An Effective Text Classifier using Machine Learning for Identifying Tweets' Polarity Concerning Terrorist Connotation

Norah AL-Harbi
Faculty of Computer Science and Information Teqnology, Taif University, Taif, Saudi Arabia
E-mail: N.motaib@tu.edu.sa

Amirrudin Bin Kamsin
Faculty of Computer Science and Information Teqnology, University of Malaya, Kuala Lumpur, Malysia
E-mail: amir@um.edu.my

Abstract: Terrorist groups in the Arab world are using social networking sites like Twitter and Facebook to rapidly spread terror for the past few years. Detection and suspension of such accounts is a way to control the menace to some extent. This research is aimed at building an effective text classifier, using machine learning to identify the polarity of the tweets automatically. Five classifiers were chosen, which are AdB_SAMME, AdB_SAMME.R, Linear SVM, NB, and LR. These classifiers were applied on three features namely S1 (one word, unigram), S2 (word pair, bigram), and S3 (word triplet, trigram). All five classifiers evaluated samples S1, S2, and S3 in 346 preprocessed tweets. Feature extraction process utilized one of the most widely applied weighing schemes tf-idf (term frequency-inverse document frequency).The results were validated by four experts in Arabic language (three teachers and an educational supervisor in Saudi Arabia) through a questionnaire. The study found that the Linear SVM classifier yielded the best results of 99.7 % classification accuracy on S3 among all the other classifiers used. When both classification accuracy and time were considered, the NB classifier demonstrated the performance on S1 with 99.4% accuracy, which was comparable with Linear SVM. The Arab world has faced massive terrorist attacks in the past, and therefore, the research is highly significant and relevant due to its specific focus on detecting terrorism messages in Arabic. The state-of-the-art methods developed so far for tweets classification are mostly focused on analyzing English text, and hence, there was a dire need for devising machine learning algorithms for detecting Arabic terrorism messages. The innovative aspect of the model presented in the current study is that the five best classifiers were selected and applied on three language models S1, S2, and S3. The comparative analysis based on classification accuracy and time constraints proposed the best classifiers for sentiment analysis in the Arabic language.

Index Terms: Twitter, machine language, Terrorism, Arabic, messaging.

1. Introduction

With the advent of the Internet, activities like sales and promotion, advertisements, e-commerce, cryptocurrency, e-learning, customer services, etc. have been revolutionized. The use of the Internet has increased exponentially in the last decade. Social media encompasses enormous applications that allow users to create and share content widely on social networking sites. Social networks like Facebook, Twitter, and YouTube are used extensively for communication and socialization. It is also used as a platform for conducting virtual, real-time meetings. Twitter is popular amongst the diverse networks available.

Twitter is an open and informal microblogging platform for people across the globe to express their views, opinions, and emotions freely. It allows the users to send and read messages about 280 characters per tweet. During the initial establishment of the Internet, the password-secured conversation boards permitted anonymous conversation amongst the affiliates. Besides, they also allowed the Islamic terror cells to propagate and talk on topics related to jihad. It served as a turning point for extremists and they started using Twitter as an avenue to publicize their posts and radicalize people. Twitter is used by radical elements as a battlefield for Jihad, a place for missionary work, a field for confronting the enemies of God [4]. Many terrorists have embraced Twitter for sharing hate speeches and gruesome beheadings to spread panic and terror in public at large [7]. Tweets come as videos, messages, and presentations.

Fewer Internet charges, low-cost portable devices, and increased social importance have made many people have a Twitter account. Huge audience, ease of use, and accessibility made twitter a dangerous platform for the propagation of

terror. It is used as an avenue by terrorists to carry out recruitment and training new members, to send orders and instructions as tweets to execute terrorist activities. The usage of Twitter for disseminating radical thoughts has been high due to the reliable results that these terror groups have achieved.

The majority of the terror groups such as Al-Qaeda and ISIS and their activities have been traced to the Arab world and Arabic language. Arabic is one of the most popular languages in the world and one of the six official languages of the UNO. Though it is a widely used language, there is little research done on retrieval or data mining. Arabic linguistic features are different from English in terms of structure and grammar. The extended attributes of the language with different characteristics, various dialects, and free writing forms used make research in the field of Arabic sentiment analysis more complicated.

