Hate Speech Detection

By Group 2: Still Processing…

Ready? If your view was correct, all strippers would be millionaires.
Reality: Strippers are dirt poor sociopaths.

Introduction

Hate Speech

"denial of the values of tolerance, inclusion, diversity and the very essence of the human rights norms and principles"
(UN, Cited May 2022)

"any form of non-acceptable language (profanity) or a targeted offence, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words"
(Zampieri et al., 2019a)

"when left unchecked, expressions of hatred can [. . .] harm social cohesion, peace and development, as it lays the ground for conflicts and tensions, wide scale human rights violations, including atrocity crimes."
(UN, Cited May 2022)

Traditional Model

Ensemble: SGD + LR + Nu-SVM + C-SVM

Neural Model

Majority vote: 5 RoBERTas

Related Research

Related Research

Surface Level Features

Bag of Words: frequency dictionary

TF-IDF: term frequency-inverse document frequency

Word n-grams

Lose syntactic and semantic context

Character n-grams

Retrieve context

(Burnap and Williams, 2016)

Overcomes typing errors & spelling variations

(Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Alorainy et al.) 2018)

Range 1-5

(Alorainy et al., 2018; MacAvaney et al., 2019; Burnap and Williams, 2016)

Related Research

Surface Level Features Speech

Capitalisation
(Burnap and Williams, 2016; MacAvaney et al., 2019)

Interpunction
(Alorainy et al., 2018; MacAvaney et al., 2019)

URL’s, @ mention, hashtags
(Davidson et al., 2017; Gambino and Pirrone, 2020)

Emojis

Related Research

Linguistic Features

Parts-of-speech
(Markov and Daelemans, 2021; Alorainy et al., 2018)

Custom POS tagger

Lemmatization
(Markov and Daelemans, 2021; Markov et al., 2021; Hee et al., 2018)

Related Research

Semantic Feature

Semantic lexicon
(Alorainy et al., 2018; Markov and Daelemans, 2021)

Vader
(Hutto and Gilbert, 2014)

AFINN
(Arup Nielsen, 2011)

Related Research

Traditional Model

Stochastic Gradient Descent Model
(Sharif et al., 2020)

Logistic Regression Model
(Alorainy et al., 2018; Davidson et al., 2017)

Support Vector Machine
(Markov and Daelemans, 2021; Burnap and Williams, 2016; MacAvaney et al., 2019)

