Detecting Hate Speech in Tweets

Team: Zuzanna Matysiak, Zachary Cahoone, Amy Zhou
Project Mentor TA: Haoyu Wang
Professor: Dinesh Jayaraman
Course: CIS 419, Applied Machine Learning
Github Link: https://github.com/zuziamatysiak/cis419-hate-speech

1) Abstract

For our project, we investigated hate speech in tweets. Hate speech detection and mitigation are extremely important– the uncontrolled spread of hate and negativity in social media has gravely damaged our society, leading to the marginalization of many people and groups. In order to create a more communicative and receptive environment on social media, it is crucial to identify hate speech occurrences and eventually enact policy to make online spaces safer and more welcoming to users. Our primary target contribution to this topic is to contribute code for this problem. Our plan is to find and optimize a model for hate speech detection for tweets. In building our models, we would like to compare different machine learning models and analyze their performances on key statistical metrics. In doing so, we also explore which features are the best to extract to train our models on. Additionally, we explore the realm of tweets considered as hate speech by diving deeper into classifications in objective offensiveness, racism, sexism, hatefulness, and abusiveness. The extent of progress so far is that we have built several models already (i.e. preliminary Linear and Logistic Regressions, followed by more complex Adaboost, K Means, Decision Trees, Random Forest models. We additionally created three different ensemble methods). Perhaps most interestingly, we find that there is significant identifiability of hate speech tweets with the Adaboost and K Means models. Also, tweets that contain a bag of words relating to hate speech are more likely to be flagged as abnormal.

2) Introduction

If we make progress towards solving this problem, we might find good algorithms that perform well in hate speech detection on tweets. If we find that some algorithms are not working well, it is also a valuable insight as it might be a sign for other people in the future to consider not using them for this problem.

Dataset	Class	Precision	Recall	F1
W. & H.	Racism	0.73	0.79	0.76
	Sexism	0.69	0.73	0.71
	Neither	0.88	0.85	0.86
W.	Racism	0.56	0.77	0.65
	Sexism	0.62	0.73	0.67
	R. & S.	0.56	0.62	0.59
	Neither	0.95	0.92	0.94
D. et al.	Hate	0.32	0.53	0.4
	Offensive	0.96	0.88	0.92
	Neither	0.81	0.95	0.87
G. et al.	Harass.	0.41	0.19	0.26
	Non.	0.75	0.9	0.82
F. et al.	Hate	0.33	0.42	0.37
	Abusive	0.87	0.88	0.88
	Spam	0.5	0.7	0.58
	Neither	0.88	0.77	0.82

Table 1: **Classifier performance**

Our project focuses on solving the problem of what is the best hate speech detection algorithm for tweets. Previous work in the research field includes making those predictions based on the logistic regression model. In our project, we decided to use the exact same inputs as they did (data from Davidson, data from Waseem, and data from Founta (we emailed the owner of the data and we were allowed to use it, however, we signed an NDA not to make the data public)— we will refer to those datasets as Davidson, Waseem, and Founta respectively both here and in our code. In order to get the Waseem dataset in a form needed we used Twitter API. Then, we selected words that we

believe could be indicators for hate speech to use as features. We also performed an investigation to see if there are other potentially predictive features to include in our models. As output we will have recall, precision and accuracy scores that we will compare to each other, as well as to the paper. On the right you can see the answers that they got in the paper. In addition to the paper's basic preliminary logistics regression model, we plan to extend upon their research and use additional and more complex machine learning models to compare their performances with those of the paper. We unfortunately could not get access to the other datasets that they used, but we believe that using those 3 datasets will be enough information to make a decision on which models can be preferable to accomplish this task. By developing models that can predict hate speech better, we will elaborate on their usability and accuracy. If some of our models and feature extraction do not perform as well as others, we plan to perform case studies, try other features, and possibly discuss reasons for poorer performance.

3) Background

Currently, there has been insufficient relevant work in the field of detecting hate speech. Perhaps some of the difficulties in the field that limit further research is the complexity of automatic detection of hate speech, perhaps through convoluted syntax in social media, or poorly written text by users. Another challenge is that classifying text lacks consistency, and there is a lack of consensus for what truly constitutes hate speech. Given that people have implicit biases of language and varying tolerances of the presence of hate, this is posed as a challenge when training models. Analyzing linguistics is a dynamic field but particularly interesting and applicable.

After performing extensive research, several papers stood out that we used as our baseline study and inspiration for our project.

- Paper Title: "Racial Bias in Hate Speech and Abusive Language Detection Datasets"
  Link: https://arxiv.org/pdf/1905.12516v1.pdf
  The authors of the baseline paper used logistic regression in order to determine that existing datasets contained bias. For our project and final deliverable, we aim to use more complex and potentially accurate models— including Adaboost and Decision Tree, along with an ensemble of other complex models. Our goal is to have a better model of detecting hate speech in Twitter data and have a better classification technique for future data points.
- Paper Title: "Automated Hate Speech Detection and the Problem of Offensive Language"
  Link: https://arxiv.org/pdf/1703.04009.pdf
  Github: https://github.com/t-davidson/hate-speech-and-offensive-language
  The authors used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. They use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We can use their basic explorations to develop more complex models and examine the results for hate speech classification
- Paper Title: "Automatic Hate Speech Detection using Machine Learning: A Comparative Study"
  Link:https://thesai.org/Downloads/Volume11No8/Paper_61-Automatic_Hate_Speech_Detection.pdf
  The aim of this paper is to compare the performance of three feature engineering techniques and eight machine learning algorithms to evaluate their performance on a publicly available dataset having three distinct classes. We can use the results of this as a baseline study for our project to detect automatic hate speech messages.

Among the above, the most relevant work to our project is "Automatic Hate Speech Detection using Machine Learning: A Comparative Study". By researching this paper's takeaways, it has provided us with a solid foundation to pursue our machine learning models. We have learned more about the process of identifying and classifying hate speech, which we utilized when we developed our own algorithms and models.

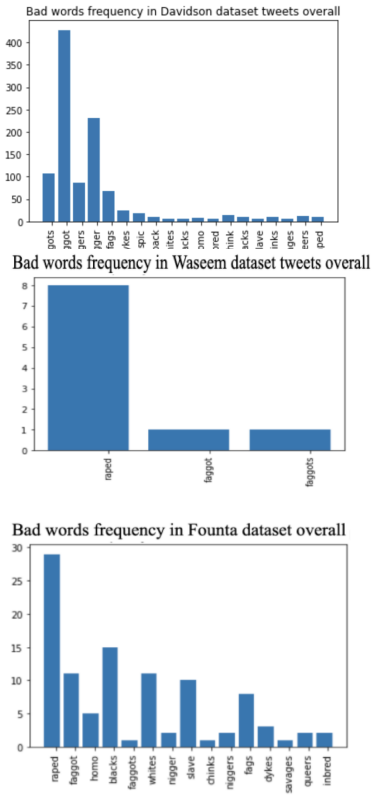
4) Summary of Our Contributions

In this project, our contributions are mainly in the form of code. We found and optimized models for hate speech detection in tweets. We followed a high-level outline of steps as follows:

1. Preprocess tweets data to extract text content and labels
2. Import relevant datasets and libraries to Colab notebook
3. Preprocess data with regex, spacing/punctuation, minimal processing
4. Tokenizing
5. Bag of Words
6. Build and Train Models
7. Running the Models
8. Confusion Matrices & Heatmaps
9. Data Analysis of Results

The notebook to our code, analysis, and contributions can be found in the github here. In the following section, we will go into more detail about each step of our process.

