data

MDPI

Article

Using Twitter to Detect Hate Crimes and Their Motivations: The HateMotiv Corpus

Noha Alnazzawi

Department of Computer Science and Engineering, Yanbu Industrial College, Royal Commission for Jubail and Yanbu, Yanbu Industrial City 41912, Saudi Arabia; alnazzawin@rcyci.edu.sa

Abstract: With the rapidly increasing use of social media platforms, much of our lives is spent online. Despite the great advantages of using social media, unfortunately, the spread of hate, cyberbullying, harassment, and trolling can be very common online. Many extremists use social media platforms to communicate their messages of hatred and spread violence, which may result in serious psychological consequences and even contribute to real-world violence. Thus, the aim of this research was to build the HateMotiv corpus, a freely available dataset that is annotated for types of hate crimes and the motivation behind committing them. The dataset was developed using Twitter as an example of social media platforms and could provide the research community with a very unique, novel, and reliable dataset. The dataset is unique as a consequence of its topic-specific nature and its detailed annotation. The corpus was annotated by two annotators who are experts in annotation based on unified guidelines, so they were able to produce an annotation of a high standard with F-scores for the agreement rate as high as 0.66 and 0.71 for type and motivation labels of hate crimes, respectively.

Keywords: text mining; corpus construction; annotation guidelines; hate crime motivation

Citation: Alnazzawi, N. Using Twitter to Detect Hate Crimes and Their Motivations: The HateMotiv Corpus. *Data* **2022**, *7*, 69. https://doi.org/10.3390/data7060069

Academic Editor: Giuseppe Ciaburro

Received: 15 March 2022
Accepted: 23 May 2022
Published: 24 May 2022

1. Introduction

Over the last decade, social media channels, such as Facebook, Twitter, etc., have become a normal part of daily life for many [1]. Twitter is an extremely popular social media platform and has been growing rapidly since its creation in 2006. It has provided a place where people can interact with each other and maintain social ties. People use Twitter to share their daily activities with their contacts, which makes it both a valuable tool and a great data source for research.

Although Twitter has become a useful way to spread information, it has introduced new modes of social discourse and produced antagonistic content that adds to the dissemination of prejudice, false information, and hostility towards people on an unprecedented scale, including refugees and other marginalized groups [2,3]. The content involving hateful messages ranges from verbal aggression to cyberbullying, offensive language, hate speech, and incitement to crime. This verbal aggression on social media platforms can cause more harm than traditional bullying because it allows users to adopt an alias [4]. Unfortunately, the reach and extent of online aggression have given such content considerable power and influence that can affect anyone, irrespective of their status, identity, and location. Thus, such incidents are not a trivial annoyance and instead have now entered the realms of criminal activity, which can affect anyone and everyone. It has been noted that hateful online content can create both mental and psychological anguish for users of social media platforms, has led some to deactivate their accounts and, in the worst cases, to commit suicide [5].

Research has shown that online victimization in the form of hatred is just one part of a broader process of harm that commences on social media platforms social media. For example, suspects in many recent hate-related terror attacks have been shown to have a comprehensive history of posting hate-related comments on social media, which would

indicate that social media could well have contributed to their radicalization [6,7]. Other studies have shown that there is a correlation between hate tweets focused on race and religion with racially and religiously aggravated offenses that occur offline in the same location (2019; 2019) [8].

A hate crime is a criminal offense that is motivated, in part or in whole, by the perpetrator's attitude towards race, religion, disability, sexual orientation, ethnicity, gender, or gender identity (2015; 2018). According to recent statistics from the FBI, the number of hate crimes is on the increase, and in 2018, the number of personal attacks motivated by personal bias hit a 16-year high (2018; 2019). In the UK, the number of hate crimes has more than doubled since 2013, and according to the Home Office data, 1605 hate crimes were identified as possible online offenses between 2017 and 2018 (2018) [9]. Europe has also witnessed a significant increase in negative xenophobic, nationalist, Islamophobic, racist, and antisemitic opinions. Their effect is not just restricted to hostile rhetoric and leads to actual crimes against groups and individuals [3]. Therefore, it is important for preventive measures to be taken to deal with abusive and aggressive behavior online [10].

