Paper No: SC-16 **Smart Computing**

Identifying religious extremism-based threats in Sri Lanka using bilingual social media intelligence

Aneesha Fernando*
Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
fernando_im14006@stu.kln.ac.lk

Thareendra Keerthi Wijayasiriwardhane
Department of Industrial Management
Faculty of Science, University of Kelaniya, Sri Lanka
thareen@kln.ac.lk

***Abstract*: Religion is one's relation to what he or she regards as holy, sacred, spiritual, or worthy of especial reverence. Religious extremism is the advocacy of extreme measures over a religion whereas religious extremists are even willing to murder as they provide sanctions for violence in the service of God. Sri Lanka has a tragic history of religious and ethnic extremism and the Easter Sunday attack coordinated by a radical Islamic group that killed over 300 and injured another several hundred can be identified as the recent climax of these events. In this modern information age, it is evident that these radical extremist groups utilize social media for spreading their extreme ideologies due to its free and unregulated nature. If there were a mechanism to even slightly identify the possibility of tragic incidents like Easter Sunday bombing, the 300 souls who had to sacrifice their lives for an unreasonable cause would be still alive happily. In this research, we propose a predictive methodology for identifying any upcoming religious extremism-based threats in Sri Lanka using social media intelligence. We aim to specifically address Sri Lanka's multi-lingual culture by analyzing all the bilingual social media posts in Sinhala and Tamil languages. A hybrid sentiment analysis methodology consisting of a Machine Learning model and a sentiment lexicon was trained on carefully chosen labelled social media text data and each text was classified as either religious-extreme or not, using Naïve Bayes, SVM, and Random Forest algorithms. When comparing their results, we were able to achieve the best results with the Naïve Bayes algorithm resulting in an accuracy of 81% for Sinhala tweets while Random Forest algorithm resulted in an accuracy of 73% for Tamil tweets proving that social media intelligence can be used to predict religious extremism-based threats.**

Keywords: *Machine learning, Religious extremism, Social media, Text analytics*

I. INTRODUCTION

Sri Lanka is a country that suffered 30 years of ethnic civil war fought between the Liberation Tigers of Tamil Eelam (LTTE) and the Sri Lankan government, from 1983 until it ended in 2009. The hope was to have peace for a long period without any more bloodshed but then, on the 21st of April 2019, 10 years after the end of the civil war, Sri Lanka experienced a series of coordinated bomb blasts targeting three main churches and international hotels, which killed over 300 and injured hundreds more, frightening all Sri Lankans. National Thowheeth Jama'ath (NTJ), a Sri Lankan radical Islamist group claimed the responsibility of the attack that appears to have characteristics in common with those of Al-Qaeda and has its roots bound to the Islamic State of Iraq and Syria (ISIS). This attack was not a singular attack which happened out of nowhere but was stimulated by a series of previous religious-based riots that happened in different parts of the country such as anti-Muslim riots of 2014 around Beruwala, resulting in four dead and 80 injured citizens, anti-Muslim riots of 2018 in Digana, Kandy with attacks on the Muslim community and mosques. After suffering 30 years of civil war, Sri Lanka has become a victim of yet another form of intimidation; religious extremism that has been developed to a very dangerous level of religious extremism-based terrorism. Religious extremism is not an isolated problem faced only by Sri Lanka. According to the Global Shapers Survey conducted by the World Economic Forum (with 31,495 participants from 180 countries), religious conflicts are the fifth of the top ten problems the world is facing today [1]. Religious extremists endorse murder because they embrace theologies that provide sanctions for violence in the service of God. They have no sympathy towards the victims of their acts as they view them as enemies of their God [2]. Further, they are ready to sacrifice their own lives in expectation of a huge and immediate afterlife reward, in return for "martyrdom" [3]. The worst part of religious extremism is its convertibility to religious terrorism. Religious terrorism is a dangerous form of terrorism that is based on motivations and goals that may have predominant religious character and influence [4].

In the age of modern information, it is identified that these religious extremist groups have utilized modern communication techniques such as the Internet and social media, to directly communicate with their worldwide audience to harass users, recruit new members and incite violence [5]. Unlike in traditional communication methods, the free and unregulated nature of social media has helped extremists to easily form online communication to disseminate their belief and training materials without the fear of getting censored. Concerning the case of Easter Bombingcoordinated by NTJ, the leader of the organization Zahran Hashim had been very active in social media, specifically on Facebook since 2017, spreading their radical ideology [6]. When the incident happened, extremists created a wave of false but flammable information and spread them through social media rapidly, provoking one religious group against another. Therefore, the Sri Lankan government had to take a drastic decision even to ban social media including Facebook, WhatsApp, Instagram, Snapchat and Viber during the time of conflict.

