We Don't Speak the Same Language: Interpreting Polarization through Machine Translation

Ashiqur R. KhudaBukhsh[*1], **Rupak Sarkar**[*2], **Mark S. Kamlet**[1], **Tom M. Mitchell**[1]
[1]Carnegie Mellon University
[2]Maulana Abul Kalam Azad University of Technology
akhudabu@cs.cmu.edu, rupaksarkar.cs@gmail.com, kamlet@cmu.edu, tom.mitchell@cs.cmu.edu

Abstract

Polarization among US political parties, media and elites is a widely studied topic. Prominent lines of prior research across multiple disciplines have observed and analyzed growing polarization in social media. In this paper, we present a new methodology that offers a fresh perspective on interpreting polarization through the lens of machine translation. With a novel proposition that two sub-communities are speaking in two different *languages*, we demonstrate that modern machine translation methods can provide a simple yet powerful and interpretable framework to understand the differences between two (or more) large-scale social media discussion data sets at the granularity of words. Via a substantial corpus of 86.6 million comments by 6.5 million users on over 200,000 news videos hosted by YouTube channels of four prominent US news networks, we demonstrate that simple word-level and phrase-level translation pairs can reveal deep insights into the current political divide – what is *black lives matter* to one can be *all lives matter* to the other.

Introduction

One mans meate is another mans poyson.
– Thomas Draxe; *Bibliotheca Scholastica*; 1616.

Polarization among US political parties (Poole and Rosenthal 1984; Layman et al. 2010; McCarty, Poole, and Rosenthal 2016; Baldwin and Lammers 2016; McConnell et al. 2017), media (Hollander 2008; Stroud 2011) and elites is a widely studied topic. Studies have shown that over the last 30 years, both Democrats and Republicans have become more negative in their views toward the opposition party (Iyengar, Sood, and Lelkes 2012). Further, behavioral studies indicate that such negative views have affected outcomes in settings as diverse as allocating scholarship funds (Iyengar and Westwood 2015), mate selection (Huber and Malhotra 2017), and employment decisions (Gift and Gift 2015). Prominent lines of prior research across multiple disciplines have observed and analyzed growing polarization in social media (Demszky et al. 2019; Darwish 2020a; Bakshy, Messing, and Adamic 2015; Darwish 2020b), and previous studies have reported substantial partisan and ideological divergence in both content and audience in major US TV news networks (Stanley 2012; Bozell 2004; Gil de Zúñiga, Correa, and Valenzuela 2012; Hyun and Moon 2016). Over the last few years, these news networks have amassed millions of subscribers in their respective YouTube channels. As a result, the overall engagement in terms of likes, views, and comments has shown a steep upward trend (see, Figure 1). Previous studies have reported news media's role in fostering partisanship (Stanley 2012; Hyun and Moon 2016). User engagement in YouTube news networks presents an excellent opportunity to study web-scale user behavior in response to mainstream news content. In this work, via a comprehensive analysis of a substantial corpus of 86.6 million user comments on over 200,000 YouTube videos hosted by four prominent US news networks, we present a novel approach to interpreting polarization using machine translation methods.

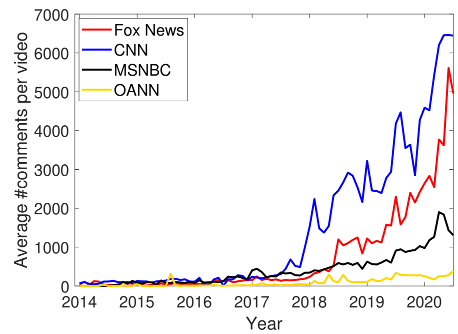

Figure 1: Temporal trend showing number of comments made about news videos on four news networks' official YouTube channels over time.

We ask the following question: *Is it possible that the two sub-communities are speaking in two different languages such that certain words do not mean the same to the liberal and conservative viewership? If yes, how do we find those words?* To do this in the context of user comments from the

*Ashiqur R. KhudaBukhsh and Rupak Sarkar are equal contribution first authors.

Republicans are the greatest threat to America	*Democrats are the greatest threat to America*
Republicans are the greatest threat to America that this nation has ever seen. They have willingly enabled a tyranny and wannabe dictator…	…Had Trump placed more restrictions on travel sooner, Democrats would have cried "racism". Democrats are the greatest threat to America today.
Republicans are traitors	***Democrats are traitors***
The Republicans are traitors. Period, full stop. All good and patriotic Americans must see this, realize it for what it is, and then begin to act accordingly…	The DEMOCRATS are TRAITORS to our country and should be rounded up and exiled to a island.
I will never vote Republican again	***I will never vote Democrat again***
What a liar! I have always voted for the man not the party but after the way the republicans have acted I will NEVER vote republican again. …	I used to vote for the democrats because they cared about poor people. Now they only care about exploitable non-american poor people, talk about being un-american. I will never vote Democrat again.
Democrats are patriots	***Republicans are patriots***
Democrats are patriots just holding on to our constitution ! McConnell and trump must have their crowns slapped off their tyranny heads	Republicans are patriots. demoRats are traitors.
Democrats are fighting for	***Republicans are fighting for***
WE ARE A NATION OF IMMIGRANTS. THAT'S WHAT MAKES AMERICA GREAT!!!…DIVERSITY IS THE CORNERSTONE OF WESTERN DEMOCRACY. THE DEMOCRATS ARE FIGHTING FOR EQUALITY AND ECONOMIC STABILITY…	Democrats are doing everything in their power to take away your power as a citizen to make choices. The Republicans are fighting for YOU as an individual. Come on Americans! Wake up!…
Vote all Democrats in	***Vote all Republicans in***
…Regardless of whether or not our candidates win in the primaries or whether we even like the Democrats we must be prepared to vote all Democrats in and all Republicans out…	We the American people are tired of these crazy dems.Hope we vote all republicans in office.

Table 1: Illustrative examples highlighting that `Democrats` and `Republicans` are used in almost mirroring contexts. Left and right column contain user comments obtained from official YouTube channels of CNN and Fox news, respectively. Our translation algorithm detects ⟨`democrats`, `republicans`⟩ as one of many translation pairs.