While most microblogging websites and social networks have banned the use of hateful language, the volume of work posted on them makes it virtually impossible to moderate everything posted. As a consequence, the need has arisen for there to be a form of automatic speech detection that can identify and remove hate speech. Thus, the dissemination of hate on Twitter constitutes a social emergency with real-life individual and social consequences [3]. Thus, the aim of this research is to build a semantically annotated corpus called HateMotiv. This was done using Twitter data with a focus on the identification of mentions denoting types of hate crimes (physical assault, verbal abuse, and incitement to hatred) and the motivation behind committing such crimes as classified by the FBI (race or ethnicity, religion, sexism, and disability). Such information could help to stimulate research on the automatic extraction of hate crime types and the motivations behind them from social media text. We also underline that we are not stating that there is any direct relationship between the use of hate speech online and hate-based actions in the physical (real) world. However, the findings of this study are pertinent to improving discrimination surveillance and mitigation efforts [11].

This research aims to use Twitter as an example of a social media platform to study hate crimes and the motivation(s) behind them. The contribution of this research is three-fold:

1. To build annotated datasets for the mentions of hate crime types and the motivation(s) behind committing such hate crimes.
2. The corpus is freely available to be used by the research community and will serve as a resource to train and evaluate text-mining (TM) tools, which in turn, can be used to automatically extract mentions of hate crimes and motivation. The TM tools can be used for the prediction of hate crime events and for better surveillance and mitigation efforts against discrimination. To the best of our knowledge, this is the first study pertaining to the investigation of Twitter data for hate crimes and the motivation(s) behind them.
3. To create a hate crime and motivation vocabulary list specifically relating to hate crimes and their motivation(s). The vocabulary list is freely available and can be used as a resource for TM techniques and named entity recognition methods.

2. Related Work

With there being such growth and proliferation of hate speech on social media platforms, there has also been considerable research on the identification of offensive language such as hate speech [12–17], cyberbullying [18,19], aggression [10], and toxic comments [20,21]. Only a nominal volume of work has specifically been centered on the detection of personal bias against specific groups that leads to hate crimes, and these efforts were limited in terms of where and when hate crimes are committed [13,22] or were about finding the associations between social media discrimination and offline hate crimes in 100 cities in the United States. However, the latter study was limited to discrimination

related to race, ethnicity, or national origin, and the classification of tweets was made at the sentence level [9,11].

Table 1 is representative of a sample of sets of public data which are available for evaluation- and training-related hate speech and offensive language training. Most of the datasets include two classes to label the text as offensive or not offensive, with a small percentage identifying racist or sexist content. Furthermore, all the mentioned datasets are annotated at the tweet level (e.g., the tweet is labeled as either racist, sexist, or both) without annotating the specific word or expression that denotes the label. This makes the dataset more suitable for developing binary classification methods as opposed to simply helping with the extraction of detailed information about any text. Annotating text at the "mentions" level gives more information about it and enhances the use of linguistic and syntactic features that can be used to train and evaluate ML methods.

Table 1. Characteristics of some of the popular hate-detection datasets.

Dataset	Source	Label	Domain	Annotation Type	Size
Davidson et al. [15]	Twitter	hate offensive neither	Hate speech	Tweet-level	25,000
Waseem and Hovy [23]	Twitter	Racism Sexism both neither	Hate speech	Tweet level	16,914
TRAC-1 [24]	Twitter and Facebook	Non-aggressive Covertly aggressive Overtly aggressive	Trolling, aggression, and cyberbullying	Tweet level	15,000
HatEval [25]	Twitter	hate not hate aggressive not aggressive	Hate speech against immigrants or women	Tweet level	19,600
OLID [26,27]	Twitter	Offensive Not offensive	Offensive language	Tweet level	14,100

