# Detecting Hate Speech on Social Media with Respect to Adolescent Vulnerability

Anna Chiu
*Department of Computer Science*
*California State University, Fullerton*
Fullerton, United States of America
anna.chiu@csu.fullerton.edu

Kanika Sood
*Department of Computer Science*
*California State University, Fullerton*
Fullerton, United States of America
kasood@fullerton.edu

Ariadne Rincon
*Department of Computer Science*
*California State University, Fullerton*
Fullerton, United States of America
arincon10912@csu.fullerton.edu

Davina Doran
*Department of Computer Science*
*California State University, Fullerton*
Fullerton, United States of America
davdoran@csu.fullerton.edu

2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) | 979-8-3503-3286-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCWC57344.2023.10099373

***Abstract*—Social media has become one of the biggest platforms for children and young adolescents to spend their free time. Due to the rising age gap between people using social media, tweens can be exposed to offensive and provocative posts made by others that are older. Twitter allows children aged 13 and up to create an account on their platform. While this is a common age restriction on most social media platforms, it can be damaging to young adolescents who are at a crucial time developing their morals and beliefs based on what they see online. Identifying hate speech within a timely matter is crucial for censoring hate speech for kids. This can impact the overall environment on social media to be less toxic and uplifting to all users. In this work, we propose to use multiple machine learning techniques: SVM, k-nearest neighbor, Naive Bayes, and soft-voting ensemble classifier.**

***Index Terms*—Machine Learning, Hate speech recognition, Classification, SVM, K-Nearest Neighbor, Naive Bayes, Ensemble Method**

## I. Introduction

As social media has become an ever-growing part of our lives, an immense amount of rude or hateful speech has surfaced without the precautions of who the reader will be. This project focuses on identifying and classifying tweets as either hateful or neutral. With posts that include suggestive offensive/harmful statements, our machine should be able to classify and censor speech from another individual's platform to the user. This work is essential because we believe in the right to free speech. Similar to YouTube having an age restriction option for its videos, we believe social media should also have a mechanism to protect both older and younger users. A harmful and neutral classification is helpful for social media sites to be able to accurately identify and censor the posts from their platforms for the intended user. This project differs from other machine learning detection systems because our project will only censor certain words for younger users aged 13-15 years old, but allow older users to remain censor free. Our proposed machine-learning technique can help companies use this algorithm to allow less censorship and termination for older accounts and safer cyberspace for tweens. We aim to detect harmful/offensive posts and censor them for a specific group of young users.

## II. Background

For the majority of the adolescent population here in the United States, social media is not only a luxury but an integral part of their lives. Social media allows everyone to share their thoughts, pictures, videos, etc. Although this gives us the feeling of connectivity to the world, we are unknowingly exposing ourselves to extreme misinformation, harassment, and toxic material [8] that also exists. Adolescents between the critical ages of 13 and 15 are by far the most affected, and we know this by observing the number of adolescent suicides resulting from bullying and harassment [8]. Livingstone and Third state that the Internet is age-blind [9]. It caters to an older demographic of users, and as a result, there is a more robust and explicit vocabulary that only the target audience should view. Our social media platforms, like Twitter, can also amplify these messages [10] to the wrong audience, our adolescents.

Our project aims to use machine learning algorithms to detect hate speech in social media. Machine learning algorithms have contributed to hate speech detection on social media analysis [1]. Social Media is integrated into a part of our daily lives. Machine learning algorithms must detect abusive messages and flag them for censorship. We research multiple works proposing different solutions for hate speech detection.

In [2], Aljero and Dimililer propose a stack ensemble approach that utilizes both the predictions from base-lever classifiers and the extracted features from the data set as input to the meta-level layer of the model to make the final predictions. They use 4 different data sets throughout their training and testing, including the same data set we've implemented in this paper. Their feature extraction phase utilizes word-to-vector and sentence-to-vector to transform their text data into a vectorized format. Of the individual base level, classifiers

979-8-3503-3286-5/23/$31.00 ©2023 IEEE

SVM [3] performs the best individually with an F1-score of .95. In contrast, their proposed stacking method performs an F1-score of .97, and consists of SVM, LR (logistic regression), and XGB(XGBoost classifier). The authors conclude the two-phase system for classification is not repetitive and therefore, justifiable given the increase in performance.

Looking into the vectorization process more, we reference the work of Anita Kumari Singh and Mogalla Shashi [4]. Their research solidifies the decision to use this method, as we would convert our data into something more meaningful for the model. The authors recommend for those seeking to use the TF-IDF [5] stick with static data. Therefore, this method would not work if we actively classify hate tweets versus neutral tweets. For our project and our research, we collect data from a static data set that would suffice to both train and test our model.