YouTube channels of different cable news networks, we begin by hypothesizing that viewers of CNN speak in what might be called "CNN-English" and viewers of Fox News speak in "Fox-News-English". We then apply modern machine translation procedures to these two "languages" in the same way English would be translated into, say, Spanish.

But, because both languages are using English words, the vast majority of words should translate into something very close to themselves. For instance, `grape` in CNN-English will very likely translate to `grape` or something highly similar in meaning to a `grape` in Fox-News-English, just as `tree` in Fox-News-English will very likely translate into `tree` in CNN-English or something close to `tree` in meaning. Recognizing this, we focus on those distinct pairs of words that translate into one another but have very different meaning and usage.

Such pairs are not hard to envision. Consider this simple word pair: ⟨`republicans`, `democrats`⟩ and illustrative examples of their appearances in CNN and Fox News YouTube user discussions (listed in Table 1) where these two terms appear in highly similar contexts. Intuitively, `republicans` will appear in largely favorable contexts in conservative discussion outlets while `democrats` will appear mostly in unfavorable contexts. Conversely, in liberal discussion outlets, their roles with be completely reversed in what appear to be virtually identical contexts.

How many such word pairs exist and what stories do they tell us? In this paper, we present a systematic approach to detect and study such word pairs developing a quantifiable framework to evaluate how *similar* or *dissimilar* web-scale discussions of two sub-communities are by offering a fresh perspective on interpreting linguistic manifestation of polarization through the lens of machine translation. With this novel proposition that two sub-communities are speaking in two different *languages*, we demonstrate that modern machine translation methods can provide a simple yet powerful and interpretable framework to understand the differences between two (or more) large-scale social media discussion data sets at the granularity of words.

Beyond promising results in quantifying ideological differences among multiple news networks, our automated method presents a compelling efficiency argument. It is infeasible to manually examine millions of social media posts (in the order of 100 million tokens) to identify and understand issue-centric differences. Our method boils down this task to manual inspection of less than a few hundred salient translation pairs that can provide critical insights into ideological differences. For example, translation pairs such as $\langle \texttt{solar}, \texttt{fossil} \rangle$ or $\langle \texttt{mask}, \texttt{muzzle} \rangle$ can provide insights to the ongoing energy debate or the debate surrounding mask and freedom of choice, and may indicate aggregate stance of a sub-community. Going beyond single-word translations, through simple phrase translations our method can reveal the current, deep political divide – what is `black lives matter` in CNN-English can be `all lives matter` in Fox-News-English.

Our General Idea

A standard machine translation system that performs single word translation takes a word in a source language as input (denoted by w_{source}) and outputs an equivalent word in a target language (denoted by w_{target}). For example, in a translation system performing *English* $\rightarrow$ *Spanish* translation, if the input word w_{source} is `hello`, the output word w_{target} will be `hola`, i.e., *translate* $(\texttt{hello})^{English \rightarrow Spanish}$ = `hola`. The distributional hypothesis of words (Harris 1954) famously stated "*You shall know a word by the company it keeps*" (Firth 1957). The "company" of a word, i.e. the set of words that tend to occur closely to it, aka its context, plays an important role in modern machine translation methods. The underlying computational intuition is that in a translation pair $\langle w_{source}, w_{target} \rangle$, the contexts in which w_{source} appears in the source language are highly similar with the contexts in which w_{target} appears in the target language.

A powerful way to operationalize the notion of words being close (or far) from one another is to employ a method which embeds each word as a vector in a high-dimensional space (referred to as an embedding) and using the proximity of any two words in that space as a measure of closeness. This approach, set forth in (Mikolov et al. 2013a), initiated a rich line of machine translation literature. (Mikolov, Le, and Sutskever 2013) first observed that continuous word embedding spaces exhibit similar structures across languages and proposed a linear mapping from a source to target embedding space. Their approach worked surprisingly well even in distant language pairs. Since then, several studies proposed improvements over this general idea of learning cross-lingual embedding spaces (Faruqui and Dyer 2014; Xing et al. 2015; Ammar et al. 2016).

In our work, we are interested in leveraging this machine translation literature to user discussions taking place at the comments section of official YouTube channels of two different news networks (e.g., CNN and Fox News). As we already mentioned, of course, both the CNN and Fox News corpora are in English. But we introduce a novel and powerful approach by treating them as two different languages: $\mathcal{L}_{cnn}$ and $\mathcal{L}_{fox}$. Given that our "languages" are actually English from different sub-communities, on most occasions, $\langle w_{source}, w_{target} \rangle$ will be identical word pairs (e.g., $\langle \texttt{grape}, \texttt{grape} \rangle$); i.e., for a given translation direction (say, $\mathcal{L}_{cnn} \rightarrow \mathcal{L}_{fox}$), *translate* $(w_{source})^{\mathcal{L}_{cnn} \rightarrow \mathcal{L}_{fox}}$ = w_{source}. The interesting cases are the pairs that include two different English words, i.e., $\langle w_{source}, w_{target} \rangle$ such that $w_{source} \neq w_{target}$. We call such word pairs *misaligned pairs*.

Category	Misaligned pairs
Political entities	$\langle \texttt{democrats}, \texttt{republicans} \rangle$, $\langle \texttt{nunes}, \texttt{schiff} \rangle$
News entities	$\langle \texttt{fox}, \texttt{cnn} \rangle$, $\langle \texttt{tapper}, \texttt{hannity} \rangle$
Derogatory	$\langle \texttt{chump}, \texttt{trump} \rangle$, $\langle \texttt{pelosi}, \texttt{pelousy} \rangle$
(Near) synonyms	$\langle \texttt{lmao}, \texttt{lol} \rangle$, $\langle \texttt{allegations}, \texttt{accusations} \rangle$
Spelling errors	$\langle \texttt{mueller}, \texttt{muller} \rangle$, $\langle \texttt{hillary}, \texttt{hilary} \rangle$
Ideological	$\langle \texttt{kkk}, \texttt{blm} \rangle$ $\langle \texttt{liberals}, \texttt{conservatives} \rangle$

Table 2: Examples of misaligned word pairs. Word pairs are presented in $\langle w_{cnn}, w_{fox} \rangle$ format where $w_{cnn} \in \mathcal{L}_{cnn}$ and $w_{fox} \in \mathcal{L}_{fox}$.

