ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

Classification and Analysis of Hate Speech Towards Asians in the COVID-19 Pandemic

Eric Shang Nan Chen[1,*], Hexuan Zhao[2], Qiwei Yang[2], Yixin Huang[3]

[1]BASIS International School Guangzhou, Guangzhou, 510633, China

[2]University of Waterloo, Waterloo N2L 3G1, Canada

[3]University of California, Berkeley, Berkeley 94720, USA

[*]Corresponding author: chen.ericsn@outlook.com

Abstract

The COVID-19 pandemic has catalyzed a surge in hate crimes and online hate speech towards East Asians. In this study, a more nuanced hate speech classifier that incorporates sentiment analysis is developed. Two successive machine learning models are used to classify a sample of 315,627 COVID-19 tweets into five categories. The first model performs a 3-category classification task with an F1 score of 0.81191, outperforming existing models. The second model performs a binary classification task with an F1 score of 0.75460. Moreover, a comparative analysis of hostile and criticism Tweets with the Whissell Dictionary of Affect reveals that hostile Tweets contain significantly more imagery (p=0.027). The nuanced model developed in this study and the incorporation of sentiment analysis may aid the development of future hate speech screening algorithms by governments and social media platforms, as well as researchers investigating the linguistic characteristics of hate speech.

Keywords

Natural Language Processing; Hate Speech; Anti-Asian Hate; Computational Social Science; Sentiment Analysis.

1. Introduction

According to the World Health Organization (2021), as of May 19th, 2021, there have been more than 163 million infections and 3.3 million deaths caused by COVID-19 [1]. Although pandemic control is the principal public concern, adverse social effects related to xenophobic sentiments and racist hate crimes warrant attention and response from policymakers and academia. In particular, hateful sentiment towards Asians, especially the Chinese, has escalated along with the outbreak's impact because of the virus's origins. As early as March 2020, the Federal Bureau of Investigation (FBI) warned of a potential increase in hate crimes against Asian Americans [2]. Between March 19th, 2020 and February 31st, 2021, the Stop-AAPI Hate reporting center received 6,603 incidents of discrimination towards Asian Americans, which consisted of verbal assaults (65.2%), shunning (18.1%), physical assault (12.6%), civil rights violations (10.3%), and online harassment (7.3%) [3]. One notable incident was the stabbing of a Burmese-American family at a Sam's Club in Texas, in which the suspect claimed that he stabbed the family because he thought they were Chinese and were spreading COVID-19 [4]. Moreover, Asian activists suggest that due to language barriers, lack of confidence in the police to take action, and "a cultural tendency to remain quiet," a large number of incidences go unreported [5]. The aforementioned incidences are significant due to their quantity and severity and the broader societal implications. According to Gover et al., individual hate crimes, combined with institutional support demonstrated by using phrases such as "Chinese Virus" by officials,

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

reinforce existing social hierarchies that place East Asians at a disadvantage [6]. While physical attacks have garnered national attention, the present paper focuses on hate speech on social media. The dissemination of online hate speech, defined by Davidson et al. as "language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group," has become a significant concern for social media platforms in recent years [7]. Beyond creating a hostile online environment, the continued spread of online hate speech normalizes othering and racist narratives, which can, in turn, incite offline acts of discrimination. Historically, hate speech has contributed to harmful social movements such as the Anti-Semiticmovement in Nazi Germany and the Indian Removal movement in the United States [8]. In 2018, a man who shot 18 people in the Tree of Life Synagogue in Pittsburgh made Anti-Semitic comments with references to the synagogue on an asocial media platform called Gab [9]. Therefore, it is evident that online hate can precipitate real-world hate crimes. In the context of the current COVID-19 pandemic, escalating hate speech towards Asians constitutes a secondary "infodemic" that spreads xenophobic sentiments. This is a recurring phenomenon, as xenophobic hate speech peaked during the previous Ebola and SARS outbreaks as well [10]. Considering the connection between online hate speech and offline acts, content moderation on social media becomes an important issue to tackle. In the present circumstance, researchers from the Alan Turing Institute have suggested that social media should be viewed as a critical infrastructure to protect during COVID-19 [11].

For the reasons stated above, hate speech on social media warrants regulation. Analysis of online hate speech can aid the creation of content moderation tools and provide insight into the dynamics of East Asian hate as a social contagion. The advent of tools such as machine learning, Natural Language Processing, and network model algorithms combined with the availability of social media data and the well-defined network structures of social media platforms allows researchers to develop automatic classifiers for hate speech and analyze the dynamics of their diffusion process [7][12]. Hence, creating a state-of-the-art classifier for hate speech and understanding the characteristics of Anti-Asian hate speech through sentiment analysis are the goals of this work. Finally, since methods of intervention for hate speech involving automatic content modulation are bound to raise debate over free speech and unethical uses of artificial intelligence, a brief discussion of hate speech ethics and the limitations of automated detection is included.

