Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Message Impartiality in Social Media Discussions

Muhammad Bilal Zafar
MPI-SWS, Germany
mzafar@mpi-sws.org

Krishna P. Gummadi
MPI-SWS, Germany
gummadi@mpi-sws.org

Cristian Danescu-Niculescu-Mizil
Cornell University, USA
cristian@cs.cornell.edu

Abstract

Discourse on social media platforms is often plagued by acute polarization, with different camps promoting different perspectives on the issue at hand—compare, for example, the differences in the liberal and conservative discourse on the U.S. immigration debate. A large body of research has studied this phenomenon by focusing on the affiliation of groups and individuals. We propose a new finer-grained perspective: studying the impartiality of individual messages.

While the notion of message impartiality is quite intuitive, the lack of an objective definition and of a way to measure it directly has largely obstructed scientific examination. In this work we operationalize message impartiality in terms of how discernible the affiliation of its author is, and introduce a methodology for quantifying it automatically. Unlike a supervised machine learning approach, our method can be used in the context of emerging events where impartiality labels are not immediately available.

Our framework enables us to study the effects of (im)partiality on social media discussions at scale. We show that this phenomenon is highly consequential, with partial messages being twice more likely to spread than impartial ones, even after controlling for author and topic. By taking this fine-grained approach to polarization, we also provide new insights into the temporal evolution of online discussions centered around major political and sporting events.

1 Introduction

As online media emerges as a popular forum for discussing political, social, and cultural issues of the day, there is a growing concern about the polarization of these discussions (Balasubramanyan et al. 2012; Bakshy, Messing, and Adamic 2015), with individuals and groups belonging to opposing ideologies expressing their biased perspectives and ignoring those of the others. Recent empirical studies have provided new insights into this phenomenon, pointing out, for example, the segregation between liberal and conservative blogs (Adamic and Glance 2005) or the echo-chambering of political discourse on online media (Conover et al. 2011b; Weber, Garimella, and Batayneh 2013; Koutra, Bennett, and Horvitz 2015).

The vast majority of empirical studies of polarization operate at the level of individuals and groups, where labeled data is often readily available. In this work, we provide a complementary perspective by focusing on a notion that operates at the finer-grained level of individual utterances: *message impartiality*.

To illustrate the notion of message impartiality, consider the following two Twitter messages announcing the imminent U.S. government shutdown of 2013:

(1) **@RepPaulTonko:** House R[epublicans]'s have America on course for #GOPShutdown. Tomorrow, hundreds of thousands of Fed employees face furloughs & many will work w/o pay

(2) **@Donna_West:** Shutdown imminent: House officials say no more funding votes tonight

While both messages are announcing the exact same event (the imminent government shutdown), message (2) is offering a more impartial account than (1) does.

While closely related, the (im)partiality of a message is crucially distinct from its author's affiliation. For example, both of the above messages were written by users with self-declared liberal affiliation. Moreover, the very same individual can author both *partial* and *impartial* messages. In particular, the author of (2) has also tweeted the following message that is certainly lacking impartiality:

(3) **@Donna_West:** #BoehnerShutdown is a temper tantrum for losing. You and your tea-publicans had this planned for over 3 yrs. You own it #JustVote

Therefore, we argue that the problem of inferring message impartiality is distinct from the task of inferring author affiliation—a task which was extensively studied in prior work (Conover et al. 2011a; Pennacchiotti and Popescu 2011; Zamal, Liu, and Ruths 2012; Wong et al. 2013).

While the notion of message impartiality is quite intuitive, there are no automated ways to *measure* impartiality, hindering large-scale empirical investigations. As impartiality is highly subjective, direct human annotation of impartiality has been found to be challenging (Yano, Resnik, and Smith 2010; Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013). Furthermore, even in the scenario where sufficient number of messages have been reliably labeled by humans (as partial or impartial) to train a supervised machine learning method, this would not be effective at predicting partiality of messages related to emerging events in new contexts not covered by the labeled data.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Against this background, in this work, we propose and validate an *operational* definition of message impartiality that allows us to automatically quantify message impartiality in a variety of contexts and emerging events, without needing to constantly regenerate impartiality labels. Our main insight is to exploit the inherent relation between impartiality of a message and the discernibility of its author's affiliation: the more impartial a message, the harder it is to infer the affiliation of its author from its content. For example, message (2) above is more impartial than (3) because it is harder to predict the affiliation of its author (which is liberal) from its content.

Building on this insight, we show how message impartiality can be successfully operationalized in terms of the *confidence with which the its author's affiliation can be inferred*, leading to a measure that aligns well with human intuition across multiple events, spanning both politics and sports. We further show that we can fully automate this framework, and that, unlike a machine learning approach trained directly on impartiality labels, our method can be used in the context of emerging events where such labels are not immediately available.

The ability to automatically quantify impartiality enables a large scale exploration of online polarization from a new fine-grained perspective, bringing new insights into this phenomenon. For example, we find that message impartiality is highly consequential, with partial tweets being twice more likely to be retweeted than impartial ones. This effect holds even when the impartial and partial messages are on the same topic and are sent by the same user.

One of the main motivations for focusing on message-level impartiality—rather than on author-level affiliation—is the observation that single individuals can vary their level of impartiality over time. To better understand this individual-level variation we study the social media discourse related to three major political and sporting events: the U.S. government shutdown of 2013, Super Bowl 2014 and Super Bowl 2015. We find that individual users do indeed change the level of impartiality of their messages over the spread of an event, and that, this variation is systematically tied to the evolution of the offline event.

To summarize, in this paper we: propose new way to operationalize the concept of message impartiality (Section 2); show that the resulting measure aligns with human judgments of impartiality (Section 3); develop a framework for automatically quantifying message impartiality (Section 4); conduct a large-scale study of message impartiality in the Twitter social network (Section 5); and provide new insights into this phenomenon (Section 6).