These different word pairs can arise for either of two very different phenomena, though both result in the same treatment by our translation algorithm, because both phenomena result in the fact that w_{source} is used by one sub-community in very similar contexts w_{target} is used by the other sub-community.

One case of misaligned pairs is where both w_{source} and w_{target} in the pair $\langle w_{source}, w_{target} \rangle$ refer to the actual same grounded entity (e.g., $\langle \texttt{pelosi}, \texttt{pelousy} \rangle$). So, for instance in "Pelosi spoke yesterday" and "Pelousy spoke yesterday", both communities are referring to Nancy Pelosi, the speaker of the United States House of Representatives. In this case, the reason that the two words appear in the same context is that the two communities are stating very similar beliefs about that entity. In this case, we can think of the two words as synonyms referring to the same entity, though the difference in the actual names can reflect important differences in attitudes toward that entity. The second case is where the word pair refers to two different entities, as in $\langle \texttt{tapper}, \texttt{hannity} \rangle$. Here, the phenomenon detected is that one sub-community makes statements about w_{source} that are very similar to the statements made by the second sub-community about word w_{target} (e.g., "Tapper is a great interviewer" vs. "Hannity is a great interviewer"). Table 2 characterizes additional phenomena that can produce word pairs, though each of the rows there correspond to one of the two phenomena above. That is, the examples under political entities, news entities, and ideological rows in Table 2 correspond to word pairs that refer to different entities. The examples under the derogatory, synonyms, and spelling errors rows correspond to the case where the two words refer to the same entity.

Our intuition is misaligned pairs may reveal useful insights into differences between the two sub-communities. For example, `solar` in $\mathcal{L}_{cnn}$ translating into `fossil` in $\mathcal{L}_{fox}$ possibly indicates that the two communities have divergent, and close to mirror image views of climate change and renewable energy. Or, `cooper` in $\mathcal{L}_{cnn}$ translating into `hannity` in $\mathcal{L}_{fox}$ possibly indicates that the CNN sub-community views Anderson Cooper favorably and Sean

Hannity unfavorably while the Fox News sub-community views the two news entities exactly the opposite way.

Quantifying similarity and dissimilarity: If two sub-communities use most words in similar contexts, the number of misaligned word pairs will be fewer than the number of misaligned word pairs if the two communities use a large number of words (e.g., entities, issues) in different contexts. We can thus construct a measure of similarity and dissimilarity between discussions in sub-communities by computing the fraction of misaligned words over the size of the source vocabulary – the larger this number, the greater the dissimilarity. Comparing across multiple corpora will require careful selection of source and target vocabulary and several other design decisions to ensure cross-corpus comparability. A format treatment of our approach follows later.

Data Set

Our data set consists of user comments posted on videos hosted by four US news networks' official YouTube channels listed in Table 3. CNN, Fox News, and MSNBC are considered to be the three leading cable news networks in the US (Statista 2020). Commensurate to their cable TV popularity, these three channels have a strong YouTube presence with millions of subscribers. Our choice of OANN, a conservative media outlet, is guided by the observation that the 45th US President shares favorable views about this network on social media platforms (Gordon 2020).

Starting from 1 January, 2014, we considered videos uploaded on or before 31 July, 2020. We used the publicly available YouTube API to collect comments from these videos. YouTube comments exhibit a two-level hierarchy. Top-level comments can be posted in response to a video and replies can be posted to these top-level comments. We collect both and for the analyses in this paper, we focus on the top-level comments. Overall, we obtain 86,610,914 million comments (50,988,781 comments and 35,622,133 replies) on 204,386 videos posted by 6,461,309 unique users. We use standard preprocessing (e.g., punctuation removal, lowercasing) for our comments.

In what follows, we present a few notable results we obtain while analyzing user engagement.

Temporal Trends of Video Likes and Dislikes: We first introduce a simple measure to evaluate viewership disagreement. For a given video v, let v_{like} and $v_{dislike}$ denote the total number of likes and dislikes v received. For each video v, we first compute the ratio $\frac{v_{dislike}}{v_{like}+v_{dislike}}$. If a video is disliked by fewer viewers and liked by a large number of viewers, this value will be close to 0. Conversely, if the video is overwhelmingly disliked, the value will be close to 1. A value close to 0.5 indicates that the opinion about the video among the viewership is divided. Formally, let $I(v, m)$ be an indicator function that outputs 1 if video v is uploaded in month m and outputs 0 otherwise. For a given channel and a particular month m^j, we compute the following viewership disagreement factor: $\frac{\Sigma_i I(v^i, m^j)\frac{v^i_{dislike}}{v^i_{dislike}+v^i_{like}}}{\Sigma_i I(v^i, m^j)}$.

Our disagreement measure has the following advantages. First, assuming $v_{like} + v_{dislike} \neq 0$, $0 \leq \frac{v_{dislike}}{v_{like}+v_{dislike}} \leq 1$. Since average of bounded variables is also bounded, our disagreement factor is also bounded within the same range [0, 1]. Since each video's disagreement measure is bounded within the range [0, 1], this measure is robust to outliers; a single heavily liked or disliked video cannot influence the overall average by more than $\frac{1}{n}$ where n is the total number of videos uploaded in that particular month[1].

News Network	#Subscribers	#Videos
CNN	10.6M	95,433
Fox News	6.09M	65,337
MSNBC	3.45M	31,732
OANN	0.84M	11,884

Table 3: List of news networks considered. Video count reflects #videos uploaded on or before 31 July 2020 starting from 1 January 2014.