5) Detailed Description of Contributions



5.1 Methods

We want to compare different machine learning models and see which one of them performs best in doing so and what features are based to extract to learn based on. We used multiple tokenization methods: Natural Language Toolkit and for CNNs Torch Text Tokenizer. Both NLT and CNNs Torch Text Tokenizer convert each word into a token but torch tokenizer preprocesses it in such a way that it can be turned into numbers that correspond to indices in the English language dictionary so that CNN models can be run on it.

Unfortunately, we do not have access to all 4 datasets that they used, hence we used the available ones only: data from Davidson, data from Waseem, and we got data from Founta through email but it is confidential. When it came to data processing, we used multiple methods. The first one that we used removed retweets, mentions, punctuation, digits, and unnecessary spaces. The purpose for this method of preprocessing is to get the tweets in order to improve the quality of the tweets and aid in dimensionality reduction. Another method of preprocessing we performed was our way of testing a control dataset. In this model, we used rudimentary techniques to remove unnecessary

spacing and digits, but kept the retweets, mentions, and punctuation. This way, when we run our models on the fully preprocessed dataset and the minimally processed dataset, we can compare the precision, recall, and f-1 scores and determine whether or not our models had a significant impact in predicting hate speech.

Following preprocessing, we also performed tokenization methods on the data using TweetTokenizer(). By breaking down the tweet strings into individuals, it allows us to better apply bags of words and run our models.

Then, we used a bag of words approach to find bad words that can be considered hate speech. We used the same bag of words they used (which ended up performing fairly poorly on the Waseem dataset but very well on the Davidson dataset). We made histograms and saw which words from that bag of words appear most frequently in our datasets (on the right we can see a histogram for the Davidson dataset).

After writing the various methods, we preprocessed the three datasets and the control. After we had our data preprocessed we used the following models to be trained on the data: Decision Tree Classifier, Adaboost Classifier, Logistic Regression, Linear Regression, Random Forest (with gini index and with entropy), K-means, 3 ensemble methods, and 5 CNN models.

After this, we tried using a different set of features to see if we could improve on our original result. Hatebase.org has a robust dataset of hateful terms that have been encountered on the internet. We constructed a set of 528 of these terms and attempted to use them as features. The hatebase API was recently taken offline, so this was done manually. Following the construction of this new "bag of words", we repeated our initial experiment, all else held equal save for the new features. We ran all of the same models as previously. The results of both of these experiments are in the appendix.

5.2 Experiments and Results

Please refer to the Appendix for our precision, recall, and F1-score values.

We used the following datasets: data from Davidson, data from Waseem, and we got data from Founta through email.

Performance metrics: precision, f-1 score, recall, we will compare it to the following metrics that [A] accomplished (Figure 1) and try to get better results than they did. We ran different machine learning algorithms, used different preprocessing methods and bags of words to see which methods perform best for this task.

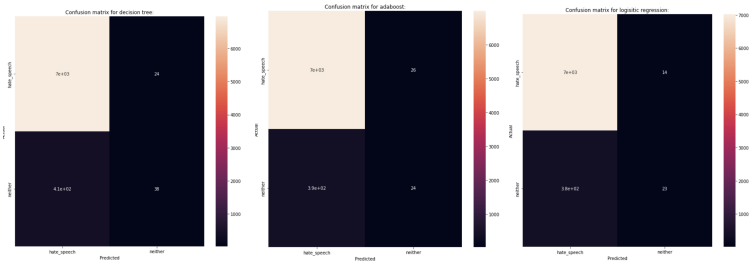
Dataset	Class	Precision	Recall	F1
W. & H.	Racism	0.73	0.79	0.76
	Sexism	0.69	0.73	0.71
	Neither	0.88	0.85	0.86
W.	Racism	0.56	0.77	0.65
	Sexism	0.62	0.73	0.67
	R. & S.	0.56	0.62	0.59
	Neither	0.95	0.92	0.94
D. et al.	Hate	0.32	0.53	0.4
	Offensive	0.96	0.88	0.92
	Neither	0.81	0.95	0.87
G. et al.	Harass.	0.41	0.19	0.26
	Non.	0.75	0.9	0.82
F. et al.	Hate	0.33	0.42	0.37
	Abusive	0.87	0.88	0.88
	Spam	0.5	0.7	0.58
	Neither	0.88	0.77	0.82

For the preprocessing method 1 and the models mentioned above we got the following results:
For our decision tree classifier, we used a max depth of 10 and min samples leaf to be 5. Our predictions for hate speech were generally higher than offensive, racism, and abusive. We also trained an adaboost classifier. Comparatively, adaboost performed better than decision tree, which makes sense because the purpose of an adaboost classifier is to boost the performance of a machine learning algorithm. The third

model we used was logistic regression. For our linear regression model, we used MSE, R2, and RMSE to assess our results. Next, we used two different random forest models. We created a Random Forest classifier with the gini index, and we also created a Random Forest classifier with entropy. Another model we used was K-means and trained each dataset on two different values of k. Then, we explored 3 different ensemble methods. For the first ensemble method, we used the Stacking Classifier. Our first layer of estimators consisted of a random forest classifier of 10 estimators and k neighbors classifier. The second layer consisted of a decision tree classifier and random forest classifier of 50 estimators. For the second ensemble method, we used the Voting Classifier. Our first layer of estimators consisted of a random forest classifier of 25 estimators and adaboost classifier. The second layer consisted of an adaboost classifier and k neighbors classifier of 5 neighbors. For the third ensemble method, we used the Bagging Classifier. Our first layer of estimators consisted of a random forest classifier of 100 estimators and a decision tree classifier. The second layer consisted of an adaboost classifier and k neighbors classifier of 5 neighbors. The reason why we chose the above combinations of ensemble models is because we wanted to combine best performing models with each other or not so well performing individually with well performing ones and see if the performance becomes better or worse (the results can be seen below).

