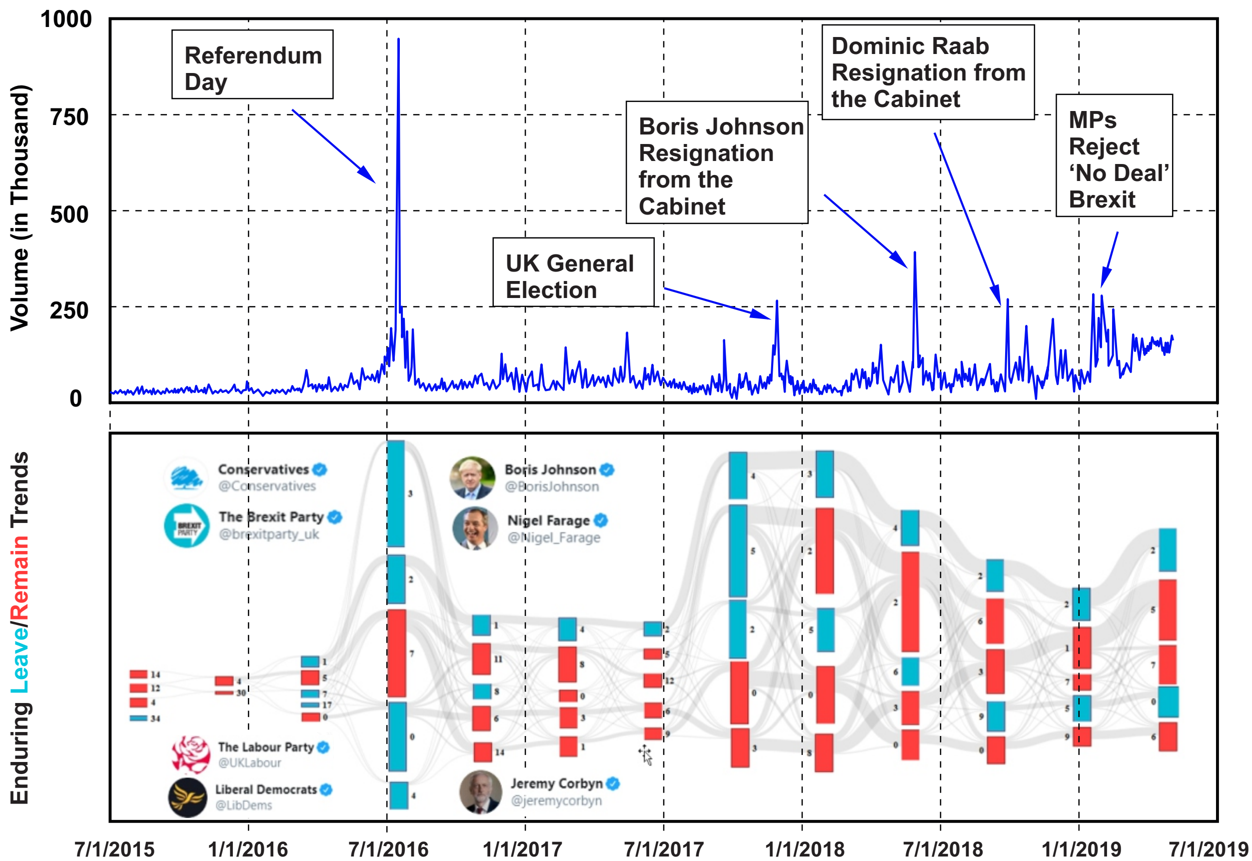


Figure 1: Daily Tweet Volumes of the UK Brexit dataset, its Leave & Remain Camps and Enduring Trends

Figure 1 which 36 communities reside in the Remain Camps and the remaining 25 in the Leave Camp with overwhelming Leave Camp community sizes during the Brexit referendum and the 2017 UK General Elections.

The label propagation algorithm is a greedy hill-climbing algorithm [18]. It is highly efficient, but it can easily converge to local optimum solutions depending on the initial label assignments and random tie breakings [11]. By running the algorithm one hundred times, out of 2.8 million users, 1.5 million get colored into the Leave Camp and the remaining 1.3 million into the Remain Camp. This result is supportive of the fact that UK not only voted to leave the EU on 23 June 2016 but they also re-elected a majority Conservative Party led House of Commons after the election on June 8, 2017.

3.2 Nord Stream 2 (NS2) Dataset

The Nord Stream 2 [23] is a controversial gas export pipeline construction project running under the Baltic Sea from Russia to Germany. In 2011, Nord Stream AG started evaluation of an expansion project consisting of two additional lines (later named Nord Stream 2) to double the annual capacity. In January 2015, Gazprom announced that the expansion project had been put on hold since the existing lines were running at only half capacity, due to EU sanctions on Russia, following its annexation of Crimea. Controversies of Nord Stream 2 include political, security and military, economic and environmental aspects.