As far as we are aware, none of the previous works have studied the problem of hate crimes as a whole with the inclusion of different personal biases that influence and motivate the occurrence of crimes regardless of if the tweet is written by the main person who committed the crime or by another party. This is mainly because there is no available annotated corpus for hate crime information. Furthermore, there is no comprehensive dictionary covering hate crime-related terms. Developing TM tools that can automatically extract hate crime-related information relies upon textual corpora where pertinent information has been specifically annotated by those experienced in the field. Building publicly available, curated datasets that identify hate crime mentions and the personal bias that motivates such crimes is essential for training machine learning (ML) techniques (2019) [26,27] and raising the bar for the effective and systematic assessment of any novel methodologies. There are several ways in which our own research varies from previous work on the subject. First, our work is devoted in particular to the detection and classification of hate crimes and the motivations at a detailed level. Second, the annotations in our work are made at the mention level of words or sequences of words that denote a hate crime (e.g., physical, verbal, etc.) and the personal bias that motivates committing the hate crime (e.g., racism, sexism, etc.). As new words are introduced frequently in expressing hateful content, annotating tweets at the mentions level will facilitate the generation of a lexical resource and the automatic augmentation of existing lexical resources. Therefore, we used our corpus to create a hate crime vocabulary list, which includes vocabulary about different hate crimes classified as different types and the personal bias motivating the crimes.

3. Material and Methods

To carry out experiments on the automatic detection of online hate crime, it is critical one has access to labeled corpora. As no benchmark exists for such a corpora for hate crimes, researchers have been forced to obtain and class data specifically for themselves [2]. As far as we know, no previous studies have been specifically centered on hate crime detection and the motivation behind them, which does not help when it comes to reaching any relevant conclusions about the most common motivations behind hatred and hate crimes. Furthermore, it is very difficult to provide recommendations on how to control bias and remove the motivation for hate-related crimes.

3.1. Corpus Construction

We retrieved examples for the HateMotiv corpus from Twitter using the TweetScraper tool. Tweets were collected from a nine-year period (between 1 January 2010, and 30 December 2019). Twitter includes a very high number of tweets related to the topic of hate crimes, and the presence of particular words, such as "hate crime", does not necessarily indicate that a tweet is related to committing a hate crime. However, the hashtag convention is predominantly utilized on Twitter as a means to connect a user's comments and points of view to an event [28]. Therefore, we used the "Hashtagify" tool (https://hashtagify.me/) [29] to find the most popular hashtags related to hate crimes. The following list of hashtags was used to collect the relevant tweets: "hate crime", "racist", "racism", "Islamophobia", "Islamophobic", "sexism", "disability", "transgender", "antisemitism", "misogyny", and "disabled". We noticed that the hashtags that were highly related to hate crime terms matched the hate crime classification of the FBI, so we used these hashtags as keywords to crawl relevant tweets. The keywords were chosen by the judge of this annotation process, who is an English teacher and who has considerable experience in annotation, and which resulted in 23,179 tweets that contained mentions of the listed hashtags in the query. Due to the cost of manual annotations in terms of time and money, the tweets were further filtered, and we randomly selected 5000 tweets to be considered for annotation by the annotators.

3.2. Annotation Process

The annotation was done through the COGITO Tech (LLC) service (https://www.cogitotech.com/about-us) [30], with which each tweet was annotated by two annotators, who are English language instructors. The tweets were annotated for mentions related to the type of hate crime and the motivation behind committing the crime through the use of the same set of guidelines applicable to the annotation. The annotation included marking up all entity mentions in the corpus related to four hate crime types and motivations, as shown in Table 2. Annotators were supported by a concise set of guidelines together with regular meetings with the judge of this annotation process to discuss any issues which arose during the annotation process and to resolve all problems and discrepancies.

Table 2. Annotated entity classes for HateMotiv corpus.