Banning social media after an incident has happened, is futile. The need to have a methodology to identify religious extremism and its trend on social media and users who spread their extreme ideologies through social media in a threatening way has become important today more than ever. In this paper, we, therefore, propose a domain-based hybrid sentiment analysis methodology for identifying any upcoming religious extremism-based threats in Sri Lanka using social media intelligence. The proposed methodology consists of two parts,

a Machine Learning model and a sentiment lexicon. Previous researches conducted in the Sri Lankan context on sentiment analysis have always focused on the Sinhala language only. Sri Lanka is a multi-religious country. 70 % of the population are Buddhists, 13 % are Hindus, 9.7 % are Muslims, 7.4 % are Christians while 0.04% fall in the "others" category. In terms of the languages prevalent in Sri Lanka, 75 % of the population with a Buddhist majority, speak Sinhala as their mother tongue while 21 % comprised mostly of Hindus and Muslims, use Tamil as their mother tongue. Another 4 % use other languages such as English and Malay. This shows the fact that there is a clear relationship between a person's religion and his/her mother tongue in the Sri Lankan context. Conducting a study on a single language would, therefore, make the results biased towards a single religion. Hence, in our study, we propose a methodology that treats both Sinhala and Tamil languages similarly. In return, we get an unbiased result fair by Sri Lanka's multi-religious culture.

On the other hand, most of the studies that introduce text classification for the Sinhala language have rarely considered the sentiment value of each word due to the unavailability of resources. When it comes to the domain of religious extremism, it is important to consider the sentiment value of words when classifying over classifying just through the term frequency of the words.

Considering the above facts, in this research, we have explored the viability of identifying religious extremism-based threats in Sri Lanka using social media intelligence through a text analytics model. Our methodology is specific to countries with multi-religious cultures in the domain of religious extremism. The accurate results we received for the Sri Lankan context have proved that this methodology could be further extended to any multi-religious country.

II. Related Work

Most researchers have conducted studies to identify terrorism, Radicalization, Anti-Black, Jihad Extremism and hate speeches using social media data. The volume of content being posted on social media platforms makes it challenging for security analysts to discover such content manually or using keyword flagging. Further, textual posts on social media websites are user-generated content, which is unstructured and informal. Therefore, they come with lots of noisy content such as incorrect grammar, misspelled words, Internet slang, abbreviations and text containing multilingual scripts [7].

Alvari et al. [8] have proposed a detection scheme that could determine whether or not a given username belongs to an extremist user. They have used a dataset from twitter and first demonstrated that extremist users on twitter tend to adopt handles that follow similar patterns, in contrast to the normal users. Then a detection framework has been proposed to identify if a given Twitter handle belongs to an extremist giving its proximity to an existing set of extremist-related handles. They have compared different supervised and semi-supervised approaches using the features from the Twitter handle, profile information, and content which are highly indicative of online extremism. To further understand the significance of the features, they have conducted a significance analysis on the features using the labeled instances and have compared their results against char-LSTM which automatically extracts features.

Balahur and Turchi [9] have introduced a hybrid technique for sentiment analysis of Twitter texts. The research has used pre-processing to normalize the texts and the linguistic peculiarities of tweets have been taken into consideration. Spelling variants, slang, special punctuation, and sentiment-bearing words from the training data have been substituted by unique labels. For example, the sentence ''I love car'' was changed to ''I like car''; according to the General Inquirer dictionary since love and like both have a positive sentiment. This approach could be used for various languages with minimal linguistic processing. The method does not require any further processing and uses tokenization only. They state that their final system should work similarly for all languages.

An interesting work [10] demonstrates an excellent methodology to identify the state of art and independent techniques for multilingual sentiment analysis. According to their work, sentiment analysis on one single language increases the risks of missing essential information in texts written in other languages. One of the main problems in multilingual sentiment analysis is a significant lack of resources. Thus, sentiment analysis in multiple languages is often addressed by transferring knowledge from resource-poor to resource-rich languages to overcome the resource unavailability problem. Wan [11] has researched sentiment analysis of Chinese reviews by translating Chinese text to English text using Google Translator and his results indicate that analysis of English reviews translated by Google Translator outperforms the analysis of original Chinese reviews.

Tan and Zhang [12] introduced an approach for sentiment classification for the Chinese language. First, POS tagging is used; the aim of using POS tagging is to parse and tag the Chinese text. After POS tagging, feature selection is used to determine discriminative terms for classification. Finally, a machine learning approach is used for sentiment classification. Their method proves that POS tagging the text in the original language before translating reduces the error of translation and increases the accuracy of the model.

Hung et al. [13] have conducted research to detect radicalization trajectories using Graph Pattern Matching Algorithms. This research has taken a different approach to track an individual's behavioral indicators of homegrown extremism, using public and law enforcement data. The intuition behind the research is to use graph pattern matching to identify suspicious trajectories and potential radicalization over a dynamic heterogeneous graph associated with the fused data from public and law enforcement.

