Detecting East Asian Prejudice on Social Media

Bertie Vidgen
The Alan Turing Institute
bvidgen@turing.ac.uk

Austin Botelho
University of Oxford

David Broniatowski
The George Washington University

Ella Guest
The Alan Turing Institute

Matthew Hall
The University of Surrey

Helen Margetts
The Alan Turing Institute

Rebekah Tromble
The George Washington University

Zeerak Waseem
University of Sheffield

Scott A. Hale
University of Oxford
Meedan

Abstract

During COVID-19 concerns have heightened about the spread of aggressive and hateful language online, especially hostility directed against East Asia and East Asian people. We report on a new dataset and the creation of a machine learning classifier that categorizes social media posts from Twitter into four classes: Hostility against East Asia, Criticism of East Asia, Meta-discussions of East Asian prejudice, and Non-related. The classifier achieves a macro-F1 score of 0.83. We then conduct an in-depth ground-up error analysis and show that the model struggles with edge cases and ambiguous content. We provide the 20,000 tweet training dataset (annotated by experienced analysts), which also contains several secondary categories and additional flags. We also provide the 40,000 original annotations (before adjudication), the full codebook, annotations for COVID-19 relevance and East Asian relevance and stance for 1,000 hashtags, and the final model.

1 Introduction

The outbreak of COVID-19 has raised concerns about the spread of Sinophobia and other forms of East Asian prejudice across the world, with reports of online and offline abuse directed against East Asian people, including physical attacks (Flanagan, 2020; Wong, 2020; Liu, 2020; Walton, 2020; Solomon, 2020; Guy, 2020). The United Nations High Commissioner for Human Rights has drawn attention to these issues, calling on UN member states to fight the 'tsunami' of hate and xenophobia (Shields, 2020).

As digital technologies become even more important for maintaining social connections, it is crucial that online spaces remain safe, accessible and free from abuse (Cowls et al., 2020)—and that people's fears and distress are not exploited during the pandemic. Computational tools, including machine learning and natural language processing, offer powerful ways of creating scalable and robust systems for detecting and measuring prejudice. These, in turn, can assist with both online content moderation processes and further research into the dynamics, prevalence, causes, and impact of abuse.

We report on the creation of a new dataset and classifier to detect East Asian prejudice in social media data. The classifier distinguishes between four primary categories: Hostility against East Asia, Criticism of East Asia, Meta-discussions of East Asian prejudice, and Non-related. It achieves a macro-F1 score of 0.83. The 20,000 tweet training dataset used to create the classifier and the annotation codebook are also provided. The dataset contains annotations for several secondary categories, including threatening language, interpersonal abuse, and dehumanization, which can be used for further research. In addition, we provide the 40,000 original annotations given by our experienced annotators, which can be used for further investigation of annotating prejudice. We also annotated 1,000 hashtags in our dataset for East Asian relevance and stance, as well as other attributes. These are also provided for other researchers.[1]

To provide insight into what types of content causes the model to fail, we conduct a ground-up qualitative error analysis. We show that 17% of errors are due to annotation mistakes and 83% due to the machine learning model. Of these, 29% are clear errors (i.e. obvious mistakes) and 54% are edge cases (i.e. more complex and nuanced cases). In the machine learning edge cases, we show that the model struggles with lexically similar content (e.g. distinguishing Hostility against East Asia from Criticism of East Asia), as well as ambiguous content (i.e. where there is uncertainty among the annotators about the correct label).

[1] All research artefacts are available at: https://zenodo.org/record/3816667

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 162–172
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Finally, we analyze the hashtags most closely associated with the primary categories in the training dataset, identifying several terms which could guide future work.

2 Background

East Asian prejudice, such as Sinophobia, can be understood as fear or hatred of East Asia and East Asian people (Billé, 2015). This prejudice has a long history in the West: in the 19th century the term "yellow peril" was used to refer to Chinese immigrants who were stereotyped as dirty and diseased and considered akin to a plague (Goossen et al., 2004). Often, the association of COVID-19 with China plays into these old stereotypes, as shown by derogatory references to 'bats' and 'monkeys' online (Zhang, 2020). In 2017 a study found that 21% of Asian Americans had received threats based on their Asian identities, and 10% had been victims of violence (Neel, 2017). Likewise, a 2009 report on the discrimination and experiences of East Asian people in the UK, described Sinophobia as a problem that was 'hidden from public view' (Adamson et al., 2009).