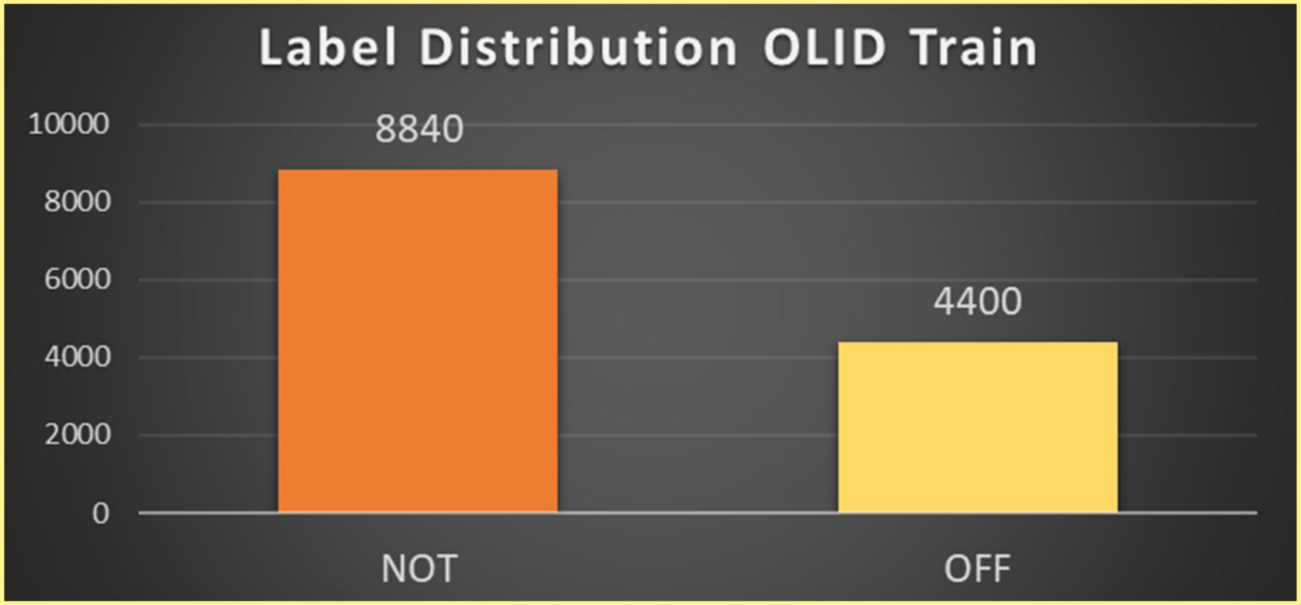
Neural Model

Hard voting: SVM + BERT + RoBERTa
Markov and Daelemans (2019)

Methodology

OLID

English tweets
Two or three annotators
Offensive or not offensive
(Zampieri et al., 2019)

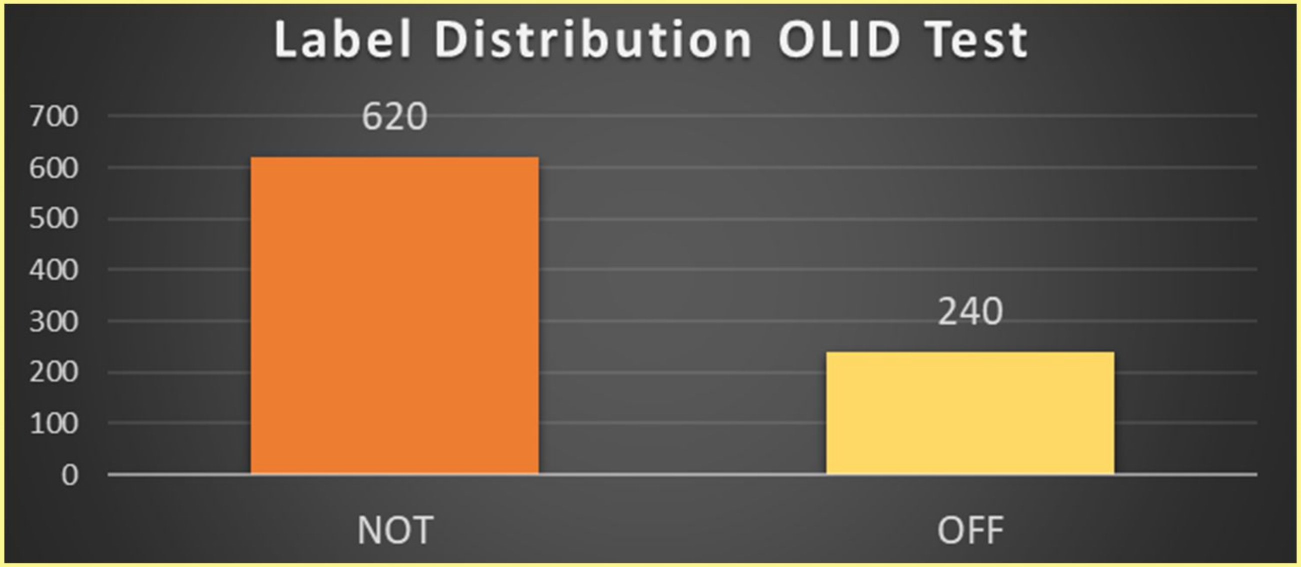


Methodology

Test Data

OLID

English tweets
Two or three annotators
Offensive or not offensive
(Zampieri et al., 2019)

