Hate Speech Detection in Incel Online Spaces

Paolo Gajo
Dept. of Interpreting and Translation
University of Bologna, Forlì
paolo.gajo@studio.unibo.it
August 29, 2022

Abstract

Automatic hate speech detection is a widely researched topic in machine learning for its potential usefulness in dealing with online toxicity and harassment. However, recognizing hate speech is challenging and its often-inconsistent definition and the use of offensive neologisms make this task even more arduous. This is especially true in spaces frequented by incels, a community which mostly comprises men unsuccessful in finding a significant other who tend to express themselves by adopting various forms of hate speech, especially against women. Their language is deeply characterized by community-specific jargon, which is created and used with unusual connotations to spread hate and attack individuals and groups of people. This paper presents a set of text classifiers developed to detect hate speech characterized by such specific terminology, trained on datasets built from posts scraped from a popular online incel forum. Additionally, the study presents a preliminary analysis of the mechanics that determine the spread of hate speech in these spaces, indicating a statistical significance in the relationship between the presence of hate speech in the original post of a thread and the percentage of hateful posts following it.

1 Introduction

The toxicity and hate speech (HS) that pervade social media platforms have been attracting the attention of an increasing number of studies. This has been the case in multiple fields of research, ranging from sociology to linguistics and, more recently, the fields of machine learning and natural language processing (e.g., Davidson et al., 2017; Mathew et al., 2021). The importance of these endeavors cannot be understated due to the concrete consequences such phenomena can bear on societies, both indirectly and directly as a result of acts of verbal and physical violence, some of which have been perpetrated by users belonging to online communities characterized by extremist ideologies. One such community is the group of people who call themselves incels, short for "involuntary celibates", which make up a subspace of the so-called "Manosphere", mostly comprising men unsuccessful in finding a sexual partner and characterized by the frequent use of racist and misogynous language (Nagle, 2017; Blommaert, 2018).

Despite the mounting effort being poured into this field of research, detecting hate speech automatically is nonetheless difficult due to its unclear definition (Davidson, 2017) and the subjective perception any individual might have when evaluating whether a particular text contains hateful content (Muti and Barrón-Cedeño, 2020). Training a neural network (NN) model capable of effective hate speech classification is therefore a challenging proposition, which holds all the more true when considering platforms where specific jargon is created and used with connotations that differ drastically from those of general English, such as online spaces frequented by incels.

Consequently, this study aims to tackle the issue of domain-specific jargon and hate speech by providing a set of original datasets and a trainable vector space model (VSM) thanks to which a convolutional neural network (CNN) classifier can be trained to detect hate speech despite the use of neologisms and opaque jargon. In particular, the datasets are specific to the environment of incel online forums, a prime example of how hate speech also involves the adoption of original terminology and the association of common words to novel meanings and connotations (Farrell et al., 2020; Gothard, 2021).

Additionally, the study presents data gathered by using text classifiers to automatically annotate forum posts for hate speech in general, and racism and misogyny in particular. The results obtained from the process are used to make statistical inferences about the nature of the spread of hate speech in these online forums, showing a statistically significant relationship between the use of hate speech in the original post (OP) of a thread and the number of hateful posts following it (i.e., containing hate speech).

2 Related Work

Many resources have been made publicly available to build upon as far as automatic hate speech detection is concerned. These include datasets published for shared tasks and competitions (Davidson et al., 2017; Basile et al., 2019; Jigsaw, 2019) which are useful starting points when developing a text classifier, due to the overhead investment involved in manually evaluating thousands of instances to build a dataset. However, datasets built by extracting content from incel platforms are rare and not necessarily appliable to the use-case of this study, either because of the source of the data only being partially compatible with the linguistic domain presently tackled (Pelzer et al., 2021) or because of the criteria according to which it was annotated (Zhou et al., 2020). As Pelzer et al. themselves find (2021), using datasets built by merging multiple sources of data is not necessarily effective, which is why a new dataset was built and annotated specifically for this study, maximizing its compatibility with the task being tackled and ensuring consistency of evaluation.

