data

MDPI

Data Descriptor

Multi-Ideology ISIS/Jihadist White Supremacist (MIWS) Dataset for Multi-Class Extremism Text Classification

Mayur Gaikwad [1], Swati Ahirrao [1,*], Shraddha Phansalkar [2] and Ketan Kotecha [3,*]

[1] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune MH 412115, India; mayurgaikwad@hotmail.com
[2] MIT Art, Design and Technology University, Pune MH 412201, India; shraddha.phansalkar@mituniversity.edu.in
[3] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune MH 412115, India
* Correspondence: swatia@sitpune.edu.in (S.A.); head@scaai.siu.edu.in (K.K.)

Abstract: Social media platforms are a popular choice for extremist organizations to disseminate their perceptions, beliefs, and ideologies. This information is generally based on selective reporting and is subjective in content. However, the radical presentation of this disinformation and its outreach on social media leads to an increased number of susceptible audiences. Hence, detection of extremist text on social media platforms is a significant area of research. The unavailability of extremism text datasets is a challenge in online extremism research. The lack of emphasis on classifying extremism text into propaganda, radicalization, and recruitment classes is a challenge. The lack of data validation methods also challenges the accuracy of extremism detection. This research addresses these challenges and presents a seed dataset with a multi-ideology and multi-class extremism text dataset. This research presents the construction of a multi-ideology ISIS/Jihadist White supremacist (MIWS) dataset with recent tweets collected from Twitter. The presented dataset can be employed effectively and importantly to classify extremist text into popular types like propaganda, radicalization, and recruitment. Additionally, the seed dataset is statistically validated with a coherence score of Latent Dirichlet Allocation (LDA) and word mover's distance using a pretrained Google News vector. The dataset shows effectiveness in its construction with good coherence scores within a topic and appropriate distance measures between topics. This dataset is the first publicly accessible multi-ideology, multi-class extremism text dataset to reinforce research on extremism text detection on social media platforms.

Dataset: https://doi.org/10.5281/zenodo.5687447.

Keywords: artificial intelligence; extremism; disinformation; ideology; propaganda; radicalization; recruitment

Citation: Gaikwad, M.; Ahirrao, S.; Phansalkar, S.; Kotecha, K. Multi-Ideology ISIS/Jihadist White Supremacist (MIWS) Dataset for Multi-Class Extremism Text Classification. *Data* **2021**, *6*, 117. https://doi.org/10.3390/data6110117

Academic Editors: Gianni Costa and Riccardo Ortale

Received: 1 September 2021
Accepted: 12 November 2021
Published: 15 November 2021

1. Summary

Extremist organizations exploit social media platforms to spread ideologies and influence youth with propaganda, radicalization, and recruitment. Multiple ideologies are coming from numerous organizations from different geographical locations. Organizations like ISIS [1] and Al Qaeda [2] have used Twitter and other social media platforms to spread propaganda and recruitment. White supremacists have also employed Twitter and websites like Stormfront [3] and Gab [4] to recruit youth. A few research works like [5] focus on the automated content restructuring of web forums for better semantic analysis on social media.

Current literature focuses on limited ideologies. Thus, it is necessary to develop an extremism text dataset containing multiple ideologies to detect extremism text. Existing

literature on online extremism detection also focuses on limited class labels. Identifying and classifying text and users into binary labels like "extremist" or "non-extremist" provides lesser insight. Researchers have classified the extremist text on the social media into major types based on the objectives of social, political, or religious nature. Classification and analysis of extremist text on social media can help to curb disinformation. This work contributes to developing a multi-ideology extremist text seed dataset that can be used for extremism detection of larger extremism text datasets collected from popular social media platforms. The dataset will also be helpful for further extremism classification into propaganda, radicalization, and recruitment [6] as seen in Figure 1.

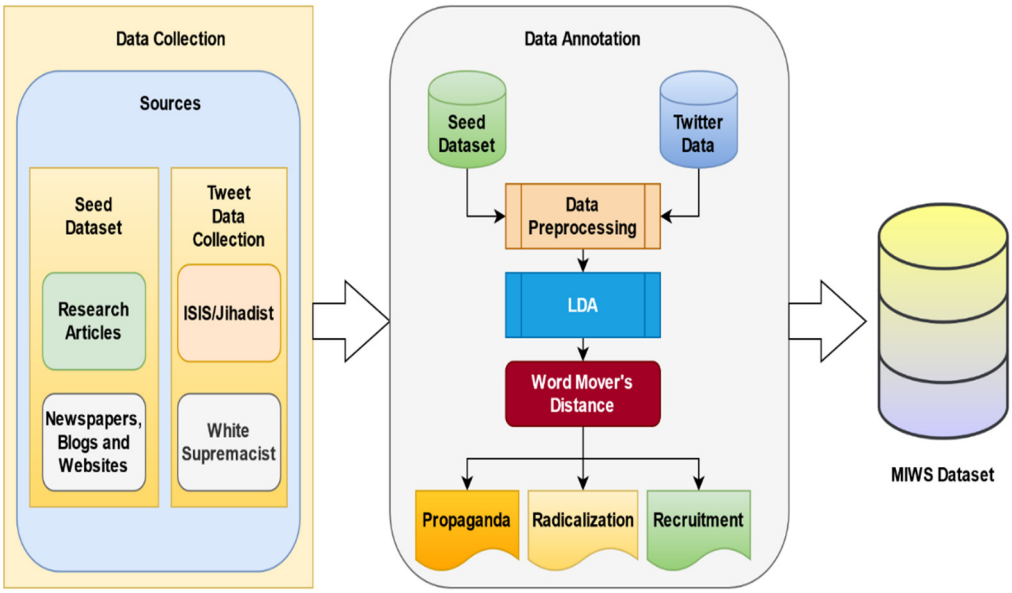

Figure 1. Creation of seed and MIWS datasets.

