Detecting East Asian Prejudice on Social Media

Bertie Vidgen[1,2], Austin Botelho[2], David Broniatowski[3], Ella Guest[1,6], Matthew Hall[1,4], Helen Margetts[1,2], Rebekah Tromble[1,3], Zeerak Waseem[5], and Scott Hale[1,2]

[1]The Alan Turing Institute
[2]The Oxford Internet Institute
[3]The George Washington University
[4]The University of Surrey
[5]University of Sheffield
[6]The University of Manchester

May 2020

Abstract The outbreak of COVID-19 has transformed societies across the world as governments tackle the health, economic and social costs of the pandemic. It has also raised concerns about the spread of hateful language and prejudice online, especially hostility directed against East Asia. In this paper we report on the creation of a classifier that detects and categorizes social media posts from Twitter into four classes: Hostility against East Asia, Criticism of East Asia, Meta-discussions of East Asian prejudice and a neutral class. The classifier achieves an F1 score of 0.83 across all four classes. We provide our final model (coded in Python), as well as a new 20,000 tweet training dataset used to make the classifier, two analyses of hashtags associated with East Asian prejudice and the annotation codebook. The classifier can be implemented by other researchers, assisting with both online content moderation processes and further research into the dynamics, prevalence and impact of East Asian prejudice online during this global pandemic.[1]

Keywords Hate speech, Sinophobia, Prejudice, Social media, Covid-19, Twitter, East Asia, Online abuse.

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 in the first few months of the pandemic, including physical attacks [1, 2, 3, 4, 5, 6]. The United Nations High Commissioner for Human Rights has drawn attention to increased prejudice against people of East Asian background and has called on UN member states to fight such discrimination [7]. Thus far, most of the academic response to COVID-19 has focused on understanding its health- and economic- impacts and how these can be mitigated [8]. There is a pressing need to also research and understand other forms of harm and danger which are spreading during the pandemic.

Social media is one of the most important battlegrounds in the fight against social hazards during COVID-19. As life moves increasingly online, it is crucial that social media platforms and other online spaces remain safe, accessible and free from abuse [9] – and that people's fears and distress during this time are not exploited and social tensions stirred up. Computational tools, utilizing recent advances in machine learning and natural language processing, offer powerful ways of creating scalable and robust models for detecting and measuring prejudice. These, in turn, can assist with both online content moderation processes and further research into the dynamics, prevalence and impact of East Asian prejudice.

In this paper we report on the creation of a classifier to detect East Asian prejudice in social media data. It distinguishes between four primary categories: Hostility against East Asia, Criticism of East Asia, Meta-discussions of East Asian prejudice and a neutral class. The classifier achieves an F1 score of 0.83. We also provide a new 20,000 tweet training dataset used to create the classifier, the annotation codebook and two analyses of hashtags associated with East Asian prejudice. The training dataset contains annotations for several secondary categories, including threatening language, interpersonal abuse and dehumanization, which can be used for further research. [2]

2 Literature Review

East Asian prejudice, such as Sinophobia, can be understood as fear or hatred of East Asia and East Asian people [10]. This prejudice has a long history in the West: in the 19th century the term "yellow peril" was used to refer to Chinese

[1] This work is a collaboration between The Alan Turing Institute and the Oxford Internet Institute. It was funded by the Criminal Justice Theme of the Alan Turing Institute under Wave 1 of The UKRI Strategic Priorities Fund, EPSRC Grant EP/T001569/1.

[2] All research materials are available at `https://zenodo.org/record/3816667`.

immigrants who were stereotyped as dirty and diseased, and considered akin to a plague [11]. The association of COVID-19 with China plays into these centuries old stereotypes, as shown by derogatory references to 'bats' and 'monkeys' [12]. Similar anti-Chinese prejudices emerged during the SARS outbreak in the early 2000s, with lasting adverse social, economic, and political impacts on East Asian diasporas globally [13].

A 2019 Pew survey examined attitudes towards China from people in 34 countries. A median of 41% of citizens had an unfavorable opinion of China. Negative opinions were particularly common in North America, Western Europe, and neighboring East Asian countries [14]. The 2019 survey, which was conducted just before the pandemic, marked a historic high in unfavorable attitudes towards China. Similarly, In 2017, a study found that 21% of Asian Americans had received threats based on their Asian identity, and 10% had been victims of violence [15]. Likewise, a 2009 report by the Universities of Hull, Leeds and Nottingham Trent reported on the discrimination and attacks that East Asian people are subjected to in the UK, describing Sinophobia as a problem that was 'hidden from public view' [16]. Official government statistics on hate crimes are not currently available for East Asian prejudice as figures for racist attacks are not broken down by type [17]). There is relatively little research into the causal factors behind East Asian prejudice, although evidence suggests that some people may feel threatened by China's growing economic and military power [18].

