Beyond a binary of (non)racist tweets: A four-dimensional categorical detection and analysis of racist and xenophobic opinions on Twitter in early Covid-19

Xin Pei¹ and Deval Mehta²

¹School of International Communications, University of Nottingham, Ningbo, China
²Independent Researcher
PEIX0001@gmail.com, deval.mehta1092@gmail.com

Abstract
Transcending the binary categorization of racist and xenophobic texts, this research takes cues from social science theories to develop a four-dimensional category for racism and xenophobia detection, namely stigmatization, offensiveness, blame, and exclusion. With the aid of deep learning techniques, this categorical detection enables insights into the nuances of emergent topics reflected in racist and xenophobic expression on Twitter. Moreover, a stage wise analysis is applied to capture the dynamic changes of the topics across the stages of early development of Covid-19 from a domestic epidemic to an international public health emergency, and later to a global pandemic. The main contributions of this research include, first the methodological advancement. By bridging the state-of-the-art computational methods with social science perspective, this research provides a meaningful approach for future research to gain insight into the underlying subtlety of racist and xenophobic discussion on digital platforms. Second, by enabling a more accurate comprehension and even prediction of public opinions and actions, this research paves the way for the enactment of effective intervention policies to combat racist crimes and social exclusion under Covid-19.

1 Introduction
The rise of racism and xenophobia has become a remarkable social phenomenon stemming from Covid-19 as a global pandemic. Especially, attention has been increasingly drawn to the Covid-19 related racism and xenophobia which has manifested a more infectious nature and harmful consequences compared to the virus itself [Wang et al., 2021]. According to BBC report, throughout 2020, anti-Asian hate crimes increased by nearly one hundred and fifty percent, and there were around three thousand eight hundred anti-Asian racist incidents. Therefore, it has become urgent to comprehend public opinions regarding racism and xenophobia for the enactment of effective intervention policies preventing the evolvement of racist hate crimes and social exclusion under Covid-19. Social media as a critical public sphere for opinion expression provides platform for big social data analytics to understand and capture the dynamics of racist and xenophobic discourse alongside the development of Covid-19.

This research agenda has drawn attention from an increasing body of studies which have regarded Covid-19 as a social media infodemic [Cinelli et al., 2020], [Gencoglu and Gruber, 2020], [Trajkova et al., 2020], [Li et al., 2020], [Guo et al., 2021]. The work in [Schild et al., 2020] made an early and probably the first attempt to analyse the emergence of Sinophobic behaviour on Twitter and Reddit platforms. Soon after [Ziems et al., 2020] studied the role of counter hate speech in facilitating the spread of hate and racism against the Chinese and Asian community. The authors in [Vishwamitra et al., 2020] attempted to study the effect of hate speech on Twitter targeted on specific groups such as the older community and Asian community in general. The work in [Pei and Mehta, 2020] demonstrated the dynamic changes in the sentiments along with the major racist and xenophobic hashtags discussed across the early time period of Covid-19. The authors in [Masud et al., 2020] explored the user behavior which triggers the hate speech on Twitter and later how it diffuses via retweets across the network. All these methods have used highly advanced computational techniques and state-of-the-art language models for extracting insights from the data mined from Twitter and other platforms.

While focusing on technical advancement, many studies tend to neglect the foundation for accurate data detection and analysis – that is how to define racism and xenophobia. Especially, the computational techniques and models tend to apply a binary definition (either racist or non-racist) to categorise the linguistic features of the texts, with limited attention paid to the nuances of racist and xenophobic behaviours. However, understanding the nuances is critical for mapping the comprehensive picture of the development of racist and xenophobic discourse alongside the evolvement of Covid-19 – whether and how the expression of racism and xenophobia may change the topics across time. More importantly, capturing these changes reflected in the online public sphere will enable a more accurate comprehension and even prediction of public opinions and actions regarding racism and xenophobia in the offline world.

Reaching this goal demands a combination of computational methods and social science perspectives, which becomes the focus of this research. With the aid of BERT
3 Method

3.1 Category-based racism and xenophobia detection

Beyond a binary categorization of racism and xenophobia, this research applies the perspective of social science to categorizing racism and xenophobia into four dimensions as demonstrated in Table 1. This basically translates into a problem of five class classification of text data, where four classes represent the racism and xenophobia categories and fifth class corresponds to the category of non-racist and non-xenophobic.

