2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-7281-8470-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLA51294.2020.00111

On Analyzing COVID-19-related Hate Speech Using BERT Attention

Nishant Vishwamitra
Clemson University
Clemson, USA
nvishwa@g.clemson.edu

Ruijia (Roger) Hu[§]
Clemson University
Clemson, USA
roger.rj.hu@gmail.com

Feng Luo
Clemson University
Clemson, USA
luofeng@clemson.edu

Long Cheng
Clemson University
Clemson, USA
lcheng2@clemson.edu

Matthew Costello
Clemson University
Clemson, USA
mjcoste@clemson.edu

Yin Yang
Clemson University
Clemson, USA
yin5@clemson.edu

***Abstract*—The emergence of COVID-19 has engendered a new wave of online hate speech in social media platforms such as Twitter. Its widespread effects range from acts of cyber-harassment towards certain ethnic communities (e.g., the Asian community), to targeting older people belonging to age groups correlated with higher mortality rates (termed infamously as "Boomer Remover"). Thus, an urgent need arises for a timely mitigation of this new wave of online hate speech. In this work, we aim to discover the hate-related keywords linked to COVID-19 in hateful tweets posted on Twitter so that users posting such keywords can be asked to reconsider posting them. We first collect a new dataset of tweets targeting older people supplementing with a dataset targeting the Asian community. Then, we develop an approach to analyze the datasets with BERT (a transformer-based model) attention mechanism and discover 186 novel keywords targeting the Asian community and 100 keywords targeting older people. Based on our study, we then propose a control mechanism wherein a user can be asked to reconsider using certain sensitive words identified by our approach. We further perform an exploratory analysis of BERT attention mechanism and find that the most high-impact, long distance attentions are learned in the earlier or later layers of the model depending on the underlying data distribution. Our study indicates that the BERT model in some cases uses a hate keyword and an associated group or individual to make predictions, a finding that is inline with existing hate-speech research, which suggests that hate-speech is often aimed at certain groups or individuals.**

***Index Terms*—hate-speech, online-hate, explanation, COVID-19, Twitter, BERT**

I. Introduction

The social and economic destabilization caused by COVID-19 has produced a range of emotions in people, including fear, anxiety, and even hostility. Notably, COVID-19-related hate speech is increasingly occurring on social media that target people based on race/ethnicity, age, social class, immigration status and political ideology. For instance, Asian Americans are frequent targets of hate speech related to COVID-19, with derogatory terms for the disease, such as "kung flu" and "chop fluey", shared more than 10,000 times on Twitter during March alone [1]. Meanwhile, the phrase "Boomer Remover", a callous nickname for COVID-19 used to mock the high mortality rate among older people infected with the disease, has been shared more than 65,000 times on Twitter [4]. Moreover, a recent report on online toxicity found a 900% increase in hate speech towards China and Chinese people on Twitter [2], and traffic to sites and posts that target Asians over COVID-19 has skyrocketed.

This recent wave of COVID-19-related hate speech has given rise to novel vocabularies and jargon that are used by Internet users to specifically target certain communities. While current social media platforms such as Twitter and Facebook are quite well equipped to detect hate-speech concerning traditional issues [3], they are not capable of addressing the new jargon related to COVID-19. Thus, there is a need to discover these novel jargon with respect to COVID-19-related hate speech. However, Internet users often find innovative ways to use such jargon [11], [18], in order to hide their true meaning (e.g., "xinpigs", "thankschina"), due to which they cannot be discovered in a straightforward manner. Thus, new strategies based on deep analysis of such texts need to be formulated to summarize such jargon by discovering the keywords that are related to them.

The detection of online hate speech should be accompanied with a strong control strategy so that Internet users can be deterred from posting such texts. User warnings and word removal recommendations [12], [20] are often used to implement such control mechanisms. However, merely asking users to remove hate-related keywords is not a strong enough control strategy, as users often come up with alternate ways to post such texts by surpassing the detection mechanisms. Moreover, the other words in a text that are semantically related to such keywords (such as names of individuals or group) can still significantly harm the targeted individuals or groups. Therefore, a control strategy that can systematically point out these semantically related words is very important for effectively controlling these instances of hate speech.

The new wave of hate speech related to COVID-19 is

[§]Intern from D.W. Daniel High School, Central, SC, USA.

DOI 10.1109/ICMLA51294.2020.00111

unique because, unlike traditional forms of hate speech that are typically rooted in deep-seated animosity, hate speech linked to the COVID-19 outbreak is spontaneous, induced by fear, anxiety, and stress resultant of a rapidly-changing reality. Previously, to understand why identity-based hate speech is becoming increasingly common online [30], sociologists and criminologists have explored the roles of strain and threat in fostering such attacks. While some works [23] theorize that deviant behavior stems from a disjuncture between culturally-valued goals, others show that financial strain, such as strain caused by unemployment/underemployment and low wages, can indeed engender harassing behavior towards immigration groups [16], [17], [29]. While fear prompted by the pandemic might trigger long-held prejudice towards certain groups, such as Asian Americans or immigrants, it is unlikely that hate-speech based on age or socio-economic status is similarly an expression of embedded bias. Thus, more information on COVID-19-related hate speech is needed to better understand its impetuses.