2. Literature Review

Social media hate speech research sits at the intersection between sociology, computer science, and data science. Thus, its evolution has involved technical advances in Natural Language Processing and theoretical advances in ethics and linguistics [13]. Of the sub-fields in hate speech research, the automatic detection of hate speech online has received considerable attention from academics. However, few studies focus on the COVID-19 context. In 2017, Davidson et al. used crowd-sourcing to label a sample of tweets collected through a crowd-sourced hate speech lexicon into three categories: those containing hate speech, those with only offensive language, and those with neither [7]. They trained a multi-class classifier to distinguish between these different categories [7]. Their work established the fundamental framework for hate speech classification research.

There have been two significant studies investigating anti East Asian hate speech during the pandemic. Vidgen et al. from the Alan Turing Institute collected more than 20,000 tweets and classified them using embedding models such as RoBERTa, LSTM, and BERT into the categories of hostility, criticism, discussion, counterhate, and neutral tweets [14]. Ultimately, their model achieved an F1 score of 0.83 [14]. In a similar study, Ziems et al. collected more than 30 million

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

tweets and trained/classified them under hate, neutral, and counter-hate labels using logistic regression models, random forest classifiers, and support vector machines (SVM), achieving an AUROC score of 0.852 [12]. One limitation of Ziems et al.'s study is that they had a limited training set; thus, they noted combining their dataset with the hand-labeled dataset of COVID-19-related hateful and counter hateful tweets by Vidgen et al. could enhance model accuracy [12]. Furthermore, Ziems et al.'s model did not account for more nuanced categories of hate and counter-hate. To improve upon these limitations, the present work aggregates Ziems et al.'s dataset with Vidgen et al.'s and develops a more nuanced classification model with the aid of sentiment analysis. Ultimately, this work aims to help policymakers and the general public raise awareness about hatred against Asians during the COVID-19 pandemic.

3. Data

The dataset used in this work is the COVID-HATE dataset of tweets collected by Ziems et al. using keywords and hashtags related to COVID-19. The dataset includes more than 33 million tweets from the timeframe of January 2020 to August 2020. First, the dataset is downloaded. Second, python scripts are utilized to strip away information other than the Tweet ID and split the data into manageable sections of 8 million tweets each. Third, Hydrator, an open-source application, converts the Twitter IDs to JSON lines format using the Twitter API. The resulting JSON lines files contain information about the tweets such as time, user, retweet count, original text, hashtags, and user information. Finally, 315,627 tweets from the dataset are randomly sampled for the application of our model (the sampling frequency was higher for Tweets created from April to August because the original dataset is smaller for those months). The Pandas library is then used to drop all but four columns of the dataset, sort the Tweets by time, and remove duplicate Tweets (determined using Tweet ID). A CSV file is obtained through this process that contains all the Tweets that will be classified with our model and required additional information for analysis. Table 1 and figure 1 exhibit essential characteristics of our dataset.

Table 1. Essential Characteristics of the sampled dataset

Essential Characteristics of the sampled dataset	
Number of Tweets	315,627
Timespan	Jan 15th to Aug 10th, 2020
Number of columns before processing	34
Number of Columns after processing	4

The distribution of tweets over time is also briefly investigated, and we discovered a large volume of tweets in March when the term "Chinese Virus" surfaced. There are also large fluctuations in tweets over time, which may be because of our random sampling process and inconsistencies in the original dataset created by Ziems et al. However, this does not significantly impact the present study because the time series analysis is not a core component.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

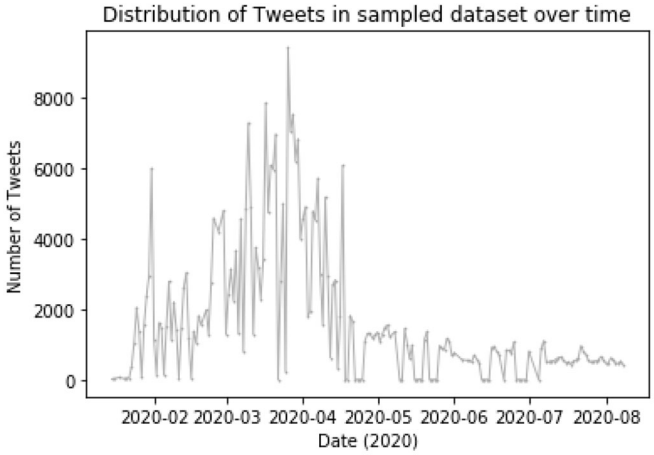

Figure 1. Line plot depicting the number of tweets over time in our sampled dataset

For a more precise analysis of the COVID-HATE dataset, Ziems et al. provide a comprehensive view in their paper. However, this work contributes additional insights by analyzing the latest data after April and developing a more nuanced classification model. In addition, two hand-labeled datasets of COVID-19 related hateful and counter-hate tweets by Ziems et al. and Vidgen et al. are utilized for model training [12] [14]. In doing so, the training set is enlarged, and the probability of overfitting our model is reduced, which was a limitation that Ziems et al. discussed.

