How do people on Twitter talk about Kamala Harris? An Exploratory Data Analysis of Tweets about Kamala Harris using LDA topic modelling and Sentiment Analysis:

Introduction:

Twitter acts as a digital public sphere for discussing political affairs and news, it also emulates culture (Hill, 2018, pp. 288). While Twitter often shows the level of polarization that occurs about political topics, it still acts as a platform for amplifying the voices of the disadvantaged and therefore has led to the prevalence of intersectionality, social justice, and identity-based ideas on social media(ibid). Many argue that 'identity politics' has become an enormous phenomenon on Twitter along with a greater push for diversity, intersectionality and feminism (Kim, 2018 pp, 1; Bennett, 2012, pp. 2). Since Kamala Harris is an American woman of Afro-Caribbean and South Asian descent (Strauss 2020), this paper aims to assess the validity of two hypotheses to see if Harris is talked about largely in relation to her 'identity' or heritage on Twitter (and whether this emulates the broader phenomenon of so-called identity politics on Twitter) as well as to see whether Harris is discussed in mostly negative, positive, or neutral terms. The paper will use topic modelling and then sentiment analysis with R.

The first hypothesis of this study is that one topic discussed in relation to Kamala Harris will be her identity as the first Black American woman to serve as Vice President of the USA (Ibid). Due to her role as the first woman of colour to serve as VP, but also the controversy surrounding her work as District Attorney and prosecutor in California, the second hypothesis of this paper is that the mean sentiment score derived from the tweet dataset about Kamala Harris will indicate predominantly neutral sentiment about Kamala Harris.

Literature Review:

Literature about public reactions to Biden and Harris' victory in the 2020 US presidential election remains scarce, due to the recency of the event. Therefore, the research presented in this paper is important as it shows us Twitter sentiment about the Vice President just a few months after the election and shortly after Biden's inauguration. Nevertheless, papers about the 2020 presidential election have studied the impact of the COVID-19 pandemic on the 2020 election, predicting that the loss of elder voters to the coronavirus would impact the voter turnout, but not enough to significantly alter the results of the presential election, had there not been a pandemic.[1] Another

[1] Johnson, A.F., Pollock, W. and Rauhaus, B., 2020. Mass casualty event scenarios and political shifts: 2020 election outcomes and the US COVID-19 pandemic. *Administrative Theory & Praxis*, *42*(2), pp.249-264.

study focused extensively on the voting behaviour and patterns of women in the 2020 US election compared to previous elections, arguing that women did not play a role as ‘swing voters’ in the election and that women’s voting patterns were ‘critical to shaping the outcomes of the 2020 election’(Ondercin, 2020). The paper also makes the note that Black women’s ‘overwhelming support’ for the Democratic party in several states was ‘key to Biden’s victory in key states such as Georgia’ (ibid). This paper is important to consider when examining public twitter content about Kamala Harris because it shows how support from other Black American women was important in her becoming the first women of colour to serve as VP of the United States – once again reinforcing the role of social identity in US political discourse on Twitter.

It is also necessary to consider literature about the depiction of Hillary Clinton before the US presential election in November 2016. As Anderson writes, social media platforms – especially microblogging platforms like Twitter and Tumblr, have become pivotal in building the perception of politicians (Anderson, 2014, pp 224). These platforms exemplify ‘postmodern political culture’ where a candidate’s image is an ‘amalgamation of image fragments generated by the individual politician, her/his campaign communication, news framing, and political pop culture’ (ibid). Anderson argues that Hillary Clinton’s image had connoted not just postfeminism and the prospects of being a woman presidential candidate, but also her tweets after the USA’s ‘handling of the terrorist attack on the U.S. consulate in Benghazi’, an event which significantly contributed to her unfavorability ratings in years prior (ibid). Anderson also writes that the postfeminist logic of US politics suggests that unfavorability of woman candidates like Clinton would be attributed to the candidates’ shortcomings rather than the fact that they are women. However, Gendered perceptions of the perceived competency of Clinton had she become president were shown to be present (Kromer, 2019, pp. 1). In applying this to Kamala Harris, this would imply that Harris’ image on Twitter would be the effect of multiple connotations – her persona on Twitter, as well as the reputation she has garnered from previous policies or events related to her work prior to participating in an election. For Hillary Clinton, her role in US foreign policy (such as Benghazi in Libya) was an image fragment that contributed to negative aspects of her reputation. Whereas for Kamala Harris, the lack of academic literature thus far means it cannot be known for sure what mishandlings may contribute to her image – but the topic modelling and sentiment analysis in this essay can help indicate what her image may consist of now. In addition, literature on women in politics describes three types of representation – descriptive, substantiative, and symbolic (Tadros, 2014, pp. 4). Descriptive representation refers to the number of women present in electoral bodies. Symbolic representation refers to “how legislators’ presence shapes the beliefs and attitudes held by elites and mass publics” (ibid). Whereas substantive representation refers to when ‘legislators pursue policy goals that are aligned with the interests of their constituents' (ibid). Using the data analysis in this paper can help to determine what kind of representation Kamala Harris is seen as embodying according to many users on Twitter.

