We have carefully addressed the comments and substantially improved our paper thanks to the feedback, thank you. We summarize in this document how we have addressed the main comments.

Answers to the summarized feedback from the metareview:

1. **Motivation could be clear:**

We have fully rewritten the introduction and modified multiple parts of the paper to clarify the motivation. We have highlighted that we conduct a constituency-level analysis where for each constituency, we compare the elected politician and its constituents. We have brought relevant literature to explain and motivate this choice. We have also changed the research questions to make our findings and motivation clearer: each research question now covers one of the dimensions that we analyze (political spectrum, margin of electoral victory, and income of the constituency).

2. **No detailed discussion and/or validation of the representativeness claim.**

We appreciate and agree with this point. Our Nextdoor data is geographically representative but we have no means to validate whether it is representative of its constituents (as we have no neighborhood-level data). We now analyze the coverage of our data in Section 2 and discuss its limitations and potential biases in Section 5.

3. **Methodological reservation concerning the sentiment and cosine similarity calculation**

Cosine similarity: For each constituency, we convert each post text into a single vector embedding using an appropriate pre-trained model. Concatenation of posts is not an option: the number of tokens per constituency exceeds the capacity of any existing model by several orders of magnitude (JINA is the model with the largest capacity and its limit is 8096 tokens per document). Instead, we follow the relevant literature (including the seminal paper on transformers and follow-up papers from its authors) and aggregate the textual embedding of all posts of a constituency with pair-wise mean-pooling aggregation. To the best of our knowledge this is the best existing solution.

We have added these and more details as well as the relevant literature. We have also included here (Fig.1) and in the paper the median token size per constituency dataset to justify our approach.

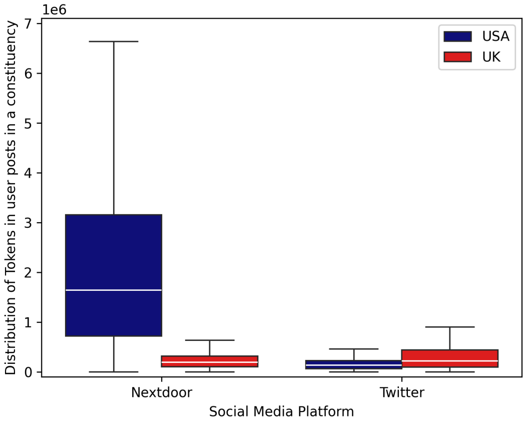

Fig. 1.

Sentiment analysis: We chose VADER because it is regarded as one of the best lexicon-based sentiment methods, it facilitates comparison with previous work (Iqbal et al. 2023 also use VADER on Nextdoor data), and outputs a single score that facilitates the comparisons required in this paper. We test the BERT-based state-of-the art Twitter-roBERTa-base (Barbieri et al. 2020) on our Twitter and Nextdoor datasets. We annotate the sentiment of posts and ask a human annotator to manually annotate the model output as correctly classified or not. Our native English-speaking human annotator takes a random sample of 2,000 posts, 500 from each model and dataset (i.e., Twitter and Nextdoor) and manually annotates them as correctly classified or not. When the human annotator and model has consensus on sentiment labels, we label that post as correctly classified. We repeat this process for all samples and use Cohen's Kappa score (K) between the model outputs and our manual annotation. We finally choose the VADER model (K=0.874) as it outperforms Twitter-roBERTa-base (K=0.771).

Individual comments not covered by our previous answers:

Reviewer 4: How does topic similarity affect the similarity in language use between left-wing citizens and right-wing politicians and vice-versa.

The main goal of this paper is to measure the similarity of politicians and their constituents and for that purpose, we focused on intra-constituency comparisons. We have now included in the paper and here (see Fig 2) comparisons between constituencies and politicians of opposing political colors: this is, we compare right-wing politicians and constituencies that elected a left-wing politician (and the other way around). Naturally, we are comparing politicians with a constituency that did not elect them. This adds completeness and it is rather insightful: it shows that the similarity between politicians and the constituencies that elect them is much greater than between such politicians and other constituencies.