2 Operationalizing impartiality

To operationalize[1] message impartiality, we need to understand *what does it mean for a message to be impartial,* and *how can we measure the impartiality of a message.* Here we propose an informal characterization of this intuitive notion and discuss its relation to the better studied notion of affiliation. Building on this discussion, we then introduce a new way to operationalize this phenomenon that will enable us to quantify and study it at scale.

2.1 Characterizing the notion of impartiality

Consider the Twitter messages (1) and (2) shown in Section 1. Although both messages share the same news—impending government shutdown—they differ dramatically in the way they convey it. That is, message (1) is, at least intuitively, less impartial than message (2). In this particular case, the difference seems to stem from the fact that message (2) announces the imminent government shutdown without leaning towards any party position, whereas, message (1) implicitly blames the Republicans while announcing the exact same news. While, this is an argument specific to this example, it does highlight some of general properties of the intuitive notion of impartiality:

1. Impartiality is a *property of the message and not of its author*. This follows the intuition that the knowledge of the affiliation of a message's author is not sufficient to determine whether the message is impartial or not. As exemplified in Section 1, authors from same affiliation can write both impartial and partial messages (examples (2) and (3) respectively) on the same topic.
2. Message impartiality *is not an absolute notion and is defined with respect to a given context*: a topic of discussion and a given set of (two or more) relevant affiliations (liberal and conservative in our running example). Importantly, a message can be considered impartial in one context and partial in another. For example, the message "Broncos is the best team in the world" is blatantly partial in the context of the Super Bowl but impartial in the context of the government shutdown.
3. Impartiality of a message is *not related to the truth value of the message*. A true statement can be partial or impartial, depending on the context in which partiality is being determined. For example, the statement "Greenhouse gases have risen", while being true, might still convey liberal partiality in the context of the global warming debate. The reason is that this statement represents the liberal point of view on the topic and not the perspective of conservatives, who generally deny the true statement (McCright and Dunlap 2000).[2]

Finally, we note that while the concepts of impartiality and sentiment seem related, high sentiment in a message does not always imply high partiality and vice versa. We discuss this further in Section 5.5.

2.2 Operational definition of impartiality

While the notion of impartiality described above is widely and intuitively understood (Gert 1995), we still lack a direct way to measure the degree of impartiality of a given message. We now propose an operational definition of this concept that allows it to be objectively quantified:

[1] Where, by operationalization, we mean making an intuitive fuzzy concept measurable, as it is common in social sciences.

[2] *"Reality has a well known liberal bias."* — Stephen Colbert (at the 2006 White House correspondents' dinner).

Given a context (e.g., U.S. government shutdown) and a set of two possible affiliations (e.g., liberal and conservative), the impartiality of a message refers to the uncertainty (or lack of confidence) in inferring the likely affiliation of its author from the content of the message. *That is, the harder it is to infer the affiliation of a message's author, the more impartial that message is.*

Applying this definition to our running examples, message (2) would be deemed impartial since the political affiliation of the author (liberal) cannot be inferred from the text. On the other hand, the content of message (1) suggests that the author was most probably a person with liberal affiliation and hence it would be deemed partial.

Note that while this definition can can be intuitively extended to events with more than two affiliations, for the sake of simplicity, in this work we *only consider events with two affiliations*. Having more affiliations may give rise to complications (e.g., in a n-party system accommodating a scenario where multiple affiliations share the same perspective) that would require a dedicated analysis. Hence, we leave it as an avenue to be explored thoroughly in future work.

3 Validating the operational definition of impartiality using human annotation

So far, our operationalization of message impartiality has been agnostic regarding *how* the confidence in discerning author affiliation of a message is estimated. In this section we will use human annotators to assess *affiliation-discernibility* in order to establish the validity of our approach. In the next section, we will propose a method for quantifying discernibility *without* relying on human labels, therefore, fully automating our method for estimating message impartiality.

3.1 Determining affiliation-discernability using human annotators

Task design (`HAT-a`). Assume a set of messages coming from authors with two different affiliations A_1 and A_2. The goal is to evaluate how impartial each message is.

For each message we ask n different human judges to guess the likely affiliation of the author of the message. We estimate the confidence in discerning author affiliation as the agreement level among the judges: the higher the agreement, the easier it is to discern the affiliation of the author and thus, the more partial the message. Formally, the agreement score for the message can be computed as:

$$\alpha = max(\alpha_1, \alpha_2), \tag{1}$$

where α_i is the number of humans that inferred the author affiliation to be A_i. For highly partial messages, we expect the human judges will have very high agreement (α close to n), whereas for impartial messages, since the messages themselves do not provide any cues about author affiliation, the judges would be making random guesses, so the agreement should be close to $\lceil \frac{n}{2} \rceil$.

Task implementation. Having specified the design of the task, we implement it to quantify message impartiality in a real-world dataset. The dataset consists of tweet messages related to the discussions of two highly polarized events: 1) U.S. government shutdown, posted by users with *known* Democratic or Republican affiliation and 2) Super Bowl 2014, posted by *known* fans of Denver Broncos or Seattle Seahawks (the two contestant teams). Details of how this data was collected are given in Section 5.1.

For each of the two events, we select 400 messages (200 random messages from each affiliation) and for each message, we ask $n = 10$ different human judges to infer the affiliation of the author of the message.