Methods and data:

This research begins by using LDA topic modelling with R for sorting the tweets into a 'bag-of-words' for each presumed topic based on how often certain words are used in co-occurrence with each other in each tweet (which acts as a document). Tweets are collected from the Twitter API (Application Programming Interface) using a Python script developed from tweepy, and appended to a CSV file, before being analysed with R. The initial sample contains 3485 tweets from the period of 28 February to 1 March 2021. However, to test for replicability of results and time bias, another set of 3500 tweets were analysed during the period of 9 May to 12 May 2021.

The R package topicmodels was used for building the topic model using a k means of 12, as was the package ggplot2 for displaying the topic model. The results are shown in figure 1 below:

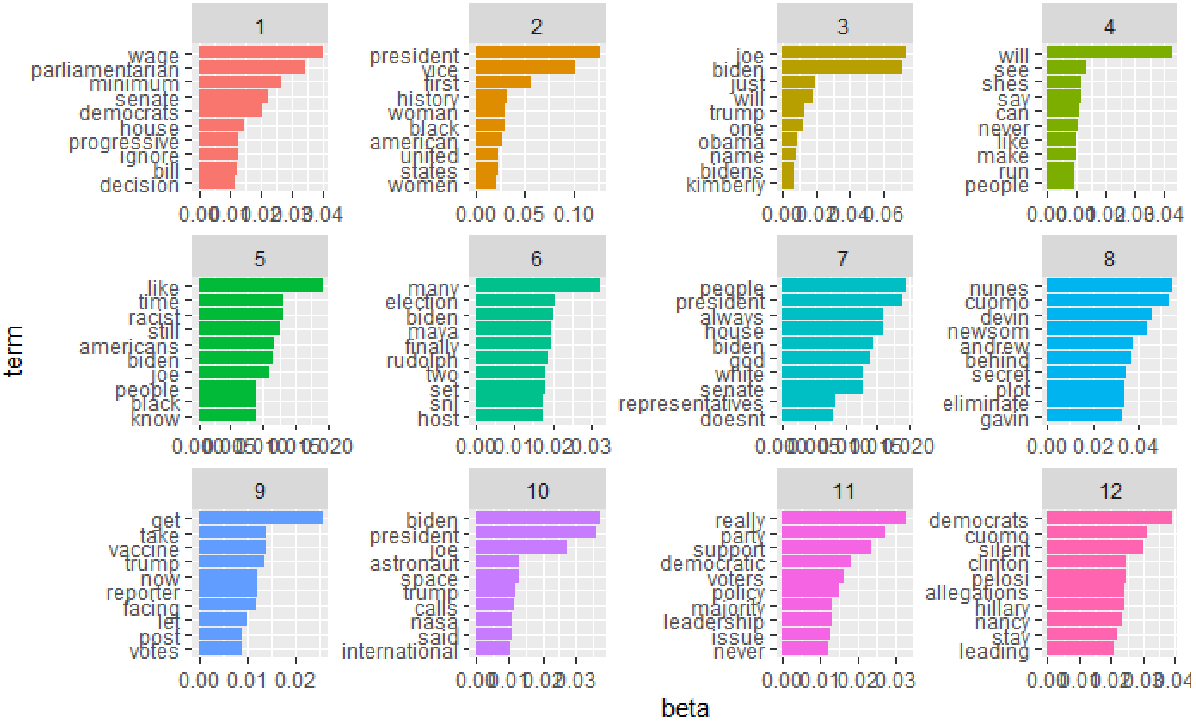

Figure 1: 'beta' in the x axis demonstrates the probability of a term occurring in a certain topic.