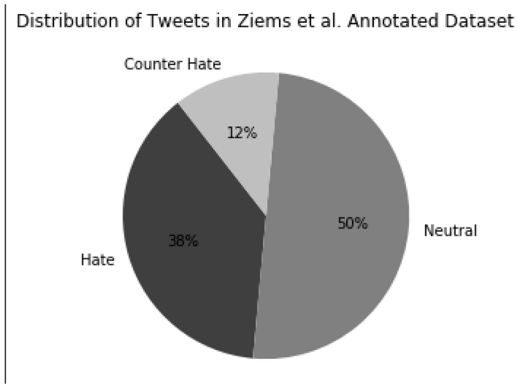

Figure 2. Distribution of annotated dataset created by Ziems et al. (2400 tweets)

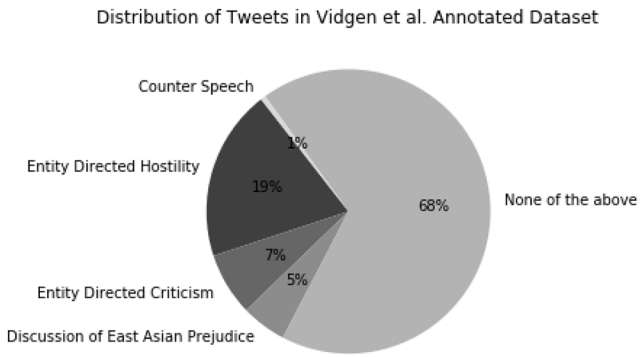

Figure 3. Introductory statistics of the annotated dataset (20,000 tweets) created by Vidgen et al. (2020)

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

4. Methodology

The proposed methodology expounds upon previous work by developing a more reliable and nuanced classifier and training it on a larger dataset. The hate speech classifier is divided into two models. The first model classifies the tweets into three categories: Hate, Neutral/Other, and Counter-hate, which are the same categories in Ziems et al.'s (2020) classifier. Then, a second model further classifies the tweets labeled as "hate" into two categories: strong hate and minor hate. Finally, the counter-hate tweets are split into aggressive counter-hate and minor counter-hate using a rule-based approach combining several metrics. The motivation behind using two models is to implement sentiment analysis in the classifier. If sentiment analysis is directly implemented in the first model, it will likely hinder the model's ability to differentiate between counter-hate and hate, as both types of tweets tend to contain negative sentiment. Therefore, a second model in which sentiment analysis can help differentiate more extreme and aggressive hate speech from less aggressive hate speech is introduced. Figure 3 shows an overview of our methodology.