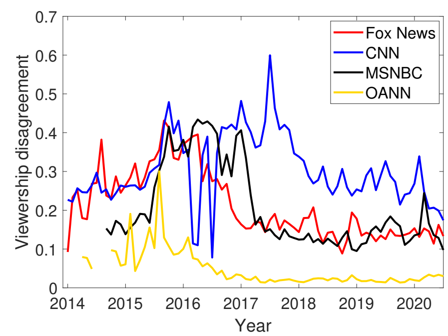

Figure 2: Temporal trend of viewership disagreement in terms of video likes and dislikes. Each point in the graph represents the monthly average of $\frac{v_{dislike}}{v_{like}+v_{dislike}}$ for each video v uploaded on the news network's official YouTube channel. We report this value only if 10 or more videos are uploaded in a given month for a specific channel.

Figure 2 presents the temporal trend of the viewership disagreement factor for four major news networks during the time period of 2014-2020. A paired t-test reveals that beyond 2017, the viewership disagreement among CNN viewers is larger than all other channels' viewership disagreement with p-value less than 0.0001. This indicates that possibly, CNN is less of an echo chamber as compared to the other three media outlets. Among these four news networks, the viewership disagreement among OANN viewers is the lowest. Our results corroborate to previous findings on the existence of echo chambers in highly conservative social media platforms (Zannettou et al. 2018; Horne, Nørregaard, and Adalı 2019). While not presented as a formal study, a parallel between Fox News and MSNBC's comparable partisanship,

[1]Figure 2 presents the temporal trend of the viewership disagreement factor for four major news networks during the time period of 2014-2020. More than 100 videos are uploaded for most of the months we considered in our analysis shown in Figure 2.

albeit for two different political views, has been reported before (Stanley 2012). Adding evidence to this observation, in Figure 2, we note that the temporal trends of viewership disagreement is similar across MSNBC and Fox News.

Related Work

Polarization and partisanship in US politics is a widely studied topic with surveys and studies focusing on diverse aspects such as congressional votes (Poole and Rosenthal 1984), response to climate change (Fisher, Waggle, and Leifeld 2013; Baldwin and Lammers 2016), polarization in media (Prior 2013) and economic decisions (McConnell et al. 2017), and partisanship in search behavior (Krupenkin et al. 2019) to name a few. Our work contrasts with recent computational social science research on polarization (Demszky et al. 2019; Darwish 2020b) along the following three main dimensions: (1) our fresh perspective on casting the task of quantifying polarization as a machine translation problem; (2) our broad treatment of the problem without focusing on specific events or type of events; and (3) our focus on YouTube data of major news networks. Unlike (Demszky et al. 2019) that focused on a specific type of events (mass-shootings) and (Darwish 2020b) that studied controversy surrounding the Kavanaugh confirmation, we consider a longer, continuous time-horizon (2017 - 2020) within which two (or more) sub-communities discuss a broad range of issues. Priors lines of research on quantifying polarization among YouTube and Facebook users have focused on characterizing user behavior in the context of scientific and conspiratorial content consumption (Bessi et al. 2016).

While work focusing on dialectical variants of English and their detection challenges exists (Blodgett, Green, and O'Connor 2016; Eisenstein, Smith, and Xing 2011), to the best of our knowledge, treating two US news networks' discussions as two distinct languages and leveraging modern machine translation literature (Smith et al. 2017) to detect mismatched translation pairs has not been explored before.

A recent work has focused on 38 prominent Indian news networks' YouTube channels leading up to 100 days of 2019 Indian General election and has used language models to mine insights (Palakodety, KhudaBukhsh, and Carbonell 2020). This study reported evidence of religious polarization in India. Our work addresses the challenge of quantifying intra-news network user discussion differences using a novel approach of machine translation.

Presence of human biases in word embedding in social media corpora is a well-established observation (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). Recent lines of work channelised considerable efforts to debias such embedding (Bolukbasi et al. 2016; Manzini et al. 2019). Our work presents a novel method to detect word pairs where two different sub-communities exhibit comparable biases for two different words across two different corpora.

Framework and Design Choices

We first describe our framework for two news networks' discussion data sets we assume to be authored in two different *languages*: $\mathcal{L}_s$ and $\mathcal{L}_t$. A more general treatment involving more than two news networks is presented later. Let $\mathcal{D}_s$ and $\mathcal{D}_t$ be two *monolingual* text corpora authored in languages $\mathcal{L}_s$ and $\mathcal{L}_t$, respectively. Let with respect to the corpora $\mathcal{D}_s$ and $\mathcal{D}_t$, $\mathcal{V}_s$ and $\mathcal{V}_t$ denote the source and target vocabularies, respectively. Let $w^{e,l}$ denote the vector representation of the word w in an embedding space trained on $\mathcal{D}_l$. A word translation scheme, $\mathcal{L}_s \rightarrow \mathcal{L}_t$, takes a word $w_{source} \in \mathcal{V}_s$ as input and outputs a single word translation w_{target} such that
(1) $w_{target} \in \mathcal{V}_t$, and
(2) $\forall w \in \mathcal{V}_t$, $dist(w^{e,s}_{source}W,\ w^{e,t}) \geq dist(w^{e,s}_{source}W,\ w^{e,t}_{target})$ where W is a transformation matrix.

We now describe and justify our design choices.

Translation algorithm: We compute W using a well-known algorithm (Smith et al. 2017)[2]. This algorithm requires two monolingual corpora and a bilingual seed lexicon of word translation pairs as inputs. First, two separate monolingual word embedding are induced using a monolingual word embedding learning model. Following (Smith et al. 2017), we use FastText (Bojanowski et al. 2017) to train monolingual embedding. Next, the bilingual seed lexicon is used to learn an orthogonal transformation matrix, which is then used to align the two vector spaces. Finally, to translate a word from the source language to the target language, we multiply the embedding of the source word with the transformation matrix to align it with the target vector space. Then, the nearest neighbour of the aligned word vector in the target vector space is selected as the translation of the source word in the target language. Following (Smith et al. 2017), we use cosine distance as our distance metric. Our choice of the translation algorithm is motivated by its (1) competitive performance (Smith et al. 2017), (2) simple and elegant design, and (3) robustness to lexicon sparsity.