Data was cleaned to remove words that may override important findings, such as stop words (like ‘the’, ‘and’ and ‘is’) as well as the terms ‘kamala’ and ‘harris’, since the tweets were collected based on a keyword search about ‘Kamala Harris’. Choosing a large k means (number of topics) for this model is valid because, as we can see in this model, Kamala Harris is talked about in relation to a variety of topics. For example, topic 6 includes words like ‘snl’, ‘maya and ‘rudolph’, referring to actress Maya Rudolph’s impersonation of Kamala Harris on American comedy show Saturday Night Live. Whereas “Cuomo” in topics 11 and 12 show that Kamala Harris is being discussed in relation to the sexual assault allegations against US senator Andrew Cuomo (Vasques, 2021; McKinley, 2021). Topic 2 includes words like ‘first’, ‘black’, ‘women’, ‘history’, ‘vice’ and ‘president’, showing that Kamala Harris *is* discussed in relation to her identity as the first black woman to serve as Vice President of the United States (Strauss, 2020). It shows us that many Twitter users acknowledge the fact that she has ‘made history’ when discussing the fact that she is the first woman vice president. However, this does not tell us whether Kamala Harris is discussed in relation to her heritage *most* often. To see whether this is true, we should plot the frequency of words rather than a topic model. Figure 2 above takes every word from the dataset with a minimum frequency of .40 and displays them in a word cloud based on their frequencies. The larger the surface area of the word in the word cloud, the greater the frequency in the dataset. As you can see, ‘president’ and ‘biden’ are the

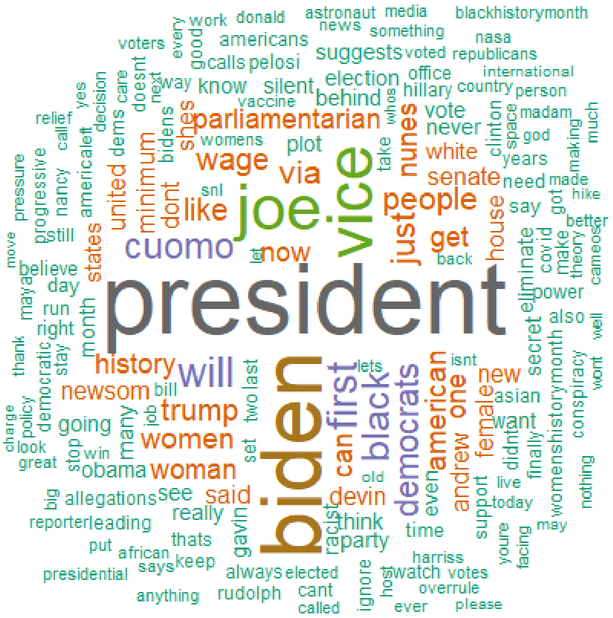

Figure 2 was built using the R package wordcloud

largest and therefore the most frequently represented words in the dataset about Kamala Harris. This is followed by the words ‘joe and ‘vice’. It would explain why ‘biden’ is present in 5 out of 12 of the topics in the topic model in figure 1. The words ‘first’ and ‘black’ are the fifth and seventh most frequently used words in the dataset. This indicates that Kamala Harris is discussed *most* frequently in relation to her associate, President Joe Biden rather than her identity.