In this regard, past studies have relied on the Pushshift Reddit API to build a corpus within the linguistic domain of inceldom (Farrell et al., 2020; Mollas et al., 2022). However, the biggest Reddit incel platforms have long since been banned from the website, making their language increasingly outdated compared to current incel discourse. As such, content from a popular active incel forum was used, as described in the next section.

3 Data[1]

3.1 Preliminary data collection

Similarly to Pelzer et al. (2021), this study also relies on incels.is[2] (formerly incels.co) to gather data, from which forum posts were downloaded and organized in an annotated dataset by using a specially made Python script. All threads found in the "Inceldom Discussion" section[3], containing the majority of threads and posts, were downloaded separately and then merged. The constructed dataset thus contains 4,281,310 posts from 211,525 different threads, totaling approximately 104M tokens. The first lines of the dataset are shown in Table 5 to exemplify its structure (see Appendix 1).

3.2 Training dataset manual construction

The full dataset was filtered in order to obtain a subset that could be evaluated by hand and thus be used to train a text classifier. The first version of the reduced dataset was built via a two-step filtering process.

First, only posts with length between 140 and 280 characters were taken into consideration, so as to resemble the length of tweets, since the majority of other studies are built around Twitter. The reasoning behind this choice was to potentially allow merging this dataset with other datasets built from Twitter in the future.

Second, the pool of posts obtained after the first filter was divided into two groups, collecting 1,000 instances from each at random. The first set contains terms from the list shown in Table 6 (see Appendix 2), while the second contains no such terminology.

The dataset was later expanded by gathering and annotating more of these filtered instances until the classifiers trained on it were able to reach an acceptable accuracy level on their test partition (>80%), thus reaching a total of 3,571 manually evaluated posts.

The lexical items listed in Table 6 were found by analyzing a subset of the constructed corpus (~10M tokens) with Sketch Engine's keyword function and were deemed to be a representative sample of terms characteristic of incel lexicon. This filtering strategy was adopted so that the models trained on the dataset could also classify effectively instances where no domain-specific language was used. If only instances containing typical incel lexicon had been taken into consideration, the models would likely learn to simply associate those words with hate speech automatically, discarding any potential nuance in the language.

[1] All datasets mentioned henceforth are available on request at: https://github.com/paolo-gajo/incel_hs
[2] https://incels.is/
[3] Up to July 31, 2022, when the scraping program was run.

3.3 Manual evaluation

Individual instances were initially assigned to 7 different classes: 0) no hate speech, 1) misogyny, 2) racism, 3) gender discrimination, 4) ablism, 5) fatphobia, 6) general threats of violence. A post was classified as containing hate speech when it attacked an individual or a group, directly or indirectly, based on one of the aforementioned categories, i.e., their gender, ethnicity, neurotypicality, etc. The posts were subsequently given a binary classification, defining them as containing hate speech whenever the previously assigned class was different from "0". Due to a quantitative lack of data, only classes 0, 1 and 2 were kept in the final version of the manually evaluated dataset, for a total of 3,466 instances.

3.4 CNN-aided dataset expansion

The manually evaluated data was used to train a preliminary 3-class CNN model with Keras and TensorFlow to automatically annotate 200,000 of the instances contained in the full dataset. The instances classified as misogynous or racist were then filtered based on whether they contained the words "woman", "women", "womens" but none of the terms listed in Table 6. The reasoning for this was that, in general, posts referring to women should not necessarily be considered hateful, which is why this strategy was adopted to quickly find instances that the model had likely classified erroneously. The lemma "woman" was specifically used as a filter because it is the most frequent lexical word in the whole corpus, according to Sketch Engine's wordlist function, and because of its prominence in being involved in occurrences of misogyny (by far the most frequent form of hate speech in this sort of platform).

