Reproducing Automated Hate Speech Detection and the Problem of Offensive Language

Sri Venkata Kameswara Naga Phanindra Kavipurapu, Jerome Roosan Chandrashekhar Meenakshisundaram, Raja Pedapudi

Macquarie University

November 2, 2023

Abstract

Our current endeavor is to reproduce the results of our source paper and code in the spirit of open science and research that is reproducible. In this report, we go into the methodology outlined in the source paper and our effort to replicate the results using the original and synthesized data. We will go into the results obtained and some of the challenges that we faced.

1 Description of Source Paper

The paper is not very hard to understand nor is the code so convoluted that we can't make sense of it. In the paper, the authors posit that one of the key challenges when it comes to hate-speech detection is making the distinction between hate speech and offensive speech. The authors start out by clarifying what they consider hate speech. Even though they acknowledge that there is no formal definition of hate speech, they state that any *language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"* is hate speech (Davidson et al., 2017).[1]

For building this model, the authors used a crowd-sourced lexicon with hate speech keywords to mine tweets. These tweets were then labeled as *"hate speech"*, *"offensive"*, and *"neither"* classes. Then, a multi-class classification model was trained to classify the tweets into the correct categories. The authors note that the fine-grained labeling that they've implemented can perform better than existing models to identify instances of hate speech on social media. They also note that this brings to light some of the challenges they face when it comes to the accuracy of the model.

1.1 Evaluation Framework

The base model is ultimately a logistic regression classifier with an L2 penalty. So, for the sake of evaluating the performance of the model, the authors have used metrics such

[1] https://arxiv.org/pdf/1703.04009v1.pdf

as precision, recall, and F-1 scores. They also visualized the performance in the form of a heatmap of the confusion matrix. This is done by simply calling the scikit-learn function *classification_report( )* on the predictions. Further details can be found in Sec. 4

1.2 Justification

At some point or the other, all members of our group have been victims of hate speech, as a result of which the subject of this paper is important to us. The methodologies outlined in this paper are rather interesting and piqued our interest. Additionally, the conference in which this paper was presented (ICWSM - International Conference on Web and Social Media) is ranked 11 on Google Scholar's Top Publications for Databases and Information Systems. The corresponding h-index is 59 and the h-5 median is 82.

Additionally, the github repository containing the code and the data has 712 stars. The paper has 147 citations and 15 references on Papers With Code[2]

2 Description of Original Data

The authors first obtained a crowd-sourced hate speech lexicon that has keywords relating to hate speech. This lexicon was then used to identify tweets with offensive or hateful speech from thirty-three thousand users.

Then the entire timelines of the users who posted these tweets were extracted, resulting in a corpus of roughly eighty-five million tweets. From this set of eighty-five million tweets, twenty-five thousand tweets were randomly sampled which were then manually labeled by workers at CrowdFlower.