Unlike typical machine translation task, we are dealing with two English corpora. For the seed lexicon of translation pairs, ideally, we would prefer words that are neutral across political beliefs with high probability. We thus construct our seed lexicon with English stopwords in the following way: $\{\langle w, w\rangle\}$, w is a stopword. We consider the default English stopword set of NLTK (Bird, Klein, and Loper 2009).

$\mathcal{V}_s$ and $\mathcal{V}_t$: We first ensure both corpora have identical size (in terms of #tokens). We next concatenate token-balanced $\mathcal{D}_s$ and $\mathcal{D}_t$ and choose the top 5,000 and top 10,000 words by frequency in the combined corpus as the source vocabulary $\mathcal{V}_s$ and target vocabulary $\mathcal{V}_t$, respectively. We exclude the stopwords while computing $\mathcal{V}_s$ and $\mathcal{V}_t$ since we use them as anchor words for our translation algorithm. Note that, here we slightly abuse the notation since we have identical $\mathcal{V}_s$ and $\mathcal{V}_t$ across both translation directions.

We token-balance our corpora to (1) enforce that the quality of the embedding is comparable across corpora and (2) ensure that both corpora have a fair influence on words that are included in $\mathcal{V}_{source}$ and $\mathcal{V}_{target}$.

***Similarity* $(\mathcal{L}_s, \mathcal{L}_t)$:** The similarity measure between two languages along a given translation direction computes the fraction of words in $\mathcal{V}_s$ that translates to itself, i.e.,

$$Similarity(\mathcal{L}_s, \mathcal{L}_t) = \frac{\Sigma_{w \in \mathcal{V}_s} I(translate(w)^{\mathcal{L}_s \rightarrow \mathcal{L}_t} = w)}{|\mathcal{V}_s|}.$$

[2]Code: https://github.com/styx97/PolarizationAAAI2021.

The indicator function returns 1 if the word translates to itself and 0 otherwise. A larger value of *Similarity* $(\mathcal{L}_s, \mathcal{L}_t)$ indicates greater similarity between a language pair.

Assigning user to a specific channel: It is not possible to unambiguously identify if a YouTube user prefers CNN over Fox News or not. We assign a user to CNN or Fox News using a simple filter. If a user has commented more on Fox News videos than on CNN videos during our period of interest[3], we assign her to Fox News and vice versa. While computing the discussion data set for a specific channel, we restrict ourselves to users assigned to that channel. We acknowledge that our filter makes certain assumptions that may not hold in the wild. It is possible that a user only comments on a video if she does not agree with its content. The qualitative nature of our analyses remains unchanged with or without this filter.

Extending to Multiple News Networks: It is straightforward to extend our method to more than two news network discussions data sets. We assume each discussion data set is authored in a distinct *language* (CNN: $\mathcal{L}_{cnn}$; Fox News: $\mathcal{L}_{fox}$; MSNBC: $\mathcal{L}_{msnbc}$; and OANN: $\mathcal{L}_{oann}$). We token-balance all corpora to identical size. Next, we concatenate all corpora and compute $\mathcal{V}_s$ and $\mathcal{V}_t$ as the top 5,000 and 10,000 words by frequency, respectively. Finally, for each translation direction, we compute pairwise similarity.

Results

We first focus on Fox News and CNN, the two most popular news networks, and present a qualitative analysis of the misaligned pairs obtained in the years 2017, 2018, 2019 and 2020.

Characterizing the misaligned pairs: Upon manual inspection, we identify the following high-level categories in the misaligned pairs listed in Table 4. Note that, we do not intend these categories to be formal or exhaustive, but rather to be illustrative of the types of misaligned pairs we encountered. Further, we realize that the misaligned pairs have the following nuance. Some of the pairs map to the same entity (e.g., ⟨liberals, libtards⟩), while the rest map to completely different entities and beliefs (e.g., ⟨nunes, schiff⟩, ⟨socialism, capitalism⟩).

We notice several misaligned pairs between political oppositions (e.g., ⟨democrats, republicans⟩) and news entities (⟨fox, cnn⟩). This result was not surprising as we have already seen in Table 1 that Republicans and Democrats are used in almost interchangeable contexts across the two news networks' user discussions. Similarly, a CNN viewer is likely to have favorable opinion toward CNN and their anchors while a Fox viewer will have positive views toward Fox News entities.

Along with a few instances of (near)-synonyms[4] and incorrect spellings[5] present in some of our misaligned pairs, we notice several derogatory terms for political and news entities (e.g., ⟨obamas, obummer⟩ or ⟨chump, trump⟩). Some of these derogatory terms could be possibly influenced by prominent public figures openly using them (e.g., ⟨schiff, schitt⟩) (Forgey 2020). We notice a misaligned pair containing the derogatory terms used to describe opposition party's fervent supporters (e.g., ⟨trumptards, snowflakes⟩).

We observe hints of the longstanding racial debate in some of the misaligned pairs (e.g., ⟨kkk, blm⟩, ⟨white, black⟩). In Table 5, we list illustrative examples of their appearances in CNN and Fox News user discussions where these two terms appear in highly similar contexts.