Class Type	Description
Hate crime type	Hate crime type refers to a type of crime classified by the FBI as one of the following: • Physical assault • Verbal abuse • Incitement to hatred
Motivation	Motivation refers to the reason for committing a hate crime, such as bias related to the following: • Racism • Religion • Disability • Sexism • Unknown

As shown in Figure 1, the most common hate crime type according to the HateMotiv corpus is physical assault, followed by incitement to hatred. The least common type of hate crime reported on Twitter is verbal abuse. Figures 2–4 show the distribution of the motivations behind committing different types of hate crimes in HateMotiv corpus. Bias

against different races and ethnicities contributed the most to committing different types of hate crimes, and it seems that people do not tolerate variation and variability in terms of different skin colors and nationalities. Disability and people's attitudes toward disabled people were the least common motivation behind committing hate crimes.

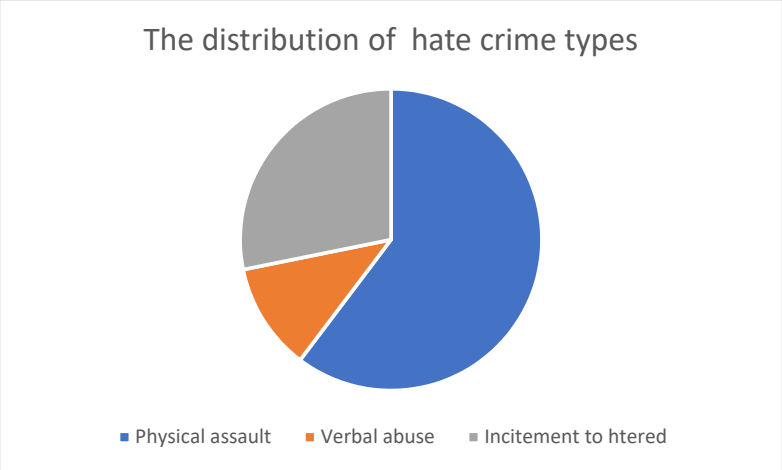

Figure 1. Distribution of the entity types in the HateMotiv corpus.

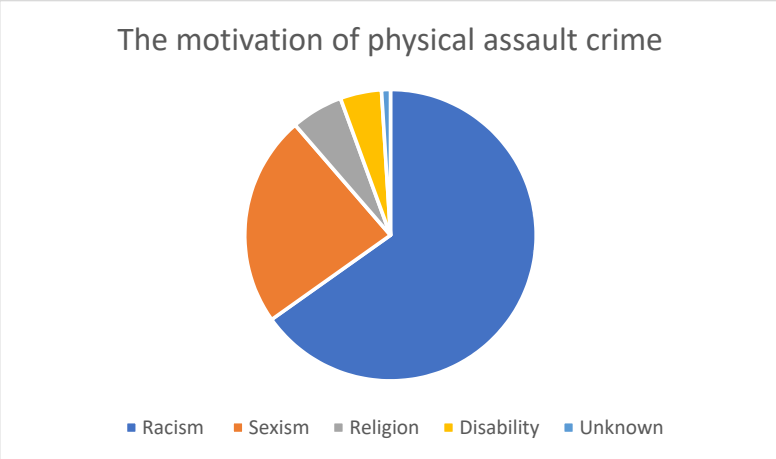

Figure 2. Distribution of the motivation for physical assault crime.