In this work, we propose a novel approach to discover new keywords linked to COVID-19-related hate speech and the word associations to effectively implement its control. We collect a new dataset (Boomer-hate dataset) of tweets targeting old people and supplement this dataset with an existing COVID-19 dataset (Asian-hate dataset) targeting Asian American community [34]. We then train a BERT (Bidirectional Encoder Representations from Transformers) model [10] to classify tweets as Hate Vs. Non-hate. Based on the analysis of BERT attention mechanism, a transformer model [32] based on attention, we develop an approach to discover new keywords (186 keywords targeting the Asian community and 100 keywords targeting older people) related to COVID-19. For implementing effective control, we develop a strategy based on the attention attributed to these keywords by other words in a tweet, so that all sensitive words in a tweet can be censored or reconsidered. We then undertake an exploratory analysis of COVID-19-related hate speech and find that most of such high-impact, long distance attentions are learned in the earlier layers of the BERT model (layers 2 to 7 for Asian-hate dataset) or later layers (layers 10 and 11 for Boomer-hate dataset) depending on the underlying data distribution. Our study also makes an important finding that in the case of Boomer-hate dataset, the BERT model makes predictions based on the association of hate keywords and targeted groups or individuals, a finding that is inline with existing hate-speech research. Our finding paves the way for deep analysis of BERT for detection of hate-speech as well as explaining BERT (known as BERTology), a largely unexplored research area concerning BERT.

Our contributions are summarized as follows:

- **New Dataset of COVID-19-related Hate Speech Against Old People**. We collect a new dataset of COVID-19-related hate speech against old people. Our Boomer-hate dataset consists of 388 hate tweets and 1358 non-hate tweets from 1401 Twitter users. We will make our dataset publicly available for further research. In our work, we supplement our own dataset with another publicly available dataset [34] pertaining to COVID-19-related Asian hate, so that our study covers a broad spectrum of hate speech witnessed during COVID-19.
- **COVID-19-related Hate Speech Keywords Discovery**. We first train a BERT model on the datasets to learn Hate Vs. Non-hate speech. We then develop an approach based on BERT attention mechanism, to discover the most attended-to keywords that are responsible for causing hate in hateful tweets. We discover 186 keywords related to Asian-hate and 100 keywords related to Boomer-hate using our approach. For effective control of hate speech, we use our approach to find the words that significantly attend to the hate keywords so that they can be presented to users for removal or reconsideration. The new keywords discovered by our approach are an important resource for further hate-speech research, and we plan to submit them to a popular online hate keywords repository [1].
- **Exploratory Findings About COVID-19-related Hate Speech**. Our exploratory findings specifically concerning BERT and hate-speech detection sheds light on the inner-workings of the BERT model, using which we can identify if the model uses specific word associations only to detect hate speech, or uses a more complex association of words. We find that the high impact attentions regarding hate speech are learned in the earlier layers of the BERT model in case of Asian-hate and later layers in case of Boomer-hate, and that BERT seems to be associating hate-related keywords and groups or individuals for hate-speech predictions for Boomer-hate.

II. Data Collection Methodology

In our study, we collect a timely dataset of tweets from Twitter related to COVID-19-related hate speech against old people. We then supplement this dataset with an existing dataset [30] of COVID-19-related hate speech against Asian American community. We use this combined dataset to study online hate speech associated with COVID-19 on Twitter.

Collection Methodology. We adopted a keyword-based approach to collect COVID-19 tweets against old people using an online Twitter data collection tool [2]. We used the keywords "boomer" with COVID-19 related keywords such as "Coronavirus" and "Covid-19" to search for such tweets. We restricted the tweet collection to English language only. Using these keywords, we collected 28,827 tweets between December 2019 and June 2020 from 1401 Twitter users. Figure 1 shows the percentage of tweets related to COVID-19 hate speech against older people and the date ranges they were searched in. Since the date ranges prior to Feb 24, 2020 yielded very low tweets, we have ignored those date ranges. It can be seen in Figure 1 that the majority of the tweets linked to COVID-19-related hate speech against old people were found

[1] https://hatebase.org/

[2] https://github.com/Jefferson-Henrique/GetOldTweets-python

in March, 2020. We note that this may be the time, during which the adverse effects of the pandemic on older individuals were brought to light that could have triggered the spike in the hate-related tweets during this time.