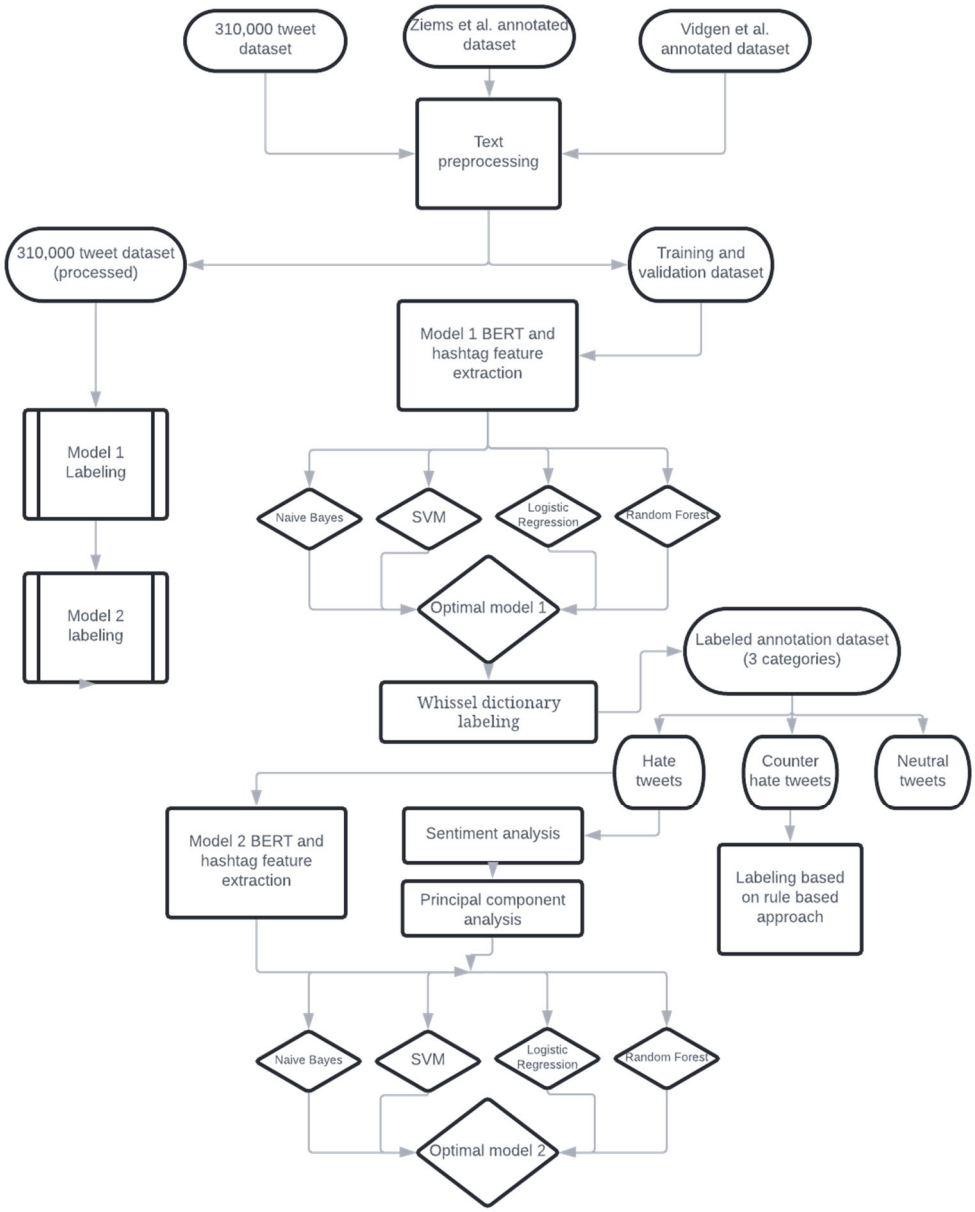

Figure 4. Flowchart overview of the main steps in our methodology.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

4.1. Tweet Category Definitions

The labels created by Ziems et al. and Vidgen et al. are converted into five different labels with detailed descriptions below:

Strong Hate Tweets: The tweets under the strong hate label are a combination of the tweets under the "Hostility against an East Asian entity" categorized by Vidgen et al. (2020) and tweets with intense negative sentiment or derogatory and malicious intent against Asian people, organization, country, or government from tweets under the "Anti-Asian COVID-19 Hate Tweets" category by Ziems et al. [12] [14].

For instance, below is an example of a tweet labeled as strong hate [12]:

"It is the Chinese virus, from China, caused by your disgusting eating habits, your cruelty. Boycott anything Chinese #kungflu #chinaliedpeopledied #covid".

Critical or Offensive Tweets: The tweets under this label are a combination of the tweets under the "Criticism of an East Asian entity" category by Vidgen et al. and tweets with minor negative sentiment against Asian people, organizations, countries, or government from tweets under the "Anti-Asian COVID-19 Hate Tweets" category by Ziems et al. [14][12].

For instance, criticism against the Chinese government is part of this category[14].

"the CCP hid information relevant to coronavirus."

Neutral Tweets: Nonhateful/counter hate tweets containing content related to COVID-19 (Ziems et al. 2020). These are tweets from the "Hate-Neutral Tweets" category by Ziems et al. (2020) and the "Neutral" category by Vidgen et al. (2020). Many tweets in this category are news-related, advertisements, or spam tweets.

For instance, below is an example of news-related neutral tweet:

"COVID-19: #WhiteHouse Asks Congress For $2.5 Bn To Fight #Coronavirus: Reports #worldpowers #climatesecurity #disobedientdss #senate #politics #news #unsc #breaking #breakingnews #wuhan #wuhanvirus https://t.co/XipNDc"

Counter-hate: These are Tweets that challenge or condemn abuse against Asian people, organizations, countries, or governments.

For instance, below is an example of a counter-hate tweet:

"you shouldn't say that, it's derogatory."

This tweet's relatively logical, mild, and corrective tone in condemning hate speech makes it fall under this category.

Strong Counter-hate: These tweets strongly challenge or condemn abuse against Asian people, organizations, countries, or governments. It may contain aggression towards a specific target.

For instance, below is an example of a strong counterhate tweet:

"@BichonSuzi @CNN @cnnhealth the idea that the virus gives a sh*t about borders is even more so. The idea that Chinese people, products, or food are riskier is dangerous and false propaganda. It denies the realities of this virus and puts Asian Americans at risk - and you engaging in it makes you an asshole."

The strong, aggressive tone of this tweet in condemning hate speech makes it fall under this category. It is also directed at specific entities and people.

4.2. Data Pre-processing

Each tweet is processed using the regex library in a similar method as Ziems et al. Preprocessing includes removing hashtags symbols (keeping the word), usernames, and links, as well as trimming extra spaces. This step ensures the proper tokenization of the tweets. In addition, the two training datasets are merged, and the existing labels are replaced with the labels defined above.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

Table 2. Label Statistics

Label Type	Label Description
Strong Hate, Hostile	Tweets with intense negative sentiment or malicious and derogatory intent against Asian individuals, organizations, governments, or countries.
Critical or offensive	Tweets containing criticism or offensive language against Asian individuals, organizations, governments, or countries.
Neutral	Non-hate or counter-hate tweets containing content related to COVID 19
Counter-hate	Tweets that challenge or condemn abuse against Asian individuals, organizations, countries, or governments.
Strong counter-hate	Tweets that challenge abuse against Asian individuals in an aggressive way, possibly directed towards an individual or specific entity.

4.3. Model 1 Creation

Hashtags: A hashtag—written with a # symbol—is used to index keywords or topics on Twitter. The features of hashtags are used as an indicator to determine the expression of a word or a sentence. A total of 60 hashtags are used in a counting system that counts the occurrence of a specific word. Some typical hashtags are “COVID-19”, “ChinaVirus,” and “RacismIsAVirus.” The occurrence of hashtags is used as one of the factors classifying the category of the tweet, as they can reveal the intended meaning of the tweet. For instance, a tweet with a hashtag of ‘ChinaVirus’ is more likely to be a hate tweet, whereas a tweet containing “RacismIsAVirus” is likely to fall under the counter-hate category [12].

