Tweeting the End of Hate Speech

Course project for CSE 6240: Web Search and Text Mining, Spring 2021

Ezekiel Day
eday30@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Andrew Hatcher
ahatcher8@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Nate Knauf
nate.knauf@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

The goal of our project was to improve upon the work done by Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar in their paper titled, "Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis" [6]. As the title of this paper suggests, anti-Asian hate and counter-hate are very prominent in social media throughout the entirety of the COVID-19 pandemic. Even prominent figures such as then President of the United States Donald Trump posted tweets containing anti-Asian rhetoric, arguably fueling additional hate on social media [4]. Our main objective through this project was to implement Aspect-Based Sentiment Analysis and compare this method with the baseline model implementations from the Ziems paper. We then created an ensemble method of the best features to create an improved model that will more reliably classify tweets as hate, counter-hate, or neutral.

We utilize two baselines, both trained and tested using the original training data of the Ziems paper [6]. The first baseline is the set of model performance metrics documented in the Ziems paper. The second baseline is the set of performance metrics generated by our partial reconstruction of the models from the Ziems paper using our own implementation of some of the features they described. Using these baselines as a point of comparison, we implemented a series of models that used various combinations of keyword aspect based sentiment analysis, group based keyword sentiment, and the baseline features used in the Ziems paper. Using this method, we found that while BERT embeddings performed the best individually, grouped keyword sentiment performed very similarly to the hashtags feature baseline, and keyword sentiment wasn't far behind, performing better than the linguistics baseline feature. Further, when creating ensemble methods, we found that when we used the baseline features and our newly created features, our AUROC results matched the best results in the Ziem's paper almost exactly. However, we noticed a drastic improvement in both Precision and Recall scores in our ensemble method over the Ziem's paper.

ACM Reference Format:

Ezekiel Day, Andrew Hatcher, and Nate Knauf. 2018. Tweeting the End of Hate Speech Course project for CSE 6240: Web Search and Text Mining, Spring 2021. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

1.1 Objective

The goal of our project is to add to and improve upon the work done by Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar in their paper titled, "Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis." [6] As the title of this paper suggests, anti-asian hate and counterhate was very prominent in social media throughout the entirety of the COVID-19 pandemic. Our main objective through this project was to implement Aspect-Based Sentiment Analysis and compare this method with the baseline model implementations from the Ziems paper. We then created an ensemble method of the best features to create an improved model that will more reliably classify tweets as hate, counter-hate, or neutral. We wish that our methods and results can be utilized to learn more about the relationship between hate and counterhate tweets as well as the effects of these interactions on social media.

1.2 Data Source

For this research, we require a large set of tweets for testing as well as a subset of previously labeled tweets as hate, counterhate, and neutral for training our models. We chose to derived our dataset from that which is used in the Ziems paper as it was curated for this specific purpose. [6]. We have 15,441,819 rows of data encompassing the accounts of 5,227,531 users. The vocabulary size was 7,810,957 unique words through all the tweets. We further calculated statistics for unique hashtags and distributions of hashtags that are discussed later.

1.3 Methodology

Our work will stem from the model used in "Racism is a Virus" [6]. The first baseline that we are utilizing is based off of the classifier and feature space used in the Ziems paper. It uses the combination of Hashtags, Linguistics, and BERT embeddings as features in the classification of hate, counterhate, and neutral tweets. The second baseline is derivative of the first baseline, but it also incorporates the overall sentiment of the tweet in the feature space. With these baselines, we were able to match the results found in the Ziems paper [6]. To improve upon the baselines, we looked to introduce aspect based sentiment analysis, both based on individual keywords, and also based on a set of grouped keywords.

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1.4 Results

Among our aspect based sentiment analysis features, we found that the addition of grouped keyword sentiment outperformed individual keyword sentiment. However, in our implemention, aspect based sentiment analysis individually slightly under performed the models that used BERT embeddings. We chose to create an ensemble model that used all of the baseline features along with our additional features, and we found almost identical AUROC scores to the ensemble baseline in the Ziems paper, both having scores of around 0.876 for the hateful class, 0.852 for counterhate, and 0.828 for neutral tweets. However, when looking at Precision and Recall, our ensemble method tends to outperform across the board. [6]

1.5 Impact

We hope that the results from this research project could be used to reliably classify hate and counterhate tweets and in turn be used to create correlations between users who write hate tweets as well as indicate the success that counterhate has in suppressing hateful rhetoric. Even broader in scale, this concept of classifier could be expanded to topics outside of just Anti-Asian hate to other forms of hate and prejudice. If applied safely, this model could be used across social media platforms to sensor hateful tweets. Overall, our model and results have the potential to reduce toxicity and create better environments for social media users.