Boomer-Hate Dataset. Since there are no ground truth labels of COVID-19-related anti old people hate tweets, we asked two experts in our research team to label the collected tweets. We first cleaned the tweets based on sentiment polarity and removed the tweets that are neutral sentiment using Python NLTK library [3]. Existing studies of hate speech from the social science literature [14], [25] have shown that hate speech is directed at an individual or group based on "an arbitrary or normatively irrelevant feature", and that it casts the target as an "undesirable presence and a legitimate object of hostility." We used a similar definition for our annotation task: (a) has one or more COVID-19-related keywords, (b) is directed towards an individual or a group of older people (Boomers), and (c) is abusive or derogatory.

The two experts labeled all the tweets in the dataset, which results in 388 hate-speech related tweets and 1358 non-hate-related or neutral tweets.

Asian-Hate Dataset. We used a publicly available dataset [30] of tweets aimed at COVID-19-related hate speech against the Asian American community. This dataset contains 2,319 labeled tweets, with 678 of them labeled as hateful tweets.

III. Background

In this paper, we focus on the BERT model [10], a large transformer [32] network. Transformers consist of multiple layers where each layer contains multiple attention heads. Each attention head takes as input a sequence of vectors $h = [h_1, ..., h_n]$ corresponding to the n tokens of the input sentence. Each vector h_i is transformed into query, key, and value vectors q_i, k_i, v_i through separate linear transformations. The head computes attention weights α between all pairs of words as softmax-normalized dot products between the query and key vectors. The output o of the attention head is a weighted sum of the value vectors, and α_{ij} represents a dot product between the query and key vectors, expressed in Equation 1 below.

$$\alpha_{ij} = \frac{exp(q_i^T k_j)}{\sum_{l=1}^{n} exp(q_i^T k_l)} \qquad o_i = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{1}$$

The attention weights can be interpreted as controlling the importance of every other token when learning the next representation of the current token.

BERT is trained using the "masked language modeling" strategy over billions of data samples, and more details about the training process can be found in [10]. An important detail about BERT training is that a special token [CLS] is added to the beginning of the text and another token [SEP] is added to the end, so that multiple sequence inputs can be trained together.

[3]https://www.nltk.org/

IV. Study Methodology

On a high level, our study is focused on studying the attention mechanism of BERT models to find important patterns about COVID-19-related hateful tweets. Since BERT is based on attention mechanism, the model learns the attentions between different tokens in all the tokens of an input sequences. This provides us a powerful tool to analyze linguistic associations in the dataset that BERT is trained on. Our work leans on the exploratory research side of BERT (known as "BERTology" [8], [24]). We first train a 12 layer, 12 attention heads "bert-base-uncased" model [32] on our dataset (we use 90% for training and 10% for testing). In the following sections, we analyze the BERT model trained on the hate datasets, spanning several layers and attention heads to formulate hate-speech control strategies and draw important observations about how BERT detects hate speech.

A. Keywords Discovery from BERT Attention Mechanism

The first objective of our work is to find new keywords of hate-speech from the two datasets (Asian-Hate and Boomer-hate datasets). In this section, we discuss our approach for discovering these keywords and our findings regarding the keywords found in the two datasets. In this experiment, we evaluate the words that are most attended to, by the fine-tuned BERT model in each layer. To achieve this, we aggregate the attention on each token of an input sequence by all attention heads in each layer, as given below in Equation 2.

$$Aggr^l(o_i) = \sum_{h \in H} o_i^h \tag{2}$$

In the Equation 2, H refers to the attention heads in each layer of BERT model and o_i refers to the attention weight of a token in an input sequence. For each layer, we take the top-k (k = 5) tokens as potential keywords. We do not consider tokens that are not split by the BERT word-piece tokenizer to reduce words normally occurring in English dictionary. We further remove those words that are not part of a sentence [4]. A summarized list of discovered keywords are depicted in Table I.

In our analysis of Table I, we found several new keywords used to propagate hate speech with respect to COVID-19-related Asian-hate and Boomer-hate. In the Asian-hate dataset, we found that BERT attributes the most attention to keywords that are a combination of word-pieces related to Asian community (e.g., "chin") and word-pieces related to the COVID-19 pandemic (e.g., "virus"), giving rise to keywords such as "chinkvirus" and "wuhanflu". In the Boomer-hate dataset, we found that certain keywords followed a similar pattern of word-pieces related to older people (e.g., "boomer") and word-pieces related to derogatory terms (e.g., "remover"), giving rise to keywords such as "boomerremover", but certain keywords did not necessarily follow any particular pattern, but seemed to be more contextual in nature (e.g., "karen", "oldaf" and "deletus"). We also found some keywords that