Tweet Embeddings: Unlike the bag-of-words style feature sets, text embedding models are widely used to incorporate word-level and sentence-level semantics, meaning they can be contextually aware to some extent. We take two candidate embedding models are taken into consideration: BERT [15] and GloVe [16]. Ziems et al. determined that the BERT model outperforms the GloVe model in the majority of the tasks by comparing the AUROC score [12]. Hence, BERT embeddings were selected to generate 768- dimensional tweet embeddings.

Model Creation: The two feature sets are concatenated together. Then four types of classifiers from the Sci-kit Learn Python library, including Naïve Bayes, Logistic Regression, Support Vector Machines, and Random Forest, are trained to classify the tweets as hate, counter-hate, and neutral (Pedregosa et al., 2011). [17]. We also use the grid search function from Sci-kit Learn to find the optimal parameters for the model, as well as cross-validation with cv = 5 to reduce overfitting [17].

4.4. Model 2 Creation

Sentiment Analysis: Sentiment analysis utilizes positive or negative classifications of textual-based opinions [18] [19] [20]. The Liu dictionary [21] and Natural Language Toolkit’s (NLTK) Vader Sentiment package are used to generate sentiment features for training [22]. The Liu dictionary consists of roughly 2,000 and 4,800 positive and negative opinion words (sentiment words), respectively, which were derived by Hu and Liu from online reviews [21]. This work uses the Liu dictionary to generate 6800-dimensional word sentiment features.

The NLTK library’s Vader Sentiment package, on the other hand, generates a continuous score from -1 to 1 for each body of text [22]. A clear distinction between Vader Sentiment and the Liu dictionary is that Vader sentiment takes into consideration factors such as capitalization and the context of the entire sentence, whereas the Liu dictionary only scores individual words.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

Thus, the Liu dictionary is not able to detect that the statement "I am not sad" is actually positive. Therefore, the scores produced by the Liu dictionary are multiplied by -1 if Vader Sentiment returns a result of the opposite sign compared to the Liu dictionary. Furthermore, the Liu dictionary word frequencies are scaled by the absolute value of the Vader Sentiment score to prevent overemphasizing the sentiment feature. In the end, the scaled word scores from the Liu dictionary and the Vader Sentiment score are used as features.

Principal Component Analysis: Since sentiment analysis produces a feature with more than 6800 dimensions, principal component analysis, a method of feature reduction based on linear algebra manipulations, is used to drastically decrease the number of features [23]. This prevents sentiment analysis from bearing too much importance in the model, reduces training time, and improves model performance.

Model Creation: The same four classifiers as model 1 (Naïve Bayes, Logistic Regression, Support Vector Machines, and Random Forest) are trained on the new feature set containing Tweet embedding and sentiment analysis features. In this second classifier, grid search and cross-validation from the Sci-Kit learn library are also implemented [17].

4.5. Emotion Analysis Using Whissell Dictionary of Affect

To further characterize the classified tweets, the Whissell Dictionary of Affect in Language is used to derive three average metrics for each category of Tweets: Pleasantness, activation, and imagery, each rated from 1 to 3 by experts [24][25]. Pleasantness and activation are affective measures, whereas imagery indicates how easy it is to form a mental picture from the text (more precise definitions are provided in the original study) [24]. The dictionary consists of 8742 words that are rated by humans on these three metrics. In the present work, words in the Tweets are matched to the dictionary to generate average scores of pleasantness, activation, and imagery for each category of Tweets.

Table 3. Whissell Dictionary Characteristics

Metric	Mean	Standard Deviation
Pleasantness	1.81	0.44
Activation	1.85	0.39
Imagery	1.94	0.63

5. Results

5.1. Model 1 Results

Upon comparing the precision, recall, and F1 scores of the four models, it is evident that support vector machines performed the best on the 3-category classification task. Support vector machines (SVM) achieved an F1 score of 0.81191, which far exceeded the other models. Although logistic regression resulted in the highest precision score, its recall score was also the lowest, making it a poor model. The results are shown in Table 4.

Table 4. Model Data

Model	Precision	Recall	F1 Score
Naïve Bayes	73.699%	74.423%	74.012%
Random Forest	78.770%	78.486%	75.082%
SVM	82.234%	82.825%	81.191%
Logistic Regression	85.488%	62.604%	71.888%

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

The normalized confusion matrix (figure 5) was analyzed to determine possible sources of error. As shown in the matrix, the model achieves great accuracy for classifying hate speech (0.94). The performance for neutral tweets is decent (0.77), but there is a significant number of tweets falsely classified as hate, which represents a false positive. This is an area for improvement because it is not appropriate or ethical to classify criticism or news as hate speech. However, the largest source of error is the counter-hate category. Only a small fraction of the counter-hate Tweets were classified correctly (0.22), and most counter-hate tweets were classified as hate or neutral.