New research related to COVID-19 has already provided insight into the nature, prevalence and dynamics of East Asian prejudice, with Schild et al. demonstrating an increase in Sinophobic language on some social media platforms, such as 4chan [19]. Analysis by the company Moonshot CVE also suggests that the use of anti-Chinese hashtags has increased substantially. 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 support or encourage violence against China during a single week in March 2020 [20]. East Asian prejudice has also been linked to the spread of COVID-19 health-related misinformation [21]. 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 [22]. During this time of heightened tension, prejudice and misinformation could exacerbate and reinforce each other, making online spaces deeply unpleasant and potentially even dangerous.

Research into computational tools for detecting, categorizing and measuring hate speech has received substantial attention in recent years, contributing to online content moderation processes in industry, and enabling new scientific research avenues [23]. However, a systematic review of hate speech training datasets conducted by Vidgen and Derczynski shows that classifiers and training datasets for East Asian prejudice are not currently available [24]. Somewhat similar datasets are available, pertaining to racism [25] Islamophobia [26] and 'hate' in general [27, 28] but they cannot easily be repurposed for East Asian prejudice detection. The absence of an appropriate training dataset (and, as such, automated detection tools) means that researchers have to rely instead on far cruder ways of measuring East Asian prejudice, such as searching for slurs and other pejoratives. These methods drive substantial errors [29] as lots of less overt prejudice is missed because the content does not contain the target keywords, and non-hateful content is misclassified just because it does contain the keywords.

However, 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 [30]. Recent studies show that many existing datasets and tools contain substantial biases, such as overclassifying African American vernacular as hateful compared with Standard American [31, 32], penalising hate against certain targets more strongly than others [33], or being easily fooled by simple spelling changes that any human can identify [34]. These issues are considerable limitations as they mean that, if used in the 'real world', computational tools for hate speech detection could not only be ineffective, they could potentially perpetuate and deepen the social injustices they are designed to address. Put simply, whilst machine learning presents many exciting research opportunities, it is no panacea and tools need to be developed in dialogue with social science research if they are to be effective [30].

3 Dataset Collection

To create a 20,000 tweet training dataset, we collected tweets from Twitter's Streaming API, using 14 hashtags which relate to both East Asia and the Virus: #chinavirus, #wuhan, #wuhanvirus, #chinavirusoutbreak, #wuhancoronavirus, #wuhaninfluenza, #wuhansars, #chinacoronavirus, #wuhan2020, #chinaflu, #wuhanquarantine, #chinesepneumonia, #coronachina and #wohan. Some of these hashtags express anti-East Asian sentiments (e.g. '#chinaflu') but others, such as '#wohan' are more neutral. Data collection ran initially from 11 to 17 March 2020, and we collected 769,763 tweets, of which 96,283 were unique entries in English. To minimize biases which could emerge from collecting data over a relatively short period of time, we then collected tweets from 1st January to 10th March 2020 using Twitter's 'Decahose', provided by a third party. We identified a further 63,037 unique tweets in English. The final database comprises 159,320 tweets.

We extracted the 1,000 most used hashtags from the 159,320 tweets and three annotators independently marked them up for: (1) whether they are East Asian relevant and, if so, (2) what Asian entity is discussed (e.g. China, Xi Jinping, South Korea), and (3) what the stance is towards the Asian entity (Very Negative, Negative, Neutral, Positive and Very Positive). Hashtags were also annotated for: (4) whether they are Coronavirus relevant and, if so, (5) what sentiment is expressed (Very Negative, Negative, Neutral, Positive and Very Positive). 97 hashtags were marked as either Negative or Very Negative towards East Asia by at least one annotator. All three annotations for hashtags are available in our data repository.

We then sampled 10,000 tweets at random from the database and a further 10,000 tweets which used one of the 97 anti-East Asia hashtags, thereby increasing the likelihood that prejudicial tweets would be identified and ensuring that our dataset is suitable for training a classifier [29, 30].

3.1 Data pre-processing

We conducted one pre-processing step before presenting the tweets to annotators: hashtag replacement. Initial qualitative inspection of the dataset showed that hashtags are often used in the middle of tweets, especially when they relate to East Asia and/or Coronavirus. Hashtags which appear in the middle of tweets often play a key role in their meaning and how prejudice is expressed. 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 the true meaning of this tweet, and almost impossible to be sure of whether there is any prejudice. However, in other cases, hashtags are less important to the meaning of the tweet but their inclusion could have the opposite effect – enabling annotators to pick up on 'prejudice' just because the hashtags express animosity against East Asia. This is problematic because it means that, in effect, we have returned to a dictionary based approach and any detection systems trained on such data would not generalize well to new contexts in which the same hashtags are not used. Note that this issue is usually less significant with Twitter data, where hashtags are usually only included at the end of tweets to identify broader conversations that a user wants to be part of.

To address this problem we implemented a hashtag replacement process. For the 1,000 most used hashtags, we had one annotator identify appropriate *thematic replacement* hashtags. We used five thematic replacements:

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

These thematic hashtag replacements were applied to all of the tweets. This means that annotators can still discern the meaning in most tweets as they are presented with the hashtags' topic but they are not unduly biased by the substantive *outlook*, stance and sentiment they 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.

For instance, the quote above would be 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.