2 LITERATURE SURVEY

We aim to utilize sentiment analysis to classify tweets based on their sentiment on different key topics. One current framework for sentiment analysis is VADER, or "Valence Aware Dictionary for sEntiment Reasoning." [5] VADER is a rule-based model that produces a numerical estimate of sentiment for input strings. VADER takes into account a lexicon of just over 7,500 features including English vocabulary, English slang such as "Yeah" and "Nah," English internet slang such as "lol," "smh," "thx," and "ty," and internet emoticons such as ":)" and "XD." In addition, VADER also incorporates the effects of relationships between words such as capitalization, punctuation, and degree modifiers such as "extremely" or "marginally." This model was developed using crowd-sourced sentiment scores from 10 independent raters recruited through Amazon's Mechanical Turk service. VADER was compared to other modern sentiment-analysis frameworks and demonstrated to be significantly more accurate than others for the task of classifying tweets.

In the "Racism is a Virus" project, researchers from Georgia Tech's CLAWS Lab developed a model to detect anti-Asian hate in COVID-19 related posts on the social media website Twitter. [6] They created a novel dataset named COVID-HATE which we discuss later on. They created a large number of features, which they grouped into three categories: Linguistic features, which included stylistic and syntactic features of tweet text as well as VADER-calculated tweet sentiment, Hashtag features, which represented the presence of COVID-19 or hate related hashtags, and finally 768-dimensional BERT embeddings of tweets. [2] It was found that the model performed best when excluding the Linguistic features to only use the Hashtag and BERT embedding features. Using Logistic Regression models that included these features, researchers were able to classify the different types of tweets with mean AUROC scores ranging from 0.828 to 0.876. In this project, we aim to improve the Linguistic features by exploring additional ways of representing tweet sentiment, as the original model only looks at the overall sentiment of each tweet.

In the "Credibility Assessment in the News" project at Georgia Tech Research Institute, researchers worked to classify news websites based on credibility and political bias. [3] Articles were scraped from various news websites using news feeds provided by the GDELT Project. A 5-bin left-right political spectrum was used to label the political content of websites, whereas three one-hot variables were used to label websites as "Fake," "Conspiracy," and "Satire." The articles were parsed using natural language tools to generate features based on article content, specifically including the relationship between publication websites and specific keywords. For website-keyword pairs, TF-IDF scores were calculated to represent the importance of the keyword to that website, and sentiment scores were calculated to represent the average sentiment or opinion of that website regarding the keyword. These textual content features were found to be effective at recovering labels of political bias on a website. In this project, we aim to use similar methods to measure the sentiment of a specific user or their tweets in relation to specific hashtags and keywords, instead of just measuring the overall sentiment of the user or their tweets.

3 DATASET DESCRIPTION

3.1 Data Preparation

Our initial dataset consisted of a list of 31 million Tweet IDs from the paper "Racism is a Virus"[6]. With these 31 million IDs, we began querying Twitter's API[1] for the Tweet text and author ID of each tweet. One hurdle we had to jump when using the API was that Twitter rate limits its API. As a result, we had to apply for multiple Academic Research API keys in order to get the tweets in a timely manner. We split the tweet IDs into 3 equal sections and individually began making requests to the API. We created a script that would automatically request new tweets at a set time interval to avoid being rate limited by the API. Through heavy use of scripting, we were able to scrape the API and de-duplicate any resultant data that might have shown up. Due to the deletion of some Tweets, we were able to gather text and Author IDs from 15,441,819 million unique tweets, which we believe is more than sufficient for this project.

After collection of the tweets was complete, we followed the data cleaning process outlined in "Racism is a Virus"[6] to ensure that we could adequately compare results. We first compiled a hashtag-vocabulary of all unique hashtags. We gathered the total count of each hashtag across all tweets as well as the individual count of each hashtag per tweet. After, following the procedure from "Racism is a Virus", we removed "urls, leading user handles, retweet indicators, emojis, white space, and pound signs from hashtags"[6].