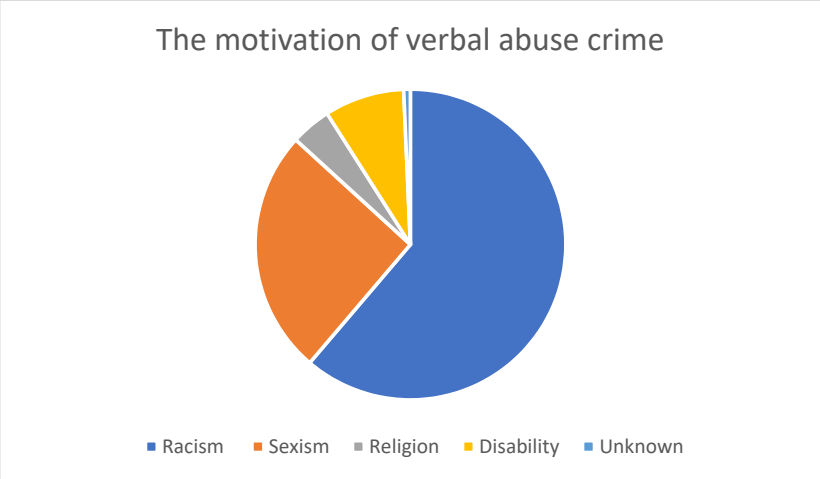

Figure 3. Distribution of the motivation for verbal abuse crime.

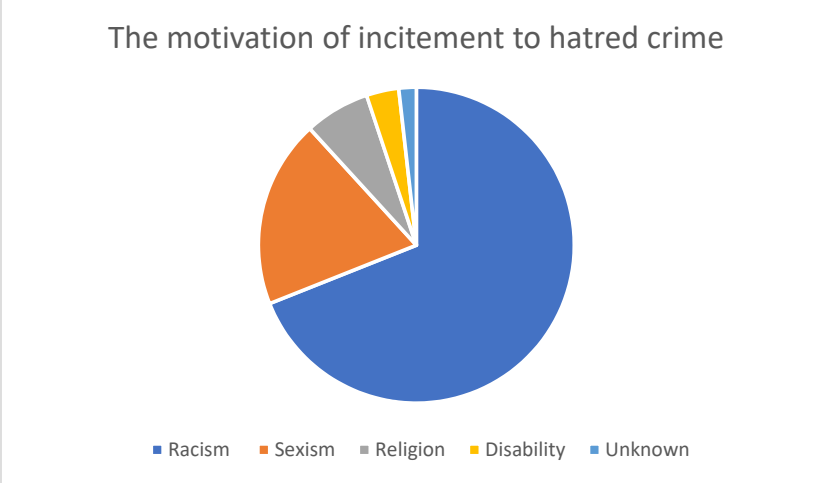

Figure 4. Distribution of the motivation for incitement to hatred crime.

Sexism and discrimination against other genders came next after racism as the second most common motivation for different types of hate crime. It should also be mentioned that hate crimes are sometimes committed without a clear motivation or reason and are just based on personal bias or mental problems. This is reflected in the HateMotiv corpus with the "unknown" motivation label. As shown in the figures, physical assault was the most common form of hate crime perpetrated without a clear reason. However, the percentage of crimes committed for unknown reason is very low compared with other motivations (only 0.011% of all hate crime types).

4. Results and Discussion

The reliability of human annotation is very important for ensuring both that ML algorithms can accurately learn the characteristics of tweets that discuss hate crimes and to provide an upper bound for the expected performance [23]. We ensured the superior quality of the generated corpus by measuring the inter-annotator agreement (IAA). If the IAA score is high, this proves that one has a reliable corpus that will be suitable for the training and testing of TM techniques and ML models. Following a number of previous studies [31–35], we used the F-score to calculate the IAA because it is the same regardless of the set of annotations utilized as the gold standard [32,36]. The F-score is calculated thus:

$$\text{F-score} = 2 \times (\text{Precision} \times \text{Recall})/\text{Precision} + \text{Recall}$$

The annotations produced by the first annotator were considered the "gold standard", and the total number of correct annotations corresponded to the number of annotated entities produced by that annotator. Based on this, we calculated the IAA based on precision, recall, and F-score. Precision (P) is the percentage of correct positive annotated entities produced by the second annotator, comparing them to the gold standard. The precision equals the ratio between the number of true positive (TP) entities and the total number of annotated entities from the second annotator, as per:

$$P = TP/TP + FP$$

Recall (R) is the percentage of correct annotations recognized by the second annotator and is calculated as the ratio between the TP and the total number of annotations in the gold standard according to the following:

$$R = TP/TP + FN$$

Tables 3 and 4 present the agreement statistics in terms of precision (P), recall®, and F-score for the annotation of hate crime types and their motivations. Overall, the precision

for hate crime types and motivation was lower than the recall. The reason for this is that the second annotator annotated mentions of incidents that are not necessarily considered hate crimes or motivations according to the annotation guidelines for this task. Moreover, the second annotator annotated every mention of hate crimes, resulting in annotating the same hate crimes more than once in a tweet. For example, consider the following tweet:

Table 3. IAA for hate crime types.