Consequently, the annotated instances were sorted by hate speech type (second column in Table 5) and 200 mistaken evaluations were corrected. The new samples were then added to the dataset (keeping 4 instances out to fit the 3-class annotation), bringing the new total to 3,662 instances. At this point, 50 samples were set aside for later model testing[4], and 8 artificial examples were added to the dataset to include specific annotated instances a classifier would almost certainly evaluate incorrectly.

Finally, this dataset containing 3,620 annotated instances was used to train a binary and a 3-class model. Both text classifiers were thus used to evaluate 20,000 instances from the full dataset, annotating each post based on whether the two classifiers agreed the text contained hate speech. The instances were then filtered according to the agreement check column (see Table 5) to find posts the two models disagreed on, and 580 of the filtered samples were corrected and appended to the dataset, bringing the total size to 4,200 instances. In this case, the length of the posts was unrestricted to improve the capability of later versions of the classifiers to handle short samples effectively.

3.5 Dataset augmentation

Lastly, the dataset was augmented automatically via substitution of specific terms. The lemma "woman" was once again used as filter on the 2,503 instances not containing hate speech, returning 295 lines in total. New lines were thus appended to the dataset after substituting the target words "woman", "women" or "womens" found in the 295 samples with terms from Table 7 (see Appendix 2).

The dataset was then balanced according to the binary and 3-class annotations, creating an undersampled version for each of them. The augmented dataset used to train the binary classifier thus has a final size of 5,006 instances with a 50/50 split, while the one used for the 3-class model contains 7,509 samples, whose classes are all balanced at one third of its size.

3.6 Dataset vectorization

Training a model capable of detecting hate speech within an online environment dedicated to the phenomenon of inceldom represents a challenging task due to various factors, such as the niche nature of the platform, the opacity of its hateful terminology, and the extremely frequent association of negative and hateful language to words that otherwise possess neutral connotation.

These issues are crucial, as the lexicon adopted by incels contains a non-negligible number of terms which cannot be found in common pre-trained VSMs, such as the ones listed on fastText's website[5].

[4] These instances were used temporarily as an external test set prior to using k-fold validation on the final versions of the classifiers.
[5] https://fasttext.cc/docs/en/english-vectors.html

Additionally, due to the pervasiveness of novel jargon, it is unviable to simply ignore out-of-vocabulary terms in this context, given also the extremely offensive nature of some of the terms, which often determine whether a post can be deemed hateful. It was therefore paramount for the text classifiers used in this study to be capable of inferring effectively on the meaning of such terms, which is why the VSM used to train the employed models needed to contain a representation of those words. To overcome this issue, a new VSM was trained with word2vec on the entirety of the posts contained in the full dataset.

Given these considerations, vectorizing the adopted datasets first involved lowercasing the text of each post. This was done because Sketch Engine's keyword function showed that case sensitivity was irrelevant for this type of content[6], even when taking into account capitalized terms such as "Stacy" and "Becky" [7]. Then, each text was tokenized using NLTK's casual tokenizer (Lane et al., 2019:48), since the tokenizer needed to be able to deal with potentially non-standard punctuation, given that user posts might contain all sorts of unusual formatting. Finally, the previously trained VSM was used to vectorize the training datasets and the full 104M-word dataset, whose instances were subsequently annotated automatically by the trained classifiers and analyzed statistically.

4 Model

4.1 Structure of the CNN

The final versions of the binary and 3-class models are both CNNs, built using the standard Sequential model in Keras with the layers listed in Table 1.

In the convolutional layer, the number of filters was set to 16 because higher values increased the computational load during training and validation without apparently improving the metrics of the classifiers. For the same reason, the kernel size was set to 3, as tests run with values ranging from 5 to 13 returned worse results. Lastly, the models performed slightly better with padding, which is why the setting was set to "same".

To prevent overfitting, especially considering the original dataset was augmented and contains many instances which only differ by a few words, a global max pooling layer and a dropout layer were added.