In [6], the authors propose a different approach for hate speech classification. They apply MANDOLA, an automated system for monitoring, detecting, analyzing, and visualizing hate speech while preserving privacy. It consists of 6 parts: data collection, data pre-processing, hate speech detection, hate topic inference, hate speech metadata storage, and a visualization dashboard. The features utilized for training are also broken down into categories consisting of simple surface, word generalization, sentiment analysis, lexical resources, and linguistic features. Their model uses a 3-layer classifier of machine learning and deep learning techniques. In their analysis, the MANDOLA method outperforms other state-of-the-art works on this topic, such as HybridCNN [7] in precision (.89), recall(.89), F1-score(.89), and AUC score(.92).

Given a sample tweet, our model should be able to accurately identify what category the post falls under. By analyzing existing tweets that are classified or labeled as either hateful or neutral, we can analyze future tweets. This will help us quickly and accurately censor hateful tweets while maintaining others' ability to post other content to their liking. Censoring offensive or harmful tweets is necessary for the modern age to keep a safe online environment. Considering that most of the world has access to the internet, it is the responsibility of social media companies to keep their platforms clean for users aged 13-15. By comparing each algorithm's results and applying an ensemble approach, we can create the best model for accurately censoring explicit tweets for adolescents.

## III. Dataset & Data Preprocessing

The dataset we used to train and test our models consisted of 8,200 data points with two possible features: 0 - hate speech / offensive language or 1 - neither. In selecting our dataset, we ensure that the dataset includes a decent amount of data points for each class label. This dataset initially contains 3 classifications: 0 - hate speech, 1 - offensive language, and 2 - neither. Although it is helpful to keep the categories separate in our project, we combine the hate speech and offensive language categories into one, as they both have the same adverse effects on social media. Reducing the features made it easier to identify the text as hateful or not. Next, it is necessary to balance out the data points that each classification contains so our dataset is not skewed. After combining both classes, we reduce each category to 4,100 tweets because class 1 results in the least number of data points, 4,163, resulting our threshold. Next, we clean up the text from each dataset, which initially includes symbols like: !, @, "; the person's Twitter handle, and any other unnecessary symbols and digits that pertains to emoji hashes. For example, if a tweet contains excessive profanity, we don't care about the filler words but focus on the words that negatively impact the tweet overall.

A novel aspect of our approach to solving this problem is using the TF-IDF, term frequency-inverse document frequency technique to vectorize our text data. Whereas similar projects tend to utilize a bag-of-words approach. This approach determines the relevancy of each word instead of the frequency of the said word; the formula is in Figure 1. This new vectorized data is used as input for all models in the following section. After completing these data preprocessing steps, we train our model.

$$TF = (\frac{\text{number of occurance of term in document}}{\text{total number of terms in document}})$$

$$IDF = log(\frac{\text{number of documents in corpus}}{\text{number of documents in corpus that contain term}})$$

$$TF_IDF = TF * IDF$$

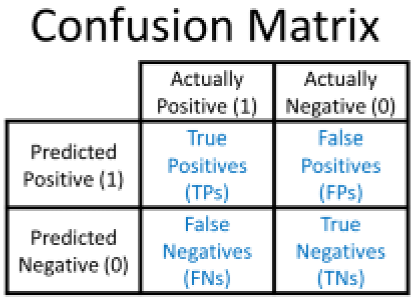
Fig. 1: Formula for TF-IDF vectorizer.

## IV. Methodology

Our approach is to individually utilize three different ML algorithms and a soft-voting classifier combining the three models. We split our processed dataset into an 80% training set and a 20% testing set. We train the vectorized training dataset from each model and test each on the same vectorized testing dataset. We relay our results using a confusion matrix alongside precision, recall, and F1-scores for analysis.

Confusion matrices compute a model's dataset prediction. By looking at the matrix, we can see the strengths and weaknesses of the model. The confusion matrix computes the following predictions: True Positives (TPs) are the number of positive examples classified as positive, and True Negative (TNs) are negative examples the model correctly classifies as negative. False Positives (FPs) are negative examples wrongly classified as positive, and False Negatives (FNs) are positive examples that the model incorrectly guesses as negative. From these measures, we can calculate the ML model's precision, recall, and F1-score with these numbers, which serve as additional performance metrics.

Accuracy differs from precision as accuracy conveys how accurate the machine learning model is correct overall. Precision measures the ability to classify positive samples in



Confusion Matrix

| | Actually Positive (1) | Actually Negative (0) |
|---|---|---|
| Predicted Positive (1) | True Positives (TPs) | False Positives (FPs) |
| Predicted Negative (0) | False Negatives (FNs) | True Negatives (TNs) |

Fig. 2: A confusion matrix (TPs, FPs, FNs, TNs) is represented.

the model. Recall, or True Positive Rate (TPR), measures the percentage of actual positives correctly identified (See figure 3). After calculating the precision and recall scores, we can calculate the F1-score. F1-scores combine the precision and recall value into a single unit to combat uneven class distributions, such as having many True Negatives. With these three scores, we can analyze which model is better for applicational use.

$$Precision = (\frac{\text{True Positive}}{\text{True Positive + False Positive}})$$

$$Recall = (\frac{\text{True Positive}}{\text{True Positive + False Negative}})$$

Fig. 3: The formula for Precision and Recall.