NS2 Data Collection. Our data collection focused on pro-NS2 Russian frames shown in Table 1 dated between October 1, 2019 and October 15, 2019. Tweet collection yielded 516,050 pro-NS2 tweets engaged by 249,798 distinct user accounts. The English and German phrases frame NS2 as inexpensive, competitive, sustainable gas, supporting clean energy transition from a reliable partner. The greatest environmental impact results from the consumption of the transported gas, if it allows more imports to the EU. That would be in conflict with decarbonization efforts for climate protection according to opponents. The #hashtag cloud corresponding to this Tweet corpus is depicted below in Figure 2. It can be inferred that this pro-NS2 campaign emphasize and target not only #ClimateChange and #ClimateAction type conversation threads but it was also active in the #ExtinctionRebellion related threads as this group was mobilizing online towards an International Rebellion starting on October 7th, for 2 weeks. Extinction Rebellion (XR) is a "global environmental movement with the stated aim of using nonviolent civil disobedience to compel government action to avoid tipping points in the climate system, biodiversity loss, and the risk of social and ecological collapse" according to Wikipedia [58].

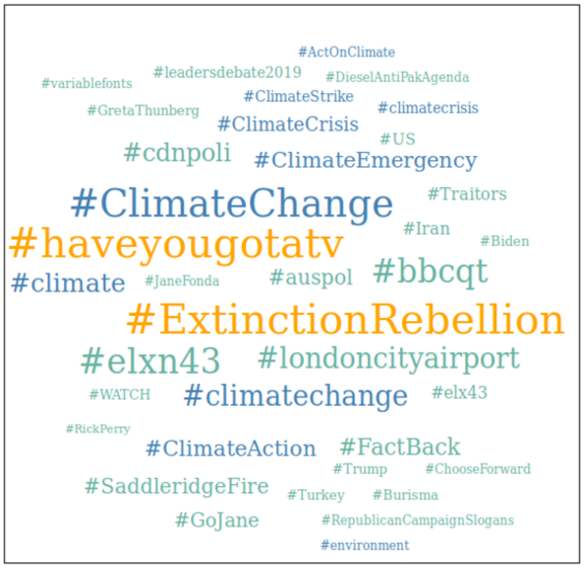


Figure 2: Nord Stream 2 (NS2) Dataset #Hashtag Cloud

4 HYPOTHESES FORMULATION AND TESTS

4.1 Two-Sample Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov goodness-of-fit test [2] is exact only when the hypothesized distribution is continuous, but later Dimitrova et. al. [17] has extended the Kolmogorov-Smirnov test to obtain a test that is fast and accurate method to compute the (complementary) CDF of the KS statistic when F(x) is discrete, and thus obtain exact p values of the KS test. We used the R implementation of the proposed method [2] in our tests.

4.2 Hypothesis 1: Negativity Bias

4.2.1 Data Preparation. Psycholinguistic studies suggest that the process of communication between humans, which tends to optimize word form [41] and message content [55] in information transfer, creates a bias towards words with positive emotional content. In [22] authors find that words with a positive emotional content are more frequently used. This lends support to Pollyanna hypothesis [10] that there should be a positive bias in human expression. Negative expressions, however, "mostly serve a different purpose, namely that of transmitting highly informative and relevant events. They are less used, but carry more information" and are more attention-arresting, consistent with the hypothesis of our "Negativity Bias" [36].We investigate the effect of Negativity Bias by first detecting sentiment of the messages as negative or non-negative using a state of the art sentiment analysis tool designed for social media (i.e. VADER [29]) augmented with an enhanced sentiment lexicon [30].

Given that we are dealing with a pair of discrete distributions (i.e. those containing Negativity Bias vs. others) in all datasets based on their retweet counts, we use a two-sample Kolmogrov-Smirnov (one-tailed K-S test) as a goodness-of-fit test [2].

[2] https://cran.r-project.org/web/packages/KSgeneral/index.html

Hypothesis 1 - Null Hypothesis: [Rejected] Popularity distribution of tweets with Negativity Bias are no greater than the popularity distribution of tweets with no Negativity Bias.

In Table 2, since the P-values are far less than the significance level that we chose 0.05 ($\alpha = 0.05$), null hypothesis is rejected for all three datasets. From the results we can infer that the popularity distribution of tweets with Negativity Bias do not follow the same popularity distribution of tweets without Negativity Bias arguments. It can also be observed in all charts in Figures 3 and 4 that messages that contain "Negativity Bias" arguments are more effective and achieve higher virality in all three datasets.