In 2018, a group of Sri Lankan researchers has conducted recent research to identify racist social media comments in Sinhala language using text analytics models with machine learning. They have shown that simple keyword spotting techniques cannot be used to identify the exact intent of a comment accurately and therefore, have built a Text Analytical Model with a Two-Class Support Vector Machine using Microsoft Azure Machine Learning Studio. In this research, they have employed the n-grams, carefully chosen and extracted from Facebook comments [14].

The survey research done by Agarwal and Sureka [7] reveals a variety of information on retrieval and machine-learning-based methods and techniques used by researchers to investigate solutions for social media extremism and online

radicalization detection. Clustering, Logistic Regression, and Dynamic Query Expansion are commonly used techniques to predict upcoming events related to social media extremism. They have observed that Named Entity Recognition (NER) is a common component in the text processing pipeline under various proposed approaches and techniques. Graph modeling is also a technique adopted by several researchers to perform event forecasting. Further, they have revealed that KNN (K Nearest Neighbor), Naïve Bayes, Support Vector Machine, Random Forest Algorithm, Rule-Based Classifier, Decision Tree, Clustering (Blog Spider), Exploratory Data Analysis (EDA), Topical Crawler/Link Analysis (Breadth-First Search, Depth First Search, Best First Search) and Keyword Based Flagging (KBF) are the most widely used techniques for online radicalization detection on social media. Their research has summarized that 90% of the studies have been successful in mining English Language text. Further, 60% of the studies target events specific to a country or region.

III. METHODOLOGY

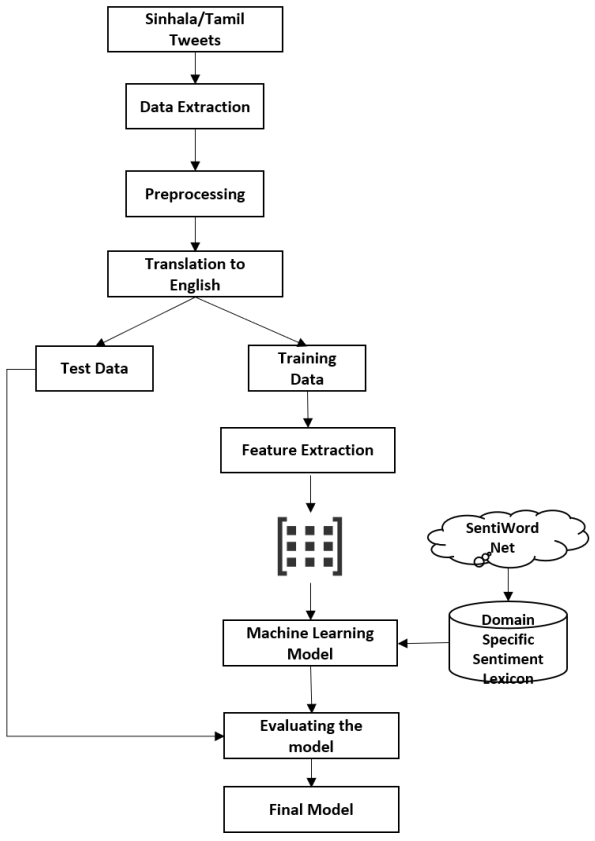

Fig 1. Research design

The proposed methodology follows a series of steps for text classification, comparing different Machine Learning algorithms for identifying the best model out of them. We use both Sinhala and Tamil tweets as inputs for our model and then evaluate the sentiment against the religious extremism of each where we finally classify them as either religious extreme or not. Figure 1 demonstrates a summary of the steps that we have performed throughout our study.

A. Data preparation

The data set was collected from Twitter using Twitter API by searching and identifying tweets based on the keywords and hashtags which contain extremism. Initially, 300 recent tweets; 200 in Sinhala and 100 in Tamil were collected based on the keywords shown in Table I. These keywords were selected by analyzing the frequency of their usage in identified extreme articles and posts on social media and with the help of language experts. Then, the extracted tweets were annotated as religious- extreme or non-religious- extreme as perceived by experts in both languages. This annotation was performed very carefully with the support of two language experts for each language. The annotations of the first expert were confirmed by the second expert to reduce the risk of misinterpretation. Though the data set collected was comparatively small, the perfect annotation increased the value of the dataset. Table II presents an example of how data were annotated.