After cleaning was complete, we began designing the features of our tweets. For our baseline features for each tweet, we calculated the average sentiment across the whole tweet, the count of each hashtag of interest, the total word count, and the total number of characters.

Overall, our dataset is sufficient to achieve our goal for a few reasons. First, the overall size of the data, while not necessarily being as large as the original paper, is still vast in size. It is approximately 2.1 GBs in memory. We have gathered most of the pertinent tweets from the time period of Jan-Aug 2020 which was when COVID dialogue was at its peak. Further, we have gathered sufficient information to determine authorship of tweets. This will allow us to accurately flag accounts as a whole that might be consistently hateful. Finally, we believe that by adding Aspect-Based Sentiment Analysis, we can outperform the original paper's model. We believe localized sentiment will be very telling as to whether a tweet is hateful or not. By building on the previous model, we are positioned to surpass the expectations of this project.

3.2 Raw Data Statistics

As stated previously, our dataset was derived from the Ziems paper's dataset[6]. We were able to retain the set of 2,319 annotated tweets as training data, as well as gather text and author information for another 15,441,819 tweet ids. We calculated key statistics among these tweets based on number of characters, number of words, and number of sentences per tweet.

We found that the average tweet in this dataset was comprised of 116 characters where the median was 95 characters, and the minimum and maximum were 1 and 300 characters respectively. The average tweet consisted of 18 words. The median number of words was 15 and the minimum and maximum were 1 and 115 respectively. Finally, there is an average of 2 sentences per tweet in our dataset. The median, unsurprisingly for tweets, was 1 sentence. Additionally, the median number of words per sentence was 9 words.

Through all of these tweets, a total of 7,810,957 unique words were identified. Finally, it should be noted that we gathered the author ids of 5,227,531 unique authors. Therefore, there were about 3 tweets captured per author in this dataset.

3.3 Data Analysis

The dataset used in the Ziems paper was originally obtained by retrieving tweets containing a hashtag from a predefined set of covid—19, hate, and counter-hate keywords[6]. Because of this, we found it pertinent to investigate the most prevalent hashtags found in the dataset. As seen in Table 1 above, among the most commonly used hashtags were terms such as "coronavirus", "covid19", and "covid—19". Each of these terms appeared in the original set of keywords and are common names for the virus, so it's no surprise that they would be present in the hashtags. However, hashtags such as "chinesevirus", "chinavirus", and "wuhanvirus" being among the most prevalent tags in the dataset gives insight as to how common this form of hate is. We also found that there were common hashtags in the dataset that weren't among the original set of keywords, including terms like "stayhome", "lockdown", "who", and "trump". The correlation of such keywords with the labels of hateful, counter-hateful, and neutral rhetoric would make for an interesting discussion point in the future.

Table 1: Most Common Hashtags

Hashtag	Occurrences
coronavirus	1448262
covid19	1309934
chinesevirus	280458
chinavirus	219902
wuhanvirus	208794
covid—19	183311
chinaliedpeopledied	173122
ccpvirus	168825
china	148972
covid_19	128340
covid	86585
covid2019	86057
corona	75394
coronavirusoutbreak	73096
wuhan	62092
ccp	53428
cina	43236
stayhome	41813
boycottchina	39252
wuhancoronavirus	38491