This dataset consists of two parts:

1. Seed dataset consisting of 400 examples collected from diverse sources and manually annotated with class labels as propaganda, radicalization, or recruitment.
2. MIWS dataset consisting of 40,000 tweets collected from Twitter and annotated with class labels as propaganda, radicalization, or recruitment from the seed dataset. Twenty thousand tweets were collected from ISIS/Jihadist ideology. Twenty thousand tweets were collected from White supremacists' ideology.

The seed and resultant MIWS dataset with multiple ideologies are statistically validated and thus can be employed to generate a robust and accurate extremism text dataset.

Research Goal of Our Datasets

The aim of the seed dataset is to classify multi-ideology extremism text into different classes such as propaganda, radicalization, and recruitment.

1. Seed dataset can be used to automatically annotate large extremism text datasets collected from social media platforms.
2. MIWS dataset is constructed and automatically annotated using seed dataset into propaganda, radicalization, and recruitment.
3. MIWS can also be used to train the classifier that detects extremism text and further classifies extremism text into propaganda, radicalization, and recruitment.
4. MIWS dataset can be further used to analyze the geographical location of extremism text to understand the spread of extremism.

2. Data Description

There are 400 records in the seed dataset collected from diverse sources such as research articles, newspapers, blogs, and websites. There are 200 records for ISIS/Jihadist ideology and 200 for White supremacist ideology in the seed dataset. In the MIWS dataset, there are 20,000 tweets of ISIS/Jihadist ideology and 20,000 tweets from White Supremacist ideology. The details can be seen in Tables 1 and 2.

Table 1. Specification of seed and MIWS dataset.

Information \ Dataset	Seed	MIWS
Subject area	NLP	NLP
Focused area	Extremism text detection and classification	Extremism text detection and classification
File type	csv	csv
Method for acquiring data	Web scraping and collection of research articles	Collection of tweets with Twitter API
No of files	2	2

Table 2. Dataset information.

Dataset	No of Examples	Sources of Data	Tweets Extracted	Ideologies Used	Attributes Used
Seed Dataset	400	Research Articles, Newspapers, Blogs, Websites	-	ISIS/Jihadist, White supremacists	Source, type of source, text, label, ideology, context location, author country affiliation, and source location
MIWS Dataset	40,000	Standard Dataset and Twitter Data	40,000 from Twitter	ISIS/Jihadist, White supremacists	tweet ID, created date, username, name, tweet, geo enabled, label.

3. Methods

Figure 2 shows the process flow for construction of seed and MIWS dataset. It contains four phases: data collection, seed data validation, data labelling and merging of data from different extremist ideology. Data collection is performed in two parts: first, seed data collection (explained in Section 3.1) and then collection of tweets or MIWS data collection (explained in Section 6.1). Seed data validation (explained in Section 5) is performed individually for each ideology. Similarly, data labelling (explained in Section 6) is done on each ideology separately. The reason behind this segregation is that the corpuses of both ideologies are different and the LDA topics (explained in Section 5) are based on the probability of keywords within the document of a particular corpus. In the last phase, merging of the ISIS/Jihadist and White supremacist labelled datasets is carried out (explained in Section 6).

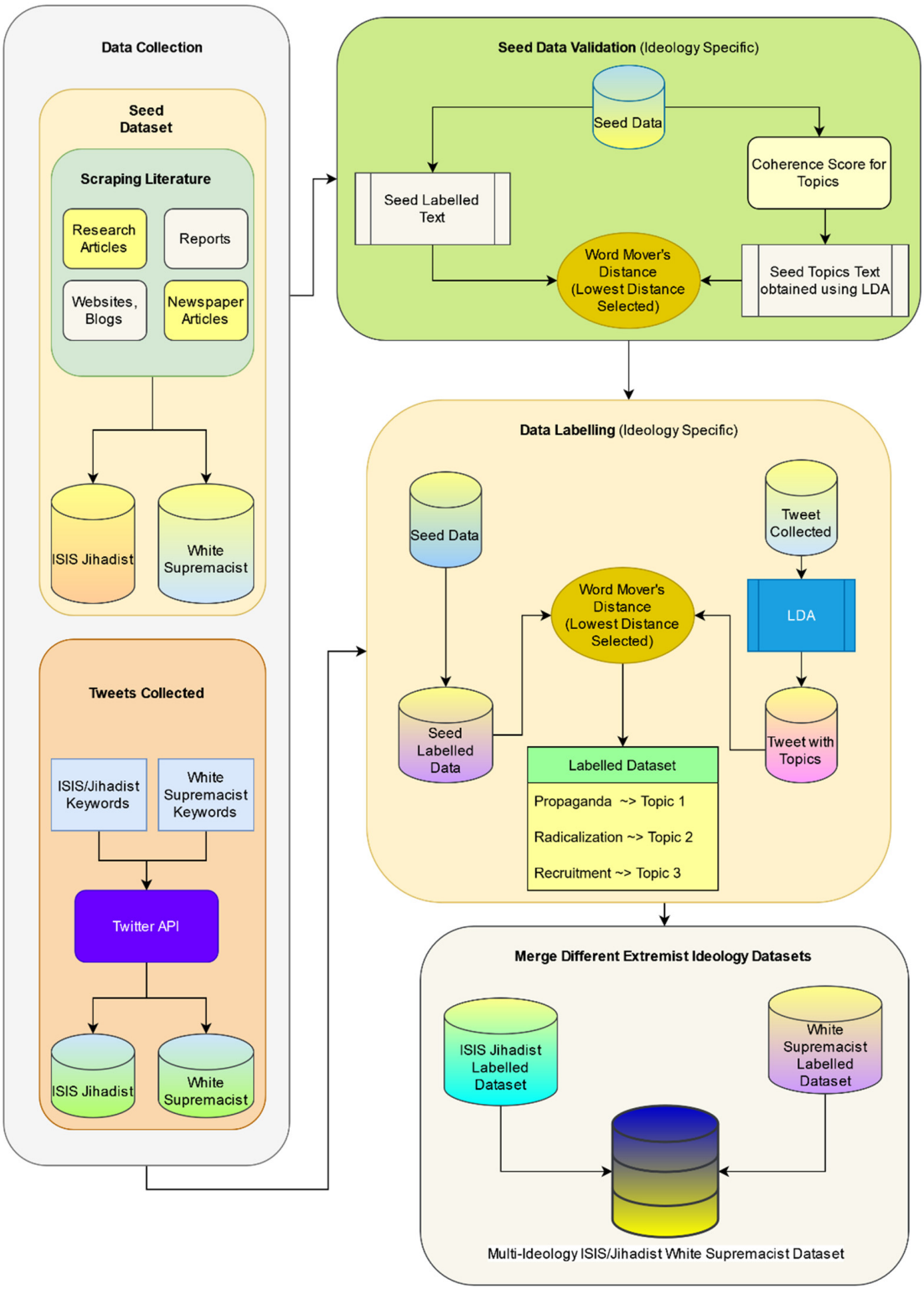

Figure 2. Process flow for construction of seed and MIWS datasets.