TABLE I. KEYWORDS

Language	Keyword
Sinhala	හලාල්, කල්ලතෝනි, තම්බි, හම්බය, තම්බියො, හම්බයෝ, තම්බියෝ, හම්බයො, බුද්ධ, බුද්ධාගම, බුදුහාමුදුරුවො, පන්සල්, හාමුදුරුවෝ, මරමු, හින්දුන්, හින්දු, කෝවිල, කතෝලික, පල්ලි, පල්ලිය, සාමය, සමගිය, ආගම්, ශ්‍රීලාංකික, සමානාත්මතාවය, සහජීවනය, ආගමික, අන්තවාදය, වහබ්වාදය, අන්තවාදීන්
Tamil	தொப்பிபிரட்டி, சோனி, பாவாடை தோட்டக்காட்டான், சக்கிலியன் பறையன், கரையான், செம்பு தூக்கி இனவாதம், மதவெறி, மதக் கலவரம், சமாதானம், மத, தீவிரவாதம், கோவில், புத்த மதம், கிரிஸ்துவர், முஸ்லீம்

TABLE II. LABELLED TWEETS

Tweet (Sinhala)	Label
බෞද්ධයො බුද්ධාගමේ චිරස්ථිතිය වෙනුවෙන් හම්බයො මරන්න, පල්ලි කඩන්න, හම්බකඩ ගිනිතියන්න	Religious Extreme
මම බෞද්ධ මනුෂ්‍යයෙක්. මට අඹමල් රේණුවක තරමින්වත් ජාතිවාදී ආගම්වාදී වෙන්න බෑ. අන්තිම හුස්ම දක්වාම සාමය ,සහජීවනය සංහිඳියාව විතරයි	Non-Religious Extreme

Out of a total of 300 tweets that were extracted, 141 were annotated as religious extreme and 159 were annotated as non-religious extreme. As identified in the study by Dias et al. [14] having more non-religious extreme tweets than religious extreme tweets makes sure that our model will minimize the false positives in prediction.

B. Pre-processing of data

The extracted and annotated data were then subject to preprocessing. Data preprocessing is an important process in this study as bilingual text data is used. Though the analysis was done on English language text, the input to the model can belong to either Sinhala or Tamil languages. Preprocessing

techniques for each language are different from each other as different languages have different morphological bases.

As suggested in the literature, preprocessing the text in the primary language before translating it to the English language increases the accuracy of the study and minimizes the negative effect on the performance of the classification, since the translation of words might not preserve the semantic orientation due to the differences between languages.

In the first phase, we performed the general preprocessing steps that apply to both languages such as removing URLs, hashtags, duplicated tweets, empty tweets and emoticon only tweets, numbers, punctuations and replacing all the emoticons with their sentiment. In the second phase, we identified the language of the text and performed preprocessing steps that are unique to the identified language. Part of Speech (POS) tagging, Tokenization and stop word removal were committed on each language. Due to the lack of language literacy, the preprocessing of Tamil language was done only using existing tools such as IndicNLP and RippleTagger. POS tagger for the Sinhala language was developed based on the assumptions of existing studies using 1300 words Sinhala wordnet developed by the University of Colombo School of Computing (UCSC)[15]. The top 30 most common words identified through the work of [16] are considered as stop words and removed from each tweet. Tokenization and POS tagging on original language were committed to preserving the semantic orientation of the context by reducing the error of translation.

In the third phase, we translated text in non-English languages; Sinhala and Tamil, to the English language to get them into a common platform for reliable and unbiased results. It also enabled the employment of SentiWordNet which is an essential and rich lexicon resource when it comes to sentiment analysis. Here, the machine translation into English was performed on tokenized and POS tagged, stop words removed – tweets, using Google Translator which is identified in the literature as an accurate tool [17], [18], [11]. By taking a randomly selected 50% from the preprocessed and translated tweets of each language to form a sample, the accuracy of the translation was calculated with the support of the language experts. The translation of Sinhala tweets had 76% accuracy while the translation of Tamil tweets had 70% accuracy.

TABLE III. FEATURE EXTRACTION

Language	Feature Extraction method	Number of Features Extracted	Min word length	Max word length
Sinhala	Count Vector	987	1	1
	Word level TF-IDF	991	1	1
	N-gram level TF-IDF	2658	2	3
Tamil	Count Vector	382	1	1
	Word level TF-IDF	387	1	1
	N-gram level TF-IDF	1026	2	3

TABLE IV. NEED FOR SENTIMENT VALUES

Original Text	Translated Tokenized Text
බෞද්ධයෝ බුද්ධාගමේ චිරස්ථිතිය වෙනුවෙන් හම්බයෝ මරන්න, පල්ලි කඩන්න, හම්බ කඩ ගිනිතියන්න	['bauddhayo', 'kill', 'hambayo', 'buddhism', 'stability', 'church', 'break', 'earned', 'burn']

The words or phrases that could not be translated directly from Google Translator were kept in its transliteration form; Singlish (Sinhala written in English alphabet) or Tanglish (Tamil written in English alphabet) format, hence, what we received at the end of the preprocessing was not pure English. Although that feature added more complexity to our research problem, it contributed to the successful identification of the sentiment of the tweets in both languages. To handle this mixed language context and then to accurately classify the data, we have introduced a novel approach.