### A. K- Nearest Neighbor

We start with a K-Nearest Neighbor (KNN) model trained on the preprocessed data. The KNN algorithm is a simple choice, as it is simultaneously easy to implement and explain. We utilize a validation set and an elbow plot to determine the most effective k-value. From figure 4, the k-value of 11 would be the most optimal number for clustering. We then train the model on our training dataset utilizing a k-value of 11.

KNN Model Analysis:

- Precision: 0.88
- Recall: 0.86
- F1-score: 0.88

### B. Naive Bayes

The second model that we train is the Naïve Bayes model. The Naïve Bayes algorithm is based on Bayes' Theorem, and the assumption is that the features are independent of each other to determine the probability of a data point belonging to each class. We decide to utilize multinomial Naïve Bayes, widely use for text classification purposes. The formula for calculating the probability of a feature belonging to a class is explained in Figure 6.

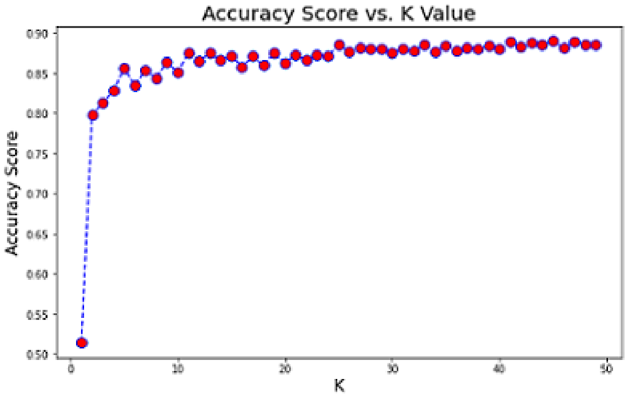


Fig. 4: Elbow plot for K-nearest neighbor.

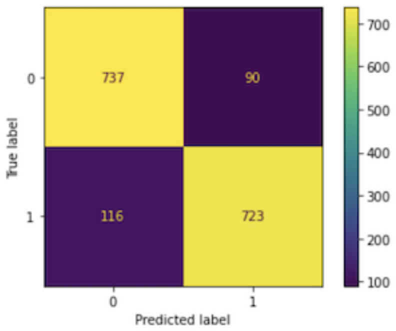


Fig. 5: Confusion Matrix for KNN.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Fig. 6: Formula for Naïve Bayes.

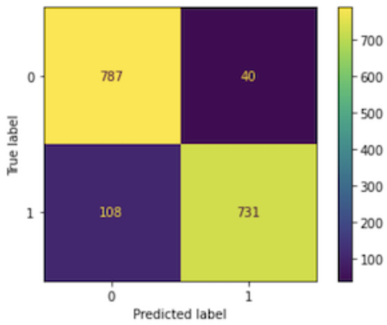


Fig. 7: Confusion Matrix for Naive Bayes.

Naive Bayes Model Analysis:

- Precision: 0.95
- Recall: 0.87
- F1-score: 0.91

### *C. SVM*

Third, we use a support vector machine model (SVM). SVM is a popular non-probabilistic model that is often used in text classification. This algorithm aims to maximize the margin between the support vectors of each class and a hyperplane separating the classes. The model is developed with a linear kernel type.

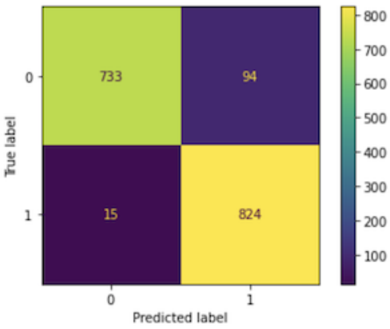


Fig. 8: Confusion Matrix for SVM.

SVM Model Analysis:

- Precision: 0.90
- Recall: 0.98
- F1-Score: 0.94

### *D. Model Ensemble*

We also use an ensemble approach to tackle this problem. We develop a soft-voting classifier utilizing the previous three models as the voters. The concept is that if all three models perform well individually, they should perform even better together. Soft-voting tends to achieve better than hard-voting since it votes on the probability of the predictions of each model. In contrast, hard voting picks the model's prediction with majority votes.