3.1. Seed Data Collection

For data collection of a seed dataset, we collected research articles from existing literature, examples from extremist identification websites, and blogs recognizing influential propagandists, radicals, and extremist recruiters [7].

3.2. Sources

The seed dataset is collected based on ISIS/Jihadist and White supremacist ideologies. The primary objective of the seed dataset is to collect text examples of propaganda, radicalization, and recruitment. Multiple newspaper articles, journal papers, book chapters, and websites are selected. Proper sources for text examples of propaganda, radicalization, and recruitment are selected using a snowballing technique [8]. Table 3 provides few examples from the seed dataset.

Table 3. Seed examples with labels and ideology.

Source	Type of Source	Text	Label	Geographical Location	Ideology
Chatfield et al. [9]	Research Article	This is so awesome. US airstrikes also by mistake hit a Shia militia convoy near Tuz 2nd one after the Jabour Strikes	Propaganda	Iraq	ISIS/Jihadist
CounterExtremism [7]	Website	IS needs few hundred fighters to establish territory in France	Recruitment	France	ISIS/Jihadist
Ray and Marsh [10]	Research Article	WE BELIEVE that the Cananite Jew is the natural enemy of our Aryan (White) race	Radicalization	USA	White Supremacist
Thompson [11]	Blog	Join your local Nazis	Recruitment	USA	White Supremacist
CounterExtremism [7]	Website	Fight the kuffar, until they pay jizjah & law of Allah rules on the land Fight Them. By permission of Allah, we will be successful	Propaganda	-	ISIS/Jihadist
No Space for Hate [12]	Website	'Diversity' is a weapon of annihilation targeting American identity	Radicalization	USA	White Supremacist
Homeland Security Today [13]	Website	Join us, or perish with the rest	Recruitment	-	White Supremacist

3.2.1. Research Articles and Reports

The seed text selected from the journal paper explicitly provides identification of extremist text as propaganda, radicalization, or recruitment [9,10,14], which are limited in numbers. Journal papers were selected from a database similar to our work in [6,15]. The search was limited to the period January 2015 to December 2020. Some older studies were also included using a snowballing technique. A total of 105 research articles and reports were surveyed, of which 18 were selected for this work.

3.2.2. Newspaper, Blogs, and Websites

A snowballing technique was also used to search and select newspaper articles, blogs, and websites for the seed dataset. Most seed examples are also chosen from newspaper articles, blogs [16], or counter-extremism websites [7,17]. Some websites classify users as propagandists or recruiters. The tweet or post of such users was considered as propaganda or recruitment. A total of 86 newspaper articles, blogs, and websites were surveyed, of which 32 were selected for this work.

3.3. Seed Data Features

The features of seed data include SOURCE, TYPE_OF_SOURCE, TEXT, LABEL, IDEOLOGY, GEOGRAPHICAL_LOCATION, and AUTHOR_COUNTRY_AFFLIATION.

1. SOURCE contains information like author name, article name, or website link of source.

2. TYPE_OF_SOURCE indicates whether a source is a research article, newspaper article, blog, or website.
3. TEXT contains actual text, tweet, or speech that is extremist provided by the source.
4. LABEL denotes whether the text is propaganda, radicalization, or recruitment as mentioned by the source.
5. IDEOLOGY mentions to which extremist ideology the text belongs.
6. GEOGRAPHICAL_LOCATION is a manually analyzed field that indicates any country mentioned in the text.
7. AUTHOR_COUNTRY_AFFLILIATION country indicates the country to which the author belongs.

4. Data Pre-Processing

The following steps were carried out for data pre-processing as seen in Figure 3:

- Removal of stopwords. Prepositions can affect the outcome of NLP algorithms, so they are removed.
- Removal of URLs. This work does not focus on the use of URLs, so regular expressions are used to remove URLs.
- Removal of emojis, hashtags, retweets, and digits. Emojis and digits are not considered in this research work, and symbols like hashtags, @, and retweets are out-of-scope for this research work. Thus, they are removed in pre-processing.
- Lemmatization and lowercase. Lemmatization is used to ensure that meaningful words get selected for analysis. The remaining documents are converted into lowercase, so the case of terms does not affect the outcome of algorithms.

Figure 3. Data pre-processing.

5. Seed Data Validation

5.1. LDA with Coherence Score

In this research work, data validation implies verifying manual annotation using the topic modeling LDA technique [18]. Topic modeling is a method to identify documents in an unsupervised way. The documents are determined based on the set of keywords that are present in the corpus. Thus, the relevance of the document can be established just by looking at those sets of keywords. Latent Dirichlet allocation (LDA) is the most popular topic modeling technique. LDA works in two parts: words belonging to a document and calculating the probability of words belonging to that topic. Thus, LDA is used to determine the importance of specific words in extremism data. We further evaluate the strength of the topic with a coherence score [19]. A coherence score is used to emphasize the semantic similarity between high-scoring words in the topic. Thus, the higher the coherence scores the more the semantic similarity within the words in the topic. Word mover's distance [20] is also used to find the relationship between LDA topics of the seed and seed labels. Thus, the empirical annotation is statistically validated. Topic coherence points to the co-occurrence of words within documents in the corpus, indicating semantic relation between the words [19]. Figure 4a,b show topic coherence for the number of topics for the seed dataset.