To get a set of reliable human judges, we use Amazon Mechanical Turk platform. We specifically use AMT 'categorization masters' who have a record of performing tasks with high accuracy. Moreover, since we are focusing on events pertaining to U.S. politics and sports, the workers were chosen to be exclusively from the U.S. Each AMT judge was shown a set of 85 tweet messages related to one of the events. For example, the question for annotating one message was:

Event: U.S. government shutdown of 2013

Tweet: We can't stop the rain but together we can stop this shutdown. Enough is enough. #JustVote to reopen the govt

Question: The user posting this tweet is:

◇ Democratic-leaning

◇ Republican-leaning

Judges were given 50 minutes to complete the task and the reward for each task was set to 5 USD. To ensure quality control, 5 out of 85 messages (placed at random points) in each set were designated as 'test messages' and were chosen such that their author affiliation was very clear. If a judge did not infer correct author affiliation for at least 4 out of these 5 test messages, we discarded all of their responses.

Table 1 shows some of the messages where all the judges inferred the same author affiliation ($\alpha = 10$) and where only around 50% judges ($5 \leq \alpha \leq 6$) inferred the same affiliation: messages where affiliations were inferred with high confidence do indeed seem partial, while ones where affiliations were inferred with low confidence appear to be impartial.

3.2 Alignment with human intuition

We now validate the methodology described above, by checking whether the resulting message impartiality scores align well with how humans intuitively perceive impartiality. To this end, we collect direct impartiality judgments using the following human annotation task.

Labeling message impartiality directly (`HAT-direct`). For all the 400 messages for which we computed impartiality scores using the method described above, we also gather "ground truth" (im)partiality ratings by asking 10 different human annotators to directly judge whether the messages are partial or impartial. Since we want these labels to reflect the human judges' intuitive perception of impartiality and not any specific interpretation, we do not define impartiality or provide illustrative examples. Instead we simply ask whether the messages are impartial or not with respect to

Event	Known author affiliation	100% of judges agreeing on the author affiliation ($\alpha = 10$)	Around 50% of judges agreeing on the author affiliation ($5 \leq \alpha \leq 6$)
Shutdown 2013	Democrat	House Republicans just delivered Americans dysfunction, partisanship, and a #GOPshutdown. RT to tell GOP to drop their extreme demands.	O'Malley says Maryland will consider tapping reserve fund to deal with federal shutdown [url]
	Republican	Its Day 4 of the Democrat's shutdown & Republicans have been hard at work finding solutions. Check out this timeline: [url]	Tonight, our elected officials—or at least their staffers—are holding some very caffeinated discussions. CNN 11pm #shutdown live now.
Super Bowl 2014	Broncos	Here we go! Let this be the ULTIMATE tale of 2 halfs! #GoBroncos #ComebackTime	Been calling a Seattle vs Denver Super Bowl for months btw.
	Seahawks	We've waited 37 years to say this: Congratulations on your Super Bowl XLVIII win @Seahawks! #GoHawks #Champions #SeattlePride	Entering #SB48, the fewest points Broncos scored w/ Peyton Manning was 17. Denver scored 8 on Sunday.

Table 1: Example messages where all/half of the judges agreed on their inferred author affiliation.

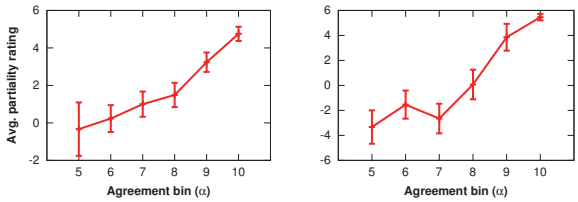
the event and the respective sides.[3] This way, each message receives a *ground truth partiality rating* ranging from –10 to 10, corresponding to the difference between the number of partial and impartial judgments it receives from different annotators.[4]

To see if the impartiality scores α based on affiliation-discernability (as obtained via `HAT-a`) align with the humans' intuition of impartiality, we bin the messages into 6 different *agreement bins* corresponding to all possible α values, ranging from 5 through 10. Essentially, these bins correspond to the degree of agreement that the annotators had in inferring the author affiliation in `HAT-a`: the agreement bin of 10 contains all the messages for which annotators inferred the same affiliation, while an agreement bin of 5 contains messages for which half of the annotators inferred one affiliation and the other half inferred the other affiliation. Figure 1 shows the average ground truth partiality ratings for each of these bins, as obtained via `HAT-direct`; for both events the impartiality scores α align well with the ground truth. [5] This validates our operational definition of impartiality, showing that the confidence with which the affiliation of the author of a message can be discerned corresponds to the intuitive perception of that message's impartiality. Building on this finding, in the next section we show how to automate the task of estimating the discernibility of author affiliation.

4 Automating affiliation-discernibility

So far we have relied on human judgments to determine whether the affiliation of the author of a message can be easily inferred, and consequently, whether that message is partial or not. However these methods are not scalable and can not be applied to real-time emerging events for which annotations are not available. We now show how to take humans out of the loop by fully automating the affiliation-discernability component of our paradigm.

Prior work has shown that messages contain cues about the affiliation of their authors (Monroe, Colaresi, and Quinn



(a) Govt. shutdown 2013 (b) Super Bowl 2014

Figure 1: Our estimation of affiliation-discernibility (obtained via `HAT-a`, x-axis) aligns well with humans' intuitive perception of message impartiality (ground truth partiality ratings obtained via `HAT-direct`, y-axis).

2008; Borge-Holthoefer et al. 2015), and that these can exploited by machine learning classifiers to infer the author affiliation automatically. While our goal is not to infer affiliation, our insight is that the posterior class probabilities that these methods assign to each particular message can be used to determine how easy it is to discern their author's affiliation. Following our operational definition, measuring message impartiality is reduced to inferring class probabilities in author affiliation classification task. Next we explore several supervised techniques for inducing class probabilities, which we train on a relatively small seed set of messages for which author affiliations are known. We note that, crucially, as opposed to message impartiality labels, author affiliation labels are often readily available in social media systems and are generally persistent across events. *This allows our paradigm to be applied in real-time to emerging events without requiring any ad-hoc annotation.*

Posterior class probabilities. We start by using two supervised classification algorithms trained to predict user affiliations: **naive bayes** and bootstrap aggregating with decision trees, or **bagged trees**. Both of these estimate the posterior class probabilities of the classified items; we will use the maximum class probability for each message as its automatically inferred partiality score (henceforth referred to as *mp-score*). Naive bayes is a popular choice for text classification since it can efficiently handle items with sparse features. Handling sparse features is specially important since tweet messages are very short—maximum of 140 characters—and

[3] We also give the annotators the option to select "can't tell" to discourage random guesses; this was rarely selected (6% of labels).