C. Feature extraction

To align with the process of using text in machine learning algorithms, we had to convert them to a numerical representation. The preprocessed dataset has many distinctive properties. In the feature extraction phase, we extracted the aspects from the preprocessed dataset into feature vectors.

Later, these aspects were used to compute the positive and negative polarity in a sentence towards religious extremism which is useful for determining the sentiment of each tweet. For experimentation purposes, we conducted our study on two feature extraction methods; Count Vectors as features and Term Frequency-Inverse Document Frequency (TF-IDF) as features. We further extracted TF-IDF feature vectors in word level (unigram) and N-gram level (Bi-gram and Tri-gram). As a result of the count vector feature extraction, we received a matrix that includes the term frequency of each word in each tokenized tweet. Unigram level TF-IDF returned a matrix representing the TF-IDF scores of every term in each tweet and N-gram level TF-IDF matrix representing the TF-IDF scores of N-grams. Table III summarizes the measures of each matrix received from data extraction.

At the end of the feature extraction phase, the preprocessed dataset was converted to its numerical representation for accurate text analysis. Feature names received from each feature extraction method were stored into a vocabulary.

D. Text classification model building and training

Text Classification models based on Sentiment analysis can be classified into three categories as corpus-based approaches, lexicon-based approaches, and hybrid approaches. Corpus-based methods use labelled data; lexicon-based methods rely on lexicons and optionally on unlabeled data; and hybrid methods are based on both labelled data and lexicons, optionally with unlabeled data. A sentiment lexicon is a collection of known sentiment terms with their positive and negative polarity values [10]. The objective of our study is to classify tweets according to their degree of religious extremism. Religious extremism is a domain that has a high

sensitivity to sentiment value. Considering only the feature vectors and their term frequencies cannot correctly identify the sentiment value of each term. Having the highest frequency does not confirm that a term has the highest sentiment. Let us take Table III as an example.

TABLE V. ASSIGNING POLARITY VALUES

Text	Sentiment
තම්බි මරන්ඩ යනවනන් මටත් කෝල් එකක් දියන්	Religious Extremist
නසිය කිව්වත් 👍 නානා කිව්වත් 👍 **තම්බි** කිව්වත් 👎 කිව්වේ ආදරයට	Non-Religious Extremist

The words ‘kill’, ‘break’ and ‘burn’ contain the highest negative polarity score though their Term Frequency or TF-IDF is less. This proves that a sentiment lexicon is essential for the validity and success of this kind of study. Therefore, a hybrid sentiment analysis model was chosen for building the text analysis model along with Machine Learning. SentiWordNet was used as the initial lexicon for the study to extract the sentiment values of each word due to its free availability for research purposes and continually updating nature.

As mentioned above, the language of the translated tokens of the tweets was not purely English; they had Singlish and Tanglish words along with pure English words. The words that were not in pure English could not be found in the SentiWordNet, but the sentiment values of those words cannot be ignored. (In the above example ‘bauddhayo’, ‘hambayo’ are in its Singlish format). However, when it comes to the domain of religious extremism, these two words definitely cannot be ignored as ‘bauddhayo’ implies ‘Buddhists’ and ‘hambayo’ implies ‘Muslims’. To address this need, a context-based algorithm was developed using the pragmatic meaning of each tweet. The polarity value of each word was extracted from the sentiment value of the labelled tweet in which the particular word is included. Then along with the training of the model, the polarity of the particular word was adjusted. As an example, the word ‘thambi’, used to address Muslims, was included in most of the tweets which spread hate towards Muslims, but it is also included in some positive tweets. Therefore, the final negative polarity value identified for the word ‘thambi’ is 0.57 which implies that the word ‘thambi’ does not imply that the tweet containing that word is always religious extreme. Table V provides a good example of this case.

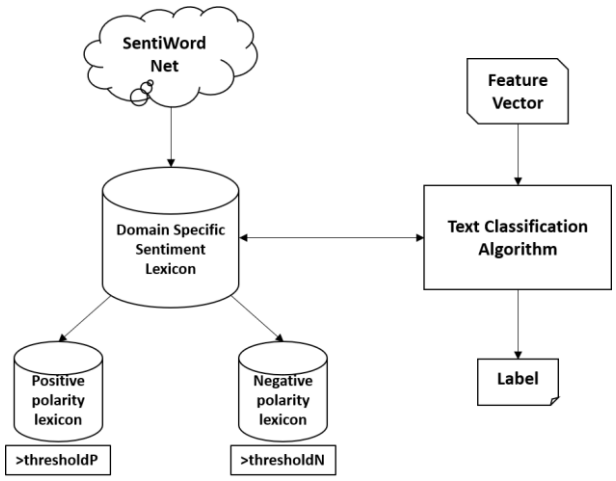

Fig 2. Text Classification Model

A Singlish lexicon consists of 389 words and a Tanglish lexicon consists of 137 words that were constructed using the developed algorithm with the calculated negative and positive polarity values. Then using SentiWordNet and the lexicons constructed, the weight of each word or N-gram in the vocabularies built during the feature extraction was calculated. By multiplying each feature vector matrix with their vocabulary’s weighted value array, the final numerical representation of the dataset was created as follows.