Layer	Parameters		Values
Convolutional 1D layer	Conv1D()	filters	16
		kernel_size	3
		padding	same
		activation	relu
		strides	1
		maxlen	100
Global max pooling layer	GlobalMaxPooling1D()		/
Dropout layer	Dropout()		0.3
Classification layer (binary)	Dense()		1
	Activation()		relu
Classification layer (3-class)	Dense()		3
	Activation()		softmax

Table 1: Layers used in the binary and 3-class CNN classifiers

4.2 Training

In combination with a dropout of 0.3, the binary classifier was trained for 2 epochs, while the 3-class model was fitted 3 times. Any additional training cycles resulted in what appeared to be overfitted models. As for the batch size, both classifiers were trained by backpropagating every four instances (batch_size = 4). This sped up the process compared to backpropagating every single instance and even provided better results compared to a batch size of 1.

Table 8 shows the performance metrics for four different combinations of models and datasets, obtained by process of 10-fold validation (see Appendix 2). As the table shows, the 3-class model trained on the augmented dataset performs sensibly better compared to the one trained on the non-augmented version. Additionally, the augmented dataset is balanced and contains a substantially higher number of instances (7,509 vs. 4,200), which makes the metrics even more reliable. As such, the full dataset was later annotated for hate speech type with the 3-class model trained on all 7,509 instances of the augmented dataset.

[6] No relevant lexical items using specific capitalization were found among the first 1,000 simple and complex terms, except for "Stacy", "Stacie", "Becky" and "Beckie", which are used as common adjectives and nouns and can therefore be lowercased.

[7] Nicknames used to rate a woman's physical appearance based on a decile rating system, wherein a "Stacy" is ranked higher than a "Becky" (Gothard, 2021:4-7).

As far as the binary models are concerned, the difference between their metrics is much less evident. Consequently, as the original binary class split in the non-augmented dataset was only 60/40, the classifier trained on the original 4,200 instances was chosen to later annotate the full dataset. The reasoning behind this choice was that the model trained on the augmented dataset containing 5,006 instances did not do well enough to warrant choosing it over the other classifier, even considering it was validated on a 50/50 split rather than on a 60/40 unbalanced set. In fact, the model trained on the augmented dataset should achieve sensibly better results despite being validated on balanced instances, given that many of them are artificial and very similar to each other.

5 Results and discussion

The chosen binary and 3-class models were used to automatically annotate all 4,281,310 posts contained in the full dataset for hate speech (binary classifier) and hate speech type (3-class classifier).

Before proceeding with further analyses, the consistency of annotation between the two classifiers was assessed based on different post lengths. Table 2 shows the agreement statistics obtained during the classification process:

Post length	Agree count	Agree ratio
length<140	2,996,443	0.94
140<length<280	568,093	0.93
length>280	424,475	0.92

Table 2: Agreement statistics of the binary and 3-class classifiers

As shown in Table 2, post length seems to be a negligible factor, and in general the agreement ratio looks promising, which arguably lends credibility to the quality of the annotations made separately by the two models.

Overall, the binary classifier annotated 13.14% of the posts as containing hate speech, while the 3-class model classified 5.00% as containing racism and 8.39% as being misogynous (see Table 3). However, the last two values are most likely underestimations of the actual figures, since the 3-class model cannot classify a post as being both racist and misogynous, which was not a rare occurrence in the original 7-class dataset (4.99% of the 3,671 manually evaluated posts).

Class annotation	Total posts	Percentage
Hate speech (binary)	562,620	13.14%
Misogyny (3-class)	359,312	8.39%
Racism (3-class)	216,240	5.00%

Table 3: Hate speech post annotation statistics

In addition, each thread was annotated with the number of posts flagged as hateful within it, which was used to calculate the percentage of posts containing hate speech in that specific thread. The first column and the second-to-last column in Table 5 (see Appendix 1) were thus paired and analyzed in R to assess whether the presence of hate speech in the OP had a statistically significant impact on the percentage of posts containing hate speech in the rest of the thread.