[4] Data available at impartiality.mpi-sws.org.

[5] The Spearman rank correlation between the impartiality scores α and the ground truth partiality ratings are 0.34 ($p < 10^{-11}$) for government shutdown and 0.45 ($p < 10^{-20}$) for Super Bowl 2014.

a single message contains only a few features. On the other side, bagged trees have been shown to produce more accurate posterior class probability estimates (Niculescu-Mizil and Caruana 2005).

Unigram method. It is known that posterior class probability estimates provided by machine learning classifiers are often skewed and hard to interpret (Niculescu-Mizil and Caruana 2005). If available, weight vectors are not always representative of the relative importance on the features (King, Tomz, and Wittenberg 2000). However, in many practical applications, understanding *why* a message is labeled as being partial is desirable (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013).

Here we propose a simple unigram based method that is specifically tailored to the task of inferring the discernibility of author affiliation. It is based on the observation that some words can serve as cues of author affiliation, and that the strength of their signal can be quantified (Monroe, Colaresi, and Quinn 2008). Messages including cues with strong affiliation-signals have easy to discern authorship, and thus are partial. In its simplicity this method has the additional advantage of high interpretability, making the link between content and (im)partiality transparent.

The unigram method for assigning *mp-score* to individual messages consists of two main steps: 1) assigning a score to all the unigrams used in all the messages in the discussion, and 2) computing the *mp-score* of a message based on the unigrams contained in it.

The process of assigning score to unigrams works as follows. Consider a discussion consisting of two different affiliations A_1 and A_2; as before, we assume to have a training corpus of messages posted by a seed set of authors with known affiliations. For each unique unigram u in the corpus, we compute the score of u towards an affiliation A_i as:

$$s_{u,A_i} = p_{u,A_i} \times log(\frac{p_{u,A_i}}{q_{u,A_i}}), \quad (2)$$

where p_{u,A_i} is the fraction of authors with known affiliation A_i using u and q_{u,A_i} is the fraction of authors with a different known affiliation ($\{A_1, A_2\} \setminus \{A_i\}$) using u. [6]

We compute the final score s_u of the unigram u as:

$$s_u = max(s_{u,A_1}, s_{u,A_2}). \quad (3)$$

We can now assign a message partiality score *mp-score* to any message related to the discussion, even if it is not uttered by authors with known affiliations in our training dataset. Consider message m that consists of z unigrams $(u_1, u_2, ..., u_z)$, and the corresponding scores of these unigrams are given as $(s_{u_1}, s_{u_2}, ..., s_{u_z})$, then the *mp-score* of the message is:

$$\textit{mp-score}_m = max(s_{u_1}, s_{u_2}, ..., s_{u_z}), \quad (4)$$

where the max function here uses the unigram with the highest score (or the most discerning unigram).

[6] Among multiple alternative methods for scoring unigrams, we choose this one for its simplicity. For example, χ^2 feature selection leads to qualitatively similar results.

Discussion	Num. of tweets
U.S. government shutdown	52,017
Sandy Hook school shooting	56,535
Zimmerman trial 2014	11,174
Immigration reform 2014	16,883
2015 SCOTUS ruling on SSM	9,316
Super Bowl 2014	100,507
Super Bowl 2015	50,042

Table 2: Number of tweets for different discussions in our dataset that are authored by users with known affiliations.

5 Evaluation

In this section, we evaluate the relative performance of our automated methods in quantifying message impartiality, as well as their effectiveness in retrieving impartial messages in the scenario of an emerging event. To this end, we gathered a dataset related to political and sports discussions on Twitter.

5.1 Dataset

Our dataset consists of Twitter discussions around five major events related to U.S. politics[7] and two major sporting events: Super Bowl 2014 and 2015 (Table 2).

For the politics-related events, we obtain a seed set of users with known Democratic or Republican affiliations by leveraging twitter lists (Ghosh et al. 2012). For the sports events, we use keyword matching on the profile descriptions of the users. For example, the two teams competing in Super Bowl 2014 were Denver Broncos and Seattle Seahawks, so we take 'broncos' or 'seahawks' in the profile of a Twitter user to mean that this user is a supporter of (or affiliated with) Denver Broncos or Seattle Seahawks respectively.

In order to collect messages related to a particular event, we select tweets posted around the time when the event occurred by authors from both affiliations and filter them using simple regular expressions. For example, we filtered tweets related to Twitter discussion on President Obama's executive action on immigration reform with the keyword pattern ('*immig*') from the tweets posted between Oct. 26, 2014 to Dec. 05, 2014. [8]

5.2 Implementation details

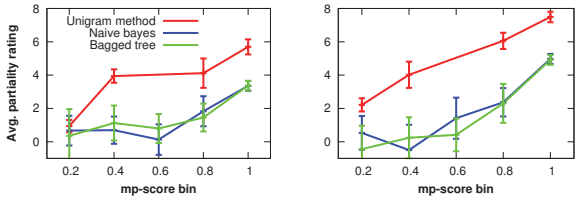
As features for the **bagged tree classifier**, we use all the unigrams that appear more than 5 times in the messages related to the discussion. As an example, for the government shutdown discussion we obtain a total of 7,994 features, and for Super Bowl 2014 discussion we get 5,336 features. For training the classifier, we use standard parameters (10 base estimators, all samples and all features used for training each base estimator, samples drawn with replacement). We train

[7] We selected these events by scanning the list of important events in a year compiled by Wikipedia editors and extracting ones with reasonably large participation from politics-related accounts.

