Quantifying Political Polarization through the Lens of Machine Translation and Vicarious Offense

Ashiqur R. KhudaBukhsh
Rochester Institute of Technology
1 Lomb Memorial Dr
Rochester, NY 14623 USA
axkvse@rit.edu

My talk will survey three related research contributions that shed light on the current US political divide:

1. a novel machine-translation-based framework to quantify political polarization (KhudaBukhsh et al. 2021, 2022);
2. an analysis of disparate media portrayal of US policing in major cable news outlets (Dutta et al. 2022); and
3. a novel perspective of vicarious offense that examines an important aspect not examined in web-toxicity literature heretofore – *can $\mathcal{A}$ predict how offensive $\mathcal{B}$ would find a given social media post when $\mathcal{A}$ and $\mathcal{B}$ belong to different identity groups?* (Weerasooriya et al. 2023)

The first part surveys a novel, machine-translation-based framework that *addresses a long-standing NLP challenge of quantifying political polarization in a multi-issue, large-scale social web setting*. I assume that two sub-communities (e.g., Fox viewers commenting on Fox News videos and CNN viewers commenting on CNN news videos) who are obviously speaking in English, are in fact speaking in two different *languages* (say, $\mathcal{L}_{cnn}$ and $\mathcal{L}_{fox}$). Next, I obtain single-word translations between these two *languages* using a well-known alignment method. In a world not fraught with polarization, any word w in $\mathcal{L}_{cnn}$ should translate to itself in $\mathcal{L}_{fox}$. However, if a word w_1 in one language translates to a different word w_2 in another, it indicates w_1 and w_2 are used in similar contexts across these two *languages* signaling (possible) political misalignment. These disagreed pairs present a quantifiable measure to compute differences between large-scale corpora. The greater the number of disagreed pairs, the farther apart the two sub-communities are.

This method has a compelling efficiency argument. It can automatically detect disagreed pairs such as $\langle \texttt{solar}, \texttt{fossil} \rangle$ or $\langle \texttt{mask}, \texttt{muzzle} \rangle$ requiring no human supervision. Furthermore, these pairs can succinctly shed light on important policy issues such as the ongoing energy debate or the debate surrounding masks and freedom of choice, and may indicate the aggregate stance of a sub-community. This framework has found applications in analyzing the Capitol riot (KhudaBukhsh et al. 2022).

Moving from the audience's side to the news creators' side, the survey's second part asks *how do major US news outlets cover politically salient events such as policing?* In recent work, I analyze the responses of three major US cable news networks to three seminal policing events in the US spanning a thirteen-month period – from the murder of George Floyd by police officer Derek Chauvin to the Capitol riot and then to Chauvin's conviction and sentencing (Dutta et al. 2022).

Our analyses reveal that across cable networks coverage of politically salient events responds quickly and dramatically to the partisan preferences of their viewership. On the methods front, I develop a novel active learning sampling strategy exploiting logical inconsistencies in text entailment.

The final part of this talk introduces the notion of *vicarious offense* in which we ask a timely and important question: *how well do Democratic-leaning users perceive what content would be deemed as offensive by their Republican-leaning counterparts or vice-versa?* Via a substantial annotation study conducted on 2,310 social media posts, our experiments reveal that (1) Republicans are the least understood political group while they also struggle the most to understand others; (2) independents are the most tolerant in terms of web censorship; and (3) hot-button issues such as reproductive rights or gun control/rights can have a profound effect on annotator disagreement on what is offensive. I conclude the talk with an outlook on how generative AI might impact this political divide and vice versa.

References

Dutta, S.; Li, B.; Nagin, D. S.; and KhudaBukhsh, A. R. 2022. A Murder and Protests, the Capitol Riot, and the Chauvin Trial: Estimating Disparate News Media Stance. In *IJCAI 2022*, 5059–5065.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2021. We Don't Speak the Same Language: Interpreting Polarization through Machine Translation. In *AAAI 2021*, 14893–14901.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2022. Fringe News Networks: Dynamics of US News Viewership following the 2020 Presidential Election. In *WebSci '22*, 269–278. ACM.

Weerasooriya, T. C.; Dutta, S.; Ranasinghe, T.; Zamperi, M.; Homan, C. M.; and KhudaBukhsh, A. R. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *EMNLP 2023*, 11648–11668.