For all CNNs however we got significantly worse accuracy results than for other models mentioned above. For Davidson: we got 0.773638 for CNNModel1, 0.773638 for CNNModel2, 0.783053 for CNNModel3, 0.773638 for CNNModel4, and 0.77 for CNNModel5. For Waseem the numbers were respectively: 0.692588, 0.003645, 0.184690, 0.303767, 0.003645. Meanwhile the above models had accuracy (not included in the data above but we also calculated that metric) in the 90s so we decided not to include CNNs in this comparison as they performed significantly worse than other algorithms that we had hence would not be concluded to be the best classifiers for this task. Additionally, another reason that Waseem may have had lower performance is that Twitter also detects and sensors many tweets labeled as abusive or hate speech, which may have led to limitations in our data.Twitter API blocks the tweets that have been marked as hate speech from being scraped. To add to that, a lot of the tweets from the original dataset have been removed either by Twitter or by the users themselves which led to way smaller sample size than they had in the original paper (smaller dataset resulted in worse predictions).

After running the models on the datasets, we compared which ones performed the best in classifying and predicting the speech. It seems that Adaboost and the K Means algorithms performed on average better across the datasets. Decision Tree performed poorer than the other models. The ensemble models performed well, with higher F1-scores. In regards to the datasets, the Waseem sexism/racism dataset performed the worst, likely due to the fact that there were only 2K data entries, as compared to the Founta dataset that had upwards of hundreds of thousands of data entries. With more data, our models have better accuracy, of course with the added natural noise of the data.



6) Compute/Other Resources Used

We used Twitter API to extract the tweet text from the tweet ID for the Founta dataset. It ended up being quite a challenging task because of access authorization issues with Twitter. We did not require any AWS resources to complete our project. All of the tasks we wanted to accomplish were able to be performed using Google Colab.

7) Conclusions

At the outset of this project, our goal was to explore whether more complex models than the ones employed in the baseline paper could more accurately predict hate speech. In our pursuit of this goal, we experienced some successes and some failures. Some of the models we trained achieved better results than those trained in the baseline paper. Details of all of our models' precision, recall, and F-1 scores can be found below.

In spite of these successes, we also learned that hate speech recognition is a much more complicated problem than choosing an optimal model with the right parameters. Feature selection is extremely important, and we experimented with multiple "bags of words", or terms that we thought might predict hate speech. Additionally, we discovered that finding a quality dataset to use for training models for this problem is non-trivial. Most existing datasets classify hate speech in one of two ways: 1) asking a group of people whether speech is hateful and classifying using the majority vote, or 2) asking a linguistics expert (or experts) to classify the speech. These methods, however, can contain bias, and one of our baseline papers even noted that some hate speech datasets themselves contain racial bias.

In future work, each of these two facets of this problem should be explored. First, how can we find the right features to predict hate speech? While the sets of features we chose enabled us to reach a good result, could we do better? Second, and perhaps more importantly, how do we ensure that the models themselves are not biased based on underlying biases in the training data? The construction of a new dataset for this problem was something we considered at the beginning of the project, and while it is not what we pursued, one of our biggest takeaways is that this is extremely important and should be explored in future research. In summary, we learned a great deal from this project. We were able to improve on some of the results in our baseline paper, which we consider a success. However, we also learned that the problem of hate speech classification is far more complex than our research or most pre-existing, and we believe that the problem is ripe for exploration.

(Exempted from page limit) Other Prior Work / References (apart from Sec 3) that are cited in the text:

- Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. 11th International Conference on Web and Social Media, ICWSM 2018.
- Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL Student Research Workshop.
- Davidson, T., Warmsley, D., Macy, M., & Weber, I. (n.d.). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the 11th International AAAI Conference on Web and Social Media.

Appendix/Supplementary Material

Precision, Recall, F1-Scores for Davidson and Waseem Data:

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.9457069	0.5116291	0.94738989	0.62025316	0.946511	0.52173913	0.9963548	0	0.9927096	0	0.99878493	0
Recall	0.99399314	0.0993228	0.99571367	0.11238532	0.99528908	0.08372093	1	0	1	0	1	0
F-1 score	0.96924901	0.16635161	0.97095089	0.19029126	0.97028738	0.14428858	0.99817407	0	0.99634146	0	0.9993921	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Adaboost		Adaboost		Adaboost		Adaboost		Adaboost		Adaboost	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.95007482	0.52380952	0.94663951	0.7	0.94922617	0.49275362	0.99878493	0	0.99878493	0	0.99392467	0
Recall	0.99430524	0.10705596	0.996997	0.11085973	0.99501921	0.08333333	1	0	1	0	1	0
F-1 score	0.97168696	0.17777778	0.9711659	0.19140625	0.97158341	0.14255765	0.9993921	0	0.9993921	0	0.99695308	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.94263399	0.66666667	0.94274912	0.62068966	0.94492441	0.48148148	0.9963548	0	0.9927096	0	0.99756987	0
Recall	0.99957161	0.01388889	0.998427	0.04072398	0.99800399	0.03087886	1	0	1	0	1	0
F-1 score	0.97026821	0.02721088	0.96978957	0.07643312	0.97073915	0.05803571	0.99817407	0	0.99634146	0	0.99878345	0

	Davidson	Davidson Hate Speech	Davidson Offensive Language	Waseem Dataset	Waseem Sexism	Waseem Racism
	Linear Regression	Linear Regression	Linear Regression	Linear Regression	Linear Regression	Linear Regression
MSE	0.04539969064	0.05000356293	0.04597649991	2.34E+22	0.004838070777	0.00242930977
R2	0.1205821785	0.1278535742	0.133496954	-4.84E+24	-0.000294762282	-0.00209618801 1
RMSE	0.2130720316	0.2236147646	0.214421314	153068317813	0.06955624183	0.04928802867