New research related to East Asian prejudice during COVID-19 has already provided insight into its nature, prevalence and dynamics with Schild et al. (2020) finding an increase in Sinophobic language on some social media platforms such as 4chan. Analysis by the company Moonshot CVE also suggests that the use of anti-Chinese hashtags has increased substantially (The New Statesman, 2020). They analysed more than 600 million tweets and found that 200,000 contained either Sinophobic hate speech or conspiracy theories, and identified a 300% increase in hashtags that supported or encouraged violence against China during a single week in March 2020. Similarly, Velásquez et al. (2020) show the existence of a Sinophobic 'hate multiverse', with hateful content following contagion patterns and clusters which are similar to the epidemiological diffusion of COVID-19 itself.

Ziems et al. (2020) argue that racism 'is a virus', and study a dataset of 30 million COVID-19 tweets using a classifier trained on 2,400 tweets. They identify 900,000 hateful tweets and 200,000 counter-hate, and find that 10% of users who share hate speech are bot accounts. Toney et al. (2020) used the Word Embedding Association Test in the context of COVID-19 to analyse anti-China sentiment on Twitter, finding substantial biases in how Asian people are viewed. East Asian prejudice has also been linked to the spread of COVID-19 health-related misinformation (Cinelli et al., 2020) and in March 2020 the polling company YouGov found that 1 in 5 Brits believed the conspiracy theory that the coronavirus was developed in a Chinese lab (Nolsoe, 2020).

Research into computational tools for detecting, categorizing, and measuring online hate has received substantial attention in recent years (Waseem et al., 2017). However, a systematic review of hate speech training datasets conducted by Vidgen and Derczynski (2020) shows that classifiers and training datasets for East Asian prejudice are not currently available. Somewhat similar datasets are available, pertaining to racism (Waseem and Hovy, 2016) Islamophobia (Chung et al., 2019) and 'hate' in general (Davidson et al., 2017; de Gibert et al., 2018) but they cannot easily be re-purposed for East Asian prejudice detection. The absence of an appropriate training dataset (and automated detection tools) means that researchers have to rely instead on less precise ways of measuring East Asian prejudice, such as using keyword searches for slurs and other pejorative terms. These methods create substantial errors (Davidson et al., 2017) as more covert prejudice is missed because the content does not contain the target keywords, and non-hateful content is misclassified simply because it does contain a keyword.

Developing new detection tools is a complex and lengthy process. The field of hate speech detection sits at the intersection of social science and computer science and is fraught with not only technical challenges but also deep-routed ethical and theoretical considerations (Vidgen et al., 2019). If machine learning tools are to be effective then they need to be developed with consideration of their social implications (Vidgen et al., 2019; Sap et al., 2019; Davidson et al., 2019; Garg et al., 2019).

3 Dataset Collection

To create our 20,000 tweet training dataset, we collected tweets from Twitter's Streaming API using 14 hashtags that relate to both East Asia and the novel coronavirus.[2] Some of these hashtags ex-

[2] We query for: #chinavirus, #wuhan, #wuhanvirus, #chinavirusoutbreak, #wuhancoronavirus, #wuhaninfluenza, #wuhansars, #chinacoronavirus, #wuhan2020, #chinaflu, #wuhanquarantine, #chinesepneumonia, #coronachina and #wohan.

press anti-East Asian sentiments (e.g. '#chinaflu') but others, such as '#wuhan' are more neutral, referring to the geographic origin of the virus. Data collection ran initially from 11 to 17 March 2020, returning 769,763 tweets, of which 96,283 were unique entries in English. To minimize biases that could emerge from collecting data over a relatively short period of time, we then collected tweets from 1 January to 10 March 2020 using the same keywords from a 10% random sample of Twitter (the 'Decahose'), provided by a third party. We identified a further 63,037 unique tweets in English, which we added to our dataset. The full dataset comprises 159,320 unique tweets.

To create a training dataset for annotation we sampled from the full dataset. To guide the sampling process, we extracted the 1,000 most used hashtags from the 159,320 tweets. Three annotators independently marked them for: (1) whether they are East Asian relevant and, if so, (2) what Asian entity is discussed (e.g., China, Xi Jinping, South Korea), (3) what the stance is towards the Asian entity (Very Negative, Negative, Neutral, Positive, or Very Positive) and also (4) whether they relate to COVID-19. 97 hashtags were marked as either Negative or Very Negative toward East Asia by at least one annotator. All annotations for hashtags are available in our data repository.

We then sampled 10,000 tweets at random from the full dataset and a further 10,000 tweets which used one of the 97 hashtags identified as Negative or Very Negative towards East Asia, thereby increasing the likelihood that prejudicial tweets would be identified and ensuring that our dataset is suitable for training a classifier (Schmidt and Wiegand, 2017; Vidgen et al., 2019). The training dataset comprises 20,000 tweets in total.