Matrix 1 = Array of weights of each term x Count Vector

Matrix 2 = Array of weights of each Unigram x TF-IDF Word Level

Matrix 3 = Array of weights of each N-gram x TF-IDF N-gram level

At the end of these processes, all the text data were converted to a numerical representation that represents the feature vectors and their sentiment values. These matrices were then inserted into the machine learning model for classification as a religious extremity or not as a religious extremity.

For experimentation purposes, several classification algorithms were selected from literature to perform text analysis and to identify the most suitable classification algorithm for our study. The algorithms selected were Naïve Bayes Classifier, Support Vector Machine (SVM) classifier, Logistic Regression, Random Forest Algorithm, Xtreme Gradient Boosting Model, Linear Classifier and Recurrent Neural Network (LSTM). Out of these seven classifiers, the present study used the classifiers which had achieved more than 70% accuracy, for further analysis to identify the most suitable algorithm to recognize religious extremism-based threats. The algorithms which had an accuracy of more than 70% were Naïve Bayes Classifier, Support Vector Machine (SVM) classifier and Random Forest Algorithm. The success of each algorithm was evaluated using 10-fold cross-validation which has been identified to be suitable for a small dataset rather than using train/test split, to reduce the risk of overfitting and to assess any such occurrence.

IV. Results and Discussion

The three feature matrices generated were used with Naïve Bayes classifier, SVM Classifier and the Random Forest Classifier for both Sinhala and Tamil language tweets. After the model was trained on training data, it was tested using the test dataset. The results were significant and deducible. Tables VI, VII, VIII below summarizes the test results of the trained machine learning models.

TABLE I. COMPARISON OF RESULTS FROM DIFFERENT CLASSIFICATION ALGORITHS WITH MATRIX 1

Sinhala				
Algorithm	***Accuracy***	***Precision***	***Recall***	***F1 -Score***
SVM	0.52	0.52	1.0	0.68
Naïve Bayes	0.81	0.72	0.96	0.82
Random Forest	0.76	0.71	0.92	0.80
SVM	0.67	0.67	1.0	0.80
Naïve Bayes	0.69	0.68	1.0	0.81
Random Forest	0.73	0.72	1.0	0.84

TABLE VIII. COMPARISON OF RESULTS FROM DIFFERENT CLASSIFICATION ALGORITHS WITH MATRIX 2

Sinhala				
Algorithm	***Accuracy***	***Precision***	***Recall***	***F1 -Score***
SVM	0.55	0.67	**1**	0.8
Naïve Bayes	0.60	0.59	0.95	0.73
Random Forest	**0.79**	**0.74**	1	**0.85**
Tamil				
SVM	0.57	0.58	**1**	0.73
Naïve Bayes	0.63	0.7	**1**	**0.82**
Random Forest	**0.7**	**0.61**	0.87	0.72

TABLE VIII. COMPARISON OF RESULTS FROM DIFFERENT CLASSIFICATION ALGORITHS WITH MATRIX 3

Sinhala				
Algorithm	***Accuracy***	***Precision***	***Recall***	***F1 - Score***
SVM	0.54	0.55	1.0	0.71
Naïve Bayes	0.73	0.65	1.0	0.79
Random Forest	0.53	0.51	1.0	0.68
Tamil				
SVM	0.62	0.62	0.69	0.65
Naïve Bayes	0.64	0.60	**0.86**	**0.71**
Random Forest	**0.74**	**1.0**	0.50	0.67

As summarized in Table VI for Matrix 1, Naïve Bayes algorithm shows the highest accuracy, precision, recall and F1 score for Sinhala language tweets while Random Forest algorithm gives the highest accuracy, precision and recall for Tamil language tweets. Table VII explains how Matrix 2 performed with each Classification algorithm. According to the results obtained, the Random Forest algorithm provides the best results for both Sinhala and Tamil languages.

According to Table VIII for matrix 3, the Naïve Bayes algorithm performed well for Sinhala language tweets while the Random forest algorithm gave the best results for Tamil language tweets. By considering all these results, we were able to finalize the best text analysis model for each language.