Although analyses using tweets from the Twitter API is seen as lacking replicability, for understanding discussion about Kamala Harris it is still necessary to see whether there would be any similarities between the first dataset, and another dataset from 9 to 12 May 2021, two months later. Therefore, another set of 3500 tweets with the keyword search ‘Kamala Harris’ were collected from the Twitter API using a python script. The results are shown below of the second topic model and word cloud are shown in figures 3 and 4 below:

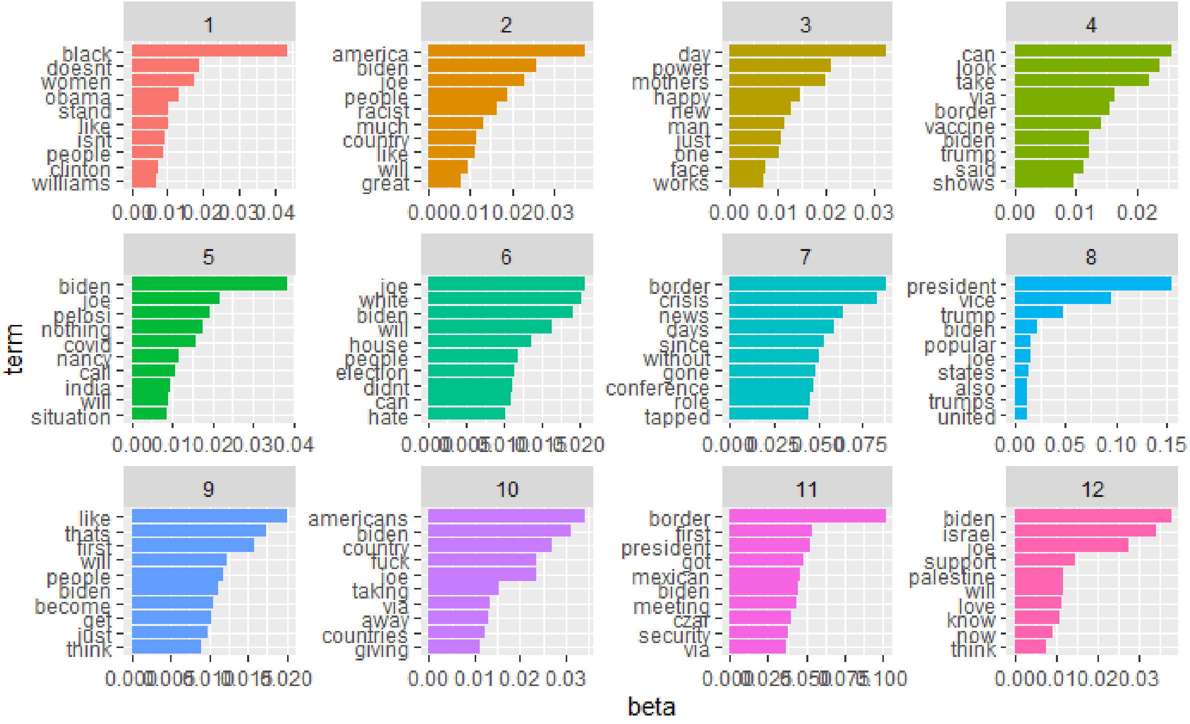

Figure 3: This figure shows some overlap with the topic model in figure 1, but there is no topic based specially on Kamala Harris’ identity, unlike topic 2 in the first topic model.