3.1 Data pre-processing for annotation

Initial qualitative inspection of the dataset showed that hashtags played a key role in how COVID-19 was discussed and how hostility against East Asia was expressed. Hashtags often appeared in the middle of tweets, especially when they related to East Asia and/or COVID-19. For example:

> its wellknown #covid19 originated from #china. Instead of #Doingtherightthing they're blaming others, typical. You cant trust these #YellowFever to sort anything out.

Without the hashtags it is difficult to discern whether this tweet expresses prejudice against East Asia. In this regard, it is important that they are seen by annotators to ensure high quality labels. However, in other cases, the inclusion of hashtags risked low quality labels. In a test round of annotation (not included in the dataset presented here) annotators over-relied on the prejudicial hashtags, marking up nearly all tweets which contained them as prejudiced against East Asia, even if they were otherwise neutral. This is problematic because we used hashtags to sample the training data so they are highly prevalent. If all of their uses were identified as prejudicial then any systems trained on the dataset would likely overfit to just a few keywords. This could severely constrain the system's generalizability, potentially leading to poor performance on new content.

To address this challenge, we performed a hashtag replacement on all tweets prior to presenting them to the annotators. For the 1,000 most used hashtags (annotated as part of the data sampling phase), we had one annotator identify appropriate *thematic replacement* hashtags. We used five thematic replacements:

- #EASTASIA: Relate only to an East Asian entity, e.g. #China or #Wuhan
- #VIRUS: Relate only to COVID-19, e.g. #coronavirus or #covid19.
- #EASTASIAVIRUS: Relate to both an East Asian entity and COVID-19, e.g. #wuhanflu.
- #OTHERCOUNTRYVIRUS: Relate to both a Country (which is not East Asian) and COVID-19, e.g. #coronacanada or #italycovid.
- #HASHTAG: Not directly relevant to COVID-19 or East Asia, e.g. #maga or #spain.

Annotators could still discern the meaning of tweets because they were presented with the hashtags' topics. However, they were not unduly biased by the substantive *outlook*, stance, and sentiment the hashtags express. All hashtags beyond our annotated list of 1,000 were replaced with a generic replacement, #HASHTAG. The 1,000 thematic hashtag replacements are available in our data repository. With this process, the quote above is transformed to:

> its wellknown #HASHTAGVIRUS originated from #HASHTAGEASTASIA. Instead of #HASHTAG they're blaming others, typical. You cant trust these #HASHTAGEASTASIAVIRUS to sort anything out.

This process, although time consuming, strikes a balance between preserving the meaning of the tweet for annotation and minimizing the risk of overfitting. It means that the text the annotators are presented with is also the same text that is fed into our final models (i.e. both annotations and the model classifications are based on the replaced hashtags). In principle, it would be easy for anyone applying our final classification model to a new dataset to update the hashtag replacement list and then apply it to their data.

4 Dataset Annotation and Taxonomy

4.1 Annotators

Annotation was completed by a team of 26 annotators, all of whom had completed at least 4 weeks of training on a previous hate speech annotation project. The annotators were all aged between 18 and 35, spoke English fluently (50% were native speakers), were 75% female and were all educated to at least an undergraduate level. 25% were studying for higher degrees. Annotators came from the United Kingdom (60%), elsewhere in Europe (30%) and South America (10%). Information about their sexuality, religious and political affiliation is not available due to their sensitivity.

Two experts were used to adjudicate decisions on the primary categories. Both experts were final year PhD students working on extreme behaviour online. One was male; one was female. They were both aged between 25 and 35, and were native English speakers.

4.2 Themes

Tweets were first annotated for the presence of two themes: (1) East Asia and (2) COVID-19. If a tweet was not East Asian relevant then no further annotations were required and it was automatically assigned to the Non-related class. Annotators then used an additional flag for how they marked up the two themes, which we call 'hashtag dependence.' For this label, annotators were asked whether they had used the hashtags to identify the themes or the themes were apparent without the hashtags.

Our approach to annotating themes and the role of hashtags required substantial training for annotators (involving one-to-one on-boarding sessions). This detailed annotation process means that we can provide insight into not only what annotations were made but also *how*, which we anticipate will be of use to scholars working on online communications beyond online prejudice.

4.3 Primary categories

Each tweet was assigned to one of five mutually exclusive primary categories.

- **Hostility against an East Asian (EA) entity**: Express abuse or intense negativity against an East Asian entity, primarily by derogating/attacking them (e.g. "Those oriental devils don't care about human life" or "Chinks will bring about the downfall of western civilization"). It also includes: conspiracy theories, claiming East Asians are a threat, and expressing negative emotions about them.