[8] The immigration reform was formally announced on Nov. 20, 2014, however, there was considerable debate on social media before and after the announcement.

Discussion	Dems/Broncos	Reps/Seahawks
Government shutdown	#GOPShutdown	#HarryReidsShutdown
	#DemandAVote	#Obamashutdown
	economy	golf
	hurting	priests
Super Bowl 2014	#TimeToRide	#GoHawks
	#GoBroncos	#12s
	orange	champs
	congratulate	thanks

Table 3: Examples of highly partial unigrams related to political/sports discussions according to the unigram method.



(a) Government shutdown (b) Super Bowl 2014

Figure 2: The automatically inferred partiality scores (*mp-score*, x-axis) align well with human intuition (ground truth partiality obtained via `HAT-direct`, y-axis). Bins with less than 10 data-points are ignored.

the bagged tree classifier for each discussion, [9] and for each message, we use the maximum predicted class probability by the trained classifier as its *mp-score*.

Since **naive bayes classifier** can handle a large number of features, we include bigram features as well. That is, for training the naive bayes classifier, we use all the unigrams and bigrams that appear more than 5 times in the training set. We obtain a total of 15,640 features for naive bayes in the case of government shutdown discussion and 9,558 features for Super Bowl 2014. Similar to the case of the bagged trees, we train the naive bayes classifier [10] and use the maximum predicted class probability of the messages as their *mp-score*.

In the **unigram method** we assign scores to all unigrams in the corpus using Equation (3). Table 3 shows some unigrams with high scores, and offers some insights into the partisanship of the discussions: the Democrats are blaming the Republicans for *hurting* the *economy*, while Republicans are blaiming *Obama* and are outraged that his favorite *golf* courses were spared from the shutdown.

5.3 Alignment with ground truth ratings

In order to measure the performance of our automated methods, we check the relation between the *mp-score* assigned by these methods and the ground truth partiality ratings. We divide messages into five different bins based on the scores assigned by each method: the first bin contains messages deemed to be most impartial (the bottom 20^{th} percentile of all *mp-scores*), while the last bin contains messages considered most partial.

Figure 2 shows that the (im)partiality scores inferred by our automatic methods align well with human perception for messages related to both the Shutdown and Super Bowl events. [11] Note that our affiliation-discernability methods do not rely on any human-generated impartiality labels. We also find that the implementations based on machine learning classifiers do not distinguish very well between messages that are impartial and those that are mildly partial, which can be attributed to the skewness in the estimation of posterior class probabilities (as previously discussed); in particular, the median class probability is 0.98 for naive bayes and 1.00 for bagged trees.

In the remainder of this paper we will use the unigram method to study message impartiality at scale. Beyond providing a better alignment with human intuition, this method also has the advantage of being transparent in terms of the content features that contribute to the impartiality score.

5.4 Applicability to emerging events

One major advantage of our framework is that it can be used to quantify impartiality of messages related to emerging events, as long as they involve the same pair of affiliations. In contrast, directly applying a supervised machine learning method would involve regenerating partiality labels for messages related to emerging events. Here we provide empirical evidence for the advantage of our methodology based on affiliation-discernability (henceforth *discernability method*) by comparing it to a supervised machine learning method trained directly on impartiality labels (henceforth *direct ML method*) to an event that was not previously used in any of our analysis: the 2015 Supreme Court ruling on same-sex marriage (henceforth *SCOTUS ruling*).

Since we treat the SCOTUS ruling as an emerging event, neither of the methods has access to any new information or labels pertaining to this event. In this scenario, we compare the performance of our affiliation-discernability method with that of an SVM classifier directly trained on human-annotated (im)partiality labels (gathered using `HAT-direct`) for messages related to a prior event, namely, the government shutdown.[12]

For evaluation, we collect ground truth partiality labels for 400 messages from the SCOTUS ruling (200 random messages from each affiliation) using `HAT-direct` and compare the two methods on the tasks of retrieving both partial and impartial messages. We rank the messages based on class probability given by the SVM in the *direct ML*

[9] The 10-fold cross validation accuracy was 80.3% for shutdown discussion and 86.7% for Super Bowl 2014.

[10] The 10-fold cross validation accuracy was 87.1% for shutdown discussion and 88.5% for Super Bowl 2014.

[11] The Spearman rank correlation between the *mp-score* assigned by these implementations and the ground truth partiality are 0.41, 0.25 and 0.20 for unigram method, naive bayes and bagged trees, respectively, for government shutdown (p-values $< 10^{-6}$); for Superbowl 2014, these are 0.45, 0.33 and 0.36 (p-values $< 10^{-11}$).

[12] The cross-validation accuracy of the SVM on the government shutdown event is 77%.

Method	Partial	Impartial
Direct ML method	0.95	0.10
Discernability method	1.00	0.60

Table 4: Precision@20 for the task of retrieving partial and impartial messages in an emerging event. Our affiliation-discernability method outperforms a machine learning method that is trained directly on impartiality labels.

method [13] and, respectively, the *mp-score* of the *discernability method*. For both methods, we compute the precision@20, a standard information retrieval metric used for evaluating retrieval tasks.[14]

The results are shown in Table 4: both methods have high precision@20 for retrieving partial messages. However, for the task of retrieving impartial messages, the *discernability method* has a much higher precision@20 than the *direct ML method*. This confirms our intuition that while affiliation-discernability transfers well from one event to another, this is not the case for direct supervised methods of measuring impartiality, as they rely on event-specific cues and thus require new impartiality-annotated data for each new event.