[4]We use Python NLTK library's POS tags

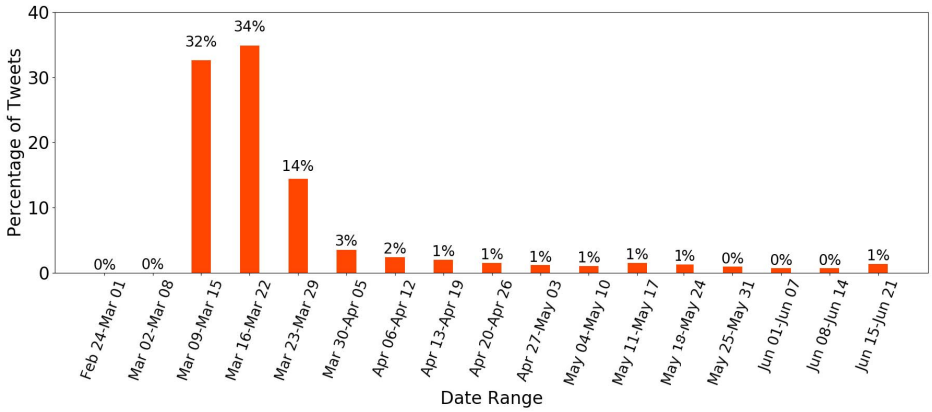


Fig. 1: Percentages of tweets collected according to date ranges. All date ranges belong to the year 2020.

TABLE I: Summarized list of sample keywords in the datasets, most attended to by BERT model.

Dataset	Top Keywords
Asian-hate Dataset	chinkvirus, wuhanflu, chinesebioterrorism, chineseviruscorona, chinaliedhdexperiencedied, wholiedpeopledied, chinamustexplain, nochinainfuenceonamerica, wuhanhealthorganisation, abioweaponslab, fuckchina, chinesebiologicalchemical, ccpvirus, prisonplanet, makechinapay, neverforgetneverforgive
Boomer-hate Dataset	boomerremover, gaslighters, corbid, 60sfolks, boomerdeath, karen, hitler , headassery, thankyouboomer, yoof, deletus, boomermoober, michiganders, entomber, boomerentomber, komekko, doubledowndonnie, boomerdoomer, coronachan, socialistremover, oldaf, immunocompromised, thintheherd

were completely new, that were simply derogatory to older individuals (e.g., "yoof" refers to the way an older person may pronounce "youth"). These findings may indicate that while users follow a particular pattern in the Asian-hate tweets, on the other hand users seem to adopt more complex and varied techniques in the Boomer-hate tweets.

Next, in order to study how these keywords are learned in each BERT layer, we analyze the attention given to these keywords by each layer of the BERT model. We recall that the BERT model used in this work has 12 layers of multi-headed attentions. In this study, we analyze the keywords that are most attended to in each BERT layer. The Table II shows the top-k (k=10) most attended keywords in each BERT layer, normalized across all attention heads. We did not find any apparent pattern which indicated that particular keywords could be receiving more attention in certain layers. Existing research in BERTology such as [8] suggest that certain layers of BERT may be focusing on different word associations. Therefore, we further analyzed the layers from this perspective. We focused on long-distance attentions in each layer based on the attention on multiple tokens, as given

TABLE II: Top-k (k = 10) keywords attended to in each layer of BERT model.

Layer #	Top-k Keywords
Layer 1	coronavirus, chinesevirus, wuhanvirus, chinavirus, ccpvirus, wuhancoronavirus, chinesevirus19, chinesecoronavirus, coronavirusoutbreak, chinaliedpeopledied
Layer 2	coronavirus, covid19, chinavirus, chinesevirus, wuhanvirus, chinaliedpeopledied, realdonaldtrump, covid2019, xijinpingvirus, chinesevirus19
Layer 3	chinaliedpeopledied, chinaliedpeopledie, fuckchina, covid19, coronavirus, wuhanvirus, chinesevirus, chinese, racismisavirus, chinavirus
Layer 4	chinaliedpeopledied, coronavirus, covid19, fuckchina, chinesevirus, chinaliedpeopledie, wuhanvirus, chinavirus, ccpvirus, chinesevirus19
Layer 5	covid19, chinaliedpeopledied, chinesevirus, coronavirus, chinavirus, wuhanvirus, chinesevirus19, ccpvirus, fuckchina, covid2019
Layer 6	chinaliedpeopledied, chinesevirus, coronavirus, chinavirus, covid19, wuhanvirus, chinaliedpeopledie, ccpvirus, fuckchina, chinesevirus19
Layer 7	chinesevirus, coronavirus, chinaliedpeopledied, wuhanvirus, chinavirus, covid19, fuckchina, ccpvirus, wuhancoronavirus, chinaliedpeopledie
Layer 8	coronavirus, chinesevirus, chinaliedpeopledied, wuhanvirus, fuckchina, chinavirus, covid19, ccpvirus, wuhancoronavirus, chinaliedpeopledie
Layer 9	chinaliedpeopledied, coronavirus, chinesevirus, fuckchina, wuhanvirus, chinavirus, covid19, ccpvirus, chinaliedpeopledie, racismisavirus
Layer 10	chinaliedpeopledied, coronavirus, fuckchina, covid19, chinesevirus, chinavirus, chinese, chinaliedpeopledie, racismisavirus, chinesevirus19
Layer 11	coronavirus, covid19, chinaliedpeopledied, fuckchina, chinesevirus, chinavirus, wuhanvirus, chinese, ccpvirus, chinaliedpeopledie
Layer 12	chinesevirus, coronavirus, chinaliedpeopledied, covid19, wuhanvirus, chinavirus, ccpvirus, chinaliedpeopledie, racismisavirus, chinesevirus19

by Equation 3.