For Sinhala language tweets, the Naïve Bayes Classification algorithm gives the best results with Count Vectors as a feature extraction method. This model achieves an accuracy of 81% that can be considered as the highest score when it comes to the classification of data in a sensitive domain like religious extremism. It also shows a precision of 0.72, recall of 0.95 and F1 Score of 0.82. Having a recall of 0.96 means that our model returns most of the relevant results. It is least likely to classify a religious extreme tweet as a non-religious extreme tweet. Having a good precision value like 0.72 proves that the developed model returns substantially more relevant results than irrelevant results. Having a high recall value and a good precision value gives a good F1 value which is an excellent measure for evaluating the model better than the accuracy. It gives the weighted average of precision and recalls enabling us to understand the success of our model and how it performs by giving expected results. By evaluating and examining all 4 measures, we can conclude that the model created with the Naïve Bayes Classification algorithm and Count Vectors as features is the best in identifying religious extremism tweets in the Sinhala language.

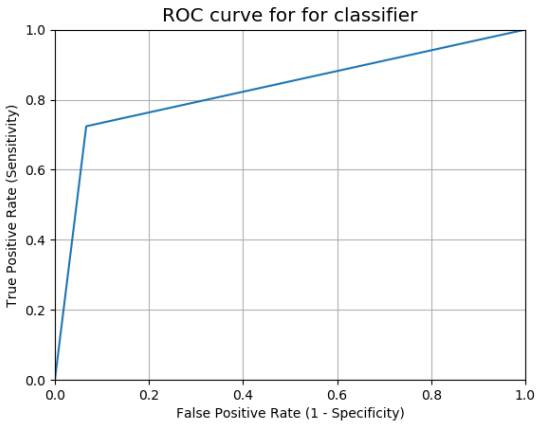

Fig 3. ROC curve of the best model for Sinhala

The Receiver Operating Characteristics Curve (ROC) of the model created using the Naïve Bayes algorithm and Matrix 1 for the Sinhala language is shown in Fig. 3. The curve has spread closer to the top left corner of the plot, depicting the better performance of our model. The calculated Area Under Curve (AUC) is 0.83 and it can be considered as a high score to indicate that our model is excellent in distinguishing between classes; religious extreme, and non-religious extreme.

For the Tamil language, the best working model is the Random Forest Algorithm with Matrix 1 which has achieved 0.73 highest accuracies along with the precision of 0.72, recall of 1.0 and F1 Score of 0.84. Having a recall of 1.0 indicates that our model always returns relevant results only.

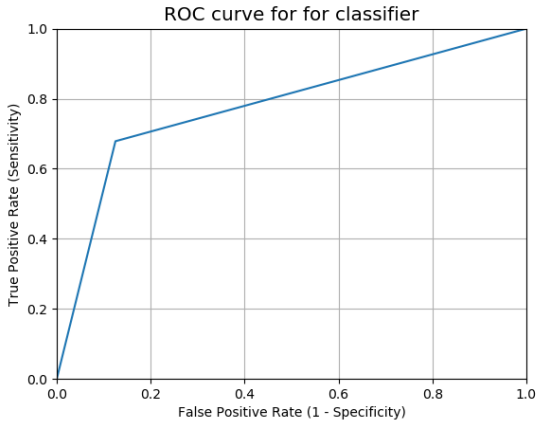

Fig 4. ROC curve of the best model for Tamil

The highest F1 Score of 0.84 gives an idea of how precise and robust the model developed is. Fig. 4. depicts the ROC Curve of the identified best performing model above for Tamil language tweets demonstrating the tradeoff between sensitivity and specificity. As the curve has spread closer to the left-hand border and then to the top border of the ROC space, it proves that our model has generated more accurate test results. The calculated AUC value of the curve is 0.78 and it can also be identified as a good score indicating that our model behaves well in classifying data between classes; religious extreme, and non-religious extreme.

Results received for both Sinhala and Tamil languages depict that our model is excellent in identifying religious extreme content in social media. Therefore, the trained and tested model can be used to classify a live stream of Sinhala and Tamil tweets to see if there is any trend of religious extremism that is nurtured through social media platforms. By frequent monitoring of social media data through this model, upcoming religious extremism-based threats in Sri Lanka could be predicted accurately, also focusing on Sri Lanka's multi-religious culture.