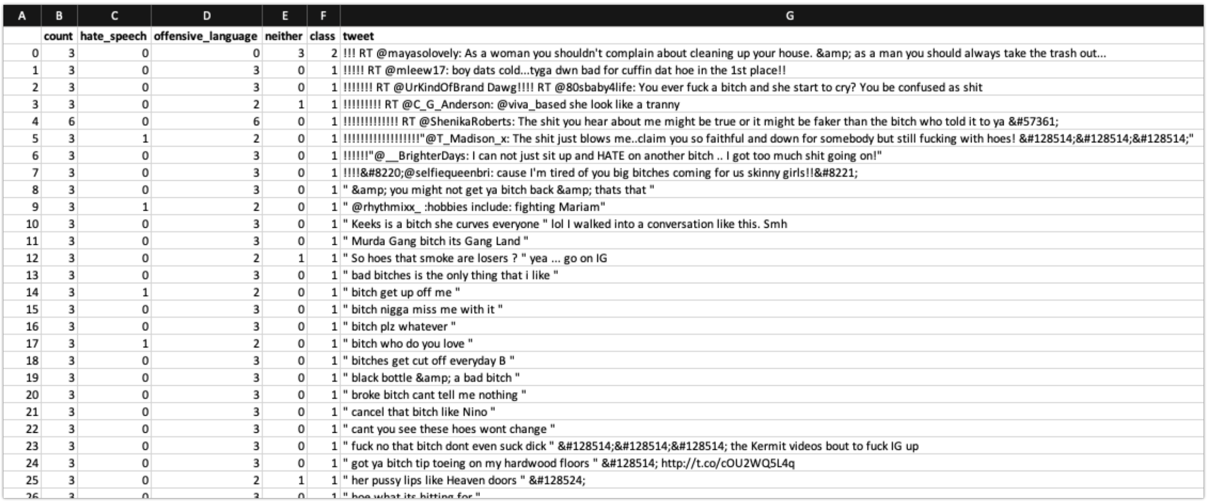
In order to maintain a high standard of labeling, the workers were provided a thorough description of what does and doesn't qualify as hate speech. Further instructions and details were provided in written form. They were specifically asked to consider the context in which the words were used and not just the words themselves. The idea they wanted to convey was that no matter how offensive a word was, only the context in which it was used matters and not the presence of the word itself.

It is worth noting that each tweet in the dataset was labeled by three or more coders (the workers performing the labeling) to account for personal biases. Then, a majority class voting was performed to assign the final label for the tweet. They noted that the intercoder-agreement was as high as 92%.

As some of the tweets did not have a majority class, they were left unlabeled. This process resulted in a corpus of roughly twenty-five thousand labeled tweets. 5% of the tweets were labeled as *"hate speech"*(label:0) by the majority and only 1.3% unanimously. A major portion of the tweets were considered *"offensive"*(label:1) with 75% of majority voting and 53% unanimous voting. The remaining tweets were labeled as *"neutral"*(label:2). Ultimately, these tweets[3] were used to construct features which were then used to train the multi-class classification model.

[2] https://paperswithcode.com

[3] https://github.com/t-davidson/hate-speech-and-offensive-language/blob/master/data/labeled_data.csv



A	B	C	D	E	F	G
	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. &amp; as a man you should always take the trash out...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya &#57361;
5	3	1	2	0	1	!!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! &#128514;&#128514;&#128514;"
6	3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
7	3	0	3	0	1	!!!!&#8220;@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!&#8221;
8	3	0	3	0	1	" &amp; you might not get ya bitch back &amp; thats that "
9	3	1	2	0	1	" @rhythmixx_ :hobbies include: fighting Mariam"
10	3	0	3	0	1	" Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh
11	3	0	3	0	1	" Murda Gang bitch its Gang Land "
12	3	0	2	1	1	" So hoes that smoke are losers ? " yea ... go on IG
13	3	0	3	0	1	" bad bitches is the only thing that i like "
14	3	1	2	0	1	" bitch get up off me "
15	3	0	3	0	1	" bitch nigga miss me with it "
16	3	0	3	0	1	" bitch plz whatever "
17	3	1	2	0	1	" bitch who do you love "
18	3	0	3	0	1	" bitches get cut off everyday B "
19	3	0	3	0	1	" black bottle &amp; a bad bitch "
20	3	0	3	0	1	" broke bitch cant tell me nothing "
21	3	0	3	0	1	" cancel that bitch like Nino "
22	3	0	3	0	1	" cant you see these hoes wont change "
23	3	0	3	0	1	" fuck no that bitch dont even suck dick " &#128514;&#128514;&#128514; the Kermit videos bout to fuck IG up
24	3	0	3	0	1	" got ya bitch tip toeing on my hardwood floors " &#128514; http://t.co/cOU2WQ5L4q
25	3	0	2	1	1	" her pussy lips like Heaven doors " &#128524;
26	3	0	3	0	1	" hoe what its hitting for "

Figure 1: Snippet of the original data used

2.1 Features

Each tweet underwent preprocessing where all of the text was lower-cased and stemmed using Porter stemming technique. Uni-gram, bi-gram, and tri-gram features, weighted by its term-frequency inverse-document-frequency, were obtained. Parts-of-speech tagging was performed to better encode the syntactic structure.