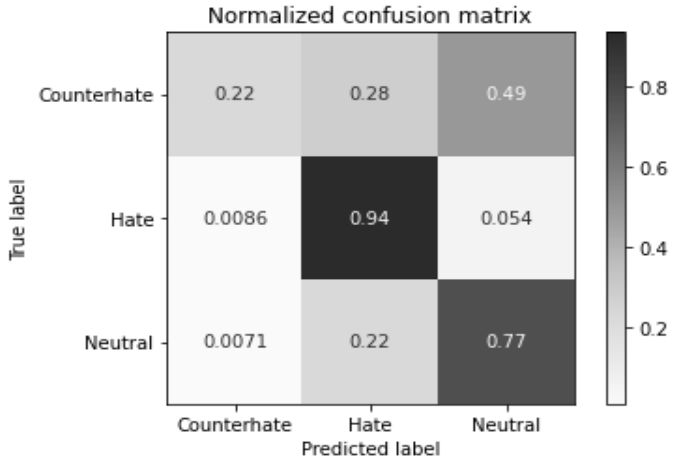

Figure 5. Confusion matrix for optimal model 1 (support vector machine).

The main reason is likely due to counter-hate being underrepresented in the training set, as counter-hate made up less than 2 percent of the dataset. In comparison with Ziems et al.'s results, the present model far exceeds their model's performance for the hate and neutral categories, but our model performed worse for counter-hate. This is likely due to our decision to the training set being a combination of the Vidgen et al. and Ziems et al. datasets, which led to more training data for hate speech but also decreased the proportion of counter-hate tweets.

5.2. Model 2 Results

Upon comparing precision, recall, and F1 scores, logistic regression emerges as the best classifier for model 2. While SVM outperforms logistic regression in terms of Precision and Recall, Logistic Regression has a higher weighted F1 score. Therefore, logistic regression is the optimal model.

Table 5. Comparison of classifiers for model 2

Model	Precision	Recall	F1 Score
Naïve Bayes	72.925%	75.445%	72.412%
Random Forest	75.298%	74.695%	66.552%
SVM	77.528%	77.320%	72.366%
Logistic Regression	75.141%	76.664%	75.460%

The F1 scores for model 2 are lower, partly due to the higher difficulty of the more nuanced classification task.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

5.3. Labelling of Larger Dataset and Analysis

The sampled dataset of more than 310,000 tweets was labeled with our two models. According to model 1, 94.09% are neutral, 4.29% contain hate, and 1.62% are counter hate.

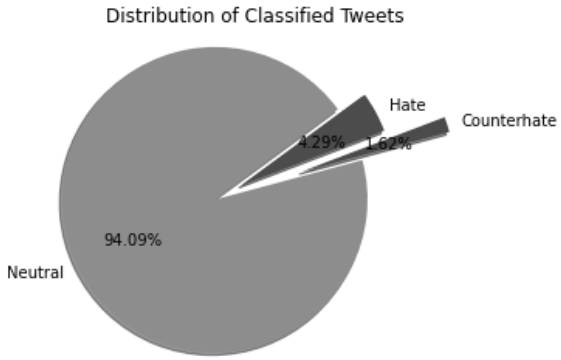

Figure 6. Distribution of classified Tweets in the test dataset

Model 2 classified 73.29% of the tweets as strong hate and 26.71% as criticism or offensive language. This demonstrates that aggressive and hostile speech is more prominent than less hostile criticism or offensive speech. The result also hints at the polarizing effect of social media on controversial topics, as hate is more prevalent than mere criticism or offensive language.

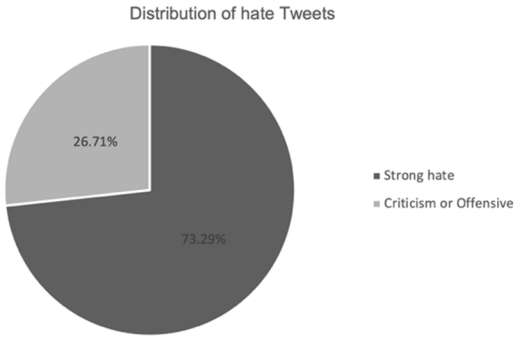

Figure 7. Distribution of hate Tweets in the test dataset

Counter-hate tweets were classified using a threshold from model one (probability higher than 0.7 is strong counter-hate) rather than model 2 due to insufficient training data. On the other hand, 47.34% of counter-hate tweets were classified as strong counter-hate, while 52.66% were classified as less aggressive counter-hate. This suggests that counter hate tweets may be more likely to be logical responses to those spreading hate rather than hostile responses.

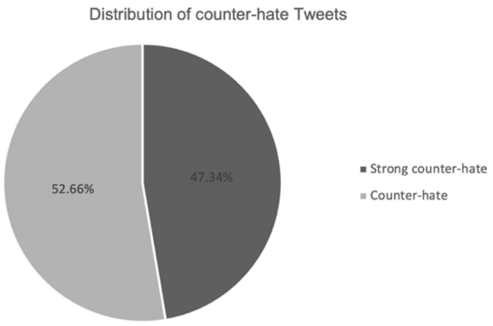

Figure 8. Distribution of counter-hate Tweets in the test dataset

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

Furthermore, time series analysis reveals that hate tweets consistently outnumber counter-hate tweets, but they follow the same patterns in spikes as the popularity of related topics fluctuates.