Hate Crime Types	P	R	F-Score
Physical assault	0.629275	0.858253	0.72614
Verbal abuse	0.592703	0.917593	0.720203
Incitement to hatred	0.624517	0.822034	0.709791
Macro-averaged	0.86596	0.718712	0.718712

Table 4. IAA for hate crime motivation.

Motivation	P	R	F-Score
Racism	0.617124	0.724858	0.666667
Religion	0.813397	0.564784	0.666667
Sexism	0.6272	0.711434	0.666667
Disability	0.665025	0.668317	0.666667
Unknown	0.399225	0.664516	0.498789
Macro-averaged	0.624394	0.666782	0.633091

> "She was killed by her brother in a hate crime. There is little doubt killing her because of her gender".

The second annotator annotated both "killed" and "killing" as physical-assault hate crimes, while the correct annotation according to the guidelines and the annotation produced by the first annotator is to only annotate the first mention of the hate crime "killed" as a physical-assault hate crime. Another example is the following tweet: "A group of whites attacked a black and Muslim man. This is racism against black people and a hate crime". The correct annotation as per the gold standard is to annotate "attacked" as a physical assault-type hate crime with the following motivations "black" as racism and "Muslim" as religion. However, the second annotator annotated extra unnecessary text e.g., the second mention of black, which should not have been included in the gold standard as a motivation since the more specific mention and description of the racism motivation "black" had already been annotated.

On the other hand, the recall was high, which means that the second annotator produced the same set of annotations compared with the gold standard.

To show the importance of our generated dataset, we compared the HateMotiv corpus with other publicly available datasets in the same domain of hate crimes, which are reported in Table 1. In particular, we compared our corpus with the dataset that used Twitter as a social media source, and they share some of the characteristics with our proposed corpus, e.g., they have been annotated for similar classes related to hate crime. As shown in Table 5, the HateMotiv corpus differs from the existing datasets in terms of the very specific domain, which is hate crimes and the motivation behind committing them, and also the level of annotations at mention level. However, the other datasets include binary or multi–class annotation at tweet level. The results of the annotation are satisfactory and are measured in terms of F-score at 0.66 and 0.71 for type and motivation labels of hate crimes, respectively.

Table 5. Comparison of HateMotiv corpus with other comparable corpora.

Dataset	Size	Text Genre	Topic	Labels	Annotation Level	#of Annotators	Agreement Measurement
HateMotiv	5000	Twitter	Hate crimes and their motivation	Hate crime types: - Physical assault - Verbal abuse - Incitement to hatred - Other Motivation: - Racism - Religion - Disability - Sexism - Unknown	mention	2 annotators	F-score 0.66 for hate crimes type 0.71 for the motivation of hate crimes
Waseem and Hovy [23]	16,914	Twitter	Hate speech	Racism Sexism Both Neither	Tweet	Crowdsource workers	Cohen's kappa = 0.57
Davidson et al. [15]	25,000	Twitter	Hate speech	Hate Offensive Neither	Tweet	Crowdsource workers	InterCoder-agreement score 92%
HatEval [25]	19,600	Twitter	Hate speech against immigrants or women	Hate Not hate Aggressive Not aggressive	Tweet	Crowdsource workers	Average IAA at 0.75
OLID [26,27]	14,100	Twitter	Offensive language	Offensive Not offensive	Tweet	Crowdsource workers	IAA at 60%