Table 2: Null Hypothesis Test for "Negativity Bias"

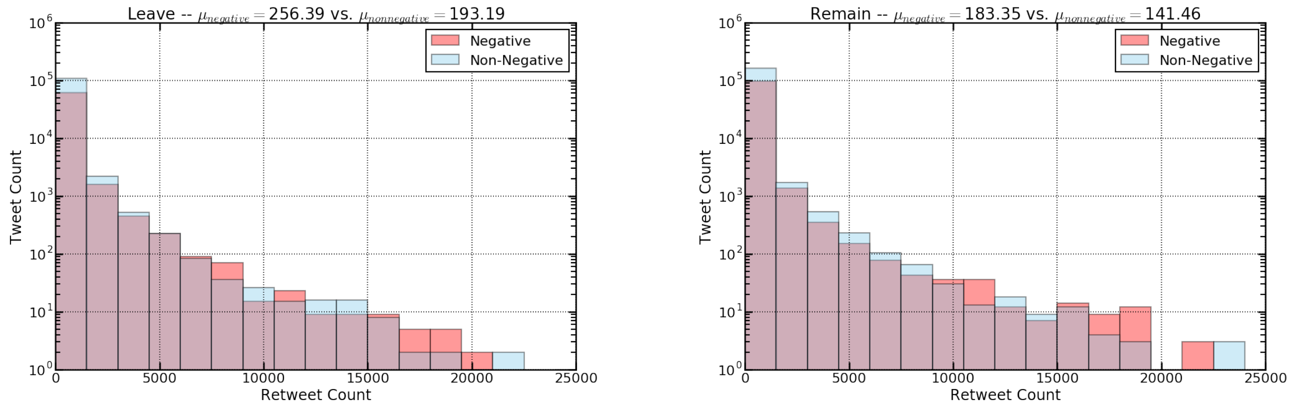
Dataset	Significance (α)	P-Value	Test Statistic
Brexit Leave	0.05	9.5931e-05(****)	2547.87
Brexit Remain	0.05	5.9370e-79(****)	3582.04
NS2	0.05	1.85e-14(****)	139.74

4.3 Hypothesis 2: Causal Arguments

4.3.1 Data Preparation. Language of cause and effect captures an essential component of the semantics of a text [19] and it is essential for extracting many semantic relationships: we think about the world around us in terms of causation, and we often consult texts about what causes, enables, hiders or prevents some phenomenon (e.g., medical symptoms, political events, or interpersonal actions). Unsurprisingly, causal language is also ubiquitous; Conrath et al. [12] found that in French, causation constituted 33% of the relations expressed between verbs. Causal language detection is an active and ongoing area of research [32, 59]. In this paper we adopt a lexicography-based matching scheme for detecting the presence of positive and negative causal language by expanding The Penn Discourse TreeBank [46] which includes several relevant relation types (i.e. primarily CAUSE and REASON). We further expand these relationships through their WordNet [21] synonyms and antonyms to develop a matcher which employs a list of causal constructions outlined in [19] in order to detect assertions about positive and negative causation that are explicit in the messages. Given that we are dealing with a pair of discrete distributions (i.e. those containing Causal Arguments vs. others) in all datasets based on their retweet counts, we use a two-sample Kolmogrov-Smirnov (one-tailed K-S test) as a goodness-of-fit test [2].

Hypothesis 2 - Null Hypothesis: [Rejected] Popularity distribution of tweets with Causal Arguments are no greater than the popularity distribution of tweets with no Causal Arguments.

In Table 3, since the P-values are far less than the significance level that we chose 0.05 ($\alpha = 0.05$), null hypothesis is rejected for all three datasets. From the results we can infer that the popularity distribution of tweets with causal arguments do not follow the same popularity distribution of tweets not involving causal arguments. It can also be observed in all charts in Figures 5 and 6 that messages that contain causal arguments are more effective and achieve higher virality in all three datasets.



(a) Negative bias on Brexit Leave Camp

(b) Negative bias on Brexit Remain Camp

Figure 3: Popularity distribution of tweets with messages involving "negative bias" vs. messages not involving negative bias in UK Brexit Leave and Remain camps (51M Tweets between June 1, 2015 and May 12, 2019).

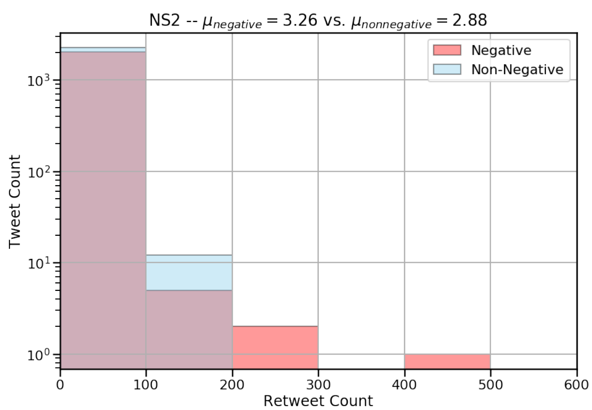


Figure 4: Popularity distribution of Tweets with negative bias vs. Tweets with No negative bias in NS2 debate Europe-wide (516,050 Tweets between 10/1/2019 and 10/15/2019)