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.94918478	0.56	0.94947025	0.5890411	0.94475363	0.52941176	0.99756987	0	0.9927096	0	0.9963548	0
Recall	0.99529848	0.10096154	0.9957265	0.10361446	0.99542334	0.08126411	1	0	1	0	1	0
F-1 score	0.97169483	0.17107943	0.97204839	0.17622951	0.96942684	0.1409002	0.99878345	0	0.99634146	0	0.99817407	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.94848733	0.53608247	0.94976239	0.52857143	0.9486446	0.55319149	0.99513973	0	0.99392467	0	0.9963548	0
Recall	0.99357602	0.12093023	0.9953045	0.09090909	0.99400514	0.12121212	1	0	1	0	1	0
F-1 score	0.97050826	0.19734345	0.97200028	0.15513627	0.97079529	0.19885277	0.99756395	0	0.99695308	0	0.99817407	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	K means k=3		K Means k=3		K Means k=3		K Means K=3		K Means k=3		K Means k=3	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.95103057	0.43478261	0.95396738	0.44871795	0.95060025	0.46835443	0.99756691	0	0.99635036	0	0.99270073	0
Recall	0.98340724	0.06756757	0.98402966	0.08293839	0.98526888	0.08352144	0.99878197	0	0.99878049	0	0.99877601	0
F-1 score	0.96694796	0.11695906	0.96876535	0.14	0.96762413	0.14176245	0.99817407	0	0.99756395	0	0.9957291	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	K means k=5		K means k=5		K means k=5		K Means k=2		K means k=2		K means k=2	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.95530493	0.58064516	0.95626822	0.37606838	0.9607461	0.44827586	0.996337	0	0.99878197	0	0.99392467	0
Recall	0.98238832	0.07982262	0.98203593	0.10451306	0.98207171	0.06388206	0.99512195	0	0.99756691	0	1	0
F-1 score	0.96865735	0.14035088	0.9689808	0.16356877	0.97129187	0.11182796	0.9957291	0	0.99817407	0	0.99695308	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.94534852	0.57377049	0.94605921	0.60526316	0.94403893	0.51351351	0.99513973	0	0.99513973	0	0.99756987	0
Recall	0.99628412	0.07990868	0.99786112	0.05450237	0.99742931	0.04387991	1	0	1	0	1	0
F-1 score	0.97014822	0.14028056	0.97126995	0.1	0.97	0.08085106	0.99756395	0	0.99756395	0	0.99695308	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.94799022	0.57746479	0.95151264	0.47959184	0.95267007	0.48404255	0.9963548	0	0.99513973	0	0.99756987	0
Recall	0.99572101	0.09669811	0.98540773	0.21123596	0.98614484	0.20967742	1	0	1	0	1	0
F-1 score	0.97126957	0.16565657	0.96816361	0.29329173	0.96911847	0.2926045	0.99817407	0	0.99756395	0	0.99878345	0

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism	
	Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3		Decision Tree	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither
Precision	0.55172414	0.0461369	0.95206702	0.41558442	0.55080214	0.04925497	0.9963548	0	0.99513973	0	0.9963548	0
Recall	0.01367132	0.81113801	0.98718314	0.15496368	0.01472691	0.80952381	1	0	1	0	1	0
F-1 score	0.02668149	0.08730779	0.96930714	0.22574956	0.02868681	0.09285993	0.99817407	0	0.99756395	0	0.99817407	0

Precision, Recall, F1-Scores of Founta and Founta Control Data:

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.80792277	0.7	0.8101072	0.81818182	0.80908002	1	0.77980432	0.83333333	0.78367455	0.76923077	0.78109876	0.6
Recall	0.99975284	0.00242047	0.99983565	0.00314575	1	0.00278164	0.99986364	0.00240906	0.99979647	0.00245218	0.99959164	0.00218182
F-1 score	0.89365933	0.00482426	0.89502722	0.00626741	0.89446571	0.00554785	0.8762286	0.00480423	0.87864063	0.00488878	0.87694053	0.00434783

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Adaboost		Adaboost		Adaboost		Adaboost		Adaboost		Adaboost	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.80907032	0.57142857	0.80415529	0.46153846	0.81158841	0.6	0.78724875	0.83333333	0.779716	0.53333333	0.78618526	0.75
Recall	0.99950638	0.00278261	0.99942067	0.00203597	0.99950787	0.00317125	0.99986493	0.00249314	0.99952277	0.00192771	0.99979713	0.00223325
F-1 score	0.89426226	0.00553825	0.89121771	0.00405405	0.89579888	0.00630915	0.88090917	0.00497141	0.87604195	0.00384154	0.88021671	0.00445324

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.80796113	0.71428571	0.81024036	0.54545455	0.80299501	0.6	0.77993407	0.8	0.78021686	0.5	0.78012336	0.66666667
Recall	0.99983526	0.0017301	0.99958929	0.00210084	0.99983426	0.00101249	0.99986368	0.00192911	1.00E+00	4.83E-04	0.99972743	0.00193097
F-1 score	0.89371572	0.00345185	0.89500975	0.00418556	0.89066883	0.00202156	0.87631052	0.00384893	0.87648903	0.00096595	0.87637765	0.00385078

	Founta Dataset	Founta Abusive	Founta Hateful	Founta Control Dataset	Founta Control Abusive	Founta Control Hateful
	Linear Regression	Linear Regression	Linear Regression	Founta Hateful	Linear Regression	Linear Regression
MSE	6.70E+24	1.92E+22	7.79E+24	1.71E-01	6.98E+23	1.75E+23
R2	-4.43E+25	-1.23E+23	-5.01E+25	7.10E-05	-4.14E+24	-1.03E+24
RMSE	258777259740	138504094215	279028847639	0.4136931472	835688268695	417820871754

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest gini		Random Forest gini		Random Forest gini	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.81065601	0.8	0.80472703	0.6	0.80962845	0.75	0.78490907	0.58333333	0.78282828	0.375	0.77917176	0.77917176
Recall	0.99975359	0.00420315	0.99966918	0.00204151	0.99975333	0.00313808	0.99966138	0.00172754	1.00E+00	7.34E-04	0.99993178	0.00144231
F-1 score	0.89532917	0.00836237	0.89166759	0.00406918	0.89470199	0.00625	0.87936376	0.00344488	0.87805605	0.00146484	0.87585527	0.00287977

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.81046883	0.57142857	0.80789456	0.57142857	0.81017378	0.81818182	0.78478723	0.44444444	0.7859916	0.53333333	0.78370906	0.4
Recall	0.99950723	0.00280308	0.99975288	0.00138408	0.99983566	0.00314685	0.99932268	0.00197336	0.99952658	0.00198413	1.00E+00	9.82E-04
F-1 score	0.89511621	0.0055788	0.89364209	0.00276148	0.89506786	0.00626959	0.87915624	0.00392927	0.87999047	0.00395355	0.87858378	0.00195982

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	K Means k=3		K Means k=3		K Means k=3		K Means k=3		K means k=3		K means k=3	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.8017178	0	0.80279627	0.75	0.80842988	0.57142857	0.78493937	0.75	0.78499654	0.85714286	0.78146435	0.77777778
Recall	0.99958492	0	0.99966838	0.00202156	0.99934151	0.00138841	0.99952597	0.00148112	0.99979686	0.00148148	0.99972791	0.00170027
F-1 score	0.88978385	0	0.89048076	0.00403226	0.89380498	0.00277008	0.87933037	0.00295639	0.87947107	0.00295785	0.87722335	0.00339312

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	K means k=4		K means k=4		K means k=4		K meanns k=4		K means k=4		K means k=4	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.808335	1	0.80908304	0.71428571	0.81013586	0.25	0.78382689	0	0.78658018	0	0.78484043	0.88888889
Recall	0.99983531	0.002079	0.999342	0.00174095	9.99E-01	3.50E-04	0.99972876	0	0.99972969	0	1.00E+00	1.97E-03
F-1 score	0.89394441	0.00414938	0.89420423	0.00347343	8.95E-01	7.00E-04	0.87871022	0	0.88043802	0	8.79E-01	3.94E-03