- **Criticism of an East Asian entity**: Make a negative judgement/assessment of an East Asian entity, without being abusive. This includes commenting on perceived social, economic and political faults, including questioning their response to the pandemic and how they are governed.

  The Hostility/Criticism distinction is crucial for addressing a core issue in online hate speech research, namely ensuring that freedom of speech is protected (Ullmann and Tomalin, 2020). The Criticism category minimizes the chance that users who engage in what has been termed 'legitimate critique' (Imhoff and Recker, 2012) will have their comments erroneously labelled as hostile.

- **Counter speech**: Explicitly challenge or condemn abuse against an East Asian entity. It includes rejecting the premise of abuse (e.g., "it isn't right to blame China!"), describing content as hateful or prejudicial (e.g., "you shouldn't say that, it's derogatory") or expressing solidarity with target entities (e.g., "Stand with Chinatown against racists").

- **Discussion of East Asian prejudice** Tweets that discuss prejudice related to East Asians but do not engage in, or counter, that prejudice (e.g., "It's not racist to call it the Wuhan

Theme	Number of Entries	Percentage
COVID-19 relevant / Both said No	2,940	14.7%
COVID-19 relevant / Both said Yes	12,255	61.3%
COVID-19 relevant / Disagreement	4,805	24.0%
East Asian relevant / Both said No	6,593	33.0%
East Asian relevant / Both said Yes	9,790	49.0%
East Asian relevant / Disagreement	3,617	18.0%

Table 1: Prevalence of themes in the dataset.

Category	Number of Entries	Percentage
Hostility	3,898	19.5%
Criticism	1,433	7.2%
Counter speech	116	0.6%
Discussion of EAP	1,029	5.1%
Non-related	13,528	67.6%
TOTAL	**20,000**	**100%**

Table 2: Prevalence of primary categories in the dataset.

Measure	Mean	Min.	Max.
Percentage agreement	78%	67%	84%
Fleiss' Kappa			
All categories	0.54	0.36	0.66
Hostility	0.53	0.22	0.66
Criticism	0.27	0.14	0.41
Counter Speech	0.33	0.11	0.61
Discussion of EAP	0.46	0.14	0.65
Non-related	0.64	0.51	0.78

Table 3: Agreement scores for primary categories.

virus”). It includes content which discusses whether East Asian prejudice has increased during COVID-19, the supposed media focus on prejudice, and/or free speech.

- **Non-related** Do not fall into any of the other categories. Note that they could be abusive in other ways, such as expressing misogyny.

The primary categories were annotated with a two step process. First, each tweet was annotated independently by two trained annotators. Second, one of two expert adjudicators reviewed cases where annotators disagreed about the primary category. Experts could decide an entirely new primary category if needed. Expert adjudication was not used for the themes and secondary categories.

Agreement is reported for each pair of annotators in Table 3, with the average, minimum, and maximum. Overall, agreement levels are moderate, with better results for the two most important and prevalent categories (Hostility and Non-related) but poorer on the less frequent and more nuanced categories (Counter Speech, Criticism and Discussion of EA prejudice). Note that if Counter Speech and Discussion of EA prejudice are combined then there is a marked improvement in overall agreement levels, with an average Kappa of 0.5 for the combined category.

Experts adjudicated 4,478 cases (22%) where annotators did not agree. Experts tended to move tweets out of Non-related into other categories, primarily Hostility. Of the 8,956 original annotations given to the 4,478 tweets they adjudicated, 34% of them were in Non-related and yet only 29% of their adjudicated decisions were in this category. This was matched by an equivalent increase in the Hostility category, from 31.6% of the original annotations to 35% of the expert adjudications. The other three categories remained broadly stable. In 347 cases (7.7%), experts choose a category that was not selected by either annotator. Of the 694 original annotations given to these 347 cases, 18.7% were for Criticism compared with 39.4% of the expert adjudications for these entries (a similar decrease can be observed for the Non-related category). The most common decision made by experts for these 347 tweets was to label a tweet as Criticism when one annotator had selected Hostility and the other selected Non-related. These results shows the fundamental ambiguity of hate speech annotation and the need for expert adjudication. With complex and often-ambiguous content even well-trained annotators can make decisions which are inappropriate.