Table 3: Null Hypothesis Test for "Causal Arguments"

Dataset	Significance (α)	P-value	Test Statistic
Brexit Leave	0.05	5.41e-75(****)	494.85
Brexit Remain	0.05	7.19e-15(****)	179.58
NS2	0.05	0.00022(***)	99.58

4.4 Hypothesis 3: Threats to Core Values

4.4.1 Data Preparation. To get an understanding regarding the effect of "threats to societal and individual core values" on message virality we look into the presence of either negativity bias or causal arguments alongside a core value of the target audiences in their messages vs. their absence. Following Figure 7 details these matching rules in the presence of positive, neutral vs. negative values, negativity bias, and positive or negative causal arguments.

Table 4: Null Hypothesis Test for "Threats to Core Values"

Dataset	Significance (α)	P-value	Test Statistic
Brexit Leave	0.05	0.05(*)	59.24
Brexit Remain	0.05	4.58e-7(****)	116.73
NS2	0.05	3.88e-09 (****)	50.41

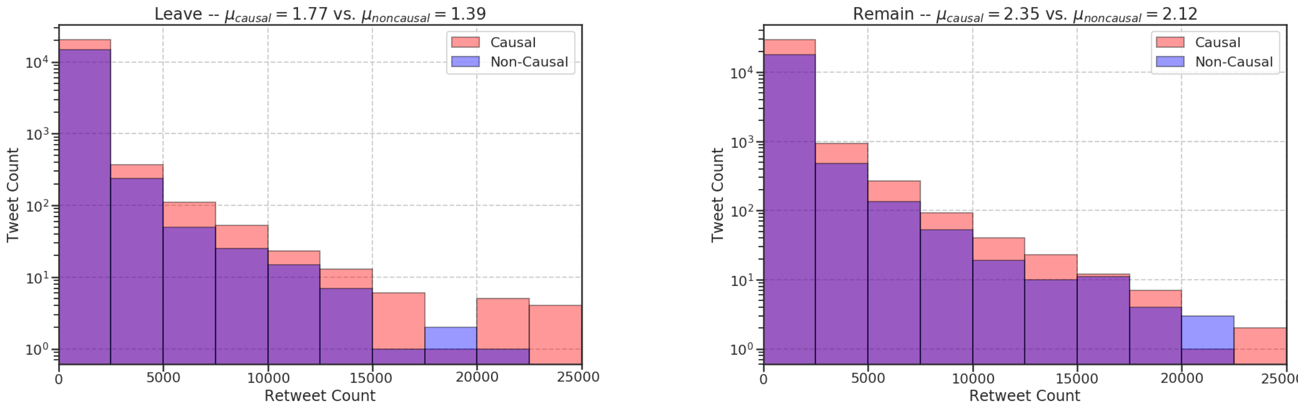
4.4.2 Values Coding. Given a dataset first we extract its list of most frequently mentioned noun phrases using a noun phrase extractor [35]. Next, a pair of social scientists scan the list of noun phrases and codes for the values. Once they achieve an acceptable intercoder reliability with Krippendorff's alpha [15], then they also code each value together as either Normative or as Sacred. Values such as family (life, bloodline), religion (freedom of expression) and nation (country, flag, and language) are considered as sacred values (i.e., immune to material tradeoffs regardless of costs and benefits, temporal and spatial discounting, and social pressure and influence) [3, 50] whereas others, such as climate, diversity, sustainability, wealth-inequality and workers' rights, etc. are coded as normative values.

We define core values as the union of normative and sacred values. Testing the hypothesis that "threats to societal and individual core values increase message virality", involves the formulation of the following null hypothesis and testing it for rejection.

Given that we are dealing with a pair of discrete distributions (i.e. those containing Threats to Core Values vs. No Threats to Core Values), in all datasets based on their retweet counts, we use a two-sample Kolmogrov-Smirnov (one-tailed K-S test) as a goodness-of-fit test [2].

Hypothesis 3 - Null Hypothesis: [Rejected] Popularity distribution of tweets with threats to core values are no greater than the popularity distribution of tweets with no threats to core values.