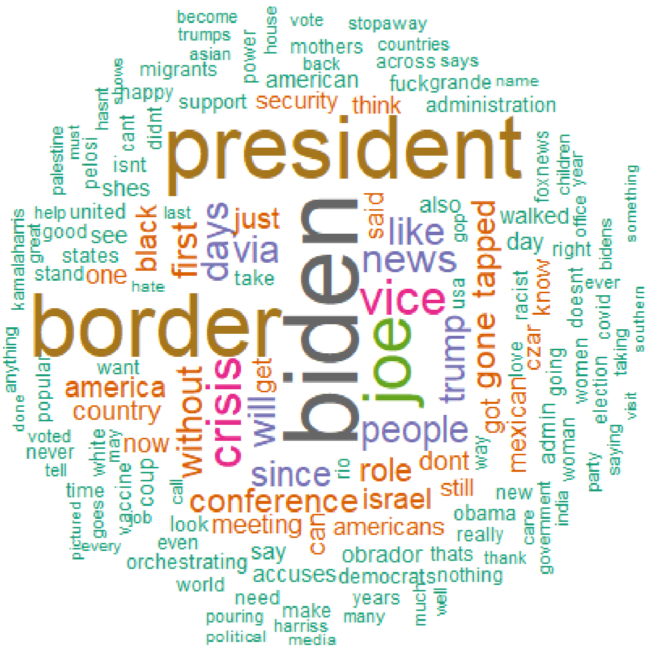

Figure 4 shows each word in the dataset with a frequency higher than .40 the words 'first' and 'black' are in orange, indicating a frequency of roughly .50.

Although 'first' 'black', 'vice' and 'president' still had a high frequency and were shown in the word cloud in figure 4, there was no topic in the model (shown in figure 3) that referred specifically to the fact that Harris' made history as the first Black woman to become vice president (as there was in topic 2 of the first model). This suggests that although Kamala Harris is often discussed in relation to her identity, this lessened during the time that the second set of tweets were collected – and other topics and words, especially Harris role in other political affairs as well as Joe Biden, are more significantly discussed. The topic modelling for both datasets shows that discussion about substantive and descriptive representation are all present on Twitter – Kamala Harris is mentioned in relation to women's history month the figure 1 and Mother's Day in figure 3, in the topic models - indicative of descriptive representation. But her role in political affairs and legislation such as news of allegations against Andrew Cuomo, coronavirus vaccinations and Israel/Palestine (topic 12 in figure 3), indicative of discussion about whether she fulfils substantive representation.

Sentiment Analysis:

The sentiment analysis with R is to test hypothesis 2 – that the mean sentiment score of tweets about Kamala Harris will indicate neutrality of Twitter sentiment about Kamala Harris. The reason for this hypothesis is due to positivity about Harris' identity, the fact that she made history as the first black woman to become vice president of the United States, but also negativity based on criticism towards her work as a prosecutor and actions taken by President Joe Biden (Strauss, 2020; Lopez, 2020). This sentiment analysis uses the commonly used R package Syuzhet for generating the sentiment scores of the tweets using four different numerical methods – syuzhet, bing, afinn and nrc (Naldi, 2019). The syuzhet method, which takes each row of the dataset and identifies words from the predefined lexicon of negative, positive, and neutral words – calculated a mean sentiment score of 0.002418.[2] The mean sentiment score is less than 0.01 above zero, which could be used to suggest that sentiment about Kamala Harris in the dataset is largely neutral. However, when using other methods in the syuzhet package, sentiment scores differ. With the bing method generating a mean sentiment score of -0.1156. While the afinn method generates a mean sentiment score of -0.05696. The bing and afinn methods appear to suggest Kamala Harris talked about in relation to negative sentiment more than positive in the dataset. On the contrary, the nrc method generates a mean sentiment score of 0.1436 – indicating slightly more positive sentiment. While the syuzhet method is considered most reliable, due to having the largest repository of predefined words in its lexicon – the difference in results generated with other methods raises concerns about the validity of such findings. The fact that different methods detect average sentiment that is only slightly negative or positive, since each of the mean sentiment scores derived are less than 1, that the sentiment about Kamala Harris is neutral.

However, the sentiment scores are still not reliable enough to verify hypothesis 2. While the standard deviation for the syuzhet score is below 1 (at 0.85202), for the bing, afinn and nrc methods they are above 1, indicating that deviation away from the mean was often high. [3]The syuzhet package is considered the most reliable, but unlike the afinn method it also does not contain internet slang words, abbreviations and acronyms from Urban Dictionary and Wiktionary that are often used on sites like Twitter (Naldi, 2019, pp. 3), so the mean score from the syuzhet method (which is the closest to zero and most indicative of neutrality) is still not enough as a standalone indicator of neutrality. On the other hand, the nrc method has a broader scale for scoring tweet sentiment, since it includes more categories for which tweets can be scored with – such as 'joy' and 'anger'. The discovery of more positive sentiment compared to negative sentiment for the nrc method, as well as the supposed prevalence of 'trust' as the most common emotion in the dataset,

[2] See figure 6 on page 8 to see scores in table.