4 Dataset Annotation

Prejudicial language takes many forms, from overt and explicit varieties of hate, such as the use of threatening language, to subtle and covert expressions, as with microaggressions and 'everyday' acts [30]. Using a binary schema (i.e. prejudiced or not) is theoretically problematic because distinct types of behaviour, with different causes and impacts, are collapsed within one category. It can also negatively impact classification performance because of substantial within-class variation [23]. We developed the taxonomy and codebook used here iteratively, moving recursively between existing theoretical work and the data to ensure the detection system can be applied by social scientists to generate meaningful insights.

The remaining subsections outline each of the annotation categories and agreement levels. For a more detailed description see our Annotation Codebook.

4.1 Annotation process

The dataset of 20,000 tweets was annotated with a three step process:

1. Each tweet was initially annotated independently by two trained annotators.
2. Cases where annotators disagreed about the primary category were identified. Then, an Expert adjudicator made a final decision after reviewing both annotators' decisions and the original tweet (Experts were able to decide a new primary category if needed). Two experts were used, both of whom are final year PhD students working on extreme political behavior online and offline with three months experience in annotating hate speech. Having developed the annotation codebook, the two experts also had a deep understanding of the guidelines.
3. The original annotations, where both annotators agreed, and the expert adjudications were combined to make a final dataset.

We deployed a large team of 26 annotators, all of whom had completed at least 4 weeks of training on a previous hate speech annotation project. The final dataset of 20,000 annotations were compiled over three weeks.

4.2 Annotations for theme

Annotations were first made for the presence of two themes: (1) COVID-19 and (2) East Asia. This is because, despite the targeted data collection process, many of the tweets do not directly relate to either COVID-19 or East Asia. If a tweet is not identified as being East Asian relevant then no further annotations are required (it is automatically assigned to the 'neutral' class).

Annotators used an additional flag for *how* they marked up the two themes, which we call 'hashtag dependence'. Annotators were asked whether they identified the themes based on the text by itself or whether they had to use thematic hashtag replacements to identify them. Our account of 'hashtag dependence' is very nuanced. Even in cases where hashtags are used, annotators should not rely *solely* on the hashtag to identify the theme. There must be a signal in the text that the annotator picks up on, which the hashtag is then used to confirm. Archetypally, hashtag dependence relates to cases where pronouns have been used ("They", "You", "These"), and the link to East Asian entities is only clear once they are taken into account. For instance:

> Love being out in the sun #VIRUS #EASTASIA

This tweet would not be considered either East Asian or COVID-19 relevant because there is not a signal in the text itself which indicates the presence of the themes.

Our approach to annotating themes and the role of hashtags required substantial training for annotators (involving one-to-one onboarding sessions). This detailed annotation process means that we can provide unparalleled 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. This initial round of annotating also helped to ensure that no tweets which were out-of-domain (i.e. not about East Asia) were accidentally annotated for categories other than Neutral.

4.3 Primary categories

Each tweet was assigned to one of five mutually exclusive categories – tweets which were not marked as East Asian relevant could only be assigned to the Neutral category.

- **Hostility against an East Asian entity:** Tweets which express abuse or intense negativity against an East Asian entity, primarily by derogating or attacking them (e.g. "Those oriental devils don't care about human life" or "Chinks will bring about the downfall of western civilization"). Common ways in which East Asian hostility is articulated include: negative representations of East Asians; conspiracy theories; claiming they are a threat; expressing negative emotions. 'Hostility' also includes animosity which is expressed more covertly.

- **Criticism of an East Asian entity** Tweets which make a negative judgment about an East Asian entity, without crossing the line into abuse. This includes questioning their response to the pandemic, how they are governed and suggesting they did not take adequate precautions and/or deploy suitable policy interventions. Note that the distinction between Hostility and Criticism partly depends upon what East Asian entity is being attacked: negativity against states tends to be criticism whilst negativity against East Asian people tends to cross into hostility.

  Drawing a clear line between Hostility and Criticism proved challenging for annotators. Often, Criticism would cross into Hostility when statements were framed normatively. For example, a criticism against the Chinese government (e.g. "the CCP hid information relevant to coronavirus") could become Hostility when turned into a derogation (e.g. "It's just like the CCP to hide information relevant to coronavirus"). However, the Hostility/Criticism distinction is crucial for addressing a core issue in online hate speech research, namely ensuring that freedom of speech is protected [35]. The Criticism category ensures that users can engage in what has been termed 'legitimate critique' [36], without their comments being erroneously labelled as hostile.

- **Counter speech** Tweets which explicitly challenge or condemn abuse against an East Asian entity. This 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, its 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 virus"). This includes content which discusses whether East Asian prejudice has increased during COVID-19, the supposed media focus on prejudice and/or free speech. Note that content which not only discussed but also expresses *support* for East Asian prejudice would cross into 'Hostility'.

- **Neutral** Tweets that do not fall into any of the other categories. Note that they could be offensive in other regards, such as expressing misogyny.