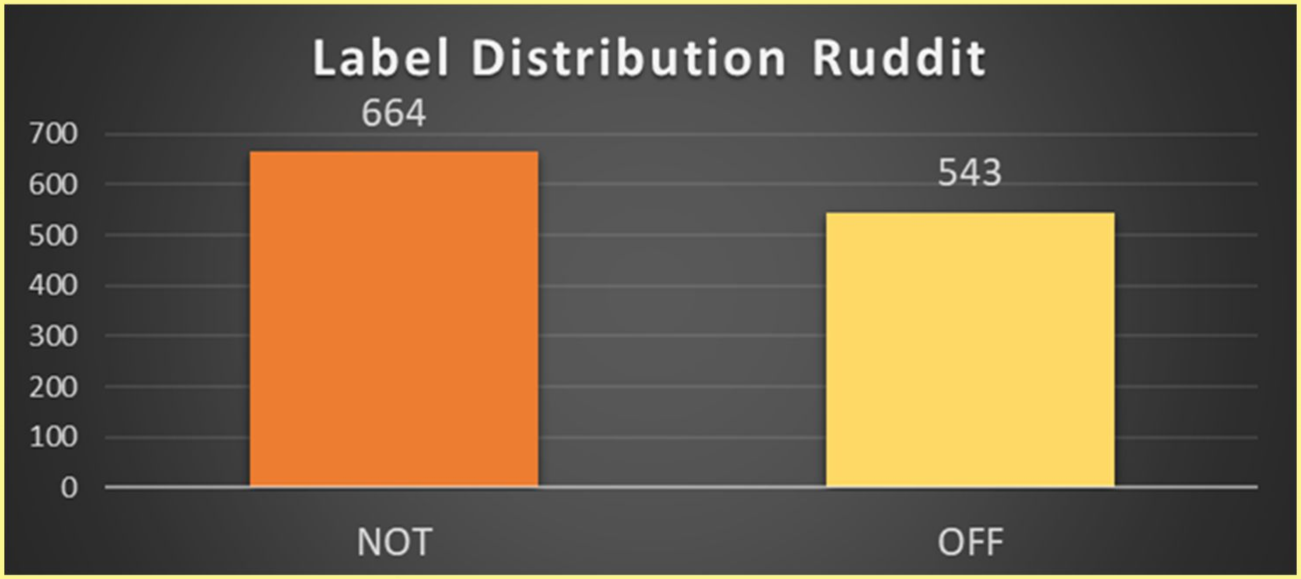


Methodology

Test Data

Ruddit

English Reddit posts
Best-worst scaling
[-1, 1] → [-0.4, 0.4]
(Hada et al., 2021)