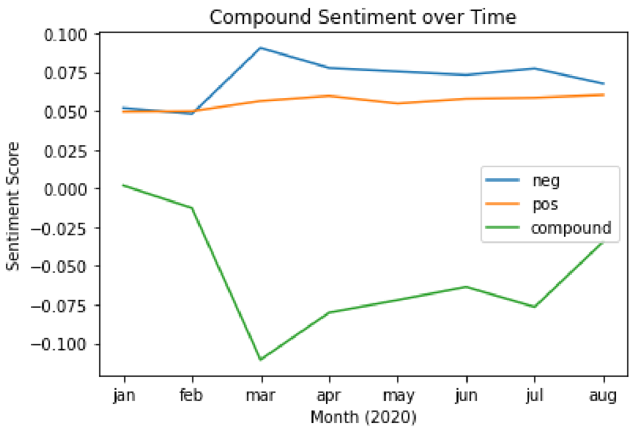

Figure 1: Compound Sentiment over Time

In our next step of analyzing our dataset, we decided it would be pertinent to chart average sentiment of a tweet over time. For this graph (See Figure 1), we randomly sampled 2000 tweets from each month of our dataset. The blue line represents the average negative sentiment (as an absolute value), the yellow line represents positive sentiment, and the green line is a compound weighted average of positive, negative, and neutral sentiment over those months. The y-axis gives a determination of average sentiment with -1 being very negative and 1 being very positive. The most important part to note is the green line – specifically the large dip in sentiment during March. This makes sense as COVID was rapidly spreading through early Spring in the US and lockdowns were beginning to be enforced during this time period. Therefore, it is not surprising that the average sentiment of tweets pertaining to COVID decreased drastically. It is also interesting to note that sentiment did not stay low for long and began to reach neutral levels by August 2020.

As our final level of analysis, we decided to plot a histogram of tweet labels (hate, counter-hate, neutral) across average sentiment

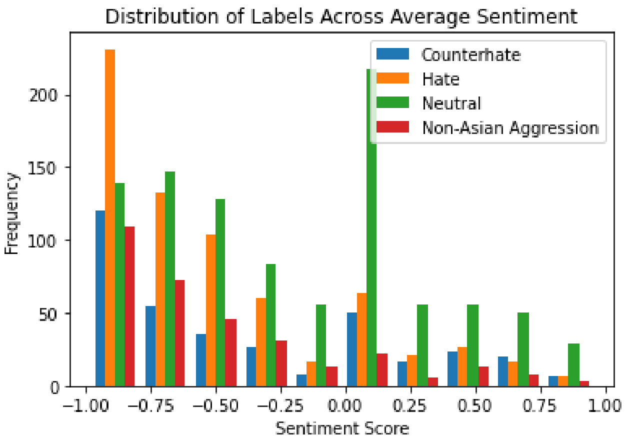

Figure 2: Distribution of Labels Across Average Sentiment

of a tweet. In Figure 2, you can see these histograms for selected average sentiment scores. An obvious first observation is that hate labels generally appear where sentiment is lowest (the largest orange bar is to the very left). Similarly, and not surprisingly, most neutral labels are centered around neutral-sentiment tweets. However, it can also be observed that most counter-hate tweets also share a very negative average sentiment. This is indicative that most counter-hate tweets tend to "fight hate with hate." Further, as the average sentiment of the tweet goes up, the number of labels in general falls. This is indicative that during this time period - with regards to COVID - the sentiment was overall negative.

4 EXPERIMENT SETTING AND BASELINES

In accordance with the research of Ziems et. al [6], our task is a three-way classification of tweets as "hateful", "counter-hateful", or "neutral." We will evaluate our results primarily by AUCROC, but we will also analyze precision, recall, and F1 score. We believe AUCROC will be the most telling metric of our model's ability to distinguish between the three classes. This is because we expect most tweets to fall into the "neutral" class. With this heavy class imbalance, a metric such as accuracy or precision alone will not be sufficient to gauge the results of our classifier. Further, we can directly compare our results to the Ziems paper's findings by using these metrics. As for the data distribution, we split our training and testing data using an 80-20 split. We shuffle the results during the process and perform 5-fold cross validation to limit over-training.

4.1 Baseline Descriptions

The first baseline we are using stems from the "Racism is a Virus" paper [6]. Their classifier used a mixture of linguistic, hashtag, and BERT embedding features to train their model. They found that a combination of hashtag and BERT embedding features performed the best compared to other combinations of features. Therefore, we will use this combination of features (hashtag and BERT embedding) as our first baseline (Emb+Hash). The hashtag feature incorporated a count of each relevant hashtag in a tweet. The embedding feature was a 768-dimensional BERT embedding. The hashtag feature is intuitive to add because most hashtags tend to summarize the main idea of the tweet [6]. While hashtags can be abused and overused, we feel it is beneficial to include this in our baseline. Further, we include the BERT embedding to apply semantic meaning to the tweet and give context. As this combination of features produced the best results in "Racism is a Virus," we felt it was a suitable baseline to use. A summary of the baseline can be found here: https://arxiv.org/pdf/2005.12423.pdf with the code repository here: http://claws.cc.gatech.edu/covid.