$$D = \frac{\sum_{i=1}^{N} \sum_{j=1}^{i} \alpha_{ij}(x) \times (i-j)}{\sum_{i=1}^{N} \sum_{j=1}^{i} \alpha_{ij}(x)} \tag{3}$$

The Equation 3 determines attention spanning across tokens,

#	Original Tweet	Keywords
1	some chinese are horrible as fuck chinaliedpeopledie boycottchina wuhanvirus	chinese, chinaliedpeopledie, boycottchina, wuhanvirus
2	itsing6 spokespersonchn fuck ccpvirus chinesevirus	fuck, ccpvirus, chinesevirus
3	h*********f j****l s**********d fuck off commie chinaliedpeopledied fucktheccp	fuck off, commie, chinaliedpeopledied, fucktheccp
4	5g does fuck u ask the kungflu	fuck, kungflu
5	it'll be the only party left come november boomerremover	boomerremover
6	magkcovid unta it incompetent NA senators they called the virus a boomer remover for a reason	magkcovid, boomer, remover

TABLE III: Samples of control strategy.

normalized by their distances (i and j are indices). Therefore, higher attention tokens farther apart would have higher distance attention. We computed this metric for each attention head in a layer and the result is depicted in Figure 2, which depicts a heat-map of the attention distance for each head in each layer for the two datasets.

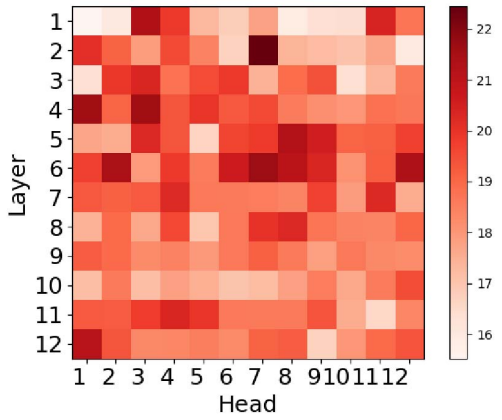
From Figure 2a which shows the results for Asian-hate dataset, we can observe that the attention distance in earlier layers (layers 2 to 7) are higher (depicted by darker color). This could indicate that the hate-related attentions for Asian-hate spanning across tokens are predominantly learned in the earlier layers of the BERT model.

On analyzing the Figure 2b which depicts the results of this experiment for Boomer-hate dataset, we observed a different result, which may indicate that in this case, the long distance attentions are learned in later layers of the BERT model, with layers 11 and 12 showing overall higher mean attention distances. This observation could be due to the fact that the hateful tweets in the Boomer-hate dataset seems to be significantly correlated to a few, specific keywords (e.g. "boomer" and "remover"). Another explanation of this observation could be that the BERT model may be dynamically learning these associations according to the underlying distribution of the training data.

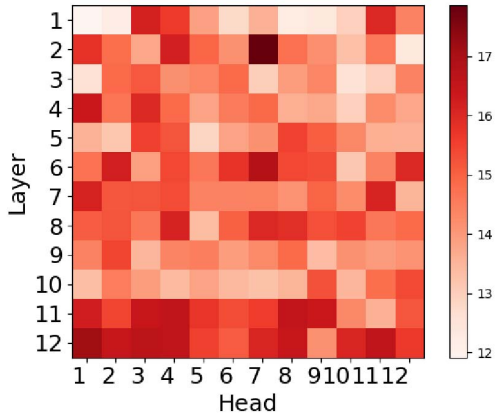
We observed that in the later layers, most attention is given to certain words or phrases, and also to the start and end tokens ("[CLS]" and "[SEP]") of the BERT tokenizer. Therefore, in COVID-19 related hate tweets, the attentions in earlier or later layers can be studied to understand the word associations in such tweets, depending on the distribution of the training data.

B. Hate Speech Control with BERT Attention

We utilize the results of the previous section to formulate a control strategy for COVID-19-related hate-speech using BERT attention mechanism. We use the attentions given to the keywords discovered in Section IV-A by other words in a sequence, in the layers found to have long distance word associations (from Figure 2a and Figure 2b). Since these other words contribute to the hateful context in an input sequence, these words must also be pointed out for removal or re-consideration. We then propose to a user to re-consider sending such words or changing these words.