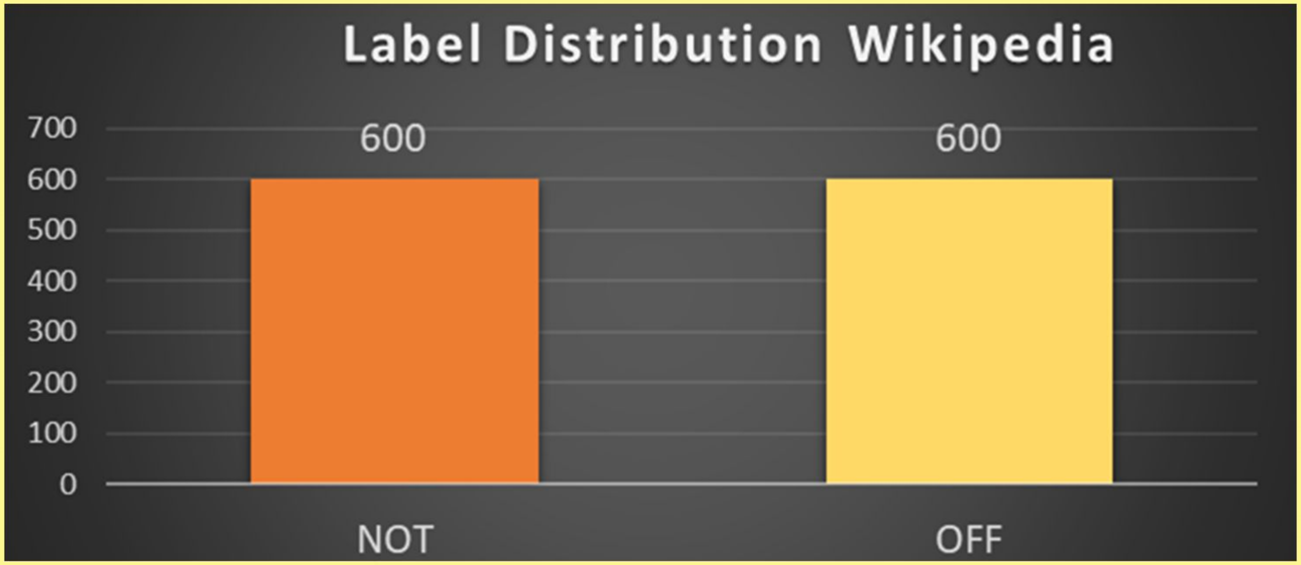


Methodology

Test Data

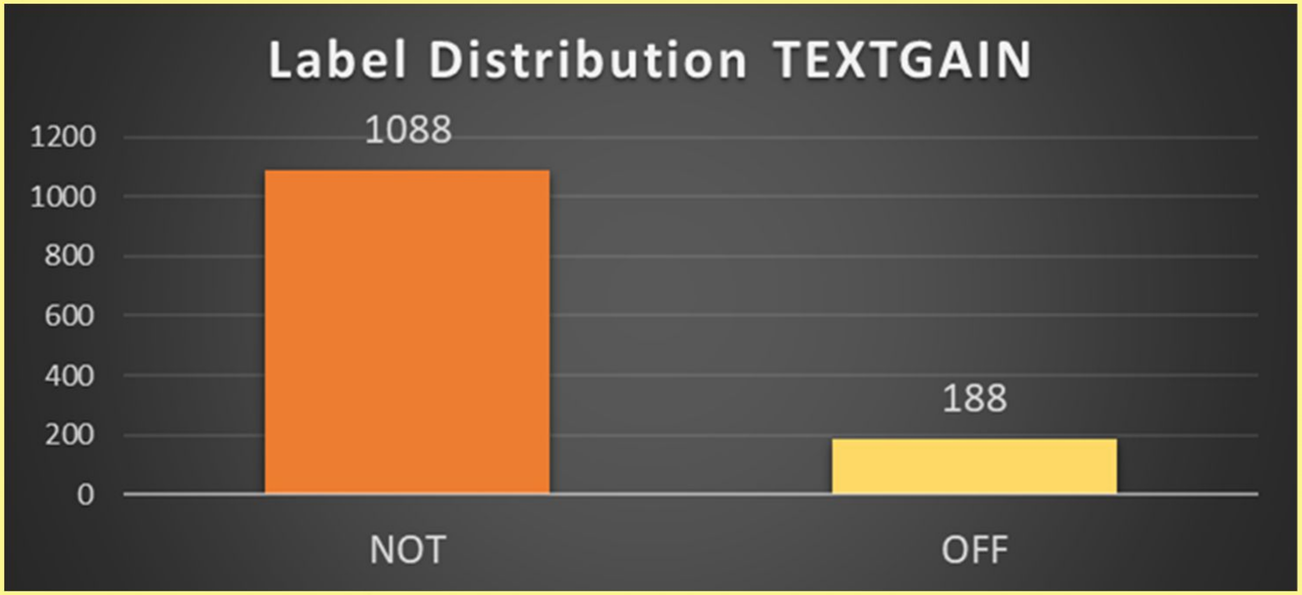
Wikipedia

English Wikipedia comments
Manually annotated
Non-toxic, moderately offensive, severely offensive
(Al, 2018)



TEXTGAIN

English football tweets
Manually labelled as offensive or not offensive



Baselines

	OLID	Ruddit	Wikipedia	Total	TEXTGAIN
Most Frequent Class	0.28	0.31	0.33	**0.30**	0.46
SpaCy BoW-Model	0.70	0.66	0.86	**0.73**	0.54
Bert	0.81	0.70	0.90	**0.81**	0.47
HateBert	0.82	0.68	0.91	**0.81**	0.48