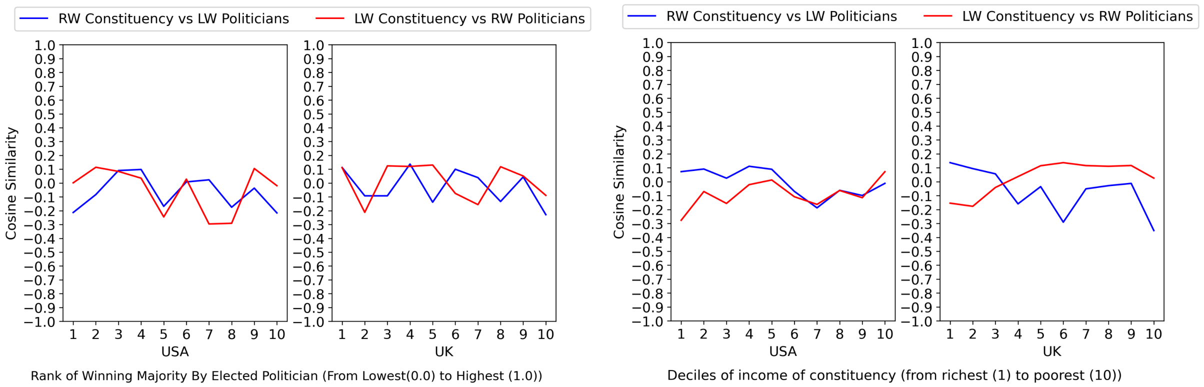

Fig. 2

Reviewer 1: The authors are trying to compare language use across Twitter (for politicians) and Nextdoor (for constituents). It's unlikely that language use will match across platforms, especially if the user bases are very different (and usually social media platforms do have very different active user bases), and so I'm not sure what comparisons across these platforms will mean. A study of both politicians and constituents on Nextdoor would have provided much more interpretable findings. This issue should have been directly discussed and addressed in some way.

We agree that Twitter and Nextdoor are different and that this poses challenges. Twitter welcomes general content and Nextdoor focuses on local issues. For instance, politicians run political campaigns on Twitter, but political campaigns are not allowed on Nextdoor. We agree that this might result in relatively low similarities between constituents and elected representatives. However, the underpinnings for this are constant (the different scope of the two platforms does not change across constituencies) and it will affect the absolute rather than relative values. Our comparisons are strictly relative: we always compare the differences of a constituency with the differences of another constituency.

We now also compute similarities between politicians and constituencies that did not elect them. We show that the similarity is consistently and substantially lower than in the intra-constituency comparisons. We have incorporated these points into the section discussing the limitations of our paper.

We also agree that comparing left and right in Nextdoor and Twitter would have been easier. We were however very keen in comparing a politician with those who elected them. Many works have compared right and left across many dimensions and with a variety of data, but we are not aware of much quantitative work that compares a politician with its direct constituents. Collecting our data and mapping constituents to elected politicians was a painstaking job. However, we believe that this is a subtle but important and exciting angle.

Reviewer 1: The LIWC analyses are difficult to assess since it sounds like the authors have reported only differences that were above some sort of threshold? And these categories also seem to be the more proprietary (also less fully explained and, in my experience anyway, very different distributions) summary ones in LIWC?

Many statistical tests will be invalid with repeated observations from the same people, unless there's something done to account for non-independence. I think there is only one formal statistical test in this paper (whether constituents on Nextdoor and politicians on Twitter use different language) and I do not see an indication that that test accounts for non-independence (I'm pretty sure the KS test relies on IID data).

We report LIWC categories that are statistically significant. We check for significant differences with a two-sample t-test on LIWC category scores from tweets and posts. We have added verifications showing that the underlying distributions are independent. To verify independence, we calculate the mutual information score for every LIWC category in the Twitter and Nextdoor datasets finding values close to 0 (varying between 0.08 and 0.13), where 0 indicates complete independence. All in all, we find that there are only five LIWC categories (i.e. Tone, Analytic, Clout, Authentic, and Linguistic) with statistically significant (i.e. $p \geq 0.05$) differences. We have now clarified this and added details and explanations of the relevant categories.

Reviewer 1: The competitiveness analysis seems interesting but should this vary in meaning across the US and UK? I would think the different party sizes and dynamics should matter.

We agree that this and other factors might make the comparison hard. We show the results side by side, and it seems revealing that despite differences, the trends and magnitudes are usually aligned. However, due to the differences between both countries, we abstain from making comparisons between the countries and focus instead in intra-country analysis,