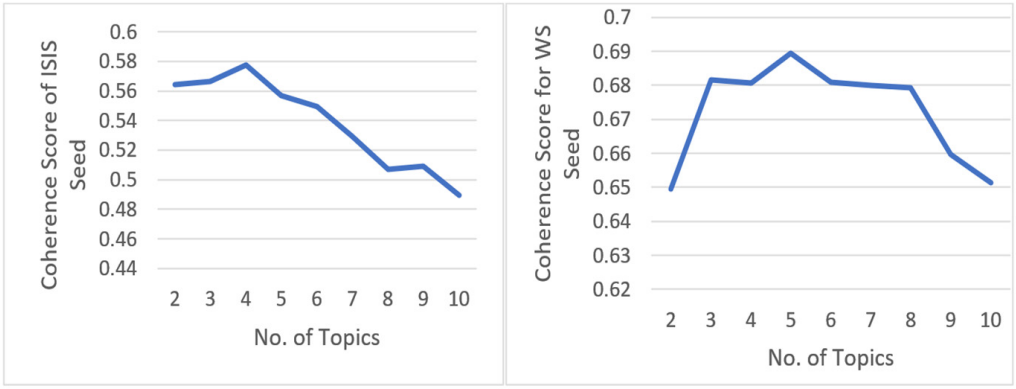

Figure 4. Coherence score vs. topics (**a**) for ISIS/Jihadist seed and (**b**) for White supremacist seed.

As seen in Figure 4a,b, a coherence score was used to determine an optimal number of topics for the seed dataset. As observed, the number for topics 3, 4, and 5 shows the highest coherence of around 0.55 for ISIS and 0.68 for White supremacists. The literature [8,9,14] indicates extremism has three main types: propaganda, radicalization, and recruitment. Hence, we chose three topics (k) within extremist speech or text. The LDA optimization is also performed using GridSearchCV with 3, 4, and 5 topics. The best LDA model found using GridSearchCV contains only three topics.

5.2. Word Mover's Distance (WMD) Using Google News Pretrained Vector

Word mover's distance is used to verify the similarity between seed labels and topics created using LDA. WMD calculates similarity or dissimilarity between documents, even if there are no words in common [18]. The intuition behind WMD is that it determines the smallest semantic distance required for one document to reach another [18]. Word embeddings like Word2Vec are necessary to calculate the semantic distance between documents. The advantages of WMD are it does not use hyperparameters, the distance between documents can be broken down to the difference between words, and it works with popular word embeddings like Word2Vec.

In this study, to calculate the WMD, the topic corpus and label corpus are compared using a Google News pretrained vector. Tables 4 and 5 show the results.

Table 4. WMD comparison of ISIS/Jihadist labels vs. ISIS/Jihadist topics from seed dataset.

ISIS Seed Labels / ISIS Seed Topics	Propaganda	Radicalization	Recruitment
Topic 0	0.8100	0.8176	0.8183
Topic 1	0.8239	0.8107	0.8162
Topic 2	0.8201	0.8198	0.7871

Table 5. WMD comparison of White supremacist labels vs. White supremacist topics from seed dataset.

WS Seed Labels / WS Seed Topics	Propaganda	Radicalization	Recruitment
Topic 0	1.0492	0.9071	0.9138
Topic 1	0.7894	0.8029	0.7988
Topic 2	0.9752	0.9618	0.9463

5.3. Inference

The comparison of seed labels and seed topics produces acceptable results. The propaganda of ISIS/Jihadist has the lowest distance of 0.8100 to topic 0 of ISIS/Jihadist. Similarly, the radicalization sub-corpus is at the lowest distance of 0.8107 from topic 1 of ISIS/Jihadist. The recruitment sub-corpus has the lowest distance of 0.7871 from topic 2 of ISIS/Jihadist.

A similar comparison is made for the White supremacist seed label and White supremacist seed topics. The propaganda sub-corpus of the WS seed is at a distance of 0.7894 from topic 1 of the WS Seed. Topic 2 of the WS seed is near to the recruitment sub-corpus at a distance of 0.9463, while Topic 0 is near radicalization at a distance of 0.9071.

6. Multi-Ideology ISIS/Jihadist White Supremacist (MIWS) Dataset

The MIWS dataset is constructed with tweets collected from Twitter for ISIS and White supremacist ideology. It can be used to train the classifier that detects extremism text and further classifies extremism text into propaganda, radicalization, and recruitment.

6.1. MIWS Data Collection

To collect relevant tweets and metadata, we constructed different search queries with different keywords. We used popular keywords that are associated with extremist ideologies like “munafiq”, “kuffar”, “white genocide”, and “anti-white”, as mentioned in Table 6. These keywords are referenced from [8,21–25]. We also used some new keywords like “kufr army”, “wesupporttaliban”, “talibanourguardians”, “globalists”, “zog” etc., to collect recent tweets. The geographical locations were found by manually searching locations from collected tweets. If no locations were present in the tweet, it was labeled as “undefined”.

Table 6. Examples of keywords and combinations used for tweet collection.

ISIS/Jihadist Keywords and Combinations	WS Keywords and Combinations
Munafiq	Antiwhite
Murtadin	White Genocide
Kufr Army	White Power
Kafir	Globalist Zog
WeStandWithTaliban	WPWW

The following are different metadata collected from tweets using Twitter API.

1. TWEET_ID: It is the unique id for a tweet.
2. CREATED_AT: Time at which tweet was created or posted.
3. USERNAME: Username of the posted tweet.
4. NAME: Name, if provided by the user.
5. TWEET: The tweet in UTF-8 format.
6. GEO_ENABLED: Boolean value for geographical data about the tweet.

Due to the Twitter data sharing policy, only Tweet_ID, Created_At, and Geo_Enabled can be shared publicly.