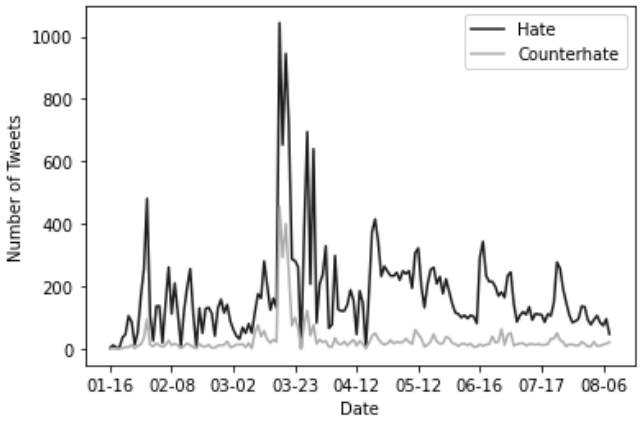

Figure 9. Number of hate Tweets and counter-hate Tweets over time

Finally, analysis of the relationship between the tweet categories and the retweet count reveals that a larger percentage of hate tweets have more than 50 retweets than counter-hate tweets (Table 6). Therefore, we conclude that hate speech may be more contagious than counter-hate, which is consistent with Ziems et al.'s findings [12].

Table 6. Number of retweets based on classifier category

Number of Retweets	0	[1,10)	(10,50]	>50
Hate	79.719%	13.378%	1.534%	5.369%
Counter-hate	85.127%	11.669%	1.249%	1.955%

5.4. Analysis with Whissell Dictionary

Labeling all classified Tweets with the Whissell dictionary resulted in the following match rates (the proportion of words that matched an entry in the dictionary). Compared to the benchmark match rates of 90% in natural language samples, the match rates for the Twitter dataset are lower due to the substantial number of non-English Tweets and misspellings.

Table 7. Whissell match rates for different categories of Tweets

Tweet Category	Match Rate
Strong Hate, Hostile	72.13%
Critical or Offensive	73.64%
Neutral	67.35%
Counter-hate	76.56%

Table 8 displays the metric scores. As shown, neutral Tweets were the most pleasant and had the most activation, whereas hostile Tweets contained the most imagery.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

Table 8. Labelled Whissell Dictionary Scores

Category	Pleasantness	Activation	Imagery
Strong Hate/Hostile	1.779	1.685	1.580
Criticism	1.778	1.678	1.539
Neutral	1.820	1.689	1.561
Counter-hate	1.788	1.652	1.462

Z-scores of each metric was also calculated for comparison to the original Whissell dictionary. The Z-scores generally fall in the lower end for Activation and Imagery. Further analysis of the linguistic characteristics of Tweets would be needed to understand the reasons.

Table 9. Z Scores Based on Original Whissell Dictionary

Category	Pleasantness	Activation	Imagery
Strong Hate/Hostile	-0.070	-0.423	-0.572
Criticism	-0.073	-0.447	-0.636
Neutral	0.023	-0.413	-0.602
Counter-hate	-0.049	-0.507	-0.759

To compare the two nuanced categories of hostile and criticism Tweets for statistically significant differences, a two-sample T-test was performed for each of the metrics, and the T-statistics were converted top-values to test for significance.

Table 10. Z Scores Based on Original Whissell Dictionary

Category	Pleasantness	Activation	Imagery
T-statistic	0.046	0.416	1.949
P-value	0.482	0.341	0.027

Using the $p < 0.05$ threshold, only the difference for the imagery category is statistically significant. This suggests that hostile Tweets usually have more words that suggest specific actions or objects that are concrete [25]. For example, the word “kill” has an imagery score of 3.

One limitation of using the Whissell dictionary is that it cannot rate words that are new and specific to COVID-19. Thus, future work may attempt to label topic-specific words based on pleasantness, activation, and imagery.

6. Discussion and Conclusion

The present study’s major contributions are combining Vidgen et al. and Ziems et al.’s dataset, developing a more accurate and nuanced model to classify hate Tweets, applying sentiment analysis to hate speech classification, and providing insight into the characteristics of anti-Asian hate Tweets and counter-hate Tweets. Augmenting Ziems et al.’s training dataset improves model performance significantly for hate speech classification. The two-model method performs slightly better than Vidgen et al.’s model in classifying strong hate tweets versus minor hate. Finally, the Tweets were labeled with Whissell Dictionary of Affect scores, and a comparison between hostile and criticism Tweets demonstrated that hostile Tweets contained more imagery by a statistically significant amount ($p=0.027$).

Most importantly, this work introduces a framework for hate speech classification that can be vastly improved in the future. For example, the scarcity of counter-hate speech creates a biased

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

dataset and leads to poor model performance. In the future, new datasets with more counter-hate tweets can be scraped from various websites to build a more robust training set. In addition, a base version of the BERT embedding model is used in this study. If fine-tuned versions of BERT or another embedding model are used in the future, it will likely result in better model performance.