A small number of tweets contained more than one primary category – but each tweet can only be assigned to one category in this taxonomy as they are mutually exclusive. To address this, we established a hierarchy of primary categories: (1) entity-directed hostility; (2) entity-directed criticism; (3) counter speech; and (4) discussions of East Asian prejudice. For example, if a single tweet contained both counter speech and hostility (e.g. "It's not fair to blame Chinese scientists, blame their lying government") then it was annotated as hostility.

4.3.1 Annotation of Primary categories

All annotations were initially completed in pairs and then any disagreements were sent to an expert adjudicator. We

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

Table 1: Prevalence of primary categories in the dataset.

calculate the agreement for each pair and then take the average, minimum and maximum over all of them. Overall, agreement levels are moderate, with better results for the two most important and prevalent categories (Hostility and Neutral) but poorer performance 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.

In the 4,478 cases (22%) where annotators did not agree, one of two experts adjudicated. Experts generally moved tweets out of Neutral into other categories; of the 8,956 original annotations given to the 4,478 cases they adjudicated, 34% of them were in Neutral and yet only 29% of their adjudicated decisions were in this category. This was matched by an equivalent increase in the Hostilit 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 Neutral category). Equally, 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 None. This 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 Counter Speech, Discussion of East Asian prejudice and Neutral, annotators did not make secondary annotations.

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 targets (largely due to miss-spellings and punctuation variations). 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). Tweets can identify multiple East Asian targets and, where relevant, intersectional characteristics were recorded (e.g. “Chinese women” or “Asian-Americans”).

4.4.1 Additional flags for Hostility

For tweets which contain 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. Interpersonal abuse is a closely related but separate challenge to prejudice, and an important focus of computational abusive language research [23].
- Use of threatening language: Content which makes a threat against an East Asian entity. This includes expressing a desire, or willingness, to inflict harm or advocating, supporting and inciting others to do so. Threats have a hugely negative impact on victims and are a key concern of legal frameworks against hate speech [37, 38].
- Dehumanization Content which describes, compares or suggests equivalences between East Asians and non-humans or sub-humans, such as insects, weeds or actual viruses. Dehumanizing content must be literal (i.e. clearly identifiable rather than requiring a 'close reading') and must indicate malicious intent from the speaker. Dehumanizing language has been linked to real acts of genocide and is widely recognized as one of the most powerful indications of extreme prejudice [39, 40].

The frequency of the additional flags is shown in Table 3. It shows the relatively low prevalence, even within this biased dataset, of the most extreme and overt forms of hate, with both Threatening language and Dehumanization appearing infrequently. 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, none of the flags 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 so did not provide any further annotations).

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

Table 2: Agreement scores for Primary categories.

Number of Entries	Threatening language	Dehumanization	Interpersonal attack
Expert Decision	105 (2.7%)	105 (2.7%)	105 (2.7%)
One annotation (No)	1,347 (34.5%)	1,439 (36.9%)	1,366 (35.0%)
Two annotations (No, No)	1,989 (51.0%)	2,270 (58.2%)	2,150 (55.1%)
Two annotations (No, Yes)	251 (6.4%)	57 (1.5%)	131 (3.4%)
One annotation (Yes)	107 (2.7%)	15 (0.4%)	88 (2.3%)
Two annotations (Yes, Yes)	99 (2.5%)	12 (0.3%)	58 (1.5%)

Table 3: Secondary categories for Hostility.

4.4.2 East Asian slurs and pejoratives

Slurs are collective nouns, or terms closely derived from collective nouns, which are pejorative (e.g. "chinks" or "Chinazi"). Pejorative terms are derogatory references that do not need to be collective nouns (e.g. "Yellow Fever"). Annotators marked up slurs and pejoratives as free text entry.

4.5 Dataset availability

The dataset described here is available in full in our Data repository. We provide it in two formats:

- **Annotations Dataset**: All 40,000 annotations provided by the annotators. The only pre-processing is the provision of cleaned targets rather than the free text targets. We recommend using this dataset for investigation of annotator agreement and to explore secondary categories.
- **Final Dataset**: The 20,000 final entries, including the expert adjudications. We recommend using this dataset for training new classifiers for East Asian prejudice.

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 Neutral.

In order to provide a baselines for future work, we finetune several contextual embedding models as well as a one-hot LSTM model with a linear input layer and tanh activation. We choose the one-hot LSTM model as a contrast to the contextual embedding models. We use contextual embeddings as they take into account the context surrounding a token when generating the embedding, providing a separate embedding for each word usage. This grounding better encodes semantic information when compared with previous recurrence based deep learning approaches [41].

The models were trained and tested using a stratified 80/10/10 training, testing and validation split. We preprocess all documents by removing URLs and usernames, lower-case the document, and replace 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 [42] as the most effective for the GLUE benchmark tasks. This included testing across learning rates 1e-5, 2e-5, 3e-5 and batch sizes 32, 64 with an early stopping regime. Performance was optimized using the AdamW algorithm [43] and a scheduler that implements linear warmup and decay. For the LSTM baseline, we conduct a hyper-parameter search over the batch-size (16, 32, 64) and learning rate $(0.1 - 0.00001)$.