4.4 Secondary categories

For the Hostility and Criticism primary categories, annotators identified what East Asian entity was targeted (e.g., “Hong Kongers”, “CCP”, or “Chinese scientists”). Initially, annotators identified targets inductively, which resulted in several hundred unique values. We then implemented a reconciliation process in which the number of unique targets was reduced to 78, reflecting six geographical areas (China, Korea, Japan, Taiwan, Singapore and East Asia in general) and several specific entities, such as scientists, women and government, including intersectional identities. For tweets identified as Hostility annotators applied three additional flags.

- Interpersonal abuse: East Asian prejudice which is targeted against an individual. Whether the individual is East Asian was not considered. (Waseem et al., 2017).
- Use of threatening language: Content which makes a threat against an East Asian entity, which includes expressing a desire/willingness to inflict harm or inciting others (The Law Commission, 2018; Weber, 2009).
- Dehumanization: Content which describes, compares or suggests equivalences between East Asians and non-humans or subhumans, such as insects, weeds, or actual viruses (Leader Maynard and Benesch, 2016; Musolff, 2015).

Note that our expert adjudicators did not adjudicate for these secondary categories. In cases where experts decided a tweet is Hostile but neither of the original annotators had selected that category then none of the secondary categories are available. In other cases, experts decided a tweet was Hostile and so only one annotation for the secondary flags is available (as the other annotator selected a different category and did not provide these secondary annotations). Future researchers can decide how to use these secondary categories.

5 Classification results

Due to their low prevalence and conceptual similarity, we combined the Counter Speech category with Discussion of East Asian Prejudice for classification. As such, the classification task was to distinguish between four primary categories: Hostility, Criticism, Discussion of East Asian Prejudice and Non-related.

Model	Macro F1	Recall	Precision
LSTM	0.76	0.67	**0.88**
AlBERT$_{xlarge}$	0.798	0.798	0.800
BART$_{large}$	0.813	0.812	0.834
BERT$_{large}$	0.823	0.823	0.827
DistilBERT$_{base}$	0.803	0.803	0.809
ELECTRA$_{large}$	0.831	0.831	0.836
RoBERTa$_{large}$	**0.832**	**0.832**	0.848
XLNet$_{large}$	0.802	0.802	0.822

Table 4: Classification performance of models on the test set.

We implemented and fine-tuned several contextual embedding models as well as a one-hot LSTM model with a linear input layer, tanh activation, and a softmax output layer. We expect contextual embeddings to perform best as they take into account the context surrounding a token when generating each embedding (Vaswani et al., 2017). We compared results against a one-hot LSTM model to test this expectation.

Models were developed with a stratified 80/10/10 training, testing, and validation split (maintaining the class distribution of the whole dataset). We processed all tweets by removing URLs and usernames, lower-casing, and replacing hashtags with either a generic hashtag-token or with the appropriate thematic hashtag-token from the annotation setup. Training was conducted using the same hyper-parameter sweep identified in Liu et al. (2019) as most effective for the GLUE benchmark tasks. This included testing across learning rates $\in \{$1e-5, 2e-5, 3e-5$\}$ and batch sizes $\in \{$32, 64$\}$ with an early stopping regime. Performance was optimized using the AdamW algorithm (Loshchilov and Hutter, 2019) and a scheduler that implements linear warmup and decay. For the LSTM baseline, we conduct a hyper-parameter search over batch sizes $\in \{$16, 32, 64$\}$ and learning rates $\{10^{-i}, i \in 1, 5$ increments$\}$.

All of the contextual embedding models outperformed the baseline in terms of macro F1. RoBERTa achieved the highest F1 score of the tested models (0.832), which is a 7-point improve-

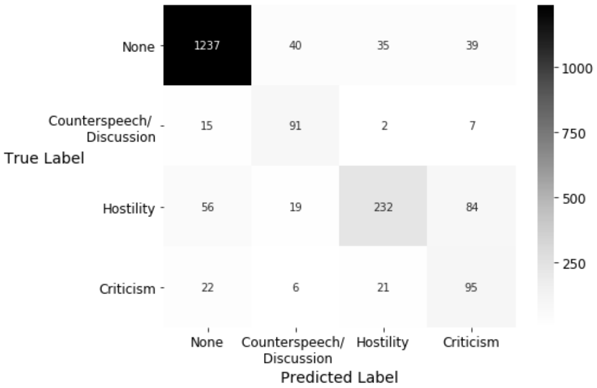

Figure 1: Confusion matrix for RoBERTa classifications on the test set.

ment over the LSTM (0.76). This model harnesses the underlying bidirectional transformer architecture of BERT (Devlin et al., 2019) but alters the training hyperparameters and objectives to improve performance. Unexpectedly, the LSTM baseline outperforms all other models in terms of precision but has far lower recall. For the best performing model (RoBERTa), misclassifications are shown in the confusion matrix. The model performs well across all categories, with strongest performance in detecting tweets in the Non-related category (Recall of 91.6% and Precision of 93%). The model has few misclassifications between the most conceptually distinct categories (e.g. Hostility versus Non-related) but has far more errors between conceptually similar categories, such as Criticism and Hostility.