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.80632069	0	0.80845539	0.5	0.80605456	0	0.78286747	0	0.78622303	0.25	0.78244234	0
Recall	1	0	0.99958841	0.00173491	1	0	1	0	1.00E+00	2.49E-04	1	0
F-1 score	0.8927769	0	0.89391932	0.00345781	0.89261374	0	0.87821162	0	8.80E-01	4.97E-04	0.87794407	0

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.6875	0.19288664	0.80949211	0.71428571	0.81459776	0.42857143	0.78443783	0.5625	0.71428571	0.21960555	0.78067843	0.6
Recall	9.07E-04	9.98E-01	0.99983557	0.00174398	0.99934641	0.00215054	1.00E+00	2.22E-03	3.40E-04	1.00E+00	0.99972765	0.00145243
F-1 score	0.00181145	0.32330449	0.89465166	0.00347947	0.89756384	0.0042796	0.87901544	0.00441393	0.00068064	0.36009414	0.87672787	0.00289785

	Founta Dataset		Founta Abusive		Founta Hateful		Founta Control Dataset		Founta Control Abusive		Founta Control Hateful	
	Ensemble 3		Ensembel 3		Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3	
	Hateful	Abusive	Abusive	Neither	Hateful	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.7	0.19141145	0.80651175	0.81818182	0.81065246	0.8	0.52941176	0.21493538	0.78608659	0.6875	0.7806136	0.63636364
Recall	5.76E-04	9.99E-01	0.99983492	0.00308748	0.99983577	0.00280505	6.09E-04	9.98E-01	0.99966182	0.0027275	0.99972761	0.00169368
F-1 score	0.00115113	0.32126495	0.89282819	0.00615174	0.89535995	0.0055905	0.00121737	0.35369803	0.88010242	0.00543344	0.87668697	0.00337838

Accuracy Scores of Davidon, Waseem, and Founta Datasets using the 5 CNN Models:

	Davidson	Waseem	Founta
	CNN Model 1	CNN Model 1	CNN Model 1
Accuracy	0.773638	0.692588	0.83726

	Davidson	Waseem	Founta
	CNN Model 2	CNN Model 2	CNN Model 2
Accuracy	0.773638	0.003645	0.83726

	Davidson	Waseem	Founta
	CNN Model 3	CNN Model 3	CNN Model 3
Accuracy	0.783053	0.18469	0.730254

	Davidson	Waseem	Founta
	CNN Model 4	CNN Model 4	CNN Model 4
Accuracy	0.773638	0.303767	0.83726

	Davidson	Waseem	Founta
	CNN Model 5	CNN Model 5	CNN Model 5
Accuracy	0.773638	0.003645	0.83726

Link to github (contains ReadMe, data sources, and colab notebooks):
https://github.com/zuziamatysiak/cis419-hate-speech

Hatebase Precision, Recall, and F-1 scores:

	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree		Decision Tree	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94570197	0.46153846	0.93847814	0.47826087	0.94160486	0.55	0.99878049	1	1	1	1	0.75	0.8252679	0.85488127	0.82462967	0.86089239	0.8215309	0.87634409
Recall	0.9990037	0.01466993	0.99827784	0.0235546	0.99871263	0.02477477	1	0.75	1	1	0.99878049	1	0.99547176	0.11234397	0.99563175	0.11322057	0.99619457	0.11080897
F-1 score	0.97162237	0.02843602	0.9674548	0.04489796	0.96931834	0.04741379	0.99938987	0.85714286	1	1	0.99938987	0.85714286	0.90241445	0.19859025	0.90209842	0.20012203	0.9004711	0.1967411
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Adaboost		Adaboost		Adaboost		Adaboost		Adaboost		Adaboost		Adaboost		Adaboost		Adaboost	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94090539	0.43859649	0.94674075	0.42857143	0.94485842	0.38888889	1	1	1	1	1	0.66666667	0.82545753	0.84715026	0.8265892	0.8625	0.82769924	0.83204134
Recall	0.99541153	0.05422993	0.9954403	0.05755396	0.99529042	0.04906542	1	1	1	1	0.99878197	1	0.99514283	0.11342352	0.99547251	0.11970854	0.99466557	0.11318102
F-1 score	0.9673913	0.0965251	0.97047996	0.10147992	0.96941896	0.08713693	1	1	1	1	0.99939062	0.8	0.90239259	0.20006118	0.90320412	0.21023766	0.90353362	0.19925743
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression		Logistic Regression	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94547908	0.16	0.94502971	0.4516129	0.9451203	0.45945946	1	1	1	1	1	1	0.81914458	0.84517766	0.8255148	0.84615385	0.82143101	0.84130982
Recall	0.99701153	0.00980392	0.99757628	0.03325416	0.99714775	0.04018913	1	1	1	1	1	1	0.99493776	0.11174497	0.99539587	0.10742937	0.99478606	0.1133356
F-1 score	0.97056175	0.01847575	0.97059232	0.0619469	0.9704372	0.07391304	1	1	1	1	1	1	0.89852357	0.19739182	0.90253084	0.19065305	0.8998353	0.19976077
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Linear Regression		Linear Regression		Linear Regression		Linear Regression		Linear Regression		Linear Regression		Linear Regression		Linear Regression		Linear Regression	
MSE	3.00E+18		35542408.94		3.02E+16		2.43E-03		0.0047677578		0.0035916471		1.45E-01		1.44E-01		1.45E-01	
R2	-5.43E+19		-637007560.7		-5.18E+17		-1.77E-03		0.0142427868		0.0110866869		7.31E-02		7.48E-02		6.25E-02	
RMSE	1730723274		5961.745461		173837541.2		0.0492800563		0.0690489522		0.0599303522		0.3803626718		0.3800071918		0.3804173009	
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini		Random Forest Gini	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94608438	0.75	0.94547908	0.56	0.94448197	0.40625	1	0.8	1	1	1	1	0.82635468	0.81400966	0.82491283	0.8337469	0.8169591	0.85117493
Recall	0.99943044	0.02912621	0.99843238	0.03349282	0.99728997	0.03066038	0.998779	1	1	1	1	1	0.99366516	0.11721739	0.99447787	0.11598205	0.99525909	0.1084137
F-1 score	0.97202604	0.05607477	0.97123449	0.06320542	0.97016789	0.05701754	0.99938913	0.88888889	1	1	1	1	0.90231967	0.20492551	0.90179372	0.20363636	0.89733783	0.19233038
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy		Random Forest entropy	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94640929	0.25925926	0.94470668	0.35	0.9463441	0.41666667	1	1	1	1	1	1	0.82649368	0.85714286	0.82622368	0.85074627	0.82821582	0.80589681
Recall	0.99715545	0.01732673	0.99814762	0.01678657	0.99700982	0.03640777	1	1	1	1	1	1	0.9954766	0.11494253	0.9950601	0.1185853	0.99351928	0.11549296
F-1 score	0.97111988	0.0324826	0.97069216	0.03203661	0.9710165	0.06696429	1	1	1	1	1	1	0.90314878	0.2027027	0.90281616	0.20815581	0.90336777	0.20203265
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	K means k=3		K Means k=3		K Means k=3		K Means K=3		K Means k=3		K Means k=3		K Means k=3		K Means k=3		K Means k=3	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.97272727	0.01813785	0.97551546	0.07411737	0.92008437	0.01385681	1	0.01992032	1	0.01509434			0.81922867	0.21929825	0.81798335	0.87937743	0.81712884	0.86486486
Recall	0.32159725	0.03348214	0.32240204	0.81074169	0.55751207	0.03053435	0.3594132	1	0.33211233	1			0.99044402	0.00864753	0.98951886	0.07758325	0.99091134	0.07652887
F-1 score	0.48338174	0.02352941	0.48463508	0.13581834	0.69431426	0.01906275	0.52877698	0.0390625	0.49862511	0.02973978			0.8967369	0.01663894	0.89561158	0.14258675	0.89566841	0.14061519
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	K means k=5		K means k=5		K means k=5		K Means k=2		K means k=2		K means k=2		K means k=4		K means k=4		K means k=4	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.98119658	0.02266895	0.98058252	0.0207433	0.98233216	0.01809409	0.98820755	0	1	0.00779221	1	0.01005025	0.8241299	0.86259542	0.82030338	0.86037736	0.8224145	0.3157894
Recall	0.08178968	0.12709832	0.08628791	0.11650485	0.07962194	0.03318584	0.51222494	0	0.53414634	1	0.51892552	1	0.98687341	0.07954945	0.98483725	0.07875648	9.85E-01	1.25E-02
F-1 score	0.15099303	0.0384755	0.15861798	0.03521643	0.14730428	0.0234192	0.6747182	0	0.6963434	0.01546392	0.68327974	0.0199005	0.89818929	0.14566549	0.8950719	0.1443038	8.96E-01	2.41E-02
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1		Ensemble 1	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.9460659	0	0.94135844	0	0.94387618	0.6	1	0.75	1	1	0.99392467	0	0.82484642	0.83421053	0.82592996	0.84659091	0.81984566	0.84496124
Recall	1	0	1	0	0.9997149	0.00714286	0.99878049	1	1	1	1	0	0.99481353	0.10995491	0.99556541	0.10445145	0.99502694	0.11028668
F-1 score	0.97228558	0	0.96979354	0	0.97099342	0.01411765	0.99938987	0.85714286	1	1	0.99695308	0	0.901892	0.19429972	0.90284863	0.18595944	0.89898158	0.1951074
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2		Ensemble 2	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94132757	0.5	0.95488197	0.44565217	0.94870748	0.32941176	1	1	1	1	1	1	0.82028689	0.85384615	0.69395018	0.18597871	0.82588429	0.81447964
Recall	0.99727872	0.0419426	0.97810846	0.27578475	0.99189189	0.0691358	1	1	1	1	1	1	9.95E-01	1.12E-01	0.0159823	0.96960057	0.9932399	0.12413793
F-1 score	0.96849572	0.07739308	0.96635567	0.34072022	0.96981919	0.11428571	1	1	1	1	1	1	0.89934846	0.19856887	0.03124499	0.31209466	0.90186391	0.21543986
	Davidson		Davidson Hate Speech		Davidson Offensive Language		Waseem Dataset		Waseem Sexism		Waseem Racism		Founta Dataset		Founta Abusive		Founta Hateful	
	Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3		Ensemble 3		Ensembel 3		Ensemble 3	
	Hate	Offensive	Hate	Neither	Offensive	Neither	Racism	Sexism	Sexism	Neither	Racism	Neither	Hateful	Abusive	Abusive	Neither	Hateful	Neither
Precision	0.94724646	0.44444444	0.94640682	0.19565217	0.94107044	0.13901345	1	1	1	1	0.99877751	0.8	0.82848162	0.82323232	0.73517787	0.18894228	0.83006581	0.81447964
Recall	0.99214734	0.10208817	0.99473684	0.02222222	0.9724889	0.06798246	1	1	1	1	0.99877751	0.8	9.94E-01	1.15E-01	0.01528223	0.97656523	0.99327373	0.12680521
F-1 score	0.96917713	0.16603774	0.96997018	0.03991131	0.95652174	0.09131075	1	1	1	1	0.99877751	0.8	0.90383182	0.20173267	0.02994205	0.31662509	0.90436536	0.21944529

Broader Dissemination Information:

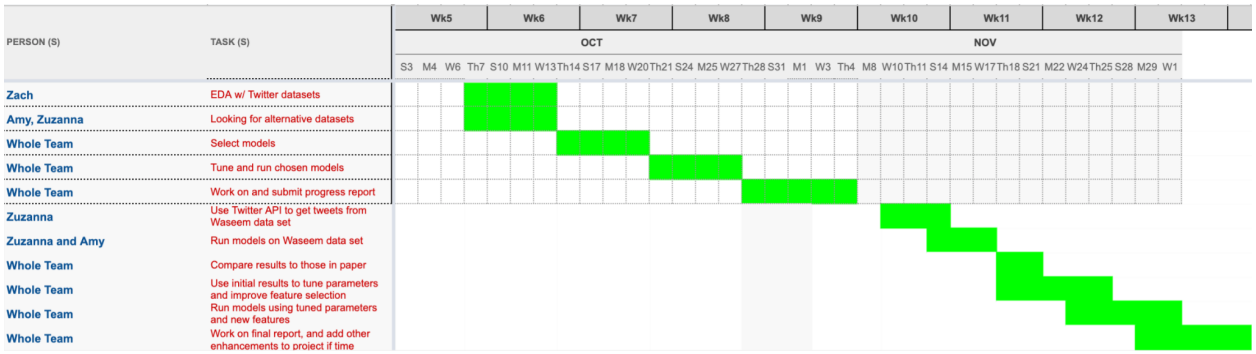
Your report title and the list of team members will be published on the class website. Would you also like your pdf report to be published?
YES / NO

If your answer to the above question is yes, are there any other links to github / youtube / blog post / project website that you would like to publish alongside the report? If so, list them here.

- <description in a few words>: <URL>
- <description in a few words>: <URL>

  ...