Annotated dataset

For this purpose, we annotate a dataset of 6000 tweets. These tweets were randomly selected from all hashtags across the three development stages, and annotated by four research assistants with inter-coder reliability reaching above 70%. The annotation followed a coding method with 0 representing stigmatization, 1 for offensiveness, 2 for blame, and 3 for exclusion in alignment with the linguistic features of the tweets. The non-marked tweets were regarded as non-racist and non-xenophobic and represented class category 4. We limit the annotation for each tweet to only one label which aligns to the strongest category. The distribution of 6000 tweets amongst the five classes is as follows - 1318 stigmatization, 1172 offensive, 1045 blame, 1136 exclusion, and 1329 non-racist and non-xenophobic.

We view the task of classification of the above-mentioned categories as a supervised learning problem and target developing machine learning and deep learning techniques for the same. We firstly pre-process the input data text by removing punctuation and URLs from a text sample and converting it to lower case before providing it to train our models. We split the data into random train and test splits with 90:10 ratio for training and evaluating the performance of our models respectively.

BERT

Recently, word language models such as Bi-directional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] have become extremely popular due to their state-of-the-art performance on natural language processing tasks. Due to the nature of bi-directional training of BERT, it can learn the word representations from unlabelled text data powerfully and enables it to have a better performance compared to the other machine learning and deep learning techniques [Devlin et al., 2018]. The common approach for adopting BERT for a specific task on a smaller dataset is to fine-tune a pre-trained BERT model which has already learnt the deep context-dependent representations. We select the “bert-base-uncased” model which comprises of 12 layers, 12 self-attention heads, a hidden size of 768 totalling 110M parameters. We fine-tune the BERT model with a categorical cross-entropy loss for the five categories. The various hyperparameters used for fine-tuning the BERT model are selected as recommended from the paper [Devlin et al., 2018]. We use the AdamW optimizer with the standard learning rate of 2e-5, a batch size of 16, and train it for 5 epochs. For selecting the maximum length of the sequences, we tokenize the whole...
| Category       | Definition                                                                 | Example                                                                 |
|----------------|----------------------------------------------------------------------------|-------------------------------------------------------------------------|
| Stigmatization | Confirming negative stereotypes for conveying a devalued social identity within a particular context [Miller and Kaiser, 2001] | "For all the #ChinaVirus jumped from a bat at the wet market"          |
| Offensiveness  | Attacking a particular social group through aggressive and abusive language [Jeshion, 2013] | "Real misogyny in communist China. #chinazi #China_is_terrorist #China_is_terrorists #FuckTheCCP" |
| Blame          | Attributing the responsibility for the negative consequences of the crisis to one social group [Coombs and Schmidt, 2000] | "These Chinese are absolutely disgusting. They spread the #ChineseVirus. Their lies created a pandemic #ChinaMustPay" |
| Exclusion      | the process of othering to draw a clear boundary between in-group and out-group members [Bailey and Harindranath, 2005] | "China deserves to be isolated by all means forever. SARS was also initiated in China, 2003 by eating anything & everything #BoycottChina" |

Table 1: Definition and example of categorization of racist and xenophobic behaviors.

![Figure 1: Density distribution of token lengths of the tweets in our dataset.](image)

The dataset using Bert tokenizer and check the distribution of the token lengths. We notice that the minimum value of token length is 8, maximum is 130, median is 37 and mean is 42. Based on the density distribution shown in Fig.1, we experiment with two values of sequence length – 64 and 128 and find that the sequence length of 64 provides a better performance.

As additional baselines, we also train two more techniques. Long Short Term Memory Networks (LSTMs) [Hochreiter and Schmidhuber, 1997] have been very popular with text data as they can learn the dependencies of various words in the context of a text. Also, machine learning algorithms such as Support Vector Machine (SVMs) [Hearst et al., 1998] have been used previously by researchers for text classification tasks. We adopt the same data pre-processing and implementation technique as mentioned earlier and train the SVM with grid search, a 5-layer LSTM (using the pre-trained Glove [Pennington et al., 2014] embeddings) and BERT model for the category detection of the racist and xenophobic tweets.