Traditional Classification

Traditional Classification

Ensemble Model

Stochastic Gradient Descent Model (SGD)
Logistic Regression Model (LR)
Linear Support Vector Machine 1 (SVM1)
Linear Support Vector Machine (SVM2)

Random state of 42

Class weights are balanced

Traditional Classification

SGD Model

SGDClassifier(
random_state=42, class_weight='balanced',
early_stopping=True, n_iter_no_change=3,
penalty='elasticnet', loss='log'
alpha=0.001)

Count: Demojised tweet column

Tfidf: POS column (demojised tweets)

Count: Lemma column (demojised tweets)

Afinn

vaderSentiment: ‘pos’, ‘neg’ and ‘compound’

F1 score of 0.71

Traditional Classification

LR Model

LogisticRegressionCV(
random_state=42, class_weight='balanced',
cv = 2, scoring='f1_macro',
penalty = 'l1', solver = 'saga',
n_jobs=-1, verbose=2, multi_class='ovr')

Tfidf: Demojised tweet column (without hashtags)

Count: POS column (demojised tweets without punctuation)

Afinn

vaderSentiment: 'neg'

F1 score of 0.71

Traditional Classification

SVM1 Model

SVC(
random_state=42, class_weight='balanced’,
kernel='linear', C=0.1, verbose = 2,
probability=True)

Tfidf: Demojised tweet column (without punctuation)

Tfidf: POS column (basic tweets)

Afinn

vaderSentiment: ‘pos’ and ‘compound’

F1 score of 0.72

Traditional Classification

SVM2 Model

NuSVC(
random_state=42, class_weight='balanced',
probability=True,
verbose=2)

Tfidf: Demojised tweet column (without hashtags)

Tfidf: POS column (basic tweets)

Afinn

vaderSentiment: 'pos' and 'compound'

Macro Average F1 score of 0.72

Traditional Classification

Ensemble Model

```
VotingClassifier(
estimators=[
('sgd2', sgd_pipe2), ('log1', log_pipe1),
('svm1', svm_pipe), ('svm2', nu_svm_pipe)
],
verbose= 10, voting='soft’,
weights=[3, 3, 2, 3])
```

Soft Voting

Macro Average F1 score of 0.74

Neural Classification

5 RoBERTa Models

Hard Majority Vote

Neural Classification

	# @User	Tokenized	Emoji	Punctuation	Lemmatized
roberta_tweet	√		√	√	
roberta_hashtag_tweet			√	√	
roberta_token_tweet		√	√	√	
roberta_token_demojize		√		√	
roberta_lemma	√		√		√

RoBERTa

```
ClassificationModel(
'roberta', 'roberta-base', num_labels=2,
args=model_args, use_cuda=True)
```

```
model_args.num_train_epochs=6
model_args.train_batch_size=64
model_args.learning_rate=1e-5
model_args.max_seq_length=128
```

Early Stopping

Pickle

Neural Classification

Majority Vote

	twt	tok_tw	tok_d	lem	hsh	sum	mv
0	1	1	1	1	1	5	1
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	1	1	1	1	5	1
6	0	1	1	1	1	4	1
7	1	1	1	1	1	5	1
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Neural Classification

F1-scores for the best performing RoBERTas + Majority Vote

	Accuracy	Macro Average
roberta_token_demojize	0.82	0.80
roberta_token_tweet	0.82	0.80
roberta_hashtag_tweet	0.81	0.79
roberta_lemma	0.80	0.78
roberta_tweet	0.81	0.79
majority vote	0.82	0.80

Neural Classification

Precision and recall for the best performing RoBERTas + Majority Vote

	Precision		Recall	
	OFF	NOT	OFF	NOT
roberta_token_demojize	0.77	0.85	0.69	0.89
roberta_token_tweet	0.75	0.85	0.71	0.87
roberta_hashtag_tweet	0.70	0.87	0.77	0.83
roberta_lemma	0.69	0.87	0.76	0.82
roberta_tweet	0.70	0.87	0.77	0.83
majority vote	0.73	0.86	0.74	0.86