All of the contextual embedding models outperformed the baseline to varying degrees. RoBERTa achieved the highest F1 score (0.83) of the tested models, which is a 7-point improvement over the LSTM (0.76). This model harnesses the underlying bidirectional transformer architecture of [46], but alters the training hyperparameters and objectives to improve performance. Curiously, we see that the LSTM baseline outperforms all other models in terms of precision.

For the best performing model (RoBERTa), sources of misclassification are shown in the confusion matrix. This shows that the model performs well across all categories, with strongest performance in detecting tweets in the Neutral category (Recall of 91.6% and Precision of 93%). The model has few misclassifications between the most conceptually distinct categories (i.e. Hostility versus Neutral or

Model	F1	Recall	Precision
LSTM	0.76	0.67	**0.88**
AlBERT$_{xlarge}$ [44]	0.80	0.80	0.80
BART$_{large}$ [45]	0.81	0.81	0.83
BERT$_{large}$ [46]	0.82	0.82	0.83
DistilBERT$_{base}$ [47]	0.80	0.80	0.81
RoBERTa$_{large}$ [42]	**0.83**	**0.83**	0.85
XLNet$_{large}$ [48]	0.80	0.80	0.82

Table 4: Classification Performance of models.

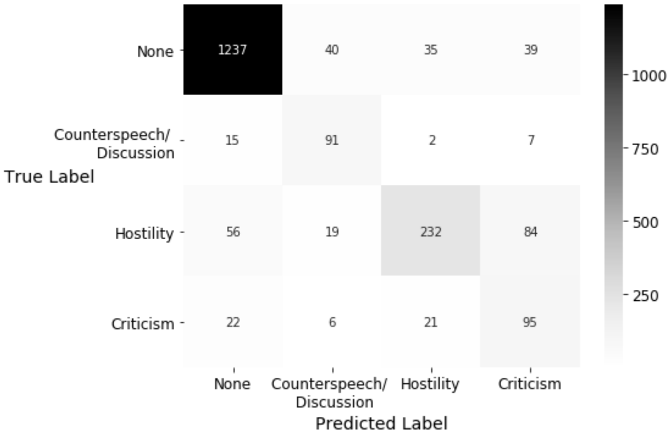

Figure 1: Confusion Matrix for RoBERTa Classifications.

Discussion of East Asian Prejudice) but has far more errors when distinguishing between the conceptually more similar categories, Criticism and Hostility.

6 Error analysis

To better understand not only the extent of misclassification but also the sources of error, we conducted a qualitative analysis of misclassified content, using a grounded theory approach [49]. This methodology for qualitative research 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 1 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%).

6.1 Annotator Errors, (17% of total)

Annotator errors are cases where the class predicted by the model may better represent the tweets' content and be more consistent with our taxonomy and guidelines. In effect, we believe that the 'wrong' classification provided by the model is right – 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'. The actual prevalence of misannotated data is likely far lower.

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 not East Asian relevant. Often, this was because annotators over-relied on hashtags for discerning East Asian relevance. These are *path dependency errors* and show the importance of annotators following the right instructions throughout the process. If a misannotation is made early on then the subsequent annotations are likely to be flawed.

6.2 Machine Learning Errors, (83% of total)

83% of the total errors were due to errors from the model. We have separated these errors into edge and non-edge cases. Non-edge cases are where the model has made an error that is easily identified by humans (comprising 37% of all errors); edge-cases are where the misclassified content contains some ambiguity and the model misclassification has some merit (comprising 46% of all errors).

6.2.1 Edge Cases

Hostility vs. Criticism Misclassifying Hostility as Criticism (and vice versa) was the largest source of error, reflecting also the high levels of annotator disagreement between these categories. The model struggled with cases where criticism was framed in a normative way (e.g. “gee, china was lying to us. what a bloody shock”). In such cases, the model misclassified the tweets as Criticism rather than Hostility.

Discussion of East Asian prejudice vs. Neutral The model misclassified Neutral 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 shows why you should blame all problems with China on Trump”).

Co-present Primary Categories In our taxonomy, annotators could only assign each tweet to a single 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. For instance, the model identified the following tweet as ’Discussion of East Asian Prejudice’: “don’t sell your mother for chinese yuan. stop being HASHTAGEASTASIA propaganda tool. calling HASHTAGEASTASIA+VIRUS where it came from, isn’t racist but truth”. It missed the criticism of China, which was also co-present and for which it had been annotated.

Ambiguous Content In some cases, the content of tweets was ambiguous, such as positively framed criticism (e.g. “china official finally admits the HASHTAGEASTASIA+VIRUS outbreaks”) or use of grammatically incorrect language.

6.2.2 Non-edge Cases

Identification Errors In several cases the misclassified tweets had clear textual signals which indicate why the tweets had been misclassified, such as the presence of signal words and phrases (e.g. ’Made in China’). This is a learned over-sensitivity to signals which are frequently associated with each primary category.