Further granular features like reading level and reading ease scores were obtained to augment the richness of features. Lastly, various measures like character count, hashtag count, syllable count, word count, etc. were recorded for each tweet.

3 Model

Once the features were obtained, the authors trained a logistic regression model using Lasso penalty. This was done to bring down the dimensionality of the data. Different models like Naive-Bayes, decision trees, random forests, and linear support vector machines (SVMs) were trained using the reduced data. Furthermore, all of the models were tested using cross-validation techniques (mainly five-fold cross-validation) coupled with a grid-search algorithm to find the best parameter values for each of the models.

After extensive testing of the different trained models via grid-search cross-validation, the authors found that the logistic regression model and the support vector classification model with a linear kernel performed the best in comparison to the other models like Naive-Bayes, decision tree, random forests, etc. In fact, they found that these two models performed significantly better than the other models mentioned. The authors decided to use an L2 penalized logistic regression model (Ridge regression model) as their final model.

This is because a logistic regression model is simpler in comparison to some of the more complex models. It is also easier to understand and explain. Furthermore, a logistic regression model with a Ridge penalty allows us to look at the class membership probabilities closely. The authors also noted that the L2 penalized logistic regression model performs well according to past papers.(Burnap and Williams, 2015; Waseem and Hovy, 2016).

The final model was built using the entire dataset as training data which was then used to predict a label for individual tweet. It should be noted a one-versus-all paradigm was used each class has a separate classifier trained. The final class label was the one with the highest probability of class membership across all the trained classifiers.

4 Results

The model that had the best performance returned an overall precision score of 0.91, an overall recall score of 0.90, and an overall F1 score of 0.90. It is worth noting that around 40% of hate speech-containing tweets are wrongly labeled as offensive or neutral. The authors note that the precision and recall scores for the *"hate"* class are 0.44 and 0.61. This indicates a bias in the model to classify tweets as less hateful in comparison to the coders at CrowdFlower.

Tweets that contained multiple hate speech keywords were more likely to be predicted as hate speech. The authors note that even replies to hate speech that are against hate speech are classified as hate speech because they themselves contain hate speech keywords. Tweets that are labeled as *"offensive"* are actually less hateful, indicating superior performance.

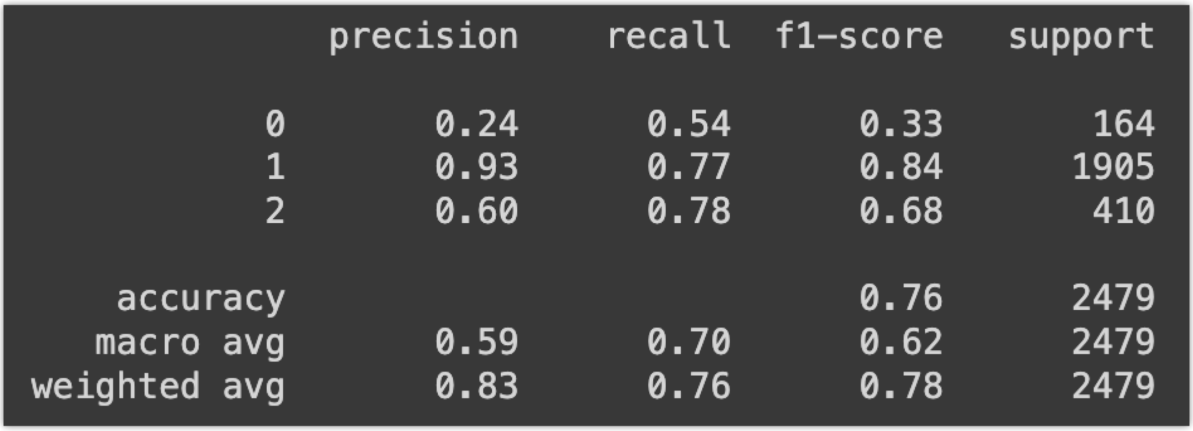
Unsurprisingly, tweets that are inaccurately labeled as *"neutral"* do not contain any hate speech keywords. The authors note that newer variations of racial slurs or novel terms of derogation tend to escape being classified as hate speech. The model performs better on tweets that have the most prevalent terms and forms of hateful speech, specifically anti-black and homophobia.