Since the statistical test was going to be configured either as a T-test or a Wilcoxon signed-rank test, the normality of the two distributions of hate speech ratios first had to be verified via Shapiro-Wilk test. The two series were found to be non-normally distributed ($p < 0.001$ for both) so they were compared by using Wilcoxon's signed-rank test. The test showed a highly significant relationship between the presence of hate speech in the OP and the percentage of hateful posts following it in the same thread ($p < 0.001$). The two distributions are displayed in the boxplot shown in Figure 1 below:

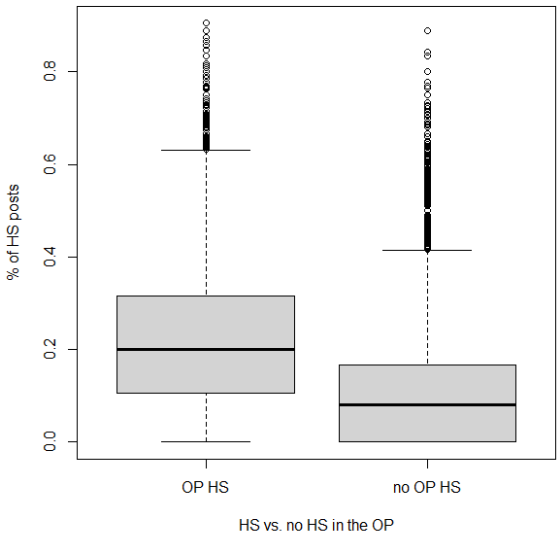

Figure 1: Distributions of the percentage of hateful posts in a thread based on the presence of hate speech in the original post

The difference shown by the graph is rather striking and is numerically explained by the fact that both the median and the mean of the left series are more than double that of the right series (see Table 4 below):

	OP HS	No OP HS
Median % of HS	0.200	0.08
Mean % of HS	0.219	0.108

Table 4: Median and mean percentage of posts containing hate speech in a thread based on the presence of hate speech in the original post

Keeping in mind that these results are likely only appliable to this type of online community, it seems two separate conclusions can be drawn from this relationship: a hateful OP will either attract specific users who are more prone to posting hate speech or incite all users in general to also post hateful replies. These hypotheses could be tested in a future study by organizing each user and their posts in a table and verifying whether they are more likely to post hateful comments based on the OP itself containing hate speech. Nonetheless, the results underline the importance of acting swiftly when moderating online platforms to prevent the spread of hate speech, as it is apparent that threads conceived as hateful from the outset tend to attract and accrue posts expressing hate speech much more prominently than unharmful OPs, effectively resulting in a snowball of hateful content.

6 Limitations of the study and further research

As already discussed in the previous section, the multi-class model is not capable of classifying instances as containing both racism and misogyny. Furthermore, due to the lack of sufficient evaluated posts, it was not possible to train the classifier to detect gender discrimination, ablism, fatphobia and general threats of violence. It is possible that by merging the used datasets with other datasets (such as the ones provided in Basile, 2019, and Davidson, 2017) these classes could also be integrated. However, the gap between the adopted definitions and the judgment of each individual human evaluator remains a difficult hurdle to overcome as far as dataset homogeneity is concerned, which ultimately affects the quality of any classifier trained on them.

Another issue concerning the classifiers is that, after training them on the constructed datasets, they quickly learn to associate individually unharmful words, such as “woman” or “black”, with hate speech. However, this was observed only while testing very short sample sentences individually (<10 words). As a matter of fact, just like in Davidson (2017), the trained classifiers tend to produce a higher rate of false negatives than false positives, i.e., the hate speech contained in a post tends to be underestimated. This is rather evident when looking at the recall and precision metrics in Table 8 (see Appendix 2), as the former is sensibly lower than the latter for all listed models. A possible way to fix this issue could be to repeat the correction procedure described in section 3.4, filtering instances classified as unhateful based on whether they contain specific keywords that are likely to be used in hateful messages. The filtered instances would thus be reclassified correctly and appended to the training dataset, which would later be used to train new and hopefully improved classifiers.