In Table 4, since the P-values are far less than the significance level that we chose 0.05 (α = 0.05), null hypothesis is rejected for



(a) Causal Argument Framing on Brexit Leave Camp

(b) Causal Argument Framing on Brexit Remain Camp

Figure 5: Popularity distribution of tweets with messages involving "causal arguments" vs. messages not involving causal arguments in the UK Brexit Leave and Remain camps (51M Tweets between June 1, 2015 and May 12, 2019)

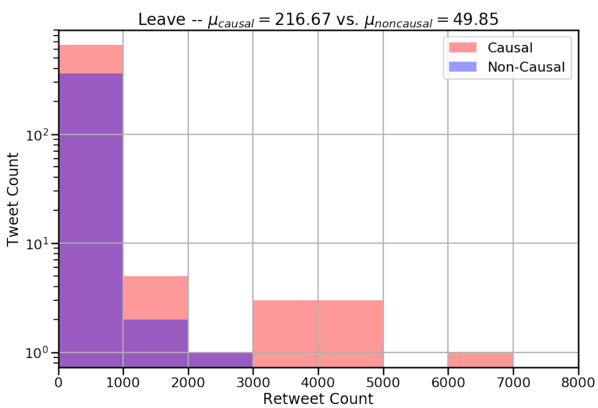


Figure 6: Popularity distribution of Tweets with causal arguments vs. Tweets without causal arguments in NS2 debate Europe-wide (516,050 Tweets between 10/1/2019 and 10/15/2019)

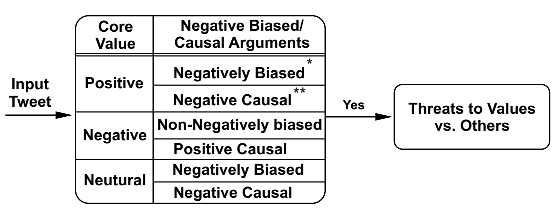


Figure 7: Threats to Core Values Data Preparation Flowchart

all three datasets. From the results we can infer that the popularity distribution of tweets with threats to core values arguments do not follow the same popularity distribution of tweets with non-threat to values arguments. It can also be observed in all charts in Figures 8 and 9 that messages that contain "threats to core values" arguments are more effective and achieve higher virality in all three datasets.

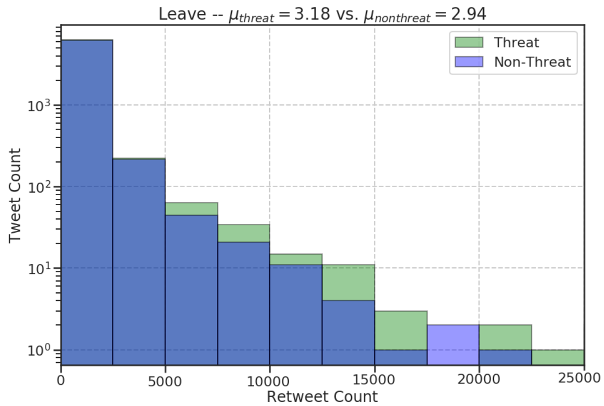
Rejection of the Null Hypothesis allows us to proceed to a second stage of hypothesis testing in order to distinguish the effects of messaging involving (threats to) sacred values versus messaging involving (threats to) normative values. Preliminary analysis suggests that normativity of values is not as predictive of virality as sacredness of values. We anticipate that hypothesis testing on Sacred Values, based on cross-culturally validated lists of Sacred Values, should continue to yield similar results; however, providing a validated machine learnable list is more complicated when factoring in multiple languages, cultures, slang and the various turns of phrase used in Social Media posts. For example, validating a word list would need to be able to account for slight variances in sentiment (e.g., threat to religion needs to be able to address the difference in saying someone is anti-Semitic vs. using the term Auschwitz).

4.5 Joint Effects

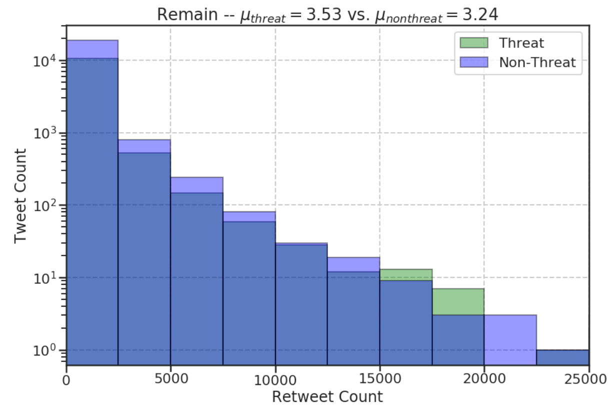
So far we studied the individual effects of negativity bias, causal arguments, and threats to core values with regards to virality on social media and we concluded that messages fitting any of these hypothesis accrues more attention and engagement compared to others. In this section we design a test to investigate the joint effects of these hypotheses. Since our re-tweet charts exhibit over dispersion and excessive amount of zeros in the retweet counts axis we will be using Zero-Inflated Negative Binomial Regression Model (ZINBR) [45] to assess the joint effects of the these hypothesis.

In this assessment, the regression dependent (target) variable is identified as the retweet count (*retweet_count*), and the independent (predictor) variables are identified as follows:

- h_1 is a weighted variable indicating a negatively biased tweet;
- h_2 is a weighted variable indicating causal arguments;
- h_3 is a weighted variable indicating threats to core values;
- *follower_count* is also a weighted variable indicating the number of followers of the tweeter who is messaging.