Certain caveats are repeatedly noted throughout the work by the authors. They note that people on social media tend to use language most prevalent in rap music with each other. Even though these words by themselves might be offensive, racist, or sexist, when put in context might not necessarily be hateful. Interestingly, tweets with a higher readability score are more likely to be classified as *"neutral"*. Neutral tweets wrongly classed as hateful or offensive tend to mention some keywords related to gender, race, or sexuality. And finally, and even more interestingly, the authors note that their model picked up on instances of hate speech that even the human coders at CrowdFlower missed.

5 Replication of Original Work

It is important that a particular piece of research be replicable as this is one of the best ways we can validate the results and claims made by the authors. This is especially the case with the rise in the spread of misinformation and the propagation of biased research pieces. This is all done in the spirit of open science and replicability in research. In our particular case, the source paper came with a GitHub repository that contains the code and the data used to produce the results used in this study.

The GitHub repository had a notebook that the authors used to obtain the results published in their study. The readme.md file suggested that we run that particular notebook to obtain the results that they obtained. As the dataset had twenty-five thousand tweets, we were successful in executing the code and obtaining the results. Below are the outputs from the replication of the source work. The code is written in Python 2.7 but we couldn't find additional information regarding the versions of the packages used. We used Jupyter Notebooks to run the code on our machines.



```
              precision    recall  f1-score   support

           0       0.24      0.54      0.33       164
           1       0.93      0.77      0.84      1905
           2       0.60      0.78      0.68       410

    accuracy                           0.76      2479
   macro avg       0.59      0.70      0.62      2479
weighted avg       0.83      0.76      0.78      2479
```

Figure 2: Precision, Recall & F1-Score of Original Source Work

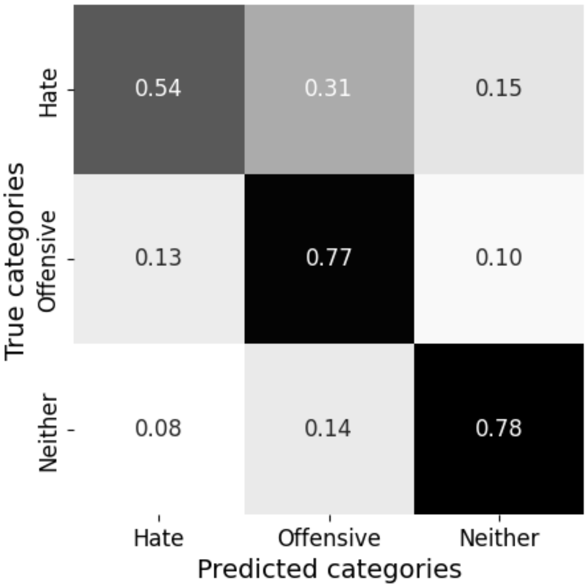


Figure 3: Confusion Matrix of Original Source work

We report that the original code ran with no hitches due to the smaller size of the data and the relative simplicity of the models used. The code for the same can be found here.[4]