Table 7 is a snapshot of collected tweets with geographical location and dominant topic. Table 8 shows that 20,000 tweets were collected for each ISIS/Jihadist and White supremacist ideology. Table 9 shows the count of tweets for some keywords. In extremist tweets, words like ‘munafiq’, ‘munafiqeen’, ‘kuffar’, and ‘white lives matter’ are frequently mentioned.

Table 7. Examples of tweets collected.

Sr no.	Tweet_ID	Created_AT	Tweet	Geographical Location	Ideology
1	1416506180336119810	2021-07-17 21:12:45	May Allah azza-wa-jall destory all these traitors of Islam in Afghanistan who joined hands with kuffar USA	['USA', 'US', 'Afghan', 'Afghanistan']	ISIS/Jihadist
2	1414276878945325059	2021-07-11 17:34:18	Taliban made a secret deal with the American Kuffar to take power, and in return Taliban protect America	America	ISIS/Jihadist
3	1415172667435487236	2021-07-14 04:53:51	Funny how your "human rights" never apply to the white working class who your Zio-shill masters work to destroy	Undefined	White Supremacist

Table 8. Count of tweets collected for particular ideology.

Ideologies	No of Tweets Collected from Tweeter
ISIS/Jihadist Tweets	20,000
White Supremacist Tweets	20,000

Table 9. Count of tweets collected for particular keyword.

Top Keyword Based Twitter Queries	Count of Tweets
Kufr	2637
Shirk	2011
Kuffar	2242
White Genocide	715
Globalists	1258
Anti-White	1820
WeStandWithTaliban	736
TalibanOurGuardians	420

6.2. Construction of MIWS Using Seed Dataset

As described in Section 6.1, tweets related to specific ideologies extracted from Twitter are merged to form the MIWS dataset as seen in Figure 5.

6.2.1. Data Pre-Processing

Data pre-processing is carried out as mentioned in Section 4.

6.2.2. Data Labeling/Annotation

The following steps are performed for data annotation:

6.2.3. LDA on Collected Tweets

As shown in Figure 5, a comparison of labeled topics from the seed dataset and topics from Twitter collected data is performed. To extract topics, the Latent Dirichlet Allocation [26] method is used. To confirm the best possible topics, GridSearchCV is used. Hyperparameter tuning is performed to select optimal parameters for the best model. Table 10 provides the best parameters for the LDA model applied to collected data based on ideology.

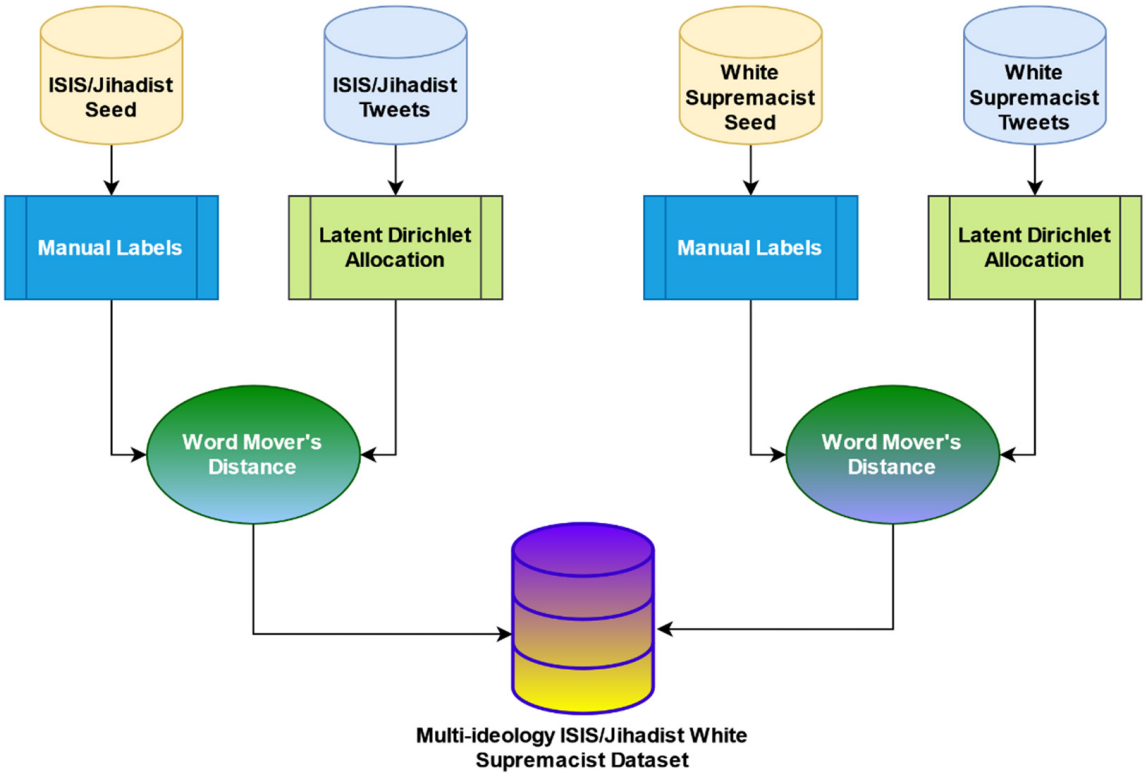

Figure 5. Data labeling/annotation.

Table 10. GridSearchCV parameters for best LDA model.

Parameters	Given Parameters	Optimal Parameters
n_components	3, 4, 5	3
learning_decay	0.8, 0.9, 0.99	0.9
max_iter	8, 9, 10	10
cv	10	10

6.2.4. Comparison between Seed Labels and Topics of Collected Tweets

It is required to compare topics based on the ideology. Labeled topics from the ISIS/Jihadist seed and White supremacist seed are compared with topics from ISIS/Jihadist and White supremacists collected tweets. This was done to maintain uniformity and accuracy across ideologies.

Word mover's distance (WMD) is used to compare collected tweets and seed labels. WMD presents semantically meaningful comparisons of words from local co-occurrences in sentences. Thus, the lower the distance the more the similarity among sentences. To leverage WMD's properties, the Word2Vec vector pretrained on Google News is used [27]. As seen from Tables 11 and 12, the lower the distance between the topic and labels the more similar they are than the others. Thus, the corresponding label is given to that topic.