(a) Value-Threat Frame on Brexit Leave Camp



(b) Value-Threat Frame on Brexit Remain Camp

Figure 8: Popularity distribution of tweets with messages involving "threats to core values" vs. messages not involving threats to core values in UK Brexit Leave and Remain camps (51M Tweets between June 1, 2015 and May 12, 2019).

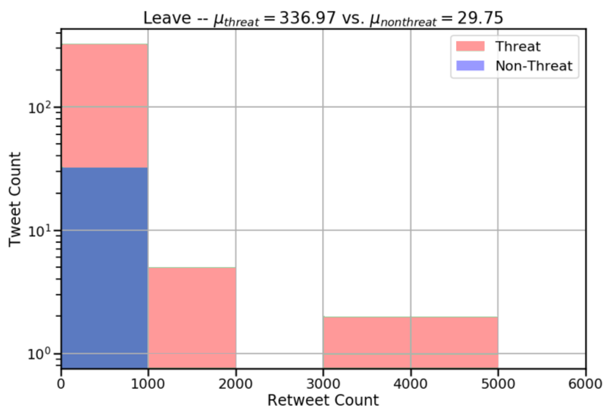


Figure 9: Popularity distribution of Tweets with Threats to Core Values vs. Tweets without Threats to Core Values in NS2 debate Europe-wide (516,050 Tweets between 10/1/2019 and 10/15/2019).

Following Tables 5 and 6 show that all of the predictors in the model are statistically significant, and each variable has a positive interaction with the others.

Table 5: Results of Running the ZINBR for Brexit Leave

Parameter Estimates for original Predictors			
Term	**Estimate**	**Std Error**	**Pr >ChiSq**
Intercept	3.00	0.015	<.0001*
h1	0.556	0.033	<0.001*
h2	0.043	0.021	<0.001*
h3	0.145	0.009	<0.001*
followers	0.00	0.00	<0.001*

Table 6: Results of Running the ZINBR for Brexit Remain

Parameter Estimates for original Predictors			
Term	**Estimate**	**Std Error**	**Pr >ChiSq**
Intercept	2.59	0.011	<.0001*
h1	0.297	0.024	<.0001*
h2	0.042	0.015	<.0001*
h3	0.138	0.007	<.0001*
followers	0.00	0.00	<.0001*

5 CONCLUSION

In this paper, we measure virality as the retweet counts of messages in a pair of real-world Twitter datasets. We find that virality of messages on social media is increased with the use of negative terms, causal arguments, and threats to core values. Next, we plan to design and follow up with real A/B testing [31] experiments in order to determine the true strength and consistency of our hypothesis test findings in the presence of other control variables. The Sankey diagram visualization of network dynamics provides recurring opportunities for detecting divisive *wedge issues [47]* and unifying *attractors [37]* alongside their content profiles (i.e. a threat to a core value, a new popular group leader or identity, or a new single-issue group structure, etc.). As future work, we plan to develop content based profiling techniques and algorithms for identifying wedges and attractors of enduring trends, as well as effective combinations of wedges that fragment certain types of target audiences and attractors that grow and coalescence them.

ACKNOWLEDGMENTS

For support we thank the US Department of Defense Minerva Initiative and the Air Force Office of Scientific Research (AFOSR Grant No. FA9550-18-1-0496). The views expressed in this work do not necessarily represent the United States Air Force, Department of Defense, or United States Government.

REFERENCES

[1] Inmaculada Adarves-Yorno, Jolanda Jetten, Tom Postmes, and S. Haslam. 2013. What Are We Fighting For?: The Effects of Framing on Ingroup Identification and Allegiance. *The Journal of social psychology* 153 (02 2013), 25–37. https://doi.org/10.1080/00224545.2012.701673

[2] Mark Edward Allen. 1976. Kolmogorov-Smirnov test for discrete distributions. https://calhoun.nps.edu/handle/10945/17830

[3] Scott Atran and Robert Axelrod. 2008. Reframing Sacred Values. *Negotiation Journal* 24, 3 (2008), 221–246. https://doi.org/10.1111/j.1571-9979.2008.00182.x

[4] Scott Atran, Robert Axelrod, Richard Davis, and Baruch Fischhoff. 2017. Challenges in researching terrorism from the field. *Science* 355, 6323 (2017), 352–354. https://doi.org/10.1126/science.aaj2037 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaj2037

[5] Scott Atran and Jeremy Ginges. 2012. Religious and Sacred Imperatives in Human Conflict. *Science* 336, 6083 (2012), 855–857. https://doi.org/10.1126/science.1216902

[6] Robert Axelrod and Rumen Iliev. 2014. Timing of cyber conflict. *Proceedings of the National Academy of Sciences* 111, 4 (2014), 1298–1303. https://doi.org/10.1073/pnas.1322638111 arXiv:https://www.pnas.org/content/111/4/1298.full.pdf