[3] See annex for standard deviation scores of each method.

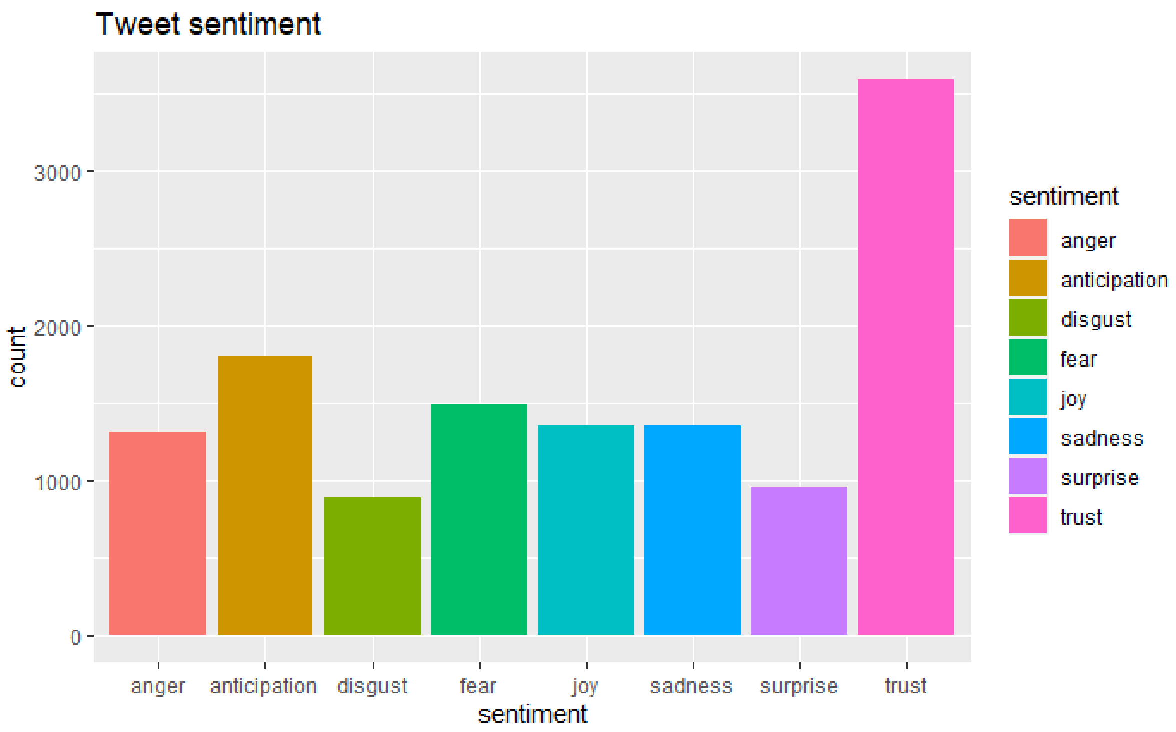

Figure 5: This is a bar plot showing the emotions detected by the NRC sentiment analysis method in the syuzhet package. Unlike the other three methods, the nrc method for the first dataset is more suggestive of positive sentiment than the other methods. With most words in the tweet set about Kamala Harris supposedly being indicative of 'trust'. Code for this plot was taken from red-gate.net (Mahtré, 2020).

makes the overall neutrality of the tweet dataset rather ambiguous. This is shown in figure 5 on the succeeding page.