(a) Attention distance by layer and head in the Asian-hate dataset.



(b) Attention distance by layer and head in the Boomer-hate dataset.

Fig. 2: Attention distance in the two COVID-29 datasets.

Existing studies on BERT attention mechanism [8], [24] suggest that the attention formulation in Equation 1 prioritizes tokens with higher dot product vectors. Hence, the attention mechanism of BERT can be used to find other words in a tweet, that attend to the hateful keywords. In this work, we

use this phenomenon to find the top other words that attend the most to the hateful keywords. Table III depicts randomly selected samples from the hate datasets with hateful words and keywords highlighted.

In a real-world system, we propose a control strategy in which a tweet posted by a user is run through our model to detect any hate content. If any hate content is detected in the tweet, keywords discovered in our work can be searched in the tweet. If any of the keywords are found, our strategy of finding other words that significantly attend to these keywords can be presented to the user for removal or reconsideration, along with the hateful keywords.

C. Is BERT Detecting Hate Speech based on Existing Definitions of Hate?

Several existing studies [15], [19], [33] suggest that hate-speech targets disadvantaged social groups in a manner that is potentially harmful to them. From a broader perspective, these disadvantaged groups could also be individuals, who could be targets of hate speech. Our objective in this experiment is to study whether the BERT model implicitly detects hate-speech based on such existing definitions of hate-speech from literature.

We first identify the words that pertain to the targets of hate-speech in both the COVID-19 datasets. We consider both groups (e.g. "Chinese", "Seniors") and individuals (e.g. "Xi Jinping") as targets for this experiment. Some samples of the chosen target words are depicted in Table IV.

Target	**Samples**
Groups	han, chinese, chinesetourists, taiwanese, libs, babyboomers, magats, muslim, jews, asians, koreans, african, africans, christians, indians
Individuals	spokespersonchn, jinping, trump, jackma, pompeo, boris, potus, chr******s, m******7, g*********8

TABLE IV: Samples of words chosen as targets. Username identifiers have been removed to preserve user identities.

Our objective is to study to what extent BERT model may be using associations between hateful keywords and such targets words to detect hate-speech. We base our study on the attention that these keywords may be attributing to these target words. If the model is learning to pay higher attention to the target words from the keywords (corresponding to higher attention weights) than the non-target words in a tweet, this could indicate that the BERT model strongly uses these associations to detect hate-speech. For each tweet in both the COVID-19 datasets, we capture the attention weights from the the hateful keywords to the target words such as the ones in Table IV. We then plot the CDF of such attention weights for certain layers for both the Asian-hate and the Boomer-hate datasets. Our results are presented in Figure 3 and Figure 4, respectively for the Asian-hate dataset and Boomer-hate dataset.

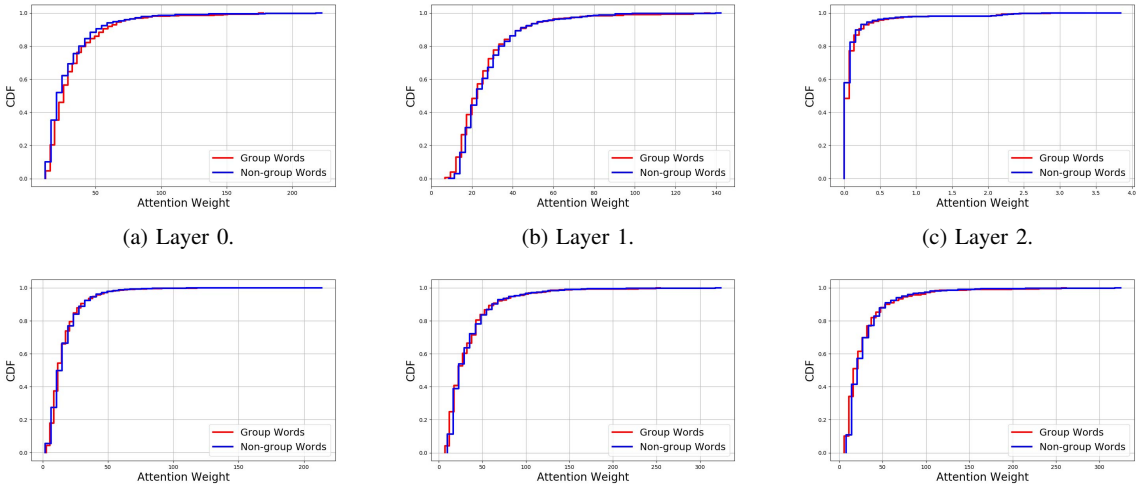
In the Asian-hate dataset results depicted in Figure 3, we plot the CDF for layers 0, 1, 2, 4, 9 and 11 for target words (depicted by red curve). We chose these layers so that we have good representation from all depth levels and also from our result from Section IV-B that for this dataset, longer distance association may be formed in the earlier layers. For comparison, we also plot the CDF for non-targets words (depicted by blue curve) occurring in the tweets, which are ordinary words. We found that for this dataset, the BERT model seems to pay similar attention for keywords and target/non-target words. While preliminarily this may indicate that BERT does not learn well to associate keywords with target words, we found that BERT learns the subtle differences between hate and non-hate tweets (e.g., "chinese get out" and "stop telling chinese to get out"), based on associations between keywords and both target words and non-target words. Our analysis of the Asian-hate dataset led to the observation that although the keywords and target words are themselves not hateful, their associations could be hateful in hate tweets. In order to make correct detection, the BERT model seems to learn the associations between these two kinds of words in conjunction with the other non-target words in the tweet to make accurate predictions. Thus, we observed that BERT does form association between hate keywords and target words, however it does not only depend on these associations to make predictions, which may be the reason why BERT is found to be more powerful than other sequence models such as recurrent neural networks.