Increasing hate crimes and hate speech towards Asians are important social issues to pay attention to amidst the COVID-19 pandemic and continue monitoring even after the pandemic. Developing a scalable, nuanced classifier for hate speech on social media platforms will be useful for semi-automatic methods of filtering hate speech, which can prevent the spread of hateful sentiment. However, there is a fine line between hate speech and acceptable free speech. Thus, any government or corporate implementation of semi-automatic hate speech removal would have to consider model imperfections and biases. For future work, an even more nuanced classifier can be developed using parts of speech tagging and context-aware embedding models to detect potential offline actions from tweets, which can help predict the occurrence of hate crimes. Finally, once there is more data about offline hate crimes and racist incidents, data analysis can reveal the correlation between online hate speech and offline hate crimes.

Acknowledgments

The authors of this paper would like to acknowledge Dr. Patrick Houlihan, adjunct professor at the University of Columbia, for his mentorship in completing this research project.

References

[1] World Health Organization, "WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard." https://covid19.who.int/ (accessed May 20, 2021).

[2] J. Margolin, "FBI warns of potential surge in hate crimes against Asian Americans amid coronavirus," ABC News, Mar. 27, 2020. https://abcnews.go.com/US/fbi-warns-potential-surge-hate-crimes-asian-americans/story?id=69831920

[3] R. Jeung, A. Yellow Horse, and C. Cayanan, "STOP AAPI HATE NATIONAL REPORT," Stop AAPI Hate, Mar. 2021. [Online]. Available: https://secureservercdn.net/104.238.69.231/a1w.90d.myftpupload.com/wp-content/uploads/2021/03/210312-Stop-AAPI-Hate-National-Report-.pdf

[4] H. Tessler, M. Choi, and G. Kao, "The Anxiety of Being Asian American: Hate Crimes and Negative Biases During the COVID-19 Pandemic," Am. J. Crim. Justice, vol. 45, no. 4, pp. 636–646, Aug. 2020, doi: 10.1007/s12103-020-09541-5.

[5] E. Yoon-Ji Kang, "Activists Say Anti-Asian Attacks Go Unreported Due To Stereotypes, Language Barriers," NPR All Things Considered, Mar. 22, 2021. https://www.npr.org/2021/03/22/980075515/activists-say-anti-asian-attacks-go-unreported-due-to-stereotypes-language-barri

[6] A. R. Gover, S. B. Harper, and L. Langton, "Anti-Asian Hate Crime During the COVID-19 Pandemic: Exploring the Reproduction of Inequality," Am. J. Crim. Justice, vol. 45, no. 4, pp. 647–667, Aug. 2020, doi: 10.1007/s12103-020-09545-1.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," ArXiv170304009 Cs, Mar. 2017, Accessed: Sep. 13, 2020. [Online]. Available: http://arxiv.org/abs/1703.04009

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

[8] A. Tsesis, Destructive messages: how hate speech paves the way for harmful social movements. New York: New York University Press, c2002.

[9] R. McIlroy-Young and A. Anderson, "From 'Welcome New Gabbers' to the Pittsburgh Synagogue Shooting: The Evolution of Gab," presented at the Thirteenth International AAAI Conference on Web and Social Media, 2019. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3264

[10] T. Karalis Noel, "Conflating culture with COVID-19: Xenophobic repercussions of a global pandemic," Soc. Sci. Humanit. Open, vol. 2, no. 1, p. 100044, 2020, doi: 10.1016/j.ssaho.2020.100044.

[11] J. Cowls, B. Vidgen, and H. Margetts, "Why content moderators should be key workers - Protecting social media as critical infrastructure during COVID-19," The Alan Turing Institute, Apr. 15, 2021.

[12] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis," ArXiv200512423 Phys., May 2020, Accessed: Aug. 16, 2020. [Online]. Available: http://arxiv.org/abs/2005.12423

[13] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts, "Challenges and frontiers in abusive content detection," in Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 2019, pp. 80–93. doi: 10.18653/v1/W19-3509.

[14] B. Vidgen et al., "Detecting East Asian Prejudice on Social Media," ArXiv200503909 Cs, May 2020, Accessed: Aug. 23, 2020. [Online]. Available: http://arxiv.org/abs/2005.03909

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," ArXiv181004805 Cs, May 2019, Accessed: Sep. 16, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[18] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd international conference on Knowledge capture, 2003, pp. 70–77.

[19] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of human language technology conference and conference on empirical methods in natural language processing, 2005, pp. 347–354.

[20] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of HTML documents," in Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 1075–1083.

[21] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[22] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Proceedings of the International AAAI Conference on Web and Social Media, 2014, vol. 8, no. 1.

[23] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202203_5(3).0052

[24] C. M. WHISSELL, "Chapter 5 - THE DICTIONARY OF AFFECT IN LANGUAGE," in The Measurement of Emotions, R. Plutchik and H. Kellerman, Eds. Academic Press, 1989, pp. 113–131. doi: https://doi.org/10.1016/B978-0-12-558704-4.50011-6.

[25] C. Whissell, "Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language," Psychol. Rep., vol. 105, no. 2, pp. 509–521, Oct. 2009, doi: 10.2466/PR0.105.2.509-521.