Neural Classification

Majority Vote

	twt	tok_tw	tok_d	lem	hsh	sum	mv
0	1	1	1	1	1	5	1
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	1	1	1	1	5	1
6	0	1	1	1	1	4	1
7	1	1	1	1	1	5	1
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Results and Error Analysis

Results

	Ruddit	OLID	Wikipedia	Total	TEXTGAIN
Traditional Ensemble	0.713	0.803	0.874	0.76	0.434
Neural RoBERTa	0.661	0.753	0.912	0.808	0.504

15 per cent offensive data

Similar format to train data

Similar format to train data

Clear division between offensive and non-offensive

Baselines

	OLID	Ruddit	Wikipedia	Total	TEXTGAIN
Most Frequent Class	0.28	0.31	0.33	0.30	0.46
SpaCy BoW-Model	0.70	0.66	0.86	0.73	0.54
Bert	0.81	0.70	0.90	0.81	0.47
HateBert	0.82	0.68	0.91	0.81	0.48
Traditional Ensemble	**0.803**	**0.713**	**0.874**	**0.76**	**0.434**
Neural RoBERTa	0.753	0.661	0.912	0.808	0.504

Results and Error Analysis

Baselines

	OLID	Ruddit	Wikipedia	Total	TEXTGAIN
Most Frequent Class	0.28	0.31	0.33	0.30	0.46
SpaCy BoW-Model	0.70	0.66	0.86	0.73	0.54
Bert	0.81	0.70	0.90	0.81	0.47
HateBert	0.82	0.68	0.91	0.81	0.48
Traditional Ensemble	**0.803**	**0.713**	**0.874**	**0.76**	**0.434**
Neural RoBERTa	**0.753**	**0.661**	**0.912**	**0.808**	**0.504**

Discussion

Discussion

Vectorizers

Tweets, lemmas
POS

N-grams

Character n-grams
Word n-grams

....

Linguistic Features

POS
Lemmas

....

Sentiment Scores

Afinn
Vader

Conclusion

NLP_Student42 @NLP_Student42 · 4 s

Wow, the group Still Processing presented so well that I don't have any questions, do you?

Works Cited

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media, 8(1):216–225.

Clayton Hutto. Cited May 2022. Vader-sentiment-analysis. https://github.com/cjhutto/vaderSentiment.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2018. Automatic detection of cyberbullying in social media text. PLoS ONE, 13(10):e0203794.

Finn Arup Nielsen. 2011. A new anew: Evaluation ˚ of a word list for sentiment analysis in microblogs. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, pages 93–98.

Giuseppe Gambino and Roberto Pirrone. 2020. Chilab@ haspeede 2: Enhancing hate speech detection with part-of-speech tagging. EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020, pages 165–170

Works Cited

Ilia Markov, Nikola Ljubesiˇ c, Darja Fi ´ ser, and Wal- ˇ ter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual crossdomain hate speech detection. Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 1–11.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 17–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171–4186.

Jigsaw/Conversation AI. 2018. Toxic comment classification challenge. https://www.kaggle.com/competitions/jigsawtoxic-comment-classification-challenge/overview.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1415–1420.

Works Cited

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

Omar Sharif, Mohammed Moshiul Hoque, A. S. M. Kayes, and Raza Nowrozy et al. 2020. Detecting suspicious texts using machine learning techniques. Applied Sciences, 10(6527):1–360.

Paula Fortuna and Sergio Nunes. 2018. A survey on ´ automatic detection of hate speech in text. ACM Computing Surveys, 51(4):85:1–85:30.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data Science, 5(11):1–15.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for english reddit comments. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2700–2717.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. PLoS ONE, 14(8):e0221152.

Works Cited

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 17–25.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1):1–4.

UN. Cited May 2022. Impact and prevention: Why tackle hate speech? https://www.un.org/en/hatespeech/impact-and-prevention/why-tackle-hatespeech.

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew Williams. 2018. The enemy among us: Detecting hate speech with threats based ’othering’ language embeddings. ArXiv, page 1801.07495.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv, page abs/1907.11692.