[7] Jonah Berger and Katherine Milkman. 2013. Emotion and Virality: What Makes Online Content Go Viral? *GfK Marketing Intelligence Review* 5 (05 2013), 18–23. https://doi.org/10.2478/gfkmir-2014-0022

[8] John Bigelow and Robert Pargetter. 1990. Metaphysics of Causation. *Erkenntnis (1975-)* 33, 1 (1990), 89–119. http://www.jstor.org/stable/20012286

[9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (Oct 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

[10] Jerry Boucher and Charles E. Osgood. 1969. The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior* 8, 1 (1969), 1–8. https://doi.org/10.1016/S0022-5371(69)80002-2

[11] Michael D. Conover, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2013. The Digital Evolution of Occupy Wall Street. *PLOS ONE* 8, 5 (05 2013), 1–5. https://doi.org/10.1371/journal.pone.0064679

[12] Juliette Conrath, Stergos Afantenos, Nicholas Asher, and Philippe Muller. 2014. Unsupervised extraction of semantic relations using discourse cues. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2184–2194. https://aclanthology.org/C14-1206

[13] S. Crowe, M. Utley, and G. Walker. 2015. A novel approach to evaluating the UK childhood immunisation schedule: estimating the effective coverage vector across the entire vaccine programme. *BMC Infectious Diseases* 15, 585 (2015). https://doi.org/10.1186/s12879-015-1299-8

[14] Richard Davis. 2016. *Hamas, Popular Support and War in the Middle East*. Routledge, London, UK.

[15] Knut De Swert. 2012. Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Center for Politics and Communication* 15 (2012).

[16] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. https://doi.org/10.1073/pnas.1517441113

[17] Dimitrina S. Dimitrova, Vladimir K. Kaishev, and Senren Tan. 2020. Computing the Kolmogorov-Smirnov Distribution When the Underlying CDF is Purely Discrete, Mixed, or Continuous. *Journal of Statistical Software* 95, 10 (2020), 1–42. https://www.jstatsoft.org/index.php/jss/article/view/v095i10

[18] Juan Du, Feida Zhu, and Ee-Peng Lim. 2013. Dynamic label propagation in social networks. In *International Conference on Database Systems for Advanced Applications*. Springer, 194–209. https://doi.org/10.1007/978-3-642-37450-0_14

[19] Jesse Dunietz, Lori S. Levin, and Jaime G. Carbonell. 2015. Annotating Causal Language Using Corpus Lexicography of Constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*. 188–196.

[20] Matthew Feinberg and Robb Willer. 2015. From Gulf to Bridge: When Do Moral Arguments Facilitate Political Influence? *Personality and Social Psychology Bulletin* 41, 12 (2015), 1665–1681. https://doi.org/10.1177/0146167215607842

[21] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

[22] David Garcia, Antonios Garas, and Frank Schweitzer. 2012. Positive words carry less information than negative words. *EPJ Data Science* 1, 1 (May 2012). https://doi.org/10.1140/epjds3

[23] Gazprom. 2019. *Nord Stream 2: A new export gas pipeline running from Russia to Europe across the Baltic Sea*. "https://www.gazprom.com/projects/nord-stream2/"

[24] G. Gigerenzer. 2010. *Rationaliy for mortals*. Oxford University Press, New York.

[25] Ángel Gómez, Lucía López-Rodríguez, Hammad Sheikh, Jeremy Ginges, Lydia Wilson, Hoshang Waziri, Alexandra Vázquez, Richard J. Davis, and Scott Atran. 2017. The devoted actor's will to fight and the spiritual dimension of human conflict. *Nature Human Behaviour* 1 (2017), 673–679.

[26] Ruocheng Guo, Elham Shaabani, Abhinav Bhatnagar, and Paulo Shakarian. 2016. Toward early and order-of-magnitude cascade prediction in social networks. *Soc. Netw. Anal. Min.* 6, 1 (2016), 64:1–64:18.

[27] Ruocheng Guo and Paulo Shakarian. 2016. A comparison of methods for cascade prediction. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2016), 591–598.

[28] Joe Hoover, Marlon Mooijman, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour* 2 (06 2018). https://doi.org/10.1038/s41562-018-0353-0

[29] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.. In *Eighth International Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press. http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14

[30] Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. San Diego, California.

[31] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7, 8 (2017), 922–929.