6 Creation of New Data

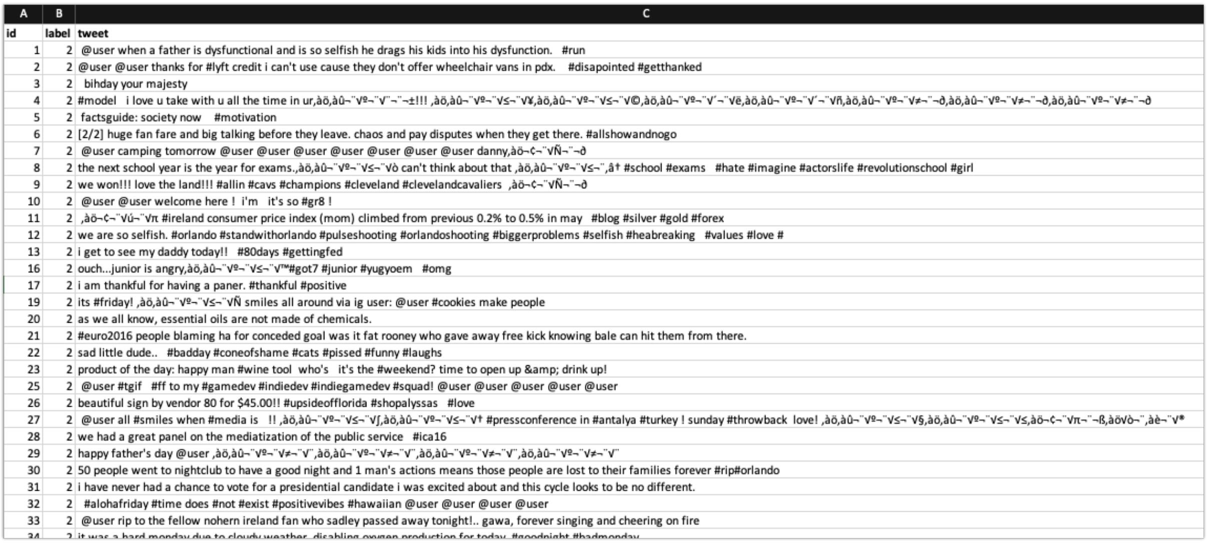
For the purpose of replicating the results on a different dataset, we obtained data from Kaggle[5]. A major reason for downloading the data directly and not obtaining it from web scraping is that the tweets data is now behind a very expensive API, since the

[4] https://github.com/jeromeroosan2023/COMP8240_Project

[5] https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech/?select=train.csv

transformation of Twitter to X. The new data already had labels but was incompatible with the source model. The labels in this data were only two-fold: offensive and not offensive. As we noted in Sec. 1, this is the exact type of naive classification we want to avoid. We need a finer level of classification. For this, we randomly sampled a thousand records from the new data and reclassified them into three classes.

Each member of the team received two-fifty tweets and was designated the task of relabeling them. Before this relabeling process, we had a discussion on how we were going to classify them. Similar to the instructions provided to CrowdFlower for labeling, we decided on a set of rules for labeling before we proceeded. We ran some test scenarios with random tweets from the dataset and found ourselves reaching a consensus in most cases. Therefore, we proceeded to relabel the thousand tweets sampled from the new data. We did not quantify the level of agreement between us when it came to labeling. The idea was that we had a pretty good idea of what we did or did not consider hate speech, etc. so we didn't find the need for multiple labels. In hindsight, we note that this is not very scientific.