Table 2: Baseline Comparison

Feature Set	Class	Precision	Recall	F$_1$*Score*	AUROC
Emb+Hash	Hate	72.3%	63.8%	67.8%	0.876
All	Hate	68.9%	64.4%	66.5%	0.867
Ling+Hash	Hate	66.9%	72.8%	56.2%	0.769
Emb+Hash	Counter	58.0%	42.8%	49.0%	0.852
All	Counter	52.8%	41.1%	46.0%	0.836
Ling+Hash	Counter	73.8%	68.4%	44.7%	0.793
Emb+Hash	Neutral	72.0%	66.5%	69.1%	0.828
All	Neutral	70.8%	65.3%	67.9%	0.820
Ling+Hash	Neutral	68.0%	68.0%	63.7%	0.732

For our second baseline, we wished to incorporate sentiment into our feature space as well. Our baseline calculates the average sentiment over the whole tweet to determine a sentiment score. This sentiment ranges from -1 to +1. Further, this baseline retains the hashtag features from Emb+Hash, which were proven to be effective in identifying hate tweets. Overall, this baseline is suitable because it introduces the notion of sentiment features for a tweet while not being overly simplistic. It still retains some complementary features from Emb+Hash to improve upon the results of sentiment alone. By adding Aspect-Based Sentiment Analysis, we hope to surpass this baseline as well as Baseline1.

4.2 Baseline results and discussion

First we will consider the Ziems baseline alone, which is comprised of Feature Sets "Emb+Hash" and "All." The metrics we will focus on for our research will stem from this "Emb+Hash" row. However, we felt it necessary to include the "All" row (which includes linguistics) as we will consider linguistics in our feature set as well.

As we can see from rows 1, 2, 4, and 5 in Table 2, the embedding and hashtag feature combination performed better for more metrics than when combined with linguistic features (All). In the Ziems paper [6], they also distinguished that BERT worked better for their embedding feature than GLoVE did. Their classifier was better able to identify hate tweets than counter-hate tweets (a difference of .024 AUROC when considering the BERT embedding and hashtag feature combination). Surprisingly, the Ziems Neutral tweet classifier had the most fluctuation across different combinations of features. When considering Precision, using solely the BERT embedding performed the best (narrowly) compared to when hashtag embeddings were also used. Recall had a tie between BERT embeddings and the BERT and hashtag combination. However, the BERT and hashtag combination still outperformed all the other feature spaces in terms of AUROC.

When comparing our baseline (Ling+Hash) to the Ziems baseline (Emb+Hash), it is clear that our baseline leaves a little to be desired. Among Hate, Counter, and Neutral labels, our baseline performed worse in terms of AUROC, which is our main metric. However, there were a few instances, especially when considering counter-hate, that our baseline did better. We generally achieved higher Recall across the labels while our Precision suffered because of this. However, there is room to build upon these baselines. It should be noted that when it comes to hate speech, it might be considered better to flag more overall tweets (higher recall) than have the flags for hate be more pure (higher precision). By iterating over these baselines and adding Aspect-Based Sentiment Analysis, we hope to surpass the Ziems baseline.

5 PROPOSED METHODS

The main novelty we brought over the baselines is the addition of Aspect-Based Sentiment Analysis. We created two groups of Aspect-Based Sentiment features. For these Aspect-Based Sentiment features, we scanned a tweet for occurrences of certain hashtags of interest. The first feature group, Keyword Sentiment, used Aspect-Based Sentiment scores for the hashtags that most frequently appear in the whole COVID-HATE dataset [5]. The second feature group, Group Sentiment, used Aspect-Based Sentiment scores for three different classes of hashtags listed in the Ziems paper: Hate hashtags, CounterHate hashtags, and COVID hashtags [6]. Records of both lists can be found inside the source code. For each hashtag that we found, sentiment scores were calculated within a 7-token-long sliding window of text centered on the hashtag. Sentiment was calculated using the vaderSentiment SentimentIntensityAnalyzer. The whole window around the hashtag was fed in as input to this analyzer.

As stated previously, we believe adding a feature for sentiment in a local area of select keywords will be more powerful than simply the overall sentiment of the tweet. When combined with BERT, hashtag, and linguistic embeddings, we found that the model performed arguably better than our baselines. Keyword sentiment allows a representation of a tweet that is more specific than the baseline tweet representation. We know that there can be mixed sentiment in any sentence of the English language. Therefore, identifying a sentiment for each keyword can help us better understand whether a tweet is truly hateful or not.