For evaluating the machine learning and deep learning approaches on our test dataset, we use the metrics of average accuracy and weighted f1-score for the five categories. The performance of the model is shown in Table 2. It can be seen from Table 2 that the fine-tuned BERT model performs the best compared to SVM and LSTM in terms of both accuracy and f1 score. Thus, we employ this fine-tuned BERT model for categorizing all the tweets from the remaining dataset. Having employed BERT on the remaining dataset, we get a refined dataset of the four categories of tweets spreaded across the three stages as shown in Table 3.

### 3.2 Topic modelling

Topic modelling is one of the most extensively used methods in natural language processing for finding relationships across text documents, topic discovery and clustering, and extracting semantic meaning from a corpus of unstructured data [Jelodar et al., 2019]. Many techniques have been developed by researchers such as Latent Semantic Analysis (LSA) [Deerwester et al., 1990], Probabilistic Latent Semantic Analysis (pLSA) [Hofmann, 1999] for extracting semantic topic clusters from the corpus of data. In the last decade, Latent Dirichlet Allocation (LDA) [Blei et al., 2003b] has become a successful and standard technique for inferring topic clusters from texts for various applications such as opinion mining [Zhai et al., 2011], social medial analysis [Cohen and Ruths, 2013], event detection [Lin et al., 2010] and consequently there have also been various developed variants of LDA [Blei and McAuliffe, 2010] and [Blei et al., 2003a].

For our research, we adopt the baseline LDA model with
### Table 4: Extracted topics and their corresponding keywords for the category of stigmatization spread across the three stages S1, S2, and S3.

| S1 | S2 | S3 |
|----|----|----|
| T1. Virus | T1. Government | T1. Government |
| virus | china | china |
| spread | world | world |
| country | spread | lie |
| travel | country | lie |
| year | travel | pay |
| control | lie | pay |
| chinese | lie | pay |
| stop | lie | communists |
| animal | lie | government |
| source | lie | ccp |
| eat | lie | make |

### Table 5: Extracted topics and their corresponding keywords for the category of offensiveness spread across the three stages S1, S2, and S3.

| S1 | S2 | S3 |
|----|----|----|
| T1. Freedom | T1. Ccp | T1. Ccp |
| world | china | china |
| stop | ccp | ccp |
| freedom | lie | lie |
| truth | lie | lie |
| spread | lie | lie |
| good | lie | lie |
| happen | lie | lie |
| good | lie | lie |
| system | lie | lie |
| security | lie | lie |
| foreign | lie | lie |
| understand | lie | lie |

### Table 6: Extracted topics and their corresponding keywords for the category of blame spread across the three stages S1, S2, and S3.

| S1 | S2 | S3 |
|----|----|----|
| T1. Government | T1. World | T1. World |
| lie | world | china |
| spread | china | country |
| virus | pay | pandemic |
| autonomy | kill | global |
| deceit | kill | economy |
| monstrosity | threat | war |
| true | threat | economy |
| horrible | true | war |
| infect | country | high |

### Table 7: Extracted topics and their corresponding keywords for the category of exclusion spread across the three stages S1, S2, and S3.