A	B	C
id	label	tweet
1	2	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	2	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
3	2	bihday your majesty
4	2	#model i love u take with u all the time in ur,àö,àû¬¨√º¬¨√¨¬¨¬±!!! ,àö,àû¬¨√º¬¨√≤¬¨√¥,àö,àû¬¨√º¬¨√≤¬¨√©,àö,àû¬¨√º¬¨√´¬¨√ë,àö,àû¬¨√º¬¨√´¬¨√ñ,àö,àû¬¨√º¬¨√≠¬¨¬ð,àö,àû¬¨√º¬¨√≠¬¨¬ð,àö,àû¬¨√º¬¨√≠¬¨¬ð
5	2	factsguide: society now #motivation
6	2	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
7	2	@user camping tomorrow @user @user @user @user @user @user @user danny,àö¬¢¬¨√Ñ¬¨¬ð
8	2	the next school year is the year for exams.,àö,àû¬¨√º¬¨√≤¬¨√ò can't think about that ,àö,àû¬¨√º¬¨√≤¬¨,â† #school #exams #hate #imagine #actorslife #revolutionschool #girl
9	2	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers ,àö¬¢¬¨√Ñ¬¨¬ð
10	2	@user @user welcome here ! i'm it's so #gr8 !
11	2	,àö¬¢¬¨√ú¬¨√π #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
12	2	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #
13	2	i get to see my daddy today!! #80days #gettingfed
16	2	ouch...junior is angry,àö,àû¬¨√º¬¨√≤¬¨√™#got7 #junior #yugyoem #omg
17	2	i am thankful for having a paner. #thankful #positive
19	2	its #friday! ,àö,àû¬¨√º¬¨√≤¬¨√Ñ smiles all around via ig user: @user #cookies make people
20	2	as we all know, essential oils are not made of chemicals.
21	2	#euro2016 people blaming ha for conceded goal was it fat rooney who gave away free kick knowing bale can hit them from there.
22	2	sad little dude.. #badday #coneofshame #cats #pissed #funny #laughs
23	2	product of the day: happy man #wine tool who's it's the #weekend? time to open up &amp; drink up!
25	2	@user #tgif #ff to my #gamedev #indiedev #indiegamedev #squad! @user @user @user @user @user
26	2	beautiful sign by vendor 80 for $45.00!! #upsideofflorida #shopalyssas #love
27	2	@user all #smiles when #media is !! ,àö,àû¬¨√º¬¨√≤¬¨√∫,àö,àû¬¨√º¬¨√≤¬¨√† #pressconference in #antalya #turkey ! sunday #throwback love! ,àö,àû¬¨√º¬¨√≤¬¨√ß,àö,àû¬¨√º¬¨√≤¬¨√≤,àö¬¢¬¨√π¬¨¬ß,àö√ò¬¨,àè¬¨√*
28	2	we had a great panel on the mediatization of the public service #ica16
29	2	happy father's day @user ,àö,àû¬¨√º¬¨√≠¬¨√¨,àö,àû¬¨√º¬¨√≠¬¨√¨,àö,àû¬¨√º¬¨√≠¬¨√¨,àö,àû¬¨√º¬¨√≠¬¨√¨
30	2	50 people went to nightclub to have a good night and 1 man's actions means those people are lost to their families forever #rip#orlando
31	2	i have never had a chance to vote for a presidential candidate i was excited about and this cycle looks to be no different.
32	2	#alohafriday #time does #not #exist #positivevibes #hawaiian @user @user @user @user
33	2	@user rip to the fellow nohern ireland fan who sadley passed away tonight!.. gawa, forever singing and cheering on fire
34	2	it was a hard monday due to cloudy weather. disabling oxygen production for today. #goodnight #badmonday