Another limitation of the four sentiment analysis methods from the syuzhet package is their treatment of negators – meaning a sentence like 'Kamala Harris is not great' would be given the same score as the sentence 'Kamala Harris is great' for the bing and syuzhet methods (Naldi, 2019, pp. 5). Each method not only uses a different scale for scoring words but also different lexicons that often fail to capture the context of words (Murthy, 2016, 2). To further test the validity of results, the tweet dataset from 9 May to 12 May 2021 was used again. Mean sentiment scores were found to be slightly more indicative of negative sentiment than the first dataset. A comparison of the results is shown in figure 6 below:

Figure 6 Mean sentiment scores

Method	Tweet Dataset 1	Tweet dataset 2
Syuzhet	0.002418	-0.07194286
Bing	0.1156	-0.1601429
Afinn	-0.05696	-0.318
Nrc	0.1436	0.07442857

The mean sentiment scores for the second dataset are lower than for dataset 1, this could be for several reasons, but timing and political discourse when the tweets were collected is likely to be the factor as to why. This shows us that sentiment can vary widely depending on the time frame of which tweets are collected, but it the fact that for each dataset the mean scores are still close to zero may suggest that general sentiment about Kamala Harris on Twitter is neutral – with neither negative nor positive sentiment outweighing each other. It is important to note that while sentiment scores about Kamala Harris were close to zero and suggested the possibility of neutrality in terms of sentiment on Twitter about VP Harris, this cannot be seen as representative of neutrality in terms of *opinion* about Kamala Harris on Twitter due to temporality, sample bias and the limitations of the sentiment analysis methods used in this paper. It is well known that Twitter is not representative of the general population – with more twitter users identifying as American, democrats and women (PEW, 2019). The Twitter API only allows tweets from (maximum) the last 14 days to be returned, and do not include tweets made from private accounts (Morstatter, 2013). As a result, tweets from more significant events liek Joe Biden's inauguration, could not be used in this paper. Thus, the tweets in this analysis still may not fully represent the views held by Twitter in general. Nevertheless, reasons for neutrality of sentiment about Kamala Harris on Twitter is not just because of the sentiment analysis methods used in this paper, but also because of discussion of topics relevant to US Politics – as seen in the topic modelling in the first half of this essay. In the topic model in figure 1, it was seen in topic 2 that Kamala Harris was often being discussed as making 'history' and being the first Black American woman to serve as VP of the United States, but at the same time there were topics in the model that were indicative of discussion about Harris' role in policymaking with Joe Biden, dealing with the coronavirus, and her relation to other key politicians like Andrew Cuomo and Nancy Pelosi. It is possible that sentiment scores about many politicians are more likely to be neutral and close to zero than sentiment about another topic because Twitter users are constantly discussing political affairs in relation to politicians, and polarization means that sentiment is unlikely to be strongly skewed to the left or right.

Conclusion:

Although there is *some* evidence to support both hypothesis 1 and 2 – the evidence is not strong enough. This is due to the results that occur if the analysis is repeated in a tweet dataset from two months later. The topic modelling showed that during 28 February to 1 March 2021, Kamala Harris was often discussed in relation to her social identity and heritage – with topic 2 in figure 1 illustrating this. However, this was likely to be due to women's history month beginning in March. In addition, repeating this topic model for a dataset collected in May 2021 had different results – with the prevalence of other topics overriding discussion of Kamala Harris' identity as the first Black

American women in history to become vice president of the USA. Therefore, the topic of Kamala Harris' identity was significantly discussed in the dataset involving women's history month – but not in a dataset from two months later – indicating that the reason is more likely to be due to temporality of tweets rather than the larger phenomenon of 'identity politics' and the push for intersectionality. These social phenomena mentioned in literature about Twitter may still be important, since mentions of Kamala Harris' race and gender are still present in the word cloud for the second dataset, but the topic models for both datasets show us that on Twitter most users are still more interested in discussing political affairs and policymaking of Kamala Harris and her partner Joe Biden. The topic modelling shows that the extent of substantive and descriptive representation is all present in discussion about Kamala Harris. The views of symbolic representation (the beliefs and attitudes held by the public) embodied by Kamala Harris on Twitter is more difficult to establish using sentiment analysis. The sentiment analysis *suggests* that the general sentiment about Kamala Harris is neutral on Twitter, but the scores do not *indicate* neutrality due to the limitations of each lexicon, the differences in measurements, and the scales employed by each method. This paper shows us that 'sentiment' as well as perceptions of politicians on social media is often difficult to summarise quantifiably regardless of the nature of the social media platform – since consensus is rare and the number of tweets is constantly increasing. My suggestion for further research would be to examine the impact that political news about politicians like Kamala Harris, has on Twitter sentiment.