Target Errors In over a third of all non-edge case errors the model identified an appropriate category – but the target was not East Asian and so the classification was wrong. For instance, tweets which expressed hostility against an invalid target (e.g. India, WHO or the American government) were routinely classified as Hostility. In such cases, an appropriate entity (i.e. China) was usually referred to but was not the object of the tweet, creating a mixed signal. Conversely, in other cases, the model failed to identify Criticism or Hostility against an East Asian entity because it required some context-specific cultural knowledge (e.g. “building firewall and great wall is the only thing government good at! HASHTAGVIRUS HASHTAGEASTASIA+VIRUS”). East Asian prejudice that targeted a well-known East Asian person, such as Xi Jinping, was also often missed.

Errors due to Tone 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”). Although the meaning is still discernible to a human reader, the model missed the important role played by tone.

6.3 Addressing Classification Errors

Classifying social media data is notoriously difficult with many different sources of error which, in turn, require many potential remedies. The prevalence of annotator errors, for example, illustrates the need for robust annotation processes and providing more support and training when applying taxonomies. Removing such errors entirely is arguably impossible, but better annotation processes and guidelines would help to limit the number of avoidable misclassifications.

Edge cases are a particularly difficult type of content for classification, and there is not an easy solution to how they should be handled. Edge cases can be expected in any taxonomy that draws distinct lines between complex and non-mutually exclusive concepts, such as hateful speech and legitimate 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 also be addressed by larger datasets, as well as more advanced machine learning architectures.

7 Hashtag analysis

Because annotators were presented with tweets where all hashtags were replaced with either a thematic hashtag replacement or just ’hashtag’ as a placeholder, we can conduct unbiased analyses on the co-occurrence of hashtags with the annotated 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 accounted for by the primary category. For brevity, only the twenty hashtags most closely associated with the Hostility category are shown here.

A small number of hashtags are highly likely to only appear in tweets that express Hostility against East Asian entities. These hashtags can be used to identify prejudiced discourses online and, in some cases, their uses may indicate intrinsic prejudice. Surprisingly, many seemingly hostile hashtags against East Asia, such as #fuckchina and #blamechina are not always associated with hostile tweets (in these cases, 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.

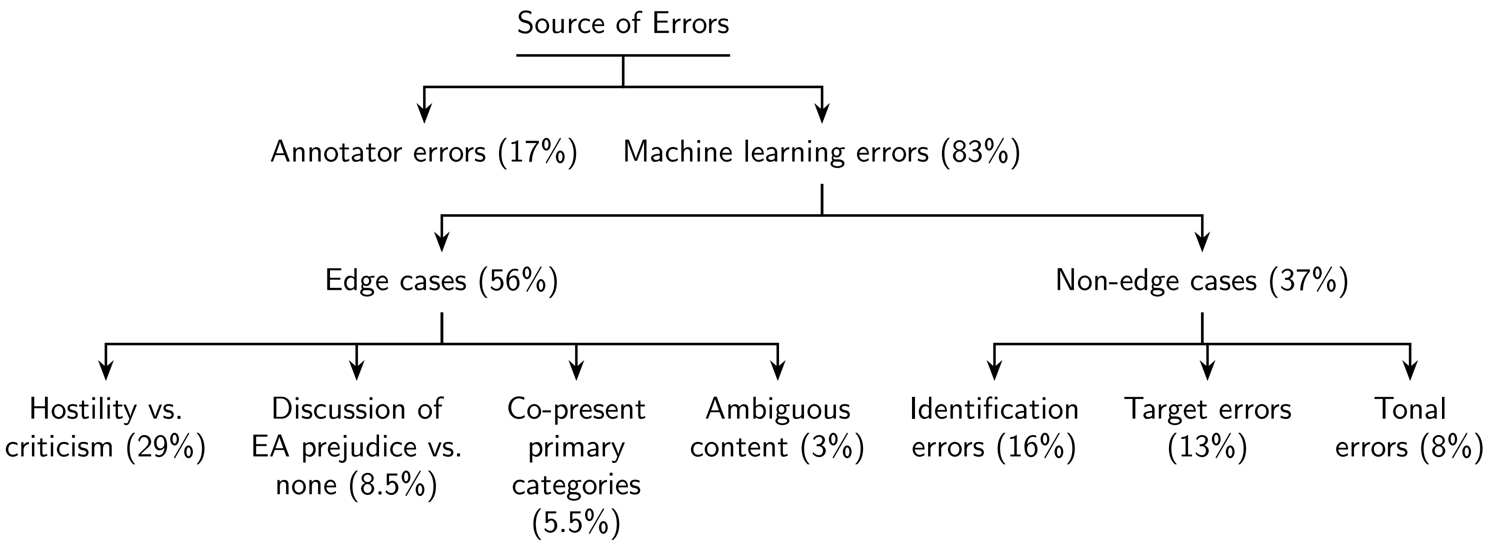
Source of Errors
Annotator errors (17%)
Machine learning errors (83%)
Edge cases (56%)
Non-edge cases (37%)
Hostility vs. criticism (29%)
Discussion of EA prejudice vs. none (8.5%)
Co-present primary categories (5.5%)
Ambiguous content (3%)
Identification errors (16%)
Target errors (13%)
Tonal errors (8%)

Figure 2: Sources of Classification error.