5.5 Limitations and discussion

Our evaluation of the three implementations of our automated methodology shows that they can measure impartiality reasonably well on real-world datasets related to political and sporting events. As a first attempt to quantify this phenomenon, this framework has some important limitations. For example, it can fail in particular cases involving the use of creative language, such as sarcasm, and there is room for introducing better linguistically informed methods for quantifying impartiality. For example, one could devise methods that leverage negations, hedges, discourse markers and conversational patterns to infer author affiliation more robustly. Similarly developing better machine learning techniques that provide more realistic class probability estimates could also lead to better performance. Since such improvements fall beyond the scope of this paper, we leave it as a venue to be explored in future work.

We showed that an advantage of our methodology stems from the fact that affiliation labels are more persistent across events, as opposed to message-level impartiality annotations. For example, there are many events related to U.S. politics (U.S. government shutdown 2013, Sandy Hook shootings, George Zimmerman's trial, etc.) for which one could use same seed set of users with known Democratic and Republican affiliations to quantify impartiality in messages related to any of these discussions. However, there might be cases where individual users change affiliations, either following a major event (2015 attacks on Charlie Hebdo offices) or gradually over time (acceptance of gay marriage in a community). In these cases, one would need to re-annotate author affiliations to quantify impartiality. Even so, annotating author affiliations for each event is still more efficient than annotating individual messages, since often in social media, the number of users is significantly less than the number of messages that they are producing.

[13] Using bagged tree and naive bayes instead of SVM in direct method also produces very similar results.

[14] We consider messages with a ground truth partiality rating greater than 2 (smaller than -2) as being partial (impartial).

Discussion	Avg. partiality rating	
	Impartial	Partial
Government shutdown	-1.2	4.7
Super Bowl 2014	-0.6	5.8

Table 5: The average ground truth partiality ratings for messages that fall in our binary (im)partiality bins.

Finally, we explore the extent to which the (im)partiality of a message is related to the sentiment it expresses. We take 100 most partial and 100 most impartial tweets—as identified by the unigram method—posted by both Democratic and Republican leaning users from the discussion related to the U.S. government shutdown 2013. We labeled the sentiment of each tweet using a state-of-the-art Twitter sentiment classifier (Gonalves et al. 2013). We find that the average sentiment score for partial and impartial tweets posted by Democrats is -0.22 and -0.25 respectively, whereas, the respective average sentiment score for partial and impartial tweets from Republicans is -0.28 and -0.20. The negative sentiment was expected, as the shutdown was generally perceived as a negative event. Similar sentiment score for both partial and impartial tweets (from both Democratic and Republican sides) suggests that, at least in this case, the degree of impartiality of a message is not tightly related to the sentiment contained within.

6 Case study: Impartiality in Twitter discussions

We now apply the methodology developed so far to characterize impartiality in Twitter discussions at scale. Specifically, we ask the following questions about the dynamics of impartiality:

- Does partial content spread more virally than impartial content?
- Does the level of message (im)partiality change during the course of a discussion?
- Do individual users change their levels of (im)partiality during the course of a discussion?

Does partial content spread more virally than impartial content? We compare the virality of partial and impartial messages in our dataset, where virality of a message is taken to be the number of times it was retweeted. [15]

So far, we have used a continuous measure of message impartiality. For the purpose of this analysis we assign binary partial/impartial labels to messages based on the distribution of the *mp-score* in the labeled part of the data. We mark all

[15] For the tweets considered here, we re-gathered their retweet counts from Twitter API at least one month after the discussion to make sure that the retweet counts have stabilized.

Discussion	# Users	Avg. of PAR	Avg. of IMP	p-value
Government shutdown	231	$0.162(\textit{22.3})$	$-0.331(\textit{14.4})$	$< 10^{-14}$
Sandy Hook Shooting	120	$0.182(\textit{20.5})$	$-0.191(\textit{12.1})$	$< 10^{-6}$
Zimmerman Trial	67	$0.143(\textit{30.4})$	$-0.149(\textit{15.1})$	< 0.01
Super Bowl 2014	339	$0.010(\textit{5.9})$	$-0.141(\textit{2.2})$	< 0.05

Table 6: Partial messages are twice more likely to be retweeted: p-values for Wilcoxon signed-rank test for the comparison of partial (PAR) and impartial (IMP) tweets. For all the events, the normalized virality (z-score) is above the per-user average for partial messages (> 0) and below per-user average for impartial messages (< 0). The un-normalized virality (shown in parentheses) for partial messages is roughly double that of impartial messages.

messages with *mp-score* < 0.1 as *impartial* and all those with *mp-score* ≥ 0.4 as *partial*. As shown in Table 5, this binary binning aligns well with human intuition.

To analyze the impact of impartiality on message virality, we need to disentangle it from the popularity of the topic in which a message appears and from characteristics of its author. For example, highly popular users (having millions of followers) are likely to be retweeted more than less popular users, regardless of what they tweet. Similarly, a highly popular discussion will inherently attract more retweets.

To control for these confounding factors, we compare the virality of partial and impartial messages posted by the same user within a given discussion. To this end, we standardize the retweet counts of all the messages of a user by converting individual retweet counts to their z-score,[16] after discarding users that did not post at least one partial and one impartial message during the discussion. This leads to a within-user paired comparison to which each user contributes equally: for each user we take the average virality z-score of their partial messages (PAR), and the average virality z-score of their impartial messages (IMP), and compare the resulting two sets using Wilcoxon signed-rank test. The results of this paired statistical test (Table 6) show that partial messages are indeed significantly more likely to be retweeted than impartial ones, even when they are sent by the same author and are on the same topic.