Category	Misaligned pairs
Political entities	⟨democrats, republicans⟩, ⟨nunes, schiff⟩, ⟨dem, republican⟩, ⟨dnc, gop⟩, ⟨kushner, burisma⟩, ⟨gop, democrats⟩, ⟨flynn, hillary⟩
News entities	⟨fox, cnn⟩, ⟨hannity, cuomo⟩, ⟨tapper, hannity⟩, ⟨tucker, cuomo⟩
Derogatory	⟨trumptards, snowflakes⟩, ⟨chump, trump⟩, ⟨liberals, libtards⟩, ⟨pelosi, pelousy⟩, ⟨obamas, obummer⟩, ⟨cooper, giraffe⟩, ⟨biden, creep⟩, ⟨schiff, schitt⟩, ⟨barr, weasel⟩
(Near) synonyms	⟨lmao, lol⟩, ⟨allegations, accusations⟩, ⟨puppet, stooge⟩, ⟨bs, bullshit⟩, ⟨potus, president⟩, ⟨hahaha, lol⟩
Spelling errors	⟨mueller, muller⟩, ⟨kavanaugh, cavanaugh⟩, ⟨hillary, hilary⟩, ⟨isreal, israel⟩
Ideological	⟨kkk, blm⟩, ⟨christianity, multiculturalism⟩, ⟨sham, impeachment⟩, ⟨antifa, nazi⟩, ⟨liberals, conservatives⟩, ⟨communism, nazism⟩, ⟨leftists, fascists⟩, ⟨liberalism, conservatism⟩, ⟨communists, nazis⟩, ⟨immigrants, illegals⟩

Table 4: Characterizing the misaligned word pairs. We consider Fox News and CNN user discussions for the years 2017, 2018, 2019, and 2020. Word pairs are presented in $\langle w_{cnn}, w_{fox}\rangle$ format where $w_{cnn} \in \mathcal{L}_{cnn}$ and $w_{fox} \in \mathcal{L}_{fox}$.

Comparing Multiple News Networks: We now perform a quantitative analysis between CNN, MSNBC and Fox News. Table 6 presents the pairwise similarity between $\mathcal{L}_{cnn}$, $\mathcal{L}_{fox}$, and $\mathcal{L}_{msnbc}$. We first note that our similarity measure is reasonably symmetric; *Similarity*$(\mathcal{L}_i, \mathcal{L}_j)$ and *Similarity*$(\mathcal{L}_j, \mathcal{L}_i)$ have comparable values across all i, j. We next note that $\mathcal{L}_{msnbc}$ is more similar to $\mathcal{L}_{cnn}$ than $\mathcal{L}_{fox}$. $\mathcal{L}_{cnn}$ is more similar to $\mathcal{L}_{msnbc}$ than $\mathcal{L}_{fox}$, and $\mathcal{L}_{fox}$ is least

[3]All our analyses are performed at the temporal granularity of a year except for 2020, where we consider the time period starting from January 1, 2020 to July 31, 2020.

[4]23% out of a randomly sampled 100 misaligned pairs.

[5]3% out of a randomly sampled 100 misaligned pairs.

KKK is a hate group	***BLM is a hate group***
...The kkk is a hate group. But drump will not call them that, he calls them very fine people...	...blm is a hate group. A group of black supremacy isn't any different than white supremacy. Defund the department of education.
KKK terrorists	***BLM terrorists***
REPUBLICANS HAVE ALWAYS BEEN NEO-NAZI'S AND KKK TERRORISTS	Step 1 - Leftist defund the police Step 2 - Antifa and BLM terrorists, looters and rioters invade neighborhoods Step 3 - Patriots (thanks to the 2nd amendment) respond to defend their families and light up the terrorists Step 4 - Anitfa and BLM call the police for help and get no answer, repeat step 3 as needed
KKK is nothing more than a	***BLM is nothing more than a***
kkk is nothing more than a low-life racist terrorist gang...	BLM is nothing more than a racist cult.

Table 5: Illustrative examples highlighting that the discovered misaligned pair $\langle \texttt{blm}, \texttt{kkk} \rangle$ are used in almost mirroring contexts. Left and right column contain user comments obtained from CNN and Fox news, respectively.

		$\mathcal{L}_{target}$		
		$\mathcal{L}_{cnn}$	$\mathcal{L}_{fox}$	$\mathcal{L}_{msnbc}$
$\mathcal{L}_{source}$	$\mathcal{L}_{cnn}$	-	90.20%	94.20%
	$\mathcal{L}_{fox}$	89.60%	-	88.70%
	$\mathcal{L}_{msnbc}$	94.10%	88.50%	-

Table 6: Pairwise similarity between languages computed for the year 2020. Each corpus has identical number of tokens. The evaluation set (5K words) is computed by concatenating all three corpora and taking the top 5K words ranked by frequency. Since stopwords are used as anchor words, stopwords are excluded in the evaluation set.

$\mathcal{L}_{cnn} \rightarrow \mathcal{L}_{fox}$	$\mathcal{L}_{cnn} \rightarrow \mathcal{L}_{msnbc}$
$\langle \texttt{trumpty}, \texttt{obummer} \rangle$	$\langle \texttt{dumpty}, \texttt{trumpty} \rangle$
$\langle \texttt{white}, \texttt{black} \rangle$	$\langle \texttt{nationalist}, \texttt{nazi} \rangle$
$\langle \texttt{pence}, \texttt{biden} \rangle$	$\langle \texttt{anderson}, \texttt{rachel} \rangle$
$\langle \texttt{supremacist}, \texttt{radical} \rangle$	$\langle \texttt{demonrats}, \texttt{demoncrats} \rangle$
$\langle \texttt{socialist}, \texttt{capitalist} \rangle$	$\langle \texttt{scientist}, \texttt{expert} \rangle$

Table 7: Discovered misaligned pairs from CNN to Fox News and MSNBC. For a translation direction $\mathcal{L}_i \rightarrow \mathcal{L}_j$, we present word pairs in $\langle w_i, w_j \rangle$ format where $w_i \in \mathcal{L}_i$ and $w_j \in \mathcal{L}_j$.

similar to $\mathcal{L}_{msnbc}$. Hence, depending on the user discussions in their respective official YouTube channels, if we seek to arrange these three news networks along a political spectrum, a consistent arrangement is the following: MSNBC, CNN and Fox News (from left to right).

One may wonder if synonymous words are causing this perception that $\mathcal{L}_{cnn}$ and $\mathcal{L}_{msnbc}$ are closer than $\mathcal{L}_{cnn}$ and $\mathcal{L}_{fox}$. We manually examine all misaligned pairs along the translation direction where $\mathcal{L}_{cnn}$ is the source. We found that even after manually removing the synonymous misaligned pairs, our conclusion still holds. Table 7 lists a random sample of unique misaligned pairs obtained along $\mathcal{L}_{msnbc} \rightarrow \mathcal{L}_{fox}$ and $\mathcal{L}_{msnbc} \rightarrow \mathcal{L}_{cnn}$ translation directions.