We conducted our investigation utilizing a Jupyter environment running Python 3.6 code. Our BERT embedding produced a 768-dimensional vector. The implementation for our BERT embedding utilized the source code from the Ziems paper. We constructed 120 keyword features for sentiment, 12 group features for sentiment, 6 linguistic features, and 42 hash features. The sentiment was computed using the vaderSentiment python package. We experimented with different combinations of features to find the optimal feature space for our Logistic classifier. Our classifier utilized the implementation provided by sklearn and specified a class_weight of 'balanced.'

6 EXPERIMENTS AND RESULTS

Tables 2 and 3 illustrate the results for the baselines and our proposed method of Aspect-Based Sentiment Analysis.

Table 3: Proposed Method Comparison

Feature Set	Class	Precision	Recall	$F_1 Score$	AUROC
Sent	Hate	71.3%	68.7%	58.3%	0.787
BERT+Sent	Hate	81.3%	75.2%	70.2%	0.874
All	Hate	80.3%	79.1%	70.1%	0.877
Sent	Counter	75.7%	61.3%	43.7%	0.777
BERT+Sent	Counter	82.1%	62.8%	52.1%	0.846
All	Counter	79.5%	71.2%	51.7%	0.852
Sent	Neutral	62.6%	68.5%	60.0%	0.707
BERT+Sent	Neutral	74.2%	72.5%	69.9%	0.822
All	Neutral	75.1%	73.5%	71.0%	0.828

First, we would like to address our Feature Sets we included in the report. As an exploratory step, we computed statistics from our dataset solely using sentiment features. This feature set is denoted "Sent" in Table 3. It is comprised of both the Keyword Sentiment feature and the Group Sentiment feature explained earlier. It is a 132-dimensional representation of a tweet. We found that sentiment alone boasted modest scores - specifically to Precision and Recall. AUROC overall was lower than most baselines. However, once we incorporated BERT embeddings along with our Sent representation, our metrics improved distinctly. We saw an average of 9% increase in Precision, 4% in Recall, and 9% in AUROC. This feature set is denoted as BERT+Sent. In the Ziems paper baseline, their BERT embedding was denoted Emb. Lastly, our "All" feature set incorporates hashtag and linguistic features on top of sentiment and BERT embedding. Improvement between BERT+Sent and All was modest but did show progress. AUROC increased in every case with the addition of hashtag and linguistic features. However, this extends the feature space to be 948-dimensional.

As for our comparison to the baselines, there are a few takeaways. First, we matched AUROC between their best feature sets and ours per class. While we were hoping to increase this metric, we are happy to meet the standards of another research team. In terms of Precision and Recall, we can provide better news. We saw a noteworthy increase in both Precision and Recall between our feature set and theirs. This is extremely encouraging. We believe that Aspect-Based Sentiment helped provide a needed boost to both.

7 CONCLUSION

We found that our biggest shortcoming, which would also be an area of further extension, is our lack of training data. There was a huge disconnect between the amount of training data available and the amount of data total. We know that hand-labeling training data is an arduous task. If further research is to be continued on this topic, we recommend expanding the training set, perhaps by using a platform akin to Amazon Mechanical Turk. An increase in training data will not only improve the generalizability of the model, it will open doors to other methods such as Deep Learning. These methods have been shown to be effective, but they also require much more data.

8 CONTRIBUTION

All team members have contributed a similar amount of effort.

REFERENCES

[1] 2020. Twitter API GET /2/tweets. https://developer.twitter.com/en/docs/twitter-api/tweets/lookup/api-reference/get-tweets#tab2

[2] M.-W.; Lee K.; Devlin, J.; Chang and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

[3] Natalie Fitch, Nate Knauf, James Fairbanks, and Erica Briscoe. 2018. Credibility Assessment in the News: Do we need to read? http://snap.stanford.edu/mis2/files/MIS2_paper_17.pdf

[4] Yulin Hswen, Xiang Xu, Anna Hing, Jared B. Hawkins, John S. Brownstein, and Gilbert C. Gee. 2021. Association of “covid19” Versus “chinesevirus” With Anti-Asian Sentiments on Twitter: March 9–23, 2020. arXiv:2021.306154

[5] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

[6] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. arXiv:2005.12423 [cs.SI]