Table 11. WMD comparison for ISIS/Jihadist seed labels vs. ISIS/Jihadist tweet topics.

ISIS Seed Labels / ISIS Tweet Topics	Propaganda	Radicalization	Recruitment
Topic 0	0.8598	0.8575	0.8591
Topic 1	0.8455	0.8588	0.8494
Topic 2	0.8490	0.8584	0.8464

Table 12. WMD comparison for White supremacist seed labels vs. White supremacist tweet topics.

WS Tweet Topics \ WS Seed Labels	Propaganda	Radicalization	Recruitment
Topic 0	0.8038	0.8039	0.8032
Topic 1	0.8028	0.8021	0.8041
Topic 2	0.7924	0.8029	0.8035

6.2.5. Inference

Similarity of ISIS/Jihadist seed labels and ISIS/Jihadist tweet topics is shown in Table 11. The propaganda of the ISIS/Jihadist seed has the lowest distance of 0.8455 to topic 1 of the ISIS/Jihadist tweet topics. Similarly, the radicalization sub-corpus is at the lowest distance of 0.8575 from topic 0 of the ISIS/Jihadist tweet topics. The recruitment sub-corpus has the lowest distance of 0.8464 from topic 2 of the ISIS/Jihadist tweet topics.

A similar comparison is made for White supremacist tweets, as seen in Table 12. The propaganda seed sub-corpus is at a distance of 0.7924 from topic 2 of the WS tweets. Topic 0 of the WS tweets is near the recruitment seed sub-corpus at a distance of 0.8032, while topic 1 is near radicalization at a distance of 0.8021.

6.2.6. Merging of Datasets

Topic 0 of ISIS/Jihadist and topic 1 of WS are labeled as radicalization and contain 10,120 tweets. Radicalization includes more politically aligned tweets. Topic 1 of ISIS/Jihadist and Topic 2 of WS are labeled as propaganda, consisting of 19,523 tweets; thus, propaganda is the largest class of all three classes. Propaganda contains religious keywords, achievements, and glorification of ideology. Topic 2 of ISIS/Jihadist and topic 0 of WS, labeled as recruitment, contains 10,893 tweets. General hate, discussion about the degradation of old or religious ways, and incitement against a particular group are observed in recruitment. Table 13 shows a few tweets and their annotated labels with ideology, while Table 14 shows the statistical summary for seed and MIWS datasets.

Table 13. Examples from MIWS dataset.

Sr No.	Tweet_ID	Tweet	Ideology	Label
1	1414277221804482560	I said shutthef up, I don't support any kufr don't ask me about USA	ISIS/Jihadist	Propaganda
2	1414646744000958464	To usher in communism, disintegration of family and the LGBTQ agenda. This is a self proclaimed marxism.	White Supremacist	Radicalization
3	1416451924736491524	You are a murtad, an enemy of lsIam and deserve to be kiIled'	ISIS/Jihadist	Recruitment

Table 14. Statistical Summary.

Information	Seed Dataset	MIWS
Time frame	2015–2020	June 2021–August 2021
No. of labels	3	3
No. of attributes	7	6
No. of words in example and tweets (min, max)	Min: 1 Max: 291	Min: 1 Max: 32
No. of records	400	40,000
Count of labels	Propaganda: 225 Recruitment: 100 Radicalization: 71	Propaganda: 19,523 Recruitment: 10,893 Radicalization: 10,120
No. of ideologies	2	2

After annotating, the data from both ideologies are merged to form a new dataset called the merged ISIS/Jihadist and White supremacist (MIWS) dataset.

6.3. Data Validation of MIWS Dataset

Data Validation is performed on the complete MIWS dataset. To validate the dataset, WMD is used. For validation, a comparison between labels is performed. This provides three different results. The WMD for propaganda and radicalization is at 5.4632, while for propaganda and recruitment, the WMD is at 3.4831. Lastly, the WMD between recruitment and radicalization is 4.6590. These results show that there is a significant difference between MIWS propaganda, radicalization, and recruitment labels.

7. Discussion and Implications

The MIWS dataset proposes multi-ideology and multi-class classification, especially in extremism types like propaganda, radicalization, and recruitment. These types of text are mostly disinformation targeted at vulnerable youth. Thus, it is vital to counter extremist text and malicious disinformation on social media. However, there are only a few standard datasets related to extremism, such as the ISIS Kaggle dataset [28], Stormfront dataset [29], and Gab dataset [30]. There are also a few custom datasets that are publicly unavailable such as Jaki et al. [31], Fraiwan et al. [32], and Ferrara et al. [33]. Most of these datasets are popular in the literature but have a few limitations. They are old, obsolete, most tweets and posts in the datasets are deleted or suspended, and classification is limited to extremist-non-extremist, hate-no hate [30,33,34].

MIWS tries to address these issues. The data collected for MIWS is recent and influenced by more recent events, writing styles, and expressions on social media. Most tweets collected are available online and more information about the extremist text can be gathered [7,35]. The dataset is annotated into propaganda, radicalization, and recruitment, which provides a clear view of conversation and topics in the extremist text. A tweet of recruitment can now be addressed distinctly from propaganda or a radical opinion which are part of disinformation spread by extremists. This helps in taking measures for effective handling of disinformation control or curbing its outreach.

This dataset is one of its kind, which also caters to multiple ideologies of extremism and tries to present the diversities of the writing and presenting styles by the activists from ISIS and White supremacist groups, which are the most popular extremist ideologies [36].

Academicians, researchers, and law enforcement agencies can use the MIWS dataset to identify and analyze extremist posts and disinformation on Twitter or any other website. Identifying extremist text as propaganda, radicalization, and recruitment can further help explore the topics and events related to extremism for adequate control of disinformation. The automatic classification that can be offered with this dataset can reduce the time taken for analysis and encourage law enforcement agencies and social media networks to take rapid action against such tweets or posts.