(Exempted from page limit) **Work Report: This may look like your GANTT chart from the midway report, with more completed steps now. Okay to modify.** (Mark completed steps in green, as shown here. For convenience, you may split into two charts, one till Nov 8, and another for after Nov 8, placed one



(Exempted from page limit) Attach your midway report here, as a series of screenshots from Gradescope, starting with a screenshot of your main evaluation tab, and then screenshots of each page, including pdf comments. This is similar to how you were required to attach screenshots of the proposal in your midway report.

QUESTION 1

Does the report follow the provided template including the 4-page limit (excluding exempted portions), with reasonable responses to all questions? **1** / 1 pt

QUESTION 2

Has feedback from the last round been effectively addressed? **2** / 2 pts

QUESTION 3

Has the team identified a clear topic and viable new target contribution, as per the project specifications provided in class? **1** / 1 pt

QUESTION 4

Has the team moved in a non-trivial way towards their target contribution? **1** / 1 pt

QUESTION 5

Has a clear and systematic work plan been formulated for the remaining weeks? **2** / 2 pts

Detecting Hate Speech in Tweets

Team: Zuzanna Matysiak, Zachary Cahoone, Amy Zhou
Project Mentor TA: Benedict Arockiaraj
Link to Colab:
https://colab.research.google.com/drive/1s_BqEavAhTcVvLULjGPadNvwyA0x0Cl7#scrollTo=GZLCLUKJ5s8t

1) Introduction

Set up of the problem:
For our project, we will be investigating hate speech in tweets. We want to compare different machine learning models and see which one of them performs best in doing so and what features are based to extract to learn based on. We plan to use Natural Language Toolkit to tokenize Twitter data and extract features. The data we will be using will be similar to the data used in the paper that was the impetus for our project. Unfortunately, we do not have access to all 4 datasets that they used, hence we will use the available ones only: data from Davidson, data from Waseem, and we will send an email to Founta to ask for his data. Then, we will use select words that we believe could be indicators for hate speech to use as features. Initially, we will use the same bag of words they used. We will also do an investigation to see if there are other potentially predictive features we can include. We plan to use various machine learning models and compare their performance to the paper which only used logistic regression. We will evaluate our results by comparing to the f-1 score, precision and recall they got in our paper. If we can find models that can predict hate speech better, it is a good sign and we should discuss why that happened. If some models and feature extraction do not perform so well, we should perform case studies, try other features, and possibly discuss what are some reasons it went wrong.

Possible add-ons to our project depending on the progress we have: using Twitter AAE model (similarly to what they did in the second part of the paper) to see if there is racial bias in classification of our models. This model takes in tweets and tells the probability they were written by black vs. white people based on the language used. We are a little skeptical about using this, as it adds more data inferred by models, but it is used in the cited paper.

Motivation
If we make progress towards solving this problem, we might find good algorithms that perform well in hate speech detection on tweets. If we find that some algorithms are not working well, it is also a valuable insight as it might be a sign for other people in the future to consider not using them for this problem.

2) How We Have Addressed Feedback From the Proposal Evaluations

In this phase of the project, we addressed feedback given to our proposal and 1-1 meetings with the TA. First, we narrowed our topic and plan to focus on detecting hate speech (rather than hate speech + racial bias) in tweets. We chose to go along this route because there would be more clarity and direction in identifying hate comments rather than implicit bias. We also received feedback from our proposal to clarify our data sources. To address this, we did some research and found interesting data from Davidson,

Waseem, Founta, and Hatebase. We will use these datasets to train and test our model and make progress toward this project. We liked Davidson's dataset because he had collected user judgements from CrowdFlower to label whether or not a tweet contained hate speech or offensive language. We want to use Waseem's dataset because it specifically identifies hate speech on the basis of cyberbullying, abusive language, online harassment, etc in online social media platforms. Hatebase contains more of a generic bag of words that we can use to detect their presence in tweets. Next, the metrics that we want to examine include precision, recall, F1, etc. We are looking for higher precision, recall, and F1 scores from the models that we build. Previous models from our literature review are simplistic, i.e. logistic regression. To build on this project, we want to train more complex models than existing that will better classify the tweets data. We are comparing the performances of decision trees, linear regression, adaboost, and cnns beyond the basic models that already exist. After continuing our work on this project, we have a much better sense of direction for the final deliverable.

3) Prior Work We are Closely Building From

The authors of the baseline paper used logistic regression in order to determine that existing datasets contained bias. For our project and final deliverable, we aim to use more complex and potentially accurate models— including Adaboost and Decision Tree, along with an ensemble of other complex models. Our goal is to have a better model of detecting hate speech in Twitter data and have a better classification technique for future data points.

- Paper Title: "Automated Hate Speech Detection and the Problem of Offensive Language"
  Link: https://arxiv.org/pdf/1703.04009.pdf
  Github: https://github.com/t-davidson/hate-speech-and-offensive-language
  The authors used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. They use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We can use their basic explorations to develop more complex models and examine the results for hate speech classification
- Paper Title: "Automatic Hate Speech Detection using Machine Learning: A Comparative Study"
  Link:https://thesai.org/Downloads/Volume11No8/Paper_61-Automatic_Hate_Speech_Detection.pdf
  The aim of this paper is to compare the performance of three feature engineering techniques and eight machine learning algorithms to evaluate their performance on a publicly available dataset having three distinct classes. We can use the results of this as a baseline study for our project to detect automatic hate speech messages.

4) What We are Contributing

1. **Contribution(s) in Code:** Finding and optimizing a model for hate speech detection for tweets. Our code progress can be found in this Colab.
2. **Contribution(s) in Application: N/A**
3. **Contribution(s) in Data: N/A**
4. **Contribution(s) in Algorithm: N/A**
5. **Contribution(s) in Analysis / Technique: N/A**

5) Detailed Description of Each Proposed Contribution, Progress Towards It, and Any Difficulties Encountered So Far

5.1 Methods

We want to compare different machine learning models and see which one of them performs best in doing so and what features are based to extract to learn based on. In terms of features extraction and tokenizing the tweets, we plan to use Natural Language Toolkit and if we do not get significantly better results BERT on AWS. The data we will be using will be similar to the data used in the paper. Unfortunately, we do not have access to all 4 datasets that they used, hence we will use the available ones only: data from Davidson, data from Waseem, and we will send an email to Founta to ask for his data. Then, we will use a bag of words approach to find bad words that can be considered hate speech. We will use the same bag of words they used. We also consider making histograms and seeing which words appear most frequently in that bag of words and possibly tweaking it a little bit and seeing how the results change. Another bag of words that we consider using is Hatebase. After that we plan to use various machine learning models and compare their performance since the paper only used logistic regression.