As a result of this research, upon the creation of the HateMotiv corpus, we were able to create a hate crime and motivation vocabulary list. As far as we know, there is no resource devoted specifically to hate crimes and their motivations. The vocabulary list is freely available and can be used as a resource for TM techniques and named entity recognition methods

5. Conclusions

We have presented the HateMotiv corpus, a new dataset with annotations of types and motivations of hate crimes. To the best of our knowledge, this is the first dataset to contain annotations of hate crime types and to target offenses on social media. The results could open up new directions for interesting research. The dataset is freely available to the research community and has the potential to stimulate investigations of the automatic detection and prediction of hate crimes and their motivation. We have also shared a vocabulary list derived from the results, which can be used as a reference for hate crimes to augment existing dictionaries or to create new specialized dictionaries. Detailed information about the motivations behind hate crimes has the potential to help control and mitigate the motivations of hate crimes in an attempt to control personal bias and reduce the number of crimes conducted based on such bias against others. In the future, we plan to use state-of-the-art ML and deep-learning techniques (e.g., neural networks) to train models to extract and recognize hate crime mentions and their motivations on social media platforms.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is freely available on Kaggle at: https://www.kaggle.com/datasets/nohaalnazzawi/the-hatemotiv-corpus.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Wang, W.; Chen, L.; Thirunarayan, K.; Sheth, A.P. Cursing in English on Twitter. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, MD, USA, 15–19 February 2014; ACM: New York, NY, USA, 2014; pp. 415–425.
2. Alorainy, W.; Burnap, P.; Liu, H.; Javed, A.; Williams, M.L. Suspended Accounts: A Source of Tweets with Disgust and Anger Emotions for Augmenting Hate Speech Data Sample. In Proceedings of the 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, China, 15–18 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 581–586.
3. Bojarska, K. *The Dynamics of Hate Speech and Counter Speech in the Social Media Summary of Scientific Research*; Centre for Internet and Human Rights: Frankfurt, Germany, 2018.
4. Sticca, F. *Bullying Goes Online: Definition, Risk Factors, Consequences, and Prevention of (Cyber) Bullying*; University of Zurich: Zürich, Switzerland, 2013.
5. Hinduja, S.; Patchin, J.W. Connecting adolescent suicide to the severity of bullying and cyberbullying. *J. Sch. Violence* **2019**, *18*, 333–346. [CrossRef]
6. Robertson, C.; Mele, C.; Tavernise, S. 11 Killed in Synagogue Massacre; Suspect Charged with 29 Counts. *The New York Times*. Available online: https://www.nytimes.com/2018/10/27/us/active-shooter-pittsburgh-synagogue-shooting.html(accessed on 20 May 2022).
7. MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate speech detection: Challenges and solutions. *PLoS ONE* **2019**, *14*, e0221152. [CrossRef] [PubMed]
8. Williams, M.L.; Burnap, P.; Sloan, L. Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *Br. J. Criminol.* **2017**, *57*, 320–340. [CrossRef]
9. Williams, M.L.; Burnap, P.; Javed, A.; Liu, H.; Ozalp, S. Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *Br. J. Criminol.* **2020**, *60*, 93–117. [CrossRef]
10. Kumar, R.; Ojha, A.K.; Malmasi, S.; Zampieri, M. Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1–11.
11. Relia, K.; Li, Z.; Cook, S.H.; Chunara, R. Race, Ethnicity and National Origin-Based Discrimination in Social Media and Hate Crimes Across 100 US Cities. In Proceedings of the International AAAI Conference on Web and Social Media, Munich, Germany, 11–14 June 2019; pp. 417–427.
12. Kwok, I.; Wang, Y. Locate the Hate: Detecting Tweets Against Blacks. In Proceedings of the AAAI'13: Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, WA, USA, 14–18 July 2013; ACM: New York, NY, USA, 2013; pp. 1621–1622.
13. Burnap, P.; Williams, M.L. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy Internet* **2015**, *7*, 223–242. [CrossRef]
14. Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; Bhamidipati, N. Hate Speech Detection with Comment Embeddings. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; ACM: New York, NY, USA, 2015; pp. 29–30.
15. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Federal Ministry of Education and Research: Bonn, Germany, 2017; pp. 512–515.
16. Malmasi, S.; Zampieri, M. Detecting hate speech in social media. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, 2–8 September 2017; INCOMA Ltd.: Seville, Spain, 2017; pp. 467–472.
17. Malmasi, S.; Zampieri, M. Challenges in discriminating profanity from hate speech. *J. Exp. Theor. Artif. Intell.* **2018**, *30*, 187–202. [CrossRef]
18. Xu, J.-M.; Jun, K.-S.; Zhu, X.; Bellmore, A. Learning from Bullying Traces in Social Media. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, QC, Canada, 3–8 June 2012; ACM: New York, NY, USA, 2012; pp. 656–666.
19. Dadvar, M.; Trieschnigg, D.; Ordelman, R.; de Jong, F. Improving Cyberbullying Detection with User Context. In Proceedings of the ECIR 2013: Advances in Information Retrieval, Moscow, Russia, 24–27 March 2013; Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 693–696.
20. Fortuna, P.; Ferreira, J.; Pires, L.; Routar, G.; Nunes, S. Merging Datasets for Aggressive Text Identification. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 128–139.
21. Georgakopoulos, S.V.; Tasoulis, S.K.; Vrahatis, A.G.; Plagianakos, V.P. Convolutional Neural Networks for Toxic Comment Classification. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence, Patras, Greece, 9–12 July 2018; ACM: New York, NY, USA, 2018; p. 35.
22. King, R.D.; Sutton, G.M. High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology* **2013**, *51*, 871–894. [CrossRef]

23. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 88–93.
24. Kumar, R.; Bhanodai, G.; Pamula, R.; Chennuru, M.R. TRAC-1 Shared Task on Aggression Identification: IIT (ISM)@ COLING'18. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 58–65.
25. Basile, V.; Bosco, C.; Fersini, E.; Debora, N.; Patti, V.; Pardo, F.M.R.; Rosso, P.; Sanguinetti, M. 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, Minneapolis, MN, USA, 6–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 54–63.
26. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 1415–1420.
27. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 75–86.
28. Burnap, P.; Williams, M.L. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* **2016**, *5*, 11. [CrossRef] [PubMed]
29. Hashtagify. Search And Find The Best Twitter Hashtags. Available online: https://hashtagify.me/ (accessed on 15 March 2022).
30. Training Data for AI, ML with Human Empowered Automation | Cogit. Available online: https://www.cogitotech.com/about-us (accessed on 15 March 2022).
31. Hripcsak, G.; Rothschild, A.S. Agreement, the f-measure, and reliability in information retrieval. *J. Am. Med. Inform. Assoc.* **2005**, *12*, 296–298. [CrossRef] [PubMed]
32. Thompson, P.; Iqbal, S.A.; McNaught, J.; Ananiadou, S. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinform.* **2009**, *10*, 349. [CrossRef] [PubMed]
33. Alnazzawi, N.; Thompson, P.; Ananiadou, S. Building a Semantically Annotated Corpus for Congestive Heart and Renal Failure From Clinical Records and the Literature. In Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), Gothenburg, Sweden, 27–30 April 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 69–74.
34. Thompson, P.; Daikou, S.; Ueno, K.; Batista-Navarro, R.; Tsujii, J.i.; Ananiadou, S. Annotation and detection of drug effects in text for pharmacovigilance. *J. Cheminform.* **2018**, *10*, 37. [CrossRef] [PubMed]
35. Alnazzawi, N. Building a semantically annotated corpus for chronic disease complications using two document types. *PLoS ONE* **2021**, *16*, e0247319. [CrossRef] [PubMed]
36. Brants, T. Inter-Annotator Agreement for a German Newspaper Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece, 31 May–2 June 2000; European Language Resources Association (ELRA): Paris, France, 2000; pp. 1435–1439.