MIWS dataset can be used in alliance with [37] to identify the high prevalence of extremism in a particular geographical area using different semantic models like SpaCy. This could extend [37] to decide about help provided to countries regarding more recent events and extremist propaganda, radicalization, and recruitment taking place in that country. MIWS dataset can be used in addition to Global Terrorism Dataset (GTD) [38] to analyze propaganda, radicalization, and recruitment leading to terrorist events and their after effects. In order for it to work with the GTD, the MIWS needs to be kept updated frequently.

8. Limitations

This research work presents two different datasets, so the limitations of each dataset are as below:

8.1. Seed Dataset

- Size of seed dataset: A limited number of studies exclusively identify extremism as propaganda, radicalization, and recruitment. This limits the size of the seed dataset.
- Limited class labels: Extremist text has multiple class labels like irrelevant, violent, or racist, but only popular class labels, i.e., propaganda, radicalization, and recruitment, are used for this study.
- Limited ideologies: There are different ideologies as well as extremist organizations with different agendas. However, we only considered popular ideologies, which are ISIS/Jihadist and White supremacist ideologies.

8.2. MIWS

- Class imbalance: There is a significant imbalance in class labels which may affect evaluation and prediction.
- Suspended tweets: As extremist tweets violate Twitter's hate speech policy, they may be removed from Twitter. Thus, recollection of tweets is an issue.
- False positives: The tweets are collected using extremist keywords. Hence, false positives like sarcastic, satirical, or critical tweets might get inadvertently selected, reducing accuracy.

9. Conclusions

The presented work contributes to the detection of extremist text on social media, characterizes the major types of extremism text, and thus contributes to curbing the spread of disinformation. This research work contributes to the construction of extremism text, multi-ideology multi-class seeds, and the MIWS dataset. To the best of our understanding, the seed dataset collected from research articles, blogs, and counter-extremism websites is the first of its kind, which can be used further to classify any extremism text into radicalization, recruitment, and propaganda.

This hypothesis is validated by constructing the MIWS dataset from recent tweets collected from Twitter and can be used to classify any extremism text into radicalization, recruitment, and propaganda. The presented seed dataset is also statistically validated using a coherence score and WMD. The MIWS dataset is validated using WMD. This makes the it statistically proven for further research on extremism text detection and analysis.

10. Future Work

There are still a few areas that can be improved:

- Size of seed dataset: The size of the seed dataset can be extended with labeled extremism text from the latest research.
- Languages: A greater number of languages such as Arabic and Urdu can be considered for extremist tweet collection.
- Removal of class imbalance: More data collection with different keywords can address an issue of class imbalance within the MIWS dataset.
- Evaluation using pre-trained networks: Pre-trained networks like BERT and ELMO can be used to evaluate and predict the MIWS dataset.
- Data validation using statistical techniques: Statistical validation can be made significant using tests like the Chi-Square test.
- Inclusion of additional ideologies and classes: More extremist ideologies with classes like neutral or irrelevant may be added to the MIWS dataset in the future.

Author Contributions: Conceptualization, S.A., S.P. and K.K.; methodology, M.G.; investigation, M.G.; resources, S.A., S.P., K.K., and M.G; data curation, M.G.; writing—original draft preparation, M.G., S.A., and S.P.; writing—review and editing, S.A., S.P., and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study is partly funded by Research Support Grant from Symbiosis International (Deemed University).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Multi-ideology ISIS/Jihadist White Supremacist Dataset is publicly available extremism dataset. The data presented in this study is openly available at https://doi.org/10.5281/zenodo.5687447.

Acknowledgments: The authors would like to thank Symbiosis International (Deemed University) for permitting us to carry out our research and to use resources to accomplish the objectives.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baele, S.; Boyd, K.; Coan, T. *ISIS Propaganda*; Oxford University Press: Oxford, UK, 2020.
2. Dornbierer, A. How al-Qaeda Recruits Online. The Diplomat, 2011. Available online: https://thediplomat.com/2011/09/how-al-qaeda-recruits-online/ (accessed on 15 May 2020).
3. Stormfront. Available online: www.stormfront.org (accessed on 20 August 2020).
4. Gab Social Media. 2020. Available online: https://gab.com/ (accessed on 10 October 2020).
5. Korobiichuk, I.; Syerov, Y.; Fedushko, S. The Method of Semantic Structuring of Virtual Community Content. In *Mechatronics 2019: Recent Advances towards Industry 4.0*; Springer International Publishing: Cham, Switzerland, 2020; pp. 11–18.
6. Gaikwad, M.; Ahirrao, S.; Phansalkar, S.; Kotecha, K. Online Extremism Detection: A Systematic Literature Review with Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools. *IEEE Access* **2021**, *9*, 48364–48404. [CrossRef]
7. ISIS Recruiters, Propagandists, and Inciters to Violence Operating on Twitter | Counter Extremism Project. *Counter Extremism*. 1 January 2021. Available online: https://www.counterextremism.com/content/isis-recruiters-propagandists-and-inciters-violence-operating-twitter (accessed on 22 March 2021).
8. Naderifar, M.; Goli, H.; Ghaljaie, F. Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. *Strides Dev. Med. Educ.* **2017**, *14*. [CrossRef]
9. Chatfield, A.T.; Reddick, C.G.; Brajawidagda, U. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *ACM International Conference Proceeding Series*; Association for Computing Machinery: Phoenix, Arizona, 2015; pp. 239–249. [CrossRef]
10. Ray, B.; Marsh, G.E. Recruitment by extremist groups on the Internet. *First Monday* **2001**, *6*. [CrossRef]
11. Thompson, A. Inside Atomwaffen As It Celebrates a Member for Allegedly Killing a Gay Jewish College Student. ProPublica, 2018. Available online: https://www.propublica.org/article/atomwaffen-division-inside-white-hate-group (accessed on 5 May 2021).
12. White Nationalist Recruitment on IU's Campus. No Space for Hate. 2020. Available online: https://nospace4hate.btown-in.org/recruitment-on-campus/ (accessed on 20 December 2020).
13. Homeland Security. The Atomwaffen Division: The Evolution of the White Supremacy Threat. Homeland Security Today, 2020. Available online: https://www.hstoday.us/subject-matter-areas/counterterrorism/the-atomwaffen-division-the-evolution-of-the-white-supremacy-threat/ (accessed on 10 October 2020).
14. Windsor, L. The Language of Radicalization: Female Internet Recruitment to Participation in ISIS Activities. *Terror. Polit. Violence* **2018**, *32*, 506–538. [CrossRef]
15. Gaikwad, M.; Ahirrao, S.; Phansalkar, S.P.; Kotecha, K. A Bibliometric Analysis of Online Extremism Detection. *Libr. Philos. Pract.* **2020**, *2020*, 1–16.
16. Johnson, B. Shared Themes, Tactics in White Supremacist and Islamist Extremist Propaganda—Homeland Security Today. Homeland Security Today, 2020. Available online: https://www.hstoday.us/subject-matter-areas/counterterrorism/shared-themes-recruitment-tactics-in-white-supremacist-and-islamist-extremist-propaganda/ (accessed on 10 March 2021).
17. Southern Poverty Law Center. Southern Poverty Law Center. 2021. Available online: https://www.splcenter.org/ (accessed on 5 May 2021).
18. Liu, H.; Xu, L.; Yang, M.; Yan, M.; Zhang, X. Predicting component failures using latent Dirichlet allocation. *Math. Probl. Eng.* **2015**, *2015*, 562716. [CrossRef]
19. Kumar, K. Evaluation of Topic Modeling: Topic Coherence. DataScience, 2018. Available online: https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/ (accessed on 10 October 2020).
20. Wu, L.; Yen, I.E.; Xu, K.; Xu, F.; Balakrishnan, A.; Chen, P.Y.; Ravikumar, P.; Witbrock, M.J. Word Mover's Embedding: From Word2Vec to Document Embedding. *arXiv* **2018**, arXiv:1811.01713.
21. Kaati, L.; Omer, E.; Prucha, N.; Shrestha, A. Detecting Multipliers of Jihadism on Twitter. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; pp. 954–960. [CrossRef]
22. Berger, J.M.; Aryaeinejad, K.; Looney, S. There and Back Again: How White Nationalist Ephemera Travels Between Online and Offline Spaces. *RUSI J.* **2020**, *165*, 114–129. [CrossRef]