[32] Dongmin Kong, Lu Shi, and Fan Zhang. 2021. Explain or conceal? Causal language intensity in annual report and stock price crash risk. *Economic Modelling* 94 (2021), 715–725. https://doi.org/10.1016/j.econmod.2020.02.013

[33] R. Kouzy, J. Jaoude, A. Kraitem, M.B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E.W. Akl, and K. Baddour. 2020. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus* 12, 3 (03 2020), 18–23. https://doi.org/10.7759/cureus.7255

[34] Anders Olof Larsson. 2017. Going viral? Comparing parties on social media during the 2014 Swedish election. *Convergence* 23, 2 (2017), 117–131. https://doi.org/10.1177/1354856515577891

[35] Steven Loria. 2018. textblob Documentation. *Release 0.15* 2 (2018).

[36] N. E. Miller. 1961. Some recent studies of conflict behavior and drugs. *American Psychologist* 16, 1 (1961), 12–24. https://doi.org/10.1037/h0048720

[37] John Milnor. 1985. On the concept of attractor. In *The theory of chaotic attractors*. Springer, 243–264.

[38] Saif M. Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

[39] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[40] Sen Pei, Lev Muchnik, José S. Andrade, Zhiming Zheng, and Hernán A. Makse. 2014. Searching for superspreaders of information in real-world social media. *Scientific Reports* 4 (2014).

[41] Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108, 9 (2011), 3526–3529. https://doi.org/10.1073/pnas.1012551108 arXiv:https://www.pnas.org/content/108/9/3526.full.pdf

[42] Geah Pressgrove, Brooke Weberling McKeever, and S. Mo Jang. 2018. What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge. *International Journal of Nonprofit and Voluntary Sector Marketing* 23, 1 (2018), e1586. https://doi.org/10.1002/nvsm.1586

[43] Clara Pretus, Nafees Hamid, Hammad Sheikh, Jeremy Ginges, Adolf Tobeña, Richard Davis, Oscar Vilarroya, and Scott Atran. 2018. Neural and Behavioral Correlates of Sacred Values and Vulnerability to Violent Extremism. *Frontiers in Psychology* 9 (2018), 2462. https://doi.org/10.3389/fpsyg.2018.02462

[44] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76, 3 (Sep 2007). https://doi.org/10.1103/physreve.76.036106

[45] Md Asif Raihan, Priyanka Alluri, Wensong Wu, and Albert Gan. 2019. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis & Prevention* 123 (2019), 303–313.

[46] Prasad Rashmi, Dinesh Nikhil, Lee Alan, Miltsakaki Eleni, Robaldo Livio, Joshi Aravind, and Webber Bonnie. 2008. The Penn Discourse TreeBank 2.0.. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (28-30), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Joseph Mariani Bente Maegaard, Jan Odijk, Stelios Piperidis, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/

[47] Martin J Riedl. 2019. Frenemies: How Social Media Polarizes America by Jaime E. Settle.

[48] T. Schelling. 1960. *The strategy of conflict*. Harvard University Press, Cambridge, MA.

[49] Mario Schmidt. 2008. The Sankey Diagram in Energy and Material Flow Management. *Journal of Industrial Ecology* 12, 1 (2008), 82–94. https://doi.org/10.1111/j.1530-9290.2008.00004.x

[50] Hammad Sheikh, Jeremy Ginges, and Scott Atran. 2013. Sacred values in the Israeli–Palestinian conflict: resistance to social influence, temporal discounting,

and exit strategies. *Annals of the New York Academy of Sciences* 1299 (09 2013), 11–24. https://doi.org/10.1111/nyas.12275

[51] Huawei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (Québec City, Québec, Canada) *(AAAI'14)*. AAAI Press, 291–297.

[52] Orowa Sikder, Robert E Smith, Pierpaolo Vivo, and Giacomo Livan. 2020. A minimalistic model of bias, polarization and misinformation in social networks. *Scientific reports* 10, 1 (2020), 1–11.

[53] H. Simon. 1997. *Models of bounded rationality*. MIT Press, Cambridge, MA.

[54] Dan Sperber and Hugo Mercier. 2019. *The Enigma of Reason*. Harvard University Press, USA.

[55] Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press, USA.

[56] J. Weedon, W. Nulan, and A. Stamos. 2017. Information operations and Facebook. (2017). https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

[57] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting Successful Memes using Network and Community Structure. *CoRR* abs/1403.6199 (2014). arXiv:1403.6199 http://arxiv.org/abs/1403.6199

[58] Wikipedia. 2018. Extinction Rebellion. https://en.wikipedia.org/wiki/Extinction_Rebellion

[59] Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting Causal Language Use in Science Findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4664–4674. https://doi.org/10.18653/v1/D19-1473

[60] Chuan Zhang, Yu-xin Tian, Ling-wei Fan, and Ying-hui Li. 2020. Customized ranking for products through online reviews: a method incorporating prospect theory with an improved VIKOR. *Applied Intelligence* (2020), 1–20.

[61] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1513–1522. https://doi.org/10.1145/2783258.2783401

[62] Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. *Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University*. http://www.scholar.google.com/url?sa=U&q=http://www.cs.cmu.edu/~zhuxj/pub/propagate.ps.gz