Next, we analyze the Boomer-hate dataset using the same procedure described above. The results of our experiment on Boomer-hate dataset is depicted in Figure 4. We found the results on this dataset to be quite different from the results in the case of Asian-hate dataset. In this case, the BERT model seemed to be associating more strongly between the hateful keywords and the target words (depicted by red curve), when compared to the non-target words (depicted by blue curve). For example, in Figures 4a and 4c, we can see clearly, the observation that association between target words and hateful keywords are given a lot more attention than the non-target words. Even in Figure 4b (a later layer with more distance associations, Section IV-B), this trend seems to be visible.

Upon further investigation, we observed that this behavior could be due to the reason that the Boomer-hate dataset is more sparsely containing hateful keywords and the target keywords. For example, in the case of Asian hate, we observed a lot of different targets ranging from groups (e.g., "chinese", "taiwanese", "asians") and keywords (e.g., "kungflu", "wuflu", "wuhanvirus"). However, in the case of Boomer-hate we found relatively fewer number of such words, as the target is mostly singular (older people only) and the hate keywords therefore, are also quite limited. Hence, we observed that in such cases, where a less varied patterns need to be learned by BERT model, it depends more on learning association between certain words than learn more subtle and varied associations.

V. Related Work

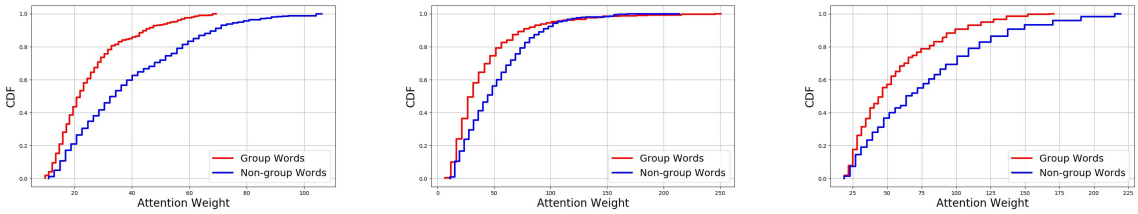
Several recent studies have emerged in the area of hate speech detection. In [13], the authors used Reddit, which is a community with a platform that shares information in the form of posts with the ability to be up voted or down voted



(a) Layer 0. (b) Layer 1. (c) Layer 2.

(d) Layer 4. (e) Layer 9. (f) Layer 11.

Fig. 3: Attentions to Target words Vs. Non-target words in case of Asian-hate.



(a) Layer 1. (b) Layer 9. (c) Layer 10.

Fig. 4: Attentions to Target words Vs. Non-target words in case of Boomer-hate.

based on the reader's opinion towards it. They used a public data set from subreddit /r/TD to collect 16,349,287 comments about the president and the presidency. They utilized TF-IDF to identify distinct hate words towards Donald Trump and used Wikipedia articles to identify nicknames for Trump. They concluded with findings about how humans used tools like bots to keep themselves entertained, but did not focus on pinpointing removing those bots, resulting in minimal research on preventing internet trolling.

The authors of [22] used Gab (gab.com) to find out the diffusion of hate speech. For the dataset, they used a Lexicon based filter to identify racial slurs, and chose non-ambiguous words to increase accuracy. They also utilized DeGroot's model of information diffusion to identify hateful users. They focused on the diffusion characteristics of hateful users, but not how to pinpoint and remove hateful comments in general. In [27], the authors used a large dataset from Reddit and Gab and narrowed it down to hate speech by using human intervention, which is inefficient because it takes a long time to label so many tweets. It is also unreliable because there are some tweets that are incorrectly labeled. They used a survey and crowdsourcing to label all the tweets, which is not reliable, takes too much time, and adds cost. They created a dataset of hate speech and used programs like Seq2Seq and VAE. These are unreliable because it only uses an input and output tags, and does not go through multiple verifications. VAE may be unreliable for such tasks because sequences are discreet (unlike continuous image signals), and does not pinpoint certain hate words.