23. Glossary of Terms and Acronyms Radicalization and Violent Extremism. Radicalization Prevention in Prisons. 2018. Available online: http://www.r2pris.org/glossary.html (accessed on 10 October 2020).
24. Habib, R.R. Taliban's Takeover of Afghanistan Should Not Be Celebrated. The Express Tribune, 2021. Available online: https://tribune.com.pk/article/97460/talibans-takeover-of-afghanistan-should-not-be-celebrated (accessed on 25 August 2021).
25. Charles, C. *(Main)streaming Hate: Analyzing White Supremacist Content and Framing Devices on YouTube*; University of Central Florida: Orlando, FL, USA, 2020.
26. Kochedykov, D.; Apishev, M.; Golitsyn, L.; Vorontsov, K. Fast and Modular Regularized Topic Modelling. In Proceedings of the 21st Conference of Open Innovations Association FRUCT, Helsinki, Finland, 6–10 November 2017; pp. 182–193. [CrossRef]
27. Rehurek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, 22 May 2010; pp. 45–50.
28. ActiveGalaxy. ISIS Related Dataset. Kaggle, 2016. Available online: https://www.kaggle.com/activegalaxy/isis-related-tweets (accessed on 10 October 2020).
29. de Gibert, O.; Perez, N.; García-Pablos, A.; Cuadros, M. Hate Speech Dataset from a White Supremacy Forum. September 2018. Available online: http://arxiv.org/abs/1809.04444 (accessed on 10 October 2020).
30. Kennedy, B.; Atari, M.; Davani, A.M.; Yeh, L.; Omrani, A.; Kim, Y.; Coombs, K.; Havaldar, S.; Portillo-Wightman, G.; Gonzalez, E.; et al. The Gab Hate Corpus: A Collection of 27k Posts Annotated for Hate Speech. Psyarxiv, 2020. Available online: https://psyarxiv.com/hqjxn/ (accessed on 10 October 2020).
31. Jaki, S.; de Smedt, T. Right-Wing German Hate Speech on Twitter: Analysis and Automatic Detection. October 2019. Available online: http://arxiv.org/abs/1910.07518 (accessed on 10 October 2020).
32. Fraiwan, M. Identification of Markers and Artificial Intelligence-Based Classification of Radical Twitter Data. *Appl. Comput. Inform.* **2020**. [CrossRef]
33. Ferrara, E.; Wang, W.-Q.; Varol, O.; Flammini, A.; Galstyan, A. *Predicting Online Extremism, Content Adopters, and Interaction Reciprocity*; Springer: Cham, Switzerland, 2016; pp. 22–39.
34. Ahmad, S.; Asghar, M.Z.; Alotaibi, F.M.; Awan, I. Detection and Classification of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques. *Hum.-Cent. Comput. Inf. Sci.* **2019**, *9*, 24. [CrossRef]
35. ADL. White Supremacists Double down on Propaganda in 2019. ADL, 2020. Available online: https://www.adl.org/blog/white-supremacists-double-down-on-propaganda-in-2019 (accessed on 10 August 2020).
36. Berger, J.M. Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS Online Social Media Networks. 2016. Available online: https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/Nazisv.ISIS.pdf (accessed on 10 October 2020).
37. Sadik-Zada, E.R. An Ode to ODA against all Odds? A Novel Game-Theoretical and Empirical Reappraisal of the Terrorism-Aid Nexus. *Atl. Econ. J.* **2021**, *49*, 221–240. [CrossRef]
38. LaFree, G.; Dugan, L. Introducing the Global Terrorism Database. *Terror. Polit. Violence* **2007**, *19*, 181–204. [CrossRef]