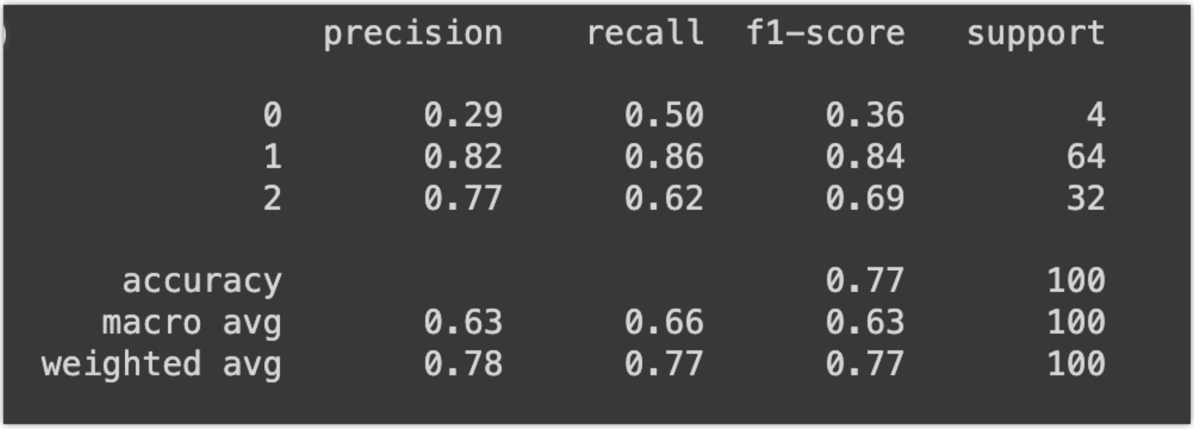
Figure 4: Snippet of the new data

7 Results on New Data

We used the thousand tweet record with the new and compatible labels to run the same code on it instead of the source data. We found that our results with the new data were very close to the source results obtained by the authors in their paper. The results can be compared to the original results in Sec. 5 We used the same evaluation framework as the source paper and used precision, recall, F1-scores, and confusion matrices to evaluate model performance.

7.1 Results on New Data

The validation results on new data is as shown below.



```
              precision    recall  f1-score   support

           0       0.29      0.50      0.36         4
           1       0.82      0.86      0.84        64
           2       0.77      0.62      0.69        32

    accuracy                           0.77       100
   macro avg       0.63      0.66      0.63       100
weighted avg       0.78      0.77      0.77       100
```

Figure 5: Precision, Recall & F1-Score of Replication with New Data

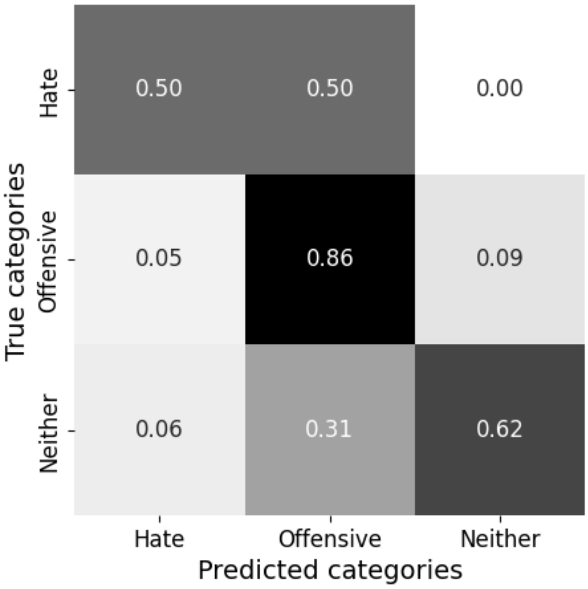


Figure 6: Confusion Matrix of Results with New Data

7.2 Issues Faced During Replication of Source Work With New Data

Our main issue with the replication of the source work with our newly labeled data was the outdated versions of the packages used. Furthermore, the version of Python used was itself old (version 2.7). We noticed that a lot of the methods and attributes were either changed or even non-existent in the up-to-date packages that we downloaded.

The issue was further exacerbated by the fact that we couldn't find any relevant information about the versions of the different packages used. We initially tried creating a virtual environment with the specific Python version (version 2.7) but failed as once again, the package versions themselves were clashing. Upon inspecting the notebooks provided by the authors, we realized that the two different notebooks (one for replicating source work with source data and the other for replicating source work with new data) were in fact similar except for a few caveats that we had to fix.

We then used the same notebook that was modified to accept our new data as input to replicate the source results with completely new data.

8 Reflections

In hindsight, we note that we should have individually labeled all thousand tweets and then taken a majority vote to obtain the final label. Having said that, we are glad to see that our results are not too far off from the results in the source paper.

References

Burnap, P. and Williams, M. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making: Machine classification of cyber hate speech. *Policy Internet*, 7.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. pages 88–93.