A recent work [34] studies the spread of hate and counter-hate during the COVID-19 pandemic. The authors collect a dataset of 2,400 tweets and train a text classifier to identify hate and counterhate tweets. The authors also find that hateful users in Twitter were less engaged in anti-Asian hate speech prior to their first anti-Asian tweet, following which such tweets turned to being more aggressive and hateful. However, a proportional rise in counterhate tweets was not observed by the authors.

Using attention mechanisms in natural language processing tasks such as classification, next sentence prediction, question answering and neural machine translation (NMT) were first introduced by [6] and [7], and most implementations are based on the models introduced in [21]. The use of attention mechanisms were broadly adapted to various NLP

tasks, often achieving then state-of-the-art performances in tasks such as reading comprehension [5] and natural language inference [26]. Multi-headed attention was first introduced by [32] for NMT and English constituency parsing and termed the model as "transformer", and further adopted for transfer learning [10], language modeling [9], [28], and semantic role labeling [31].

VI. Conclusion

In this work, we have studied the recent phenomena of hate speech triggered by the COVID-19 pandemic. We have focused our study on the hate-speech in Twitter against Asian community and old people. We have trained a BERT-based model to detect hate-speech based on the datasets in this work and used the multi-headed attention mechanism of BERT to discover novel keywords (186 keywords targeting the Asian community and 100 keywords targeting older people) using our strategy. Further, we have discussed how BERT could be learning longer distance attentions based on the underlying distribution of training data, and found that such attentions are learned in the earlier layers for the Asian-hate dataset and later layers for the Boomer-hate dataset. We have introduced a strategy to study whether BERT is learning hate-speech detection based on existing definitions of hate-speech. We have learned that in the case of Asian-hate dataset, BERT focuses on varied attention between several words, whereas in the case of the Boomer-hate dataset, BERT focuses on certain word associations to detect hate-speech.

Acknowledgment

This work is supported in part by National Science Foundation (NSF) under the Grant No. 2031002.

References

[1] Macguire, E., Anti-Asian Hate Continues to Spread Online Amid COVID-19 Pandemic, in Al-Jazeera. https://www.aljazeera.com/news/2020/04/anti-asian-hate-continues-spread-online-covid-19-pandemic-200405063015286.html, 2020.

[2] Mehta, I., Twitter Sees 900% Increase in Hate Speech Towards China – because Coronavirus, in The Next Web. https://thenextweb.com/world/2020/03/27/twitter-sees-900-increase-in-hate-speech-towards-china-because-coronavirus, 2020.

[3] Perspective API. https://www.perspectiveapi.com//home, 2020.

[4] Whalen, A., What is Boomer Remover and Why is it Making People So Angry?, in Newsweek. https://www.newsweek.com/boomer-remover-meme-trends-virus-coronavirus-social-media-covid-19-baby-boomers-1492190, 2020.

[5] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

[6] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.

[7] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[8] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.

[12] Karmen Erjavec and Melita Poler Kovačič. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920, 2012.

[13] Claudia I Flores-Saviaga, Brian C Keegan, and Saiph Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[14] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[15] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[16] David Gadd, Bill Dixon, and Tony Jefferson. Why do they do it? racial harassment in north staffordshire. *Centre for Criminological Research, Keele University*, 2005.

[17] Donald P Green, Dara Z Strolovitch, and Janelle S Wong. Defended neighborhoods, integration, and racially motivated crime. *American journal of sociology*, 104(2):372–403, 1998.

[18] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[19] James B Jacobs, Kimberly Potter, et al. *Hate crimes: Criminal law & identity politics*. Oxford University Press on Demand, 1998.

[20] Jennifer L Lambe. Who wants to censor pornography and hate speech? *Mass Communication & Society*, 7(3):279–299, 2004.

[21] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[22] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182. ACM, 2019.

[23] Robert King Merton and Robert C Merton. *Social theory and social structure*. Simon and Schuster, 1968.

[24] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024, 2019.

[25] Bhikhu Parekh et al. Is there a case for banning hate speech? *The content and context of hate speech: Rethinking regulation and responses*, pages 37–56, 2012.

[26] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

[27] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*, 2019.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[29] Larry J Ray and David Smith. Hate crime, violence and cultures of racism. 2002.

[30] Ashley Reichelmann, James Hawdon, Matt Costello, John Ryan, Catherine Blaya, Vicente Llorent, Atte Oksanen, Pekka Räsänen, and Izabela Zych. Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, pages 1–12, 2020.

[31] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[33] Samuel Walker. *Hate speech: The history of an American controversy*. U of Nebraska Press, 1994.

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