Hashtag	Frequency in Hostile tweets	Percentage of all uses	Number of total uses
#rule2	20	87.0%	23
#rule3	17	85.0%	20
#rule1	22	84.6%	26
#makechinapay	18	72.0%	25
#hkgovt	22	71.0%	31
#fuckchina	54	67.5%	80
#blamechina	23	60.5%	38
#batsoup	15	60.0%	25
#hkairport	11	55.0%	20
#huawei	16	53.3%	30
#boycottchina	185	52.9%	350
#communismkills	14	51.9%	27
#communistchina	34	50.7%	67
#chinaisasshoe	41	50.6%	81
#chinapropaganda	10	50.0%	20
#china_is_terrorist	168	48.7%	345
#xijingping	17	48.6%	35
#chinashouldapologize	14	48.3%	29
#madeinchina	39	47.6%	82
#ccp	395	46.5%	850

Table 5: Hashtags in Hostile tweets.

8 Conclusion

East Asian prejudice is a deeply concerning problem linked with COVID-19, reflecting the huge social costs that the pandemic has inflicted on society, as well as the health-related costs. In this paper we have reported on development of several research artefacts which we hope will enable future research into this source of online harm. We make these available in our repository:

1. A classifier for detecting East Asian prejudice.
2. A 20,000 tweet training dataset used to create the East Asian prejudice classifier.
3. A 40,000 annotations training dataset, which contains the full annotations made by each annotator.
4. A list of hashtag 'replacements', where COVID-19 specific hashtags are swapped out with thematic replacements.
5. Three sets of annotations for the 1,000 most used hashtags in the original database of COVID-19 related tweets. Hashtags were annotated for COVID-19 relevance, East Asian relevance and stance.
6. The full codebook used to annotate the tweets. This contains extensive guidelines, information and examples for hate speech annotation.

References

[1] Ryan Flanagan. Canada's top doctor calls out 'racism and stigmatizing comments' over coronavirus, January 2020. URL `https://www.ctvnews.ca/canada/canada-s-top-doctor-calls-out-racism-and-stigmatizing-comments-over-coronavirus-1.4790762`.

[2] Tessa Wong. Sinophobia: How a virus reveals the many ways China is feared - BBC News. URL `https://www.bbc.co.uk/news/world-asia-51456056`.

[3] Yuebai Liu. Coronavirus prompts 'hysterical, shameful' Sinophobia in Italy, February 2020. URL `https://www.aljazeera.com/news/2020/02/coronavirus-prompts-hysterical-shameful-sinophobia-italy-200218071444233.html`.

[4] Kate Walton. Wuhan Virus Boosts Indonesian Anti-Chinese Conspiracies. URL `https://foreignpolicy.com/2020/01/31/wuhan-coronavirus-boosts-indonesian-anti-chinese-conspiracies/`.

[5] Salem Solomon. Coronavirus Brings 'Sinophobia' to Africa | Voice of America - English. URL `https://www.voanews.com/science-health/coronavirus-outbreak/coronavirus-brings-sinophobia-africa`.

[6] Jack Guy. East Asian student assaulted in 'racist' coronavirus attack in London, March 2020. URL `https://www.cnn.com/2020/03/03/uk/coronavirus-assault-student-london-scli-intl-gbr/index.html`.

[7] Michael Shields. U.N. asks world to fight virus-spawned discrimination. *Reuters*, February 2020. URL `https://www.reuters.com/article/us-china-health-rights-idUSKCN20L16B`.

[8] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N Kamel Boulos, and Jon Crowcroft. Leveraging Data Science To Combat COVID-19 : A Comprehensive Review. *Pre-Print*, pages 1–19, 2020.

[9] Josh Cowls, Bertie Vidgen, and Helen Margetts. Why content moderators should be key workers Protecting social media as critical infrastructure during COVID-19, apr 2020.

[10] Franck Billé. *Sinophobia: anxiety, violence, and the making of Mongolian identity.* 2015. ISBN 978-988-8208-28-9. OCLC: 904372882.

[11] Tam Goossen, Jian Guan, and Ito Peng. Yellow Peril Revisited: Impact of SARS on the Chinese and Southeast Asian Canadian Communities June, 2004 Project Coordinator and Author: Carrianne Leung. 2004.

[12] Zhang. Pinning coronavirus on how chinese people eat plays into racist assumptions, January 2020. URL `https://www.eater.com/2020/1/31/21117076/coronavirus-incites-racism-against-chinese-people-and-their-diets-wuhan-market`.

[13] Kevin Lee. SARS and Its Resonating Impact on the Asian Communities. *Lehigh Review*, 21(1):49–55, 2013. URL `http://preserve.lehigh.edu/cas-lehighreview-vol-21{%}0Ahttp://preserve.lehigh.edu/cas-lehighreview-vol-21/24`.

[14] Laura Silver, Kat Devlin, and Christine Huang. People around the globe are divided in their opinions of China, December 2019. URL `https://www.pewresearch.org/fact-tank/2019/12/05/people-around-the-globe-are-divided-in-their-opinions-of-china/`.