6 Error analysis

To better understand classification errors, we conducted a qualitative analysis of misclassified content (from the best performing model, RoBERTa), using a grounded theory approach (Corbin and Strauss, 1990). This qualitative methodology is entirely inductive and data-driven. It involves systematically exploring themes as they emerge from the data and organizing them into a taxonomy—refining and collapsing the categories until 'saturation' is reached and the data fits neatly into a set of mutually exclusive and collectively exhaustive categories. Figure 2 shows the error categories within our sample of 340 misclassified tweets from the 2,000 (10%) validation split. The errors broadly fit within two branches: annotator errors (17%) and machine learning errors (83%). In future work, these errors could be addressed through creating a larger and more balanced dataset, more sophisticated machine learning architectures and reannotation of data.

6.1 Annotator errors (17% of total)

Annotator errors are cases where the classification from the model better captures the tweets' content and is more consistent with our taxonomy and guidelines. In effect, we believe that the 'wrong' classification provided by the model is correct—and that a mistake may have been made in the annotation process. Approximately 17% (N=58) of the errors were due to this. Note that this does not mean that 17% of the dataset is incorrectly annotated as this sample is biased by the fact that it has been selected precisely because the model made an 'error'.

36 of the annotator errors were clear misapplications of primary categories. The other 22 were cases where annotators made detailed annotations for tweets which were incorrectly marked as East Asian relevant. These are *path dependency errors* and show the importance of annotators following the right instructions throughout the process. If an incorrect annotation is made early on then the subsequent annotations are likely to be flawed.

6.2 Prediction errors (83% of total)

83% of the total errors were due to errors from the model. We have separated these into clear errors and edge cases. Clear errors are where the model has made an error that is easily identified by humans (accounting for 29% of all errors). Edge-cases are where the misclassified content contains some ambiguity and the model misclassification has some merit (accounting for 54% of all errors).

Clear error (Lexical similarity), 16% In several cases the misclassified tweets were clearly assigned to the wrong class. This suggests possible overfitting as the tweets were often lexically similar to tweets which did belong in the category (e.g., they contained phrases such as 'Made in China' and were mistaken for Hostility). This is most likely a learned over-sensitivity and could only be addressed through using a far larger dataset.

Clear error (Target confusion), 13% The model sometimes identified tweets which were not East Asian relevant as Criticism, Hostility, or Discussion of East Asian prejudice. Aside from this, the classifications were correct, i.e. the tweets expressed hostility against another identity, such as

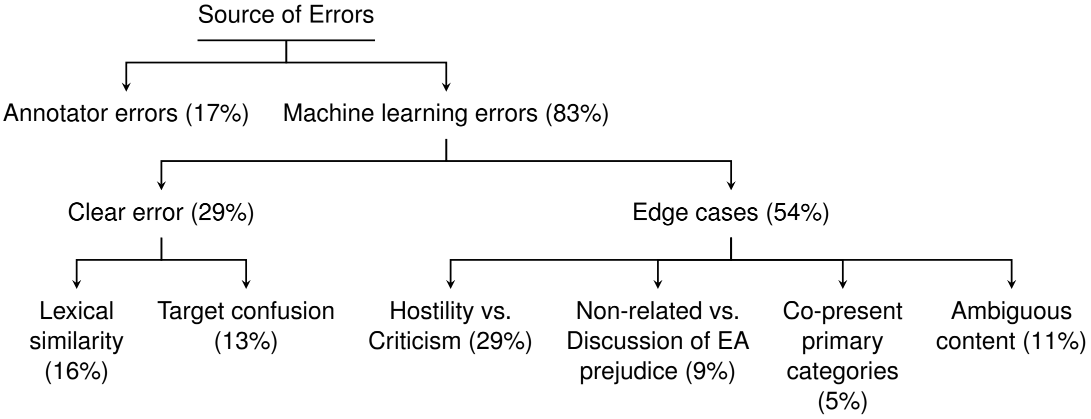

Figure 2: Sources of classification error.

women or gay people. Furthermore, in many cases, a relevant East Asian entity (e.g. China) was usually referred to but was not the object of the tweet, creating a mixed signal for the model.

Edge case (Hostility vs. Criticism), 29% Misclassifying Hostility as Criticism (and vice versa) was the largest source of error. The model particularly struggled with cases where criticism was framed normatively or expressed with nuanced linguistic expressions, e.g., "gee, china was lying to us. what a bloody shock".