As far as the adopted VSM is concerned, the one employed in this study did outperform fastText’s 600B-token pre-trained word vectors for the specific use-case at hand; however, 104M tokens are hardly enough for training a satisfactory VSM. In future studies, the corpus used to train the VSM could therefore be expanded by scraping content from other similar websites and by integrating data from incel subreddits via Reddit’s Pushshift API.

Regarding the statistical relationship displayed in Figure 1, a similar study could be carried out by taking into account racism and misogyny individually and running a separate statistical test for each of them. Given a much larger dataset, this could also be done for all other previously mentioned classes. Additionally, the tests could potentially be carried out after discarding the most extreme outliers, as Figure 1 seems to show a large quantity of them.

Finally, as previously mentioned at the end of section 5, the study could be expanded to verify whether users in general become more likely to post hateful content in a thread whenever the OP is hateful, or if an OP containing hate speech simply attracts users who post hateful content more frequently than others on average. If the former hypothesis were to be confirmed, the importance of moderating hateful threads as soon as possible would become all the more significant, as it would indicate a capacity of the OP to influence the language of other users.

Appendix 1

HS	Type	Username	Text	Dataset	Thread	Post number	Op index	Agree check	HS thread count	HS thread ratio	Post length
0	0	Lv99_BixNood	View: https://old.reddit.com/r/dating_advice/comments/w8r2ya/guy_lied_about_his_height_pretended_to_be_shorter/ View: https://old.reddit.com/r/dating_advice/comments/w8r2ya/guy_lied_about_his_height_pretended_to_be_shorter/ihrq01v/ View: https://old.reddit.com/r/dating_advice/comments/w8r2ya/guy_lied_about_his_height_pretended_to_be_shorter/ihr2qh7/ If it was a 5'5 manlet pretending he's 6'2 they would've blown him to shreds for lying and deceiving women, tallfags can't lose	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	0	0	1	8	0.2162	37
1	2	Deadgeneration	Being mogged by tallchad faggots makes me angery	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	1	0	1			8
0	0	ThreeInchesLong	These same reddit cucks will be whinging like there's no tomorrow when they see a picture of someone in 1952 doing blackface. These hypocrites cannot make independent judgements at all, and only react the way the soy media tells them to. That guy should have known better than to use manlet heights as a costume without experiencing the dicrimination manlets face.	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	2	0	1			65
0	0	Rice Rice Baby	Heightpocracy at it again	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	3	0	1			4
1	0	SlayerSlayer	you can't make this shit up. AT THE END OF THE DAY, HE IS A LIAR, A FUCKING LIAR. HE FALSELY PRESENTED HIMSELF AS SOMEONE WHO HE IS NOT. WHY CANT THESE FUCKING TOILETS HOLD TALLFAGS TO THE SAME STANDARDS AS MANLETS	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	4	0	0			47
0	0	Gamblord	It's like able-bodied man using a wheelchair. Why?	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	5	0	1			10
1	1	Zhou Chang-Xing	6'2 is Turbomanlet height in today's society, major cope from the toilet. He's probably around 6'7~6'9 but she just calls it 6'2~6'4 because she can't tell height for shit.	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	6	0	1			46
0	0	Ehwhatever	You and the heightpill threads Hypocrisy aside, you know he's gotta be good looking at 5'5 and still getting matches. I didn't want to think my personality was that bad but the evidence has become irrefutable. Or maybe it was his tallfag genes that subconsciously triggered her tingles despite him lying about his height. You can roughly gauge someone's height from a pic so she could be full of shit or not even be aware of the fact.	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	7	0	1			85
0	0	MasterRaceRice	Having a long femur bone equals to a good personality	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	8	0	1			10
0	0	LeFrenchCel	Tallfags problem	NOFILTER_CRONODESC_no_markers_1_100_wUsernames_merged_file.txt	!s1	9	0	1			2