V. Conclusion and Future Work

In this paper, we presented a model that can be used to identify religious extremism using social media intelligence. As was described in the literature review, previous studies on sentiment analysis have barely focused on the domain of religious extremism though they have covered areas like racism, terrorism, and radicalization on social media platforms. Those domains can be easily differentiated from the domain of religious extremism since religion is a more sensitive area that is based on someone's faith and belief for a journey of self-cultivation and self-realization. In this stage, we have considered the sentiment value of each word and term without relying only on the frequency of each word or term in the data corpus. Methods used in other similar domains like racism and radicalization were not apt to be used here. Therefore, we have introduced a novel method of identifying religious extremism by engaging a context-based algorithm that addresses the pragmatic meaning of each tweet. On the other hand, we have developed this model to match with the Sri Lankan context by enabling the analysis of both Sinhala and Tamil language social media posts and it has achieved successful results for both languages. None of the sentiment analysis researches in Sri Lanka have addressed Sri Lanka's multilingual culture and most of them have only focused on Sinhala language data. As explained, there is a clear relationship between a person's religion and his/her mother tongue in the Sri Lankan context. Therefore, conducting a study on a single language would make the results more biased to a single religion. Thus, in the present study, we have proposed a methodology that treats both Sinhala and Tamil languages equally. In return, we get unbiased results that are fair by Sri Lanka's multi-religious culture. Using the results, we have identified that Naïve Bayes Classifier is the best algorithm for the Sinhala language with an accuracy of 81% while Random Forest Classifier is the best algorithm for the Tamil language with an accuracy of 73% in identifying the religious extremism using social media intelligence. For future work, we expect to extend this analysis with more social media data from other platforms like Facebook and YouTube. Doing so, we believe that we will be able to get more data on the Tamil language for better accuracy. Additionally, we are hoping to interact more with language experts in both languages to increase the accuracy of the model by understanding the special properties related to each language., We expect to generalize our model further, to analyze religious extremism in other multireligious, multilingual countries.

References

[1] T. L. Jackson, Abby, The 10 most critical problems in the world, according to millennials. Business Insider. Available: https://www.businessinsider.com/world-economic-forum-world-biggest-problems-concerning-millennials-2016-8 [Accessed: 16-Dec-2019].

[2] L. R. Iannaccone and E. Berman, "*Religious extremism: The good, the bad, and the deadly,"* Public Choice, vol. 128, no. 1, pp. 109–129, Jul. 2006.

[3] B. Caplan, *"Terrorism: The relevance of the rational choice model,"* Public Choice, vol. 128, no. 1, pp. 91–107, Jul. 2006.

[4] Brannan, David. "Understanding Terrorism: A Social Science View on Terrorism – CHDS Self-Study Courses." Available: https://www.chds.us/selfstudy/courses/understanding-terrorism/. [Accessed: 16-Dec-2019].

[5] J. S. closeJoanna S. correspondent covering S. AsiaEmailEmailBioBioFollowFollow, "The remote Sri Lankan enclave that produced the mastermind of a massacre," Washington Post. Available: https://www.washingtonpost.com/world/asia_pacific/the-remote-sri-lankan-enclave-that-produced-the-mastermind-of-a-massacre/2019/04/26/9358f48c-6830-11e9-a698-2a8f808c9cfb_story.html. [Accessed: 16-Dec-2019].

[6] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in Social Informatics, Cham, 2016, pp. 22–39.

[7] S. Agarwal and A. Sureka, "Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats," arXiv:1511.06858 [cs], Nov. 2015.

[8] Alvari, Hamidreza, S. Sarkar, and P. Shakarian. *"Detection of Violent Extremists in Social Media."* 2019 2nd International Conference on Data Intelligence and Security (ICDIS), 2019, pp. 43–47.

[9] A. Balahur and M. Turchi, *"Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data,"* in Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria, 2013, pp. 49–55.

[10] K. Dashtipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cogn Comput, vol. 8, no. 4, pp. 757–771, Aug. 2016.

[11] X. Wan. *"Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis".* In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics; 2008, pp. 553–61.

[12] Tan, S. and Zhang, J., *"An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications",* vol. 34, no. 4, pp. 2622-2629, 2008

[13] B. W. K. Hung, A. P. Jayasumana, and V. W. Bandara, *"Detecting radicalization trajectories using graph pattern matching algorithms,"* 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 313–315, 2016.

[14] D. S. Dias, M. D. Welikala, and N. G. J. Dias, *"Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning,"* in 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), 2018, pp. 1–6.

[15] "Lexical Resources – Local Language Portal." [Online]. Available: https://www.language.lk/en/resources/lexical-resources/. [Accessed: 16-Feb-2020].

[16] S. Gallege, Analysis of Sinhala Using Natural Language Processing Techniques. CiteSeerX. 2010.

[17] Remus, Robert, Uwe Quasthoff, and Gerhard Heyer. "SentiWS - A Publicly Available German-Language Resource for Sentiment Analysis." In LREC, 2010.

[18] H. Ghorbel and D. Jacot, "Sentiment Analysis of French Movie Reviews," in Advances in Distributed Agent-Based Retrieval Tools, V. Pallotta, A. Soro, and E. Vargiu, Eds. Berlin, Heidelberg: Springer, 2011, pp. 97–108.