[15] Joe Neel. Poll: Asian-Americans See Individuals' Prejudice As Big Discrimination Problem, December 2017. URL `https://www.npr.org/2017/12/06/568593799/poll-asian-americans-see-individuals-prejudice-as-big-discrimination-problem`.

[16] Sue Adamson, Bankole Cole, Gary Craig, Basharat Hussain, Luana Smith, Ian Law, Carmen Lau, Chak-Kwan Chan, and Tom Cheung. Hidden from public view? Racism against the UK Chinese population. Technical report, 2009.

[17] Home Office. Hate crime, England and Wales, 2018 to 2019. *Home Office Statistical Bulletin*, (24):21, 2019. URL http://www.report-it.org.uk/files/hate{_}crime{_}operational{_}guidance.pdf.

[18] Tim Kelly. Japan lists China as bigger threat than nuclear-armed North Korea. *Reuters*, September 2019. URL https://www.reuters.com/article/us-japan-defence-idUSKBN1WC051.

[19] 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:2004.04046 [cs]*, April 2020. URL http://arxiv.org/abs/2004.04046. arXiv: 2004.04046.

[20] The New Statesman. Covid-19 has caused a major spike in anti-Chinese and anti-Semitic hate speech, apr 2020.

[21] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 Social Media Infodemic. *arXiv:2003.05004 [nlin, physics:physics]*, March 2020. URL http://arxiv.org/abs/2003.05004. arXiv: 2003.05004.

[22] Eir Nolsoe. COVID-19: Bogus claims fool Britons | YouGov. URL https://yougov.co.uk/topics/health/articles-reports/2020/03/30/covid-19-bogus-claims-fool-britons.

[23] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *1st Workshop on Abusive Language Online*, pages 78–84, 2017. ISBN 0141-0296. doi: 10.1080/17421770903114687. URL http://arxiv.org/abs/1705.09899.

[24] Bertie Vidgen and Leon Derczynski. Directions in Abusive Language Training Data: Garbage In, Garbage Out. *Arxiv:2004.01670v2*, 1(1):1–26, 2020. URL http://arxiv.org/abs/2004.01670.

[25] Zeerak Waseem and Dirk Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL-HLT*, pages 88–93, 2016. doi: 10.18653/v1/n16-2013.

[26] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 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, 2019. doi: 10.18653/v1/p19-1271.

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

[28] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the Second Workshop on Abusive Language Online*, pages 11–20, 2018. URL http://arxiv.org/abs/1809.04444.

[29] Anna Schmidt and Michael Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. In *International Workshop on NLP for Social Media*, pages 1–10, Valencia, Spain, 2017. ISBN 9781945626425. doi: 10.18653/v1/w17-1101.

[30] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3509. URL https://www.aclweb.org/anthology/W19-3509.

[31] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, Noah A Smith, and Paul G Allen. The Risk of Racial Bias in Hate Speech Detection. In *ACL Proceedings*, pages 1668–1678, 2019. URL www.figure-eight.com.

[32] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *ACL Proceedings*, pages 1–11, 2019. URL http://arxiv.org/abs/1905.12516.

[33] Sahaj Garg, Ankur Taly, Vincent Perot, Ed H. Chi, Nicole Limtiaco, and Alex Beutel. Counterfactual fairness in text classification through robustness. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019. doi: 10.1145/3306618.3317950.

[34] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All You Need is "Love": Evading Hate-speech Detection. *arXiv:1808.09115v2*, pages 1–11, 2018.

[35] Stefanie Ullmann and Marcus Tomalin. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80, 2020. ISSN 15728439. doi: 10.1007/s10676-019-09516-z. URL https://doi.org/10.1007/s10676-019-09516-z.

[36] Roland Imhoff and Julia Recker. Differentiating Islamophobia: Introducing a New Scale to Measure Islamoprejudice and Secular Islam Critique. *Political Psychology*, 33(6):811–824, 2012. doi: 10.1111/j.1467-9221.2012.00911.x.

[37] The Law Commission. *Abusive and Offensive Online Communications: A scoping report*. The Law Commission, London, 2018. ISBN 9781528608480.

[38] Anne Weber. *Manual on Hate Speech*. Council of Europe, Strasbourg, 2009. ISBN 9789287166135.

[39] Jonathan Leader Maynard and Susan Benesch. Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention. *Genocide Studies and Prevention*, 9(3):70–95, 2016. ISSN 1911-0359. doi: 10.5038/1911-9933.9.3.1317.

[40] Andreas Musolff. Dehumanizing metaphors in UK immigrant debates in press and online media. *Journal of Language Aggression and Conflict*, 3(1):41–56, 2015. ISSN 2213-1272. doi: 10.1075/jlac.3.1.02mus.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

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

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[44] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[45] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[47] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.

[49] Juliet Corbin and Anselm Strauss. Grounded Theory Research: Procedures, Canons and Evaluative Criteria. *Qualitative Research*, 13(1):3–21, 1990. ISSN 0974360X.