Table 5: Head of the dataset extracted from incels.is

Appendix 2

foid worship	racepill	hapa	slut
gold digger	currycel	foid	cunt
gold diggers	cumskin	femoid	jb
jbw	blackcel	whore	normie
ricecel	noodlewhore	landwhale	chad
whitecel	curryland	roastie	chadlite
shitskin	deathnik	gf	gigachad
slavs	deathnic	Stacie	tyrone
curry	aryan	Stacy	normalfag

Table 6: List of terms characteristic of incel lexicon

Singular synonym	**Plural synonym**	**Classification**
cunt	cunts	misogyny
femoid	femoids	misogyny
foid	foids	misogyny
gold digger	gold diggers	misogyny
landwhale	landwhales	misogyny
noodlewhore	noodlewhores	misogyny
roastie	roasties	misogyny
slut	sluts	misogyny
whore	whores	misogyny
cumskin	cumskins	racism
curry	curries	racism
currycel	currycels	racism
deathnic	deathnics	racism
deathnik	deathniks	racism
ricecel	ricecels	racism
shitskin	shitskins	racism
Tyrone	Tyrones	racism

Table 7: Offensive terms typical of incel lexicon used to augment the dataset via substitution

Model	**Binary**				**3-Class**			
Training dataset	**Non-augmented (4,200 instances)**		**Augmented (5,006 instances)**		**Non-augmented (4,200 instances)**		**Augmented (7,509 instances)**	
Metric	Avg.	St. dev.	Avg.	St. dev.	Avg.	St. dev.	Avg.	St. dev.
Loss	0.244	0.048	0.235	0.034	0.272	0.037	0.218	0.051
Accuracy	0.899	0.027	0.906	0.016	0.901	0.018	0.921	0.016
F_1	0.872	0.044	0.902	0.019	0.900	0.017	0.922	0.016
Precision	0.895	0.063	0.914	0.031	0.911	0.016	0.929	0.016
Recall	0.868	0.082	0.898	0.046	0.889	0.018	0.916	0.016

Table 8: 10-fold cross validation data for different classifiers

References

Basile, V. et al. 2019. 'SemEval 2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter', in *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*.

Blommaert, J. 2018. 'Online-offline modes of identity and community: Elliot Rodger's twisted world of masculine victimhood', in Hoondert, M., Mutsaers, P. and Arfman, W. (eds.), *Cultural practices of victimhood*. Abdingdon: Routledge, pp. 193-213.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. 2017. 'Automated Hate Speech Detection and the Problem of Offensive Language', in *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 512-515).

Farrell, T., Araque, O., Fernandez, M. and Alani, H. 2020. 'On the use of Jargon and Word Embeddings to Explore Subculture within the Reddit's Manosphere', in *12th ACM Conference on Web Science*. New York: Association for Computing Machinery, pp. 221-230.

Gothard, K.C. 2021. *The Incel Lexicon: Deciphering the Emergent Cryptolect of a Global Misogynistic Community*. Burlington: The University of Vermont and State Agricultural College.

Hapke, H., Howard, C. and Lane, H. 2019. *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon and Schuster.

Jigsaw. 2019. *Toxic comment classification challenge: Identify and classify toxic online comments*. Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P. and Mukherjee, A. 2021. 'Hatexplain: A benchmark dataset for explainable hate speech detection', in *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 17, pp. 14867-14875).

Mollas, I., Chrysopoulou, Z., Karlos, S. and Tsoumakas, G. 2022. 'ETHOS: a multi-label hate speech detection dataset', in *Complex & Intelligent Systems*, pp. 1-16.

Nagle, A. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. Winchester: Zero Books.

Pelzer, B., Kaati, L., Cohen, K. and Fernquist, J. 2021. 'Toxic language in online incel communities', in *SN Social Sciences*, *1*(8), pp. 1-22.

Zhou, L., Caines, A., Pete, I. and Hutchings, A. 2022. 'Automated hate speech detection and span extraction in underground hacking and extremist forums', in *Natural Language Engineering*. Cambridge University Press, pp. 1–28.