Primetime Comedies: *Which language do viewers of prime time comedies speak?* We construct a data set of 4,099,081 comments from official YouTube channels of well-known comedians focusing on political comedies (Trevor Noah, Seth Meyers, Stephen Colbert, Jimmy Kimmel, and John Oliver). Table 8 shows that the language of YouTube primetime comedy consumers, $\mathcal{L}_{comedy}$, is farthest from $\mathcal{L}_{fox}$ and closest to $\mathcal{L}_{cnn}$. Two interesting misaligned pairs along the translation direction $\mathcal{L}_{comedy} \rightarrow \mathcal{L}_{fox}$ include $\langle \texttt{orange}, \texttt{dotard} \rangle$ and $\langle \texttt{blue}, \texttt{red} \rangle$.

All Four News Networks: Data set size is one of the most important contributing factors to ensure the quality of word embedding (Mikolov et al. 2013b). A large data set presents a word in richer contexts, ensuring that the embedding captures more semantic information. As shown in Figure 1, of all the four channels we consider, OANN has the least user engagement in terms of comments. After adding OANN in our comparison framework and sub-sampling all other corpora to match with $\mathcal{D}_{oann}$'s size, we observe that the pairwise similarity between all channels reduced. However, if we ignore $\mathcal{L}_{oann}$ and just focus on the three languages, the qualitative conclusions: (1) $\mathcal{L}_{cnn}$ is closer to $\mathcal{L}_{msnbc}$ than $\mathcal{L}_{fox}$; (2) $\mathcal{L}_{fox}$ is farthest from $\mathcal{L}_{msnbc}$; and (3) $\mathcal{L}_{msnbc}$ is closer to $\mathcal{L}_{cnn}$ and farthest from $\mathcal{L}_{fox}$, remain unaffected.

As shown in Table 9, $\mathcal{L}_{oann}$ is farthest from $\mathcal{L}_{msnbc}$ and closest to $\mathcal{L}_{fox}$. However, $\mathcal{L}_{fox}$, a well-known conservative outlet, is closer to other two mainstream media outlets than $\mathcal{L}_{oann}$. In fact, a notable misaligned pair along the transla-

		$\mathcal{L}_{target}$			
		$\mathcal{L}_{cnn}$	$\mathcal{L}_{fox}$	$\mathcal{L}_{msnbc}$	$\mathcal{L}_{comedy}$
$\mathcal{L}_{source}$	$\mathcal{L}_{cnn}$	-	88.7%	90.3%	83.2%
	$\mathcal{L}_{fox}$	88.7%	-	85.7%	75.0%
	$\mathcal{L}_{msnbc}$	90.3%	85.8%	-	78.4%
	$\mathcal{L}_{comedy}$	81.9%	74.6%	78.0%	-

Table 8: Pairwise similarity between languages computed for the year 2019.

		$\mathcal{L}_{target}$			
		$\mathcal{L}_{cnn}$	$\mathcal{L}_{fox}$	$\mathcal{L}_{msnbc}$	$\mathcal{L}_{oann}$
$\mathcal{L}_{source}$	$\mathcal{L}_{cnn}$	-	61.1%	62.0%	42.2%
	$\mathcal{L}_{fox}$	60.1%	-	53.2%	52.7%
	$\mathcal{L}_{msnbc}$	63.0%	52.8%	-	41.9%
	$\mathcal{L}_{oann}$	43.3%	54.8%	42.5%	-

Table 9: Pairwise similarity between languages computed for the year 2020.

tion direction is $\mathcal{L}_{fox} \rightarrow \mathcal{L}_{oann}$ is $\langle \texttt{mask}, \texttt{muzzle} \rangle$. We further note that if we arrange all channels along a political spectrum, a consistent arrangement is the following: MSNBC, CNN, Fox and finally, OANN (from left to right).

Translating Trigrams: Similar to the translation retrieval task presented in (Smith et al. 2017), we conduct a translation retrieval task focused on high-frequency trigrams. Consistent with our earlier design choices, our source and target phrase vocabulary consist of 5,000 and 10,000 high-frequency trigrams of the combined 2020 Fox News and CNN corpora, respectively. Of the misaligned pairs we observe, the most notable is $\langle$`black lives matter`,`all lives matter`$\rangle$, `black lives matter` $\in \mathcal{L}_{cnn}$ and `all lives matter` $\in \mathcal{L}_{fox}$.

Conclusions and Future Work

In this paper, we provide a novel perspective on analyzing political polarization through the lens of machine translation. Our simple-yet-powerful approach allows us to both gather statistical aggregates about large-scale discussion data sets, and at the same time zoom into nuanced differences in view points at the level of specific word pairs. Future lines of research include: (1) looking into the possibility of dialects (e.g., languages spoken by centrist Democrats and their more liberal counterpart); (2) further leverage unsupervised machine translation to contrast political beliefs at the level of phrases and sentences; and (3) robustness analysis on other social media platforms and countries.

Ethics Statement

We present a new approach to analyze social media that can help detect differences in viewpoints in particular group of people and issues. While we do not discuss how to react to these discoveries, they can help uncover biases and differences of view points from large-scale data. Some of the misaligned pairs our work found are deeply disturbing. Our method is finding these word pairs because our method detects what words (or phrases) different sub-communities use in similar contexts. `Black lives matter` and `all lives matter` being used in similar contexts across two sub-communities indicates a worrisome, deep, political divide. Our method is effective in detecting many such word pairs that indicate similar deep, political divides. Ascribing any equivalence between the concepts and philosophies these words embody, is certainly not our intention. But our intention is to bring to light these real differences which do occur across these sub-communities.

References

Ammar, W.; Mulcaire, G.; Tsvetkov, Y.; Lample, G.; Dyer, C.; and Smith, N. A. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925* .