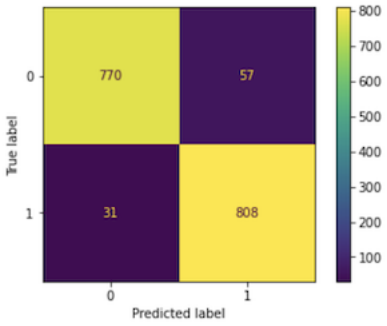


Fig. 9: Confusion Matrix for Soft Voting Ensemble Method.

Model Ensemble Analysis:

- Precision: 0.93
- Recall: 0.96
- F1- Score: 0.95

## V. Results

With our results, we gather information about the different machine-learning techniques performed. While all the models perform well, there are slight variations in different areas depending on the model. The soft-voting classifier ensemble model performs the best overall, with an accuracy score of 0.95, a precision score of 0.93, a recall score of 0.96, and an F1-score of 0.95. While this model did perform the best overall, the SVM model has the highest recall score at 0.98, so it is doing the best at calculating true negatives. The Naive Bayes model has the highest precision score, which means it performs best in calculating the true positives.

## VI. Conclusion and Future Work

Comparing the precision results of all the machine learning techniques we apply in our project, we see that the soft voting classifier ensemble model method yields the best results. With an accuracy of 95 % and a precision of 93 %, we believe that given a tweet as an input in our model, we will correctly identify if that tweet should be classified as hateful and/or offensive or neutral. Knowing that the ensemble technique trains the best model, we would feel confident implementing the model in a real-case scenario if an application like Twitter was looking to modify their platform in a way where their users aged 13-15 are shown more restricted content while the rest are exposed to un-monitored content. Our model would aid that decision of when a tweet should be classified as hateful/offensive or not.

Freedom of speech is valued on social media, where thoughts and opinions can be openly discussed without censorship. Our project has gained more importance with recent changes. Twitter is now under new management and has been evolving to provide ”free speech” to every user. However, the levels of hateful and offensive tweets may increase now that fewer consequences exist. It is important to create a safe environment for tween users aged 13-15 while not restricting older users.

## References

[1] “Advances in machine learning algorithms for hate speech detection in Social Media: A Review,” IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9455353. [Accessed: 14-Nov-2022].

[2] Aljero, Mona Khalifa A., and Nazife Dimililer. ”A Novel Stacked Ensemble for Hate Speech Recognition.” Applied Sciences 11, no. 24 (2021): 11684.

[3] Agarwal S, Sureka A. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology. Springer, 2015. pp. 431–42. https://doi.org/10.1007/978-3-319-14977-6_47. [Accessed: 14-Nov-2022].

[4] Anita Kumari Singh, and Mogalla Shashi. 2019. ”Vectorization of Text Documents for Identifying Unifiable News Articles”. International Journal of Advanced Computer Science and Applications (IJACSA).

[5] Arroyo-Fernández, Ignacio, et al. “Unsupervised Sentence Representations as Word Information Series: Revisiting TF–IDF.” Computer Speech & Language, vol. 56, 2019, pp. 107–29, https://doi.org/10.1016/j.csl.2019.01.005. [Accessed: 14-Nov-2022].

[6] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. ACM Trans. Internet Technol. 20, 2, Article 11 (May 2020), 21 pages. https://doi-org.lib-proxy.fullerton.edu/10.1145/3371276

[7] Mohan, Alkha, and M. Venkatesan. “HybridCNN Based Hyperspectral Image Classification Using Multiscale Spatiospectral Features.” Infrared Physics & Technology, vol. 108, 2020, p. 103326–, https://doi.org/10.1016/j.infrared.2020.103326.

[8] Sheth, A., Shalin, V. L., & Kursuncu, U. (2021, December 2). Defining and detecting toxicity on social media: Context and knowledge are key. Neurocomputing. Retrieved February 28, 2023, from https://www.sciencedirect.com/science/article/abs/pii/S0925231221018087?casa_token=Ek4hqNLRtfUAAAAA%3ACaJtRWb0WP29kZ3_ghbsansFpT3_42w32QDTc_7H01EEK__rvJxEMRS8e3aR93pATYbTNgdMihY

[9] Livingstone, S., &amp; Third, A. (n.d.). Children and young people’s rights in the digital age: An emerging ... Retrieved March 1, 2023, from https://journals.sagepub.com/doi/abs/10.1177/1461444816686318

[10] Chetty, N., & Alathur, S. (2018, May 4). Hate speech review in the context of online social networks. Aggression and Violent Behavior. Retrieved February 28, 2023, from https://www.sciencedirect.com/science/article/pii/S1359178917301064?casa_token=q-eg1xLFdFAAAAAA%3A1Lw00iYRj2nmE2UI5tn2p059doDl9ZEtPpkSszNSrejcADM3vL6vAufvZjNHjrNtDifGnoVaZRo