Edge case (Non-related vs. Discussion of EA prejudice), 9% The model misclassified Non-related as Discussion of East Asian prejudice in several tweets. These were usually cases where the virus was named and discussed but *prejudice* was not discussed explicitly, e.g. "corona just shows why you should blame all our problems with China on Trump".

Edge case (Co-present primary categories), 5% In our taxonomy, annotators could assign each tweet to only one primary category. However, in some cases this was problematic and the model identified a co-present category rather than the primary category which had been annotated.

Edge case (Ambiguous content), 11% The model often misclassified content which was ambiguous. This is content where the true meaning is not immediately discernible to a human without careful analysis. For instance, positively framed criticism, e.g. "so glad that china official finally admits the HASHTAGEASTASIA+VIRUS outbreaks". In some tweets, complex forms of expression were used, such as innuendo or sarcasm, e.g. "I think we owe you china, please accept our apologies to bring some virus into your great country".

6.3 Addressing Classification Errors

Classifying social media data is notoriously difficult. There are many different sources of error, each of which, in turn, require different remedies. Annotator errors, for example, illustrate the need for robust annotation processes and providing more support and training when taxonomies are applied. Removing such errors entirely is unlikely, but the number of obvious misclassifications could be minimized.

Machine learning errors are where the bulk of the errors fall (83%). Edge cases are a particularly difficult type of content for classification. They can be expected in any taxonomy that draws distinct lines between complex and non-mutually exclusive concepts, such as Hostility and Criticism. Nonetheless, larger and more balanced datasets (with more instances of content in each category) would help in reducing this source of error. Equally, the frequency of non-edge case machine learning errors (i.e. cases where the model made a very obvious mistake) could be addressed by larger datasets as well as more advanced machine learning architectures.

7 Hashtag analysis

We analysed the 20,000 tweet training dataset to understand which hashtags were associated with which primary categories. For each category, we filtered the data so only hashtags which appeared in at least ten tweets assigned to that category were included. Then, we ranked the hashtags by the percentage of their uses which were in the primary category—our goal being to understand which hashtags are most closely associated with that category. For brevity, only the twenty hashtags most closely associated with the Hostility category are shown in Table 5. This analysis is only possible because all hashtags were replaced in the tweets which were presented to annotators (either with a generic hashtag token or thematic replacement), letting us conduct meaningful analysis of the co-occurrence of hashtags with the annotated primary categories.

A small number of hashtags are highly likely to only appear in tweets that express Hostility against East Asian entities. These hashtags could be used to filter for prejudiced discourses online and, in some cases, their uses may intrinsically indicate prejudice. Surprisingly, many seemingly hostile hashtags against East Asia, such as #fuckchina and #blamechina are not always associated with hostile tweets (Hostility accounts for 67.5% and 60.5% of their total use, respectively). This shows the importance of having a purpose-built machine learning classifier for detecting East Asian hostility, rather than relying on hashtags and keywords alone.

Hashtag	# in Hostile Tweets	% of All Uses	# of Total Uses
#rule2	20	87%	23
#rule3	17	85%	20
#rule1	22	85%	26
#makechinapay	18	72%	25
#hkgovt	22	71%	31
#fuckchina	54	68%	80
#blamechina	23	61%	38
#batsoup	15	60%	25
#hkairport	11	55%	20
#huawei	16	53%	30
#boycottchina	185	53%	350
#communismkills	14	52%	27
#communistchina	34	51%	67
#chinaisasshoe	41	51%	81
#chinapropaganda	10	50%	20
#china_is_terrorist	168	49%	345
#xijingping	17	49%	35
#chinashould apologize	14	48%	29
#madeinchina	39	48%	82
#ccp	395	47%	850

Table 5: Hashtags in Hostile tweets.

8 Conclusion

Prejudice of all forms is a deeply concerning problem during COVID-19, reflecting the huge social costs and difficulties that the pandemic has inflicted. In this paper we have reported on development of several research artefacts that we hope will enable further research into East Asian prejudice, including a trained classifier, a training dataset (20,000 entries), annotations dataset (40,000 entries), 1,000 annotated hashtags, a list of hashtag replacements, a list of hashtags associated with hostility against East Asia, and the full codebook (with extensive guidelines, information and examples).

One concern with any model is that it will not generalize to new settings and platforms, limiting its utility for real-world applications. Our hashtag replacement method was adopted to increase the generalizability of the final model, maximizing the likelihood that it can be applied to new contexts—as it would pick up on the semantic features of hostile tweets rather than specific tokens. Nonetheless, we caution that any re-use of the model should be accompanied by additional testing to understand its performance on different data.