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239): 1130–1132.

Baldwin, M.; and Lammers, J. 2016. Past-focused environmental comparisons promote proenvironmental outcomes for conservatives. *PNAS* 113(52): 14953–14957.

Bessi, A.; Zollo, F.; Del Vicario, M.; Puliga, M.; Scala, A.; Caldarelli, G.; Uzzi, B.; and Quattrociocchi, W. 2016. Users polarization on Facebook and Youtube. *PloS one* 11(8): e0159641.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 EMNLP*, 1119–1130.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *TACL* 5: 135–146.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS 2016*, 4349–4357.

Bozell, L. B. 2004. *Weapons of mass distortion: The coming meltdown of the liberal media*. National Review.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.

Darwish, K. 2020a. Quantifying Polarization on Twitter: the Kavanaugh Nomination. *CoRR* abs/2001.02125. URL http://arxiv.org/abs/2001.02125.

Darwish, K. 2020b. Quantifying Polarization on Twitter: the Kavanaugh Nomination. *CoRR* abs/2001.02125. URL http://arxiv.org/abs/2001.02125.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Shapiro, J.; Gentzkow, M.; and Jurafsky, D. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of NAACL-HLT 2019*, 2970–3005. ACL.

Eisenstein, J.; Smith, N. A.; and Xing, E. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 2011 ACL:HLT*, 1365–1374.

Faruqui, M.; and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 2014 EACL*, 462–471.

Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* .

Fisher, D. R.; Waggle, J.; and Leifeld, P. 2013. Where does political polarization come from? Locating polarization within the US climate change debate. *American Behavioral Scientist* 57(1): 70–92.

Forgey, Q. 2020. Trump blasts 'little Adam Schitt' on Twitter. URL https://www.politico.com/story/2018/11/18/trump-adam-schiff-schitt-twitter-1002708. Online; accessed 3-Sep-2020.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS* 115(16): E3635–E3644.

Gift, K.; and Gift, T. 2015. Does politics influence hiring? Evidence from a randomized experiment. *Political Behavior* 37(3): 653–675.

Gil de Zúñiga, H.; Correa, T.; and Valenzuela, S. 2012. Selective exposure to cable news and immigration in the US: The relationship between FOX News, CNN, and attitudes toward Mexican immigrants. *Journal of Broadcasting & Electronic Media* 56(4): 597–615.

Gordon, D. 2020. Trump's Favorite TV Network Is Post-parody. URL https://www.theatlantic.com/politics/archive/2020/05/trumps-favorite-tv-network-post-parody/611353/. Online; accessed 3-Sep-2020.

Harris, Z. S. 1954. Distributional structure. *Word* 10(2-3): 146–162.

Hollander, B. A. 2008. Tuning out or tuning elsewhere? Partisanship, polarization, and media migration from 1998 to 2006. *Journalism & Mass Communication Quarterly* 85(1): 23–40.

Horne, B. D.; Nørregaard, J.; and Adalı, S. 2019. Different spirals of sameness: A study of content sharing in mainstream and alternative media. In *ICWSM 2019*, volume 13, 257–266.

Huber, G. A.; and Malhotra, N. 2017. Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics* 79(1): 269–283.

Hyun, K. D.; and Moon, S. J. 2016. Agenda setting in the partisan TV news context: Attribute agenda setting and polarized evaluation of presidential candidates among viewers of NBC, CNN, and Fox News. *Journalism & Mass Communication Quarterly* 93(3): 509–529.

Iyengar, S.; Sood, G.; and Lelkes, Y. 2012. Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly* 76(3): 405–431.

Iyengar, S.; and Westwood, S. J. 2015. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science* 59(3): 690–707.

Krupenkin, M.; Rothschild, D.; Hill, S.; and Yom-Tov, E. 2019. President Trump stress disorder: partisanship, ethnicity, and expressive reporting of mental distress after the 2016 election. *Sage open* 9(1): 2158244019830865.

Layman, G. C.; Carsey, T. M.; Green, J. C.; Herrera, R.; and Cooperman, R. 2010. Party polarization, party commitment, and conflict extension among American party activists. *American Political Science Review* 104(2): 324–346.

Manzini, T.; Yao Chong, L.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *2019 NAACL-HLT*, 615–621.

McCarty, N.; Poole, K. T.; and Rosenthal, H. 2016. *Polarized America: The dance of ideology and unequal riches*. MIT Press.

McConnell, C.; Margalit, Y.; Malhotra, N.; and Levendusky, M. 2017. Research: Political polarization is changing how Americans work and shop. *Harvard Business Review* .

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013a. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 3111–3119.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS 2013*, 3111–3119.

Palakodety, S.; KhudaBukhsh, A. R.; and Carbonell, J. G. 2020. Mining Insights from Large-scale Corpora Using Fine-tuned Language Models. In *Proceedings of the ECAI 2020*, 1890–1897.

Poole, K. T.; and Rosenthal, H. 1984. The polarization of American politics. *The journal of politics* 46(4): 1061–1079.

Prior, M. 2013. Media and political polarization. *Annual Review of Political Science* 16: 101–127.

Smith, S. L.; Turban, D. H. P.; Hamblin, S.; and Hammerla, N. Y. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017*.

Stanley, A. 2012. How MSNBC Became Fox's Liberal Evil Twin. URL https://www.nytimes.com/2012/08/31/us/politics/msnbc-as-foxs-liberal-evil-twin.html. Online; accessed 01-September-2020.

Statista. 2020. Leading cable news networks in the United States in April 2020, by number of primetime viewers. URL https://www.statista.com/statistics/373814/cable-news-network-viewership-usa/. Online; accessed 3-Sep-2020.

Stroud, N. J. 2011. *Niche news: The politics of news choice*. Oxford University Press on Demand.

Xing, C.; Wang, D.; Liu, C.; and Lin, Y. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *2015 NAACL-HLT*, 1006–1011.

Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Kwak, H.; Sirivianos, M.; Stringini, G.; and Blackburn, J. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, 1007–1014.