Possible add-ons to our project depending on the progress we have: using Twitter AAE model (similarly to what they did in the second part of the paper) to see if there is racial bias in classification of our models. That model basically takes in tweets and tells the probability they were written by black vs white people based on the language used. We are a little sceptical about it as it adds more data inferred by models but it is something they also used in paper later on and we think it is an interesting add on. Another thing we consider adding are various graphs and histograms to understand the data better and to make comparisons between different models.

5.2 Experiments and Results

We will be using the following datasets: data from Davidson, data from Waseem, and we will send an email to Founta to ask for his data.

Performance metrics: precision, f-1 score, recall, we will compare it to the following metrics that [A] accomplished (Figure 1) and try to get better results than they did. We plan to run different machine learning algorithms and see which one performs best.

Dataset	Class	Precision	Recall	F1
W. & H.	Racism	0.73	0.79	0.76
	Sexism	0.69	0.73	0.71
	Neither	0.88	0.85	0.86
W.	Racism	0.56	0.77	0.65
	Sexism	0.62	0.73	0.67
	R. & S.	0.56	0.62	0.59
	Neither	0.95	0.92	0.94
D. et al.	Hate	0.32	0.53	0.4
	Offensive	0.96	0.88	0.92
	Neither	0.81	0.95	0.87
G. et al.	Harass.	0.41	0.19	0.26
	Non.	0.75	0.9	0.82
F. et al.	Hate	0.33	0.42	0.37
	Abusive	0.87	0.88	0.88
	Spam	0.5	0.7	0.58
	Neither	0.88	0.77	0.82

Figure 1

So far, we have tokenized the Twitter data from the Davidson using nltk tokenize as mentioned before. Then, we used the bag of words that was used in the paper as features.

We have run Decision Tree, AdaBoost, and Logistic Regression on the Davidson data set with the following results. Using a weighted average of precision, recall, and F-1 score in order to account for class imbalance, our results are shown in Figure 2.

Interestingly, our models do a better job of predicting Hate, but do worse predicting the other two classes. Additionally, we see very high recall

Decision Tree			
	Precison	Recall	F-1 Score
Hate	0.622	0.779	0.6917
Offensive	0.1133	0.9946	0.2034
Neither	0.1917	0.8737	0.3144

AdaBoost			
	Precison	Recall	F-1 Score
Hate	0.5484	0.7784	0.6435
Offensive	0.0798	0.9948	0.1477
Neither	0.1393	0.8734	0.2403

Logistic Regression			
	Precison	Recall	F-1 Score
Hate	0.5574	0.7753	0.6485
Offensive	0.0774	0.9953	0.1436
Neither	0.136	0.8716	0.2353

Figure 2

scores for the Offensive class. We will investigate the causes of these phenomena between now and our final report.

We also created correlation matrices & heatmaps, a frequency distribution of bad words, and analyzed a word cloud. These visualizations and descriptions can be found in our Colab here.

6) Risk Mitigation Plan

How will you build a minimum viable project within the remaining time? We want to run more models for our final deliverable, ensemble them together, and analyze the resulting metrics. We can try different features extraction with bag of words. Next, we plan to attempt the BERT language model. Moving forward, we have the following specific plan of action: 1) pull tweets from the Waseem data set using Twitter API. 2) run the same models we ran on the Davidson data set on this data set. 3) compare our initial results to those in the paper. 4) use these initial results to inform parameter tuning and feature selection. 5) run models and report final results after parameter tuning and improved feature selection. These five steps will result in a minimum viable project.

Will you start with a simplified setting where you can get some early results, so that you still have time to pivot if needed? We have already done this, running three different models on the Davidson hate speech Twitter data set. After examining our preliminary results, we do not believe that pivoting will be necessary.

If your approach doesn't work, how will you still turn in a useful project report? If our approach does not work, we will discuss why the models we ran did not provide significant contributions to the problem. This could be a sign to other researchers in the future that following our path is not the right way to proceed. Additionally, we will discuss areas for future research, and other things that should be tried based on the results of our experiments.

What if you find that you need too much compute? If we find that we need too much compute, we will dial back the types / amount of various models we are running. However, given our current plan, we do not anticipate this problem. If we do have time remaining, it would be interesting to extend this project to explore racial divisions and biases in tweets.

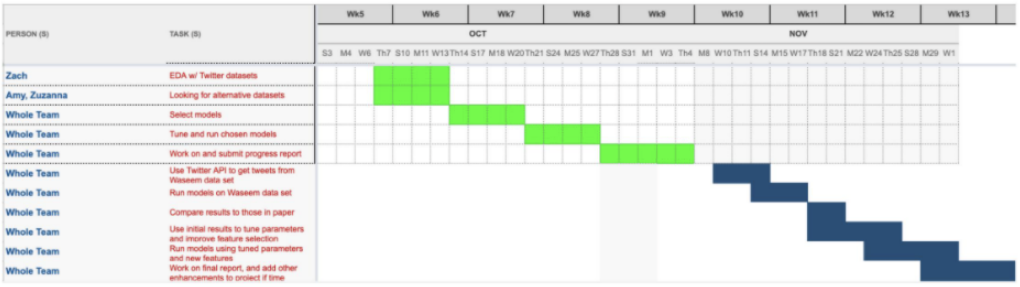
Will you evaluate the reasons for failure, or find specific examples where your approach works better, even if it does not do better on the full dataset? Given the limited amount of time remaining to build more models, draw meaningful conclusions, and finish the project by the deadline, we are aware that we could run out of time to meet all of our objectives. Even if our models do not perform as well compared to our original paper sources, we still think it is worthwhile to pursue, not only to make progress on detecting hate speech, but also to get more experience with the machine learning process. We are optimistic that we will have substantial takeaways from this project from an educational and personal perspective.

Will you try your algorithm on different, simpler data, such as a "toy" synthetic dataset you generated? No, as we do not come up with novel new algorithms and use the known models to solve this problem.

(Exempted from page limit) Other Prior Work / References (apart from Sec 3) that are cited in the text:

- Paper Title: "Automated Hate Speech Detection and the Problem of Offensive Language"
  Link: https://arxiv.org/pdf/1703.04009.pdf
  Github: https://github.com/t-davidson/hate-speech-and-offensive-language
- Paper Title: "Automatic Hate Speech Detection using Machine Learning: A Comparative Study"
  Link: https://thesai.org/Downloads/Volume11No8/Paper_61-Automatic_Hate_Speech_Detection.pdf

(Exempted from page limit) **Full Work Plan, including the previous work plan with completed/incomplete steps (okay to modify from the proposal), and the remaining steps:** (create additional columns with deadlines for steps towards the final report, assigning responsibilities to individual team members to the extent possible. The GANTT chart you used in the proposal will be a good starting point. Mark completed steps in green, as shown here. For convenience, you can split into two charts, one till Nov 8, and another for after Nov 8, placed one below the other.)



(Exempted from page limit) Supplementary Materials if any (but not guaranteed to be considered during evaluation):