Bibliography:

Anderson, K.V. and Sheeler, K.H., 2014. Texts (and Tweets) from Hillary: Meta-Meming and Postfeminist Political Culture. *Presidential Studies Quarterly*, *44*(2), pp.224-243.

Bennett, W.L., 2012. The personalization of politics: Political identity, social media, and changing patterns of participation. *The annals of the American academy of political and social science*, *644*(1), pp.20-39.

Hill, M.L., 2018. "Thank you, Black Twitter": State violence, digital counterpublics, and pedagogies of resistance. *Urban Education*, *53*(2), pp.286-302.

Johnson, A.F., Pollock, W. and Rauhaus, B., 2020. Mass casualty event scenarios and political shifts: 2020 election outcomes and the US COVID-19 pandemic. *Administrative Theory & Praxis*, *42*(2), pp.249-264.

Kim, D., Russworm, T.M., Vaughan, C., Adair, C., Paredes, V. and Cowan, T.L., 2018. Race, gender, and the technological turn: A roundtable on digitizing revolution. *Frontiers: A Journal of Women Studies*, *39*(1), pp.149-177.

Kromer, M. and Parry, J.A., 2019. The Clinton Effect? The (Non) Impact of a High-Profile Candidate on Gender Stereotypes. *Social Science Quarterly*, *100*(6), pp.2134-2147.

Lopez, German. 2020. "Kamala Harris' controversial record on criminal justice, explained." *Vox.* Published 12 August 2020. Available at: https://www.vox.com/future-perfect/2019/1/23/18184192/kamala-harris-president-campaign-criminal-justice-record (Accessed 19 May 2021)

Mahtre, Sanil. 2020. "Text Mining and Sentiment Analysis: Analysis with R - Simple Talk (red-gate.com)." red-gate.com. viewed 11 May 2021, https://www.red-gate.com/simple-talk/sql/bi/text-mining-and-sentiment-analysis-with-r/

McKinley, Jesse. 2021. "Cuomo faces new claims of Sexual Harassment from Current Aide." *New York Times.* Published 19 March 2021. Available at: https://www.nytimes.com/2021/03/19/nyregion/alyssa-mcgrath-cuomo-harassment.html. (Accessed: 19 May 2021)

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. 2013. "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, no. 1.

Murthy, D., 2017. The ontology of tweets: Mixed methods approaches to the study of Twitter. *The SAGE handbook of social media research methods*, pp.559-572.

Naldi, M., 2019. A review of sentiment computation methods with R packages. *Department of Civil Engineering and Computer Science. arXiv preprint arXiv:1901.08319*.

Ondercin, H.L., 2020, December. Marching to the Ballot Box: Sex and Voting in the 2020 Election Cycle. In *The Forum* (Vol. 18, No. 4, pp. 559-580). De Gruyter.

Strauss, Daniel. 2020. "Kamala Harris makes history as first woman of colour to be elected US Vice President". *The Guardian.* Published 7 November 2020. Available at: https://www.theguardian.com/us-news/2020/nov/07/kamala-harris-first-woman-of-color-us-vice-president. (Accessed 19 May 2021).

Tadros, M. ed., 2014. *Women in Politics: Gender, Power and Development*. Zed Books Ltd..

Vasquez, Zach., 2020. "Saturday Night Live: Maya Rudolph's Kamala Harris as the Real President? Bad idea. *The Guardian.* Published 20 December 2020. Available at: https://www.theguardian.com/tv-and-radio/2020/dec/20/saturday-night-live-maya-rudolph-kamala-harris-joe-biden-kristen-wiig. (Accessed 19 May 2021)

Wojcik, S. and Hughes, A., 2019. Sizing up Twitter users. *PEW research center*, *24*.

Annex

Table of standard deviation scores for each sentiment analysis:

Method	**Dataset 1**	**Dataset 2**
Syuzhet	0.852022	0.8451918
bing	1.001276	0.9802653
Afinn	2.417186	2.534971
Nrc	1.066268	1.060684