Acknowledgments

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Criminal Justice System" theme within that grant, and The Alan Turing Institute. Scott Hale was supported in part by the Volkswagen Foundation.

References

Sue Adamson, Bankole Cole, Gary Craig, Basharat Hussain, Luana Smith, Ian Law, Carmen Lau, Chak-Kwan Chan, and Tom Cheung. 2009. *Hidden from public view? Racism against the UK Chinese population*. Sheffield Hallam University.

Franck Billé. 2015. *Sinophobia: anxiety, violence, and the making of Mongolian identity*. University of Hawaii Press.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 2819–2829.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 Social Media Infodemic. *arXiv:2003.05004*.

Juliet Corbin and Anselm Strauss. 1990. Grounded Theory Research: Procedures, Canons and Evaluative Criteria. *Qualitative Research*, 13(1):3–21.

Josh Cowls, Bertie Vidgen, and Helen Margetts. 2020. Why content moderators should be key workers protecting social media as critical infrastructure during covid-19. *The Alan Turing Institute*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 1–11.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th ICWSM*, pages 1–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, page 4171–4186.

Ryan Flanagan. 2020. Canada's top doctor calls out 'racism and stigmatizing comments' over coronavirus. *CTVNews*.

Sahaj Garg, Ankur Taly, Vincent Perot, Ed H. Chi, Nicole Limtiaco, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the Second Workshop on Abusive Language Online (ACL)*, pages 11–20.

Tam Goossen, Jian Guan, and Ito Peng. 2004. Yellow Peril Revisited: Impact of SARS on the Chinese and Southeast Asian Canadian Communities June, 2004 Project Coordinator and Author: Carrianne Leung. *Resources for Feminist Research*, 33(1-2):135–150.

Jack Guy. 2020. East Asian student assaulted in 'racist' coronavirus attack in London. *CNN*.

Roland Imhoff and Julia Recker. 2012. Differentiating Islamophobia: Introducing a New Scale to Measure Islamoprejudice and Secular Islam Critique. *Political Psychology*, 33(6):811–824.

Jonathan Leader Maynard and Susan Benesch. 2016. Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention. *Genocide Studies and Prevention*, 9(3):70–95.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Yuebai Liu. 2020. Coronavirus prompts 'hysterical, shameful' Sinophobia in Italy. *Al Jazeera*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR 2019*, pages 1–18.

Andreas Musolff. 2015. Dehumanizing metaphors in UK immigrant debates in press and online media. *Journal of Language Aggression and Conflict*, 3(1):41–56.

Joe Neel. 2017. Poll: Asian-Americans See Individuals' Prejudice As Big Discrimination Problem. *NPR*.

Eir Nolsoe. 2020. COVID-19: Bogus claims fool Britons. *YouGov*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, Noah A Smith, and Paul G Allen. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 1668–1678.

Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. 2020. "Go eat a bat, Chang!": An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. *arXiv:2004.04046*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.

Michael Shields. 2020. U.N. asks world to fight virus-spawned discrimination. *Reuters*.

Salem Solomon. 2020. Coronavirus Brings 'Sinophobia' to Africa. *VOA News*.

The Law Commission. 2018. *Abusive and Offensive Online Communications: A scoping report*. The Law Commission.

The New Statesman. 2020. Covid-19 has caused a major spike in anti-Chinese and anti-Semitic hate speech. *The New Statesman*.

Autumn Toney, Akshat Pandey, Wei Guo, David Broniatowski, and Aylin Caliskan. 2020. Pro-Russian Biases in Anti-Chinese Tweets about the Novel Coronavirus. *arXiv:2004.08726*.

Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 1–11.

N. Velásquez, R. Leahy, N. Johnson Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. Jha, B. Goldberg, and N.F. Johnson. 2020. Hate multiverse spreads malicious covid-19 content online beyond individual platform control. *arXiv:2004.00673*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data: Garbage In, Garbage Out. *arXiv:2004.01670*, pages 1–26.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online (ACL)*, pages 80–93.

Kate Walton. 2020. Wuhan Virus Boosts Indonesian Anti-Chinese Conspiracies. *Foreign Policy*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL-HLT*, pages 88–93.

Anne Weber. 2009. *Manual on Hate Speech*. Council of Europe.

Tessa Wong. 2020. Sinophobia: How a virus reveals the many ways China is feared. *BBC News*.

Zhang. 2020. Pinning coronavirus on how chinese people eat plays into racist assumptions.

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