Does the level of message (im)partiality change during the course of a discussion? To answer this question, we analyze the average daily message partiality in the discussion of U.S. government shutdown of 2013—an event that went through several offline developments. The results, presented in Figure 3, show that the discussion stays somewhat impartial untill Sep. 26, 2014. However, we see a sharp rise in partiality on Sep. 27, corresponding to the U.S. Senate amending a bill related to Affordable Care Act, a move which would eventually lead to the Oct. 1 shutdown. Similarly, we see an increase in partiality immediately after the start of shutdown. Further fluctuations in partiality can also be mapped to offline developments. Finally, after the shutdown ended on Oct. 16, we can see gradual decrease in partiality suggesting that the discussions returns to a relatively impartial point after the issue has been resolved.

[16]Z-score of a sample indicates the signed number of standard deviations that the sample is away from the population mean. Converting a retweet count to z-score ensures that popular users with high retweet counts do not skew the analysis.

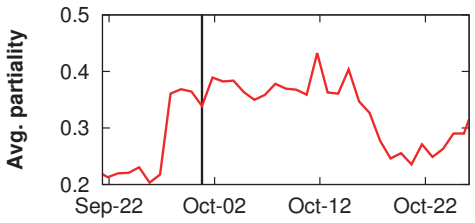


Figure 3: Fluctuations in partiality (*mp-score*) over the development of the Shutdown 2013 event; solid line indicates October 1st, the day the government shut down.

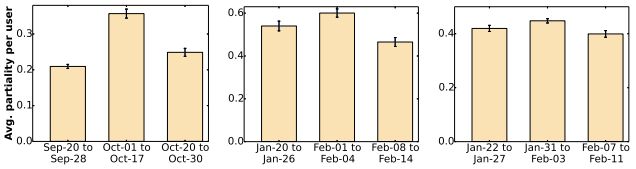
Do individual users change their levels of (im)partiality during the course of a discussion? As exemplified in the introduction, an author with a given affiliation can post both partial and impartial messages. Here, we study this phenomenon more generally and investigate whether individual users in Twitter vary their levels of (im)partiality over time.

In order to study the change in individual users' (im)partiality over time, we consider three discussions: U.S. government shutdown of 2013, Super Bowl 2014 and Super Bowl 2015. We divide these discussions into three different time intervals: before the event, during the event and after the event. For users with known affiliations (Democrats / Republicans, Broncos / Seahawks, Patriots / Seahawks) *who have posted messages in all three intervals*, we calculate their average partiality (the average *mp-score* of the messages that they posted) during each interval.

Figure 4 shows the changes in average user partiality during the three events. It can be seen that in all three discussions, the average partiality is high around the time when the event was taking place. This confirms our intuition that even users with set affiliations can vary their level of partiality during the course of a discussion.

7 Related work

A large body of research has focused on detecting polarization in various offline and online contexts. For example, studies have analyzed congressional votes in order to understand polarization in the context of U.S politics (McCarty, Poole, and Rosenthal 2006, inter alia). Similarly, other studies have shown that political discussions in online world occur within highly polarized communities of users sharing similar political leanings or ideologies, with relatively little exchange of information between commu-



(a) Shutdown 2013 (b) Super Bowl 2014 (c) Super Bowl 2015

Figure 4: Change in user partiality (*mp-score*) around major events; each bin reflects the partiality of messages posted by the same group of users during the three intervals. Consistently across the three events, users post more partial messages around the time the event takes place (middle bars).

nities (Adamic and Glance 2005; Conover et al. 2011b; Weber, Garimella, and Batayneh 2013). Some prior studies have explored the bias of media sources using human annotations (Yano, Resnik, and Smith 2010; Budak, Goel, and Rao 2015) or unsupervised methods (Niculae et al. 2015b). In this work, we provide a finer-grained perspective on polarization by designing a framework that operates at the granularity of individual messages.

Another line of research has focused on detecting the political leaning of individual users in social media (Conover et al. 2011a; Pennacchiotti and Popescu 2011; Zamal, Liu, and Ruths 2012; Wong et al. 2013). As discussed in the introduction, the problem of measuring the impartiality of an individual message is crucially different from that of detecting its author affiliation.

A number of studies have used difference in language usage to detect polarity. For example, our unigram implementation is directly inspired by a comparison of methods capturing partisanship-inducing words in senate speeches (Monroe, Colaresi, and Quinn 2008). Another study pointed out that difference in usage of a hashtag by two parties can be used to measure political leaning (Weber, Garimella, and Teka 2013). In the context of Egyptian politics language differences are used to classify tweets into pro and anti-military categories (Borge-Holthoefer et al. 2015). We use insights from these studies to propose an automated way to quantify message impartiality.

The problem of detecting impartiality in a message is also different from detecting sentiment (Pang, Lee, and Vaithyanathan 2002) and subjectivity (Wiebe et al. 2004). While earlier works have tried to use sentiment analysis to quantify political polarity of messages (Wong et al. 2013, inter alia), and showed that news articles with high sentiment tend to attract more popularity (Reis et al. 2015), we show that the sentiment of a message does not necessarily correspond to its partiality. Similarly, the lack of any subjectivity cues in our introductory example (1) illustrates how the problem of quantifying impartiality is distinct from detecting subjectivity, a point that is addressed in more detail in (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013).

Finally, earlier studies focusing on impartial language (Yano, Resnik, and Smith 2010; Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013) have relied on labeled data, such as the Wikipdia NPOV corpus. In this work, we devise a framework that is applicable in settings where impartiality labels are not available, such as emerging events, enabling a large scale investigation of the phenomenon.

8 Conclusion and future work

In this paper, we made the case for a finer-grained approach to polarization in social media by analyzing the impartiality of individual messages. While the notion of message impartiality is intuitive, a lack of automated ways to measure it has hindered scientific inquiry so far. Against this background, we proposed an operationalization of the intuitive notion of message impartiality based on affiliation-discernability. We validated that the resulting measure aligns with human judgments of impartiality and then proposed ways in which the methodology can be automated. Since this method relies on a seed set of known author affiliations, which are generally more persistent and more easily obtainable than direct message partiality labels, our framework can be applied in real time to emerging events without requiring new annotations. Having an automated measure for judging message impartiality enabled us to conduct a large-scale study of message impartiality in Twitter discussions. This provided new insights not only into how impartiality of a message impacts its chances of spreading, but also on how impartiality of social media discussions (and of single users) varies over time.