| S1 | S2 | S3 |
|----|----|----|
| T1. Government | T1. Lie | T1. Lie |
| support | lie | lie |
| product | spread | chinese |
| world | virus | coronavirus |
| stop | death | government |
| govt | lie | protect |
| join | die | ccp |
| people | day | human |
| evil | day | country |
| time | order | china |
| tag | day | wuhan |
| stand | day | cover |
| sanction | day | day |
| government | day | body |
| economic | day | thing |
| inflation | day | care |

```
Variational Bayes sampling from Gensim\textsuperscript{2} and the LDA Mal-let model [McCallum, 2002] with Gibbs sampling for extracting the topic clusters from the text data. Before passing the corpus of data to the LDA models, we perform data pre-processing and cleaning which include the following steps. Firstly, we remove any new line characters, punctuations, URLs, mentions and hashtags. Later we tokenize the texts in the corpus and also remove any stopwords using the Gensim utility of pre-processing and stopwords defined in the NLTK\textsuperscript{3} corpus. Finally, we make bigrams and lemmatize the words in the text.

After employing the above pre-processing for our corpus, we employ topic modelling using LDA from Gensim and LDA Mallet. We perform experiments by varying the number of topics from 5 to 25 at an interval of 5 and checking the corresponding coherence score of the models as was done in [Fang et al., 2016]. We train the models for 1000 iterations with varying number of topics, optimizing the hyperparameters every 10 passes after each 100 pass period. We set the values of $\alpha$, $\beta$ which control the distribution of topics and the vocabulary words amongst the topics to the default settings of 1 divided by the number of topics. We notice from our experiments that LDA Mallet has a higher coherence score (0.60-0.65) compared to the LDA model from Gensim (0.49-0.55) and thus we select LDA Mallet model for the task of topic modelling on our corpus of data.

The above strategy is employed for each racist and xenophobic category and for every stage individually. We find the highest coherence score corresponding to a specific number of topics for each category and stage. To analyse the results, we reduce the number of topics to 5 by clustering closely related topics using equation 1.

\[
T_c = \frac{\left( \sum_{i=1}^{N} \sum_{j=1}^{M} p_j x_{ij} \right)}{N} \tag{1}
\]

where $N$ refers to the number of topics to be clustered, $M$ represents the number of keywords in each topic, $p_j$ corresponds to the probability of the word $x_i$ in the topic, and $T_c$ is the resultant topic containing the average probabilities of all the words from the $N$ topics. We then represent the top 10 highest probability words in the resultant topic for every category and stage as is shown in Tables 4 to 7.

4 Findings

Table 4, 5, 6 and 7 demonstrate the ten most salient terms related to the generated five topics for each stage (S1, S2, and S3) of four categories, and we summarize each topic through the correlation between the ten terms. We put a question mark 

\footnote{\url{https://pypi.org/project/gensim/}}
\footnote{\url{https://pypi.org/project/nltk/}}

For instance, as demonstrated in the findings, the topics falling under the category of offensiveness are more likely to be associated with sensitive political issues around China rather than virus in stage 1 and stage 2. Therefore, how to split the discussion of virus from the association of virus with other political topics should draw attention from government of different countries, and this agenda should be incorporated into the official media coverage from the government. Another example is from the category of blame. As shown in the findings, blame usually targets at the transparency of the information from the government (Chinese government especially in early Covid-19). Consequently, it is critical for government of different countries to work on effective and prompt communication with the public under Covid-19. We believe the contribution of this research can be generated beyond the context of Covid-19 to provide insights for future research on racism and xenophobia on digital platforms.
References

[Bailey and Harindranath, 2005] Olga Guedes Bailey and Ramaswami Harindranath. Racialised ‘othering’. Journal of Social Issues, 57(1):73–92, 2001.

[Pei and Mehta, 2020] Xin Pei and Deval Mehta. # coronavirus or# chinesevirus?!: Understanding the negative sentiment reflected in tweets with racist hashtags across the development of covid-19. arXiv preprint arXiv:2005.08224, 2020.

[Pennington et al., 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[Schild et al., 2020] Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. “go eat a bat, chang!”: An early look on the emergence of sinophobic behavior on web communities in the face of covid-19. arXiv preprint arXiv:2004.04046, 2020.

[Trajkova et al., 2020] Milka Trajkova, Francesco Cafaro, Sanika Vedak, Rashmi Mallappa, Sreekanth R Kankara, et al. Exploring casual covid-19 data visualizations on twitter: Topics and challenges. In Informatics, volume 7, 2020.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.

[Sanika Vedak, Rashmi Mallappa, Sreekanth R Kankara, et al. Exploring casual covid-19 data visualizations on twitter: Topics and challenges. In Informatics, volume 7, 2020.
[Vishwamitra et al., 2020] Nishant Vishwamitra, Ruijia Roger Hu, Feng Luo, Long Cheng, Matthew Costello, and Yin Yang. On analyzing covid-19-related hate speech using bert attention. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 669–676. IEEE, 2020.

[Wang et al., 2021] Simeng Wang, Xiabing Chen, Yong Li, Chloé Luu, Ran Yan, and Francesco Madrisotti. ‘i’m more afraid of racism than of the virus!’: racism awareness and resistance among chinese migrants and their descendants in france during the covid-19 pandemic. European Societies, 23(sup1):S721–S742, 2021.

[Zhai et al., 2011] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Constrained lda for grouping product features in opinion mining. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 448–459. Springer, 2011.

[Ziems et al., 2020] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv preprint arXiv:2005.12423, 2020.