Our work also opens numerous opportunities for future studies. First, having a reliable distinction between partial and impartial messages can be used to inform social media users about the degree of impartiality of the content they are consuming. This distinction could be leveraged for the task of balanced news reporting.

The performance of our automated method could be improved further by better integrating linguistic information and by using machine learning approaches that are better suited for measuring affiliation-discernability. To make our approach more general, we envision unsupervised methods that can jointly learn to detect message impartiality and author affiliations. To this end, we are distributing the annotated data to encourage future work on this phenomenon.[17]

In future work we plan to explore the nature of the interplay between impartiality and related concepts like disagreement (Allen, Carenini, and Ng 2014; Wang and Cardie 2014), credibility (Gupta et al. 2014) and rationality (Liu and Weber 2014). Understanding the role that impartiality plays in the outcome of consequential interactions, such as teamork (Niculae et al. 2015a) and debates (Romero et al. 2015; Tan et al. 2016; Zhang et al. 2016) is also an interesting avenue for future work.

Acknowledgments We thank the anonymous reviewers for their helpful suggestions and Evgeniy Gabrilovich, Ravi Kumar, Alexandru Niculescu-Mizil, Noah Smith and Arthur Spirling for their insightful comments. This work was supported in part by a Google Faculty Research Award.

[17] Available at impartiality.mpi-sws.org.

References

Adamic, L., and Glance, N. 2005. The political blogosphere and the 2004 US election: Divided they blog. In *Proc. LinkKDD*.

Allen, K.; Carenini, G.; and Ng, R. T. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proc. EMNLP*.

Bakshy, E.; Messing, S.; and Adamic, L. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*.

Balasubramanyan, R.; Cohen, W. W.; Pierce, D.; and Redlawsk, D. P. 2012. Modeling polarizing topics: When do different political communities respond differently to the same news? In *Proc. ICWSM*.

Borge-Holthoefer, J.; Magdy, W.; Darwish, K.; and Weber, I. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proc. CSCW*.

Budak, C.; Goel, S.; and Rao, J. M. 2015. Fair and balanced? Quantifying media bias through crowdsourced content analysis. In *Proc. ICWSM*.

Conover, M. D.; Goncalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011a. Predicting the political alignment of Twitter users. *IEEE Xplore*.

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011b. Political polarization on Twitter. In *Proc. ICWSM*.

Gert, B. 1995. Moral impartiality. *Midwest studies in Philosophy*.

Ghosh, S.; Sharma, N.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Cognos: Crowdsourcing search for topic experts in microblogs. In *Proc. SIGIR*.

Gonalves, P.; Arajo, M.; Benevenuto, F.; and Cha, M. 2013. Comparing and combining sentiment analysis methods. In *Proc. COSN*.

Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014. TweetCred: A real-time web-based system for assessing credibility of content on Twitter. In *Proc. SocInfo*.

King, G.; Tomz, M.; and Wittenberg, J. 2000. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*.

Koutra, D.; Bennett, P.; and Horvitz, E. 2015. Events and controversies: Influences of a shocking news event on information seeking. In *Proc. WWW*.

Liu, Z., and Weber, I. 2014. Is Twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In *Proc. SocInfo*.

McCarty, N. M.; Poole, K. T.; and Rosenthal, H. 2006. *Polarized America: The dance of ideology and unequal riches*. MIT Press Cambridge.

McCright, A. M., and Dunlap, R. E. 2000. Challenging global warming as a social problem: An analysis of the conservative movement's counter-claims. *Social Problems*.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*.

Niculae, V.; Kumar, S.; Boyd-Graber, J.; and Danescu-Niculescu-Mizil, C. 2015a. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proc. ACL*.

Niculae, V.; Suen, C.; Zhang, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015b. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proc. WWW*.

Niculescu-Mizil, A., and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proc. ICML*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. EMNLP*.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to Twitter user classification. In *Proc. ICWSM*.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *Proc. ACL*.

Reis, J.; Benevenuto, F.; Olmo, P.; Prates, R.; Kwak, H.; and An, J. 2015. Breaking the news: First impressions matter on online news. In *Proc. ICWSM*.

Romero, D. M.; Swaab, R. I.; Uzzi, B.; and Galinsky, A. D. 2015. Mimicry is presidential. *Personality and Social Psychology Bulletin*.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proc. WWW*.

Wang, L., and Cardie, C. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proc. ACL*.

Weber, I.; Garimella, V. R. K.; and Batayneh, A. 2013. Secular vs. Islamist polarization in Egypt on Twitter. In *Proc. ASONAM*.

Weber, I.; Garimella, V. R. K.; and Teka, A. 2013. Political hashtag trends. In *Proc. ECIR*.

Wiebe, J.; Wilson, T.; Bruce, R.; Bell, M.; and Martin, M. 2004. Learning subjective language. *Computational Linguistics*.

Wong, F. M. F.; Tan, C. W.; Sen, S.; and Chiang, M. 2013. Quantifying political leaning from Tweets and Retweets. In *Proc. ICWSM*.

Yano, T.; Resnik, P.; and Smith, N. A. 2010. Shedding (a thousand points of) light on biased language. In *Proc. NAACL HLT*.

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proc. ICWSM*.

Zhang, J.; Kumar, R.; Ravi, S.; and Danescu-Niculescu-Mizil, C. 2016. Conversational Flow in Oxford-style Debates. In *Proc. NAACL*.