Twitter contains an enormous amount of data, and is emerging as a real-time repository of knowledge that can be exploited by researchers and applications. Sentiment analysis has become one of the dynamic research fields in NLP (Natural Language Processing). It can be used to extract information from unstructured data. It has become a crucial element for decision-makers and business leaders for making essential investments in measuring public opinion about their products or services. It is an area of data mining where the opinion of people from the text data of tweets can be analyzed and classified into positive, negative, or neutral. It evaluates the polarity in various domains.

The research study is conducted to find a viable model with proposed features and evaluate its performance to detect radical content in Arabic on Twitter.

1.1. Problem Statement

Tweet classification and detection of tweet polarity have been an area of interest in the field of computer science because a wider user community posts tweets on Twitter each day. The frequency of words, causes, motives, and promotions come into the limelight in the form of Twitter trends. The users also misuse this feature and propagate hate speech and terrorist agenda by using provocative hashtags and keywords. The Arabic language has an entirely different structure than the English language. It has 28 alphabets divided into 14 sun and 14 moon letters. The script is written right to left, and letters may have different forms based on the preceding or following letters. Therefore, parsing and classifying tweets in the Arabic language is far more complicated than in English. The study of Gamal et al. [14] also confirmed that studies on Arabic sentiment analysis on Twitter are very limited when a comparison is made with English sentiment analysis. Gamal et al. [14] attributed this situation to the unique difficulties and complexities associated with the Arabic language. According to Gama et al. [14], it has led to a reduction in Arabic datasets that are used for sentiment analysis of opinions available on Twitter in Arabic. Zaib et al. [16] highlighted the need for using aggression detection techniques on Twitter so that the misuse of this networking platform could be avoided. According to Zaib et al. [16], users may exhibit varied and unpredictable behaviors on social media, and therefore, the best approach to track those behaviors is social media mining and behavior analytics. Akther et al. [17] suggested that cyberbullying should also be evaluated in relation to terror messages and the study of Akther et al. [17] presented a training algorithm that classifies the content of cyberbullying into three sub-categories of harassment, racism, and shaming.

Twitter is an American company headquartered in California. Therefore, there is not the same level of focus and commitment in classifying Arabic text messages as in English. The local and regional Arabic users, on the other hand, are attracted more to the Arabic messages, and hence, it has also become a strategic tool for the terrorists to advance their hate agendas and false motives. This study recognizes the need for developing a classification scheme for detecting Arabic terror messages. The existing solutions have shown promising results in English language message detection. The study capitalizes on the findings of these machine learning algorithms and presents a comprehensive scheme for detecting Arabic terrorism messages.

1.2. Research Objectives

The following research objectives were formulated by considering the focus of this study on using machine learning algorithms and detecting Arabic terrorism messages.

- To identify different classification algorithms that are available in the contemporary context for tweets' classification
- To evaluate how the tweets' classification algorithms can be optimized and mapped to detecting Arabic terrorism messages on Twitter

2. Related Work

Studies by [13,5] have clearly demonstrated that terrorist's accessibility and connections to online materials and resources have played a significant role in their radicalization process and procedure. These terrorist groups primarily focus on destabilizing people by causing unease, fear, and terror among the majority of harmless citizens through their dissemination of ferocious audios and videos displaying fights and the massacring process of individuals who have different ideologies from the terrorist groups [10]. The activities and practices of terrorists have been discovered and

there is evidence that they are rampant in most Arab countries [2]. The majority of these terror groups such as al-Qaeda, ISIS, and Al-Shabaab have been traced to the Arab world. Moreover, Twitter has been discovered to be one of the most vital tools adopted by these terrorist groups to disseminate threats, enlist members, and serve as avenues for the training of new members [8].

One of the key techniques to help society is by tracking and suspending accounts used by radical groups to disseminate information about recruitment and activities, propaganda, hate speeches, and other terrorist and extremist tendencies [11]. However, the method requires the activities of human analysts to read and scrutinize the vast amount of data on social media carefully and physically. Also, with the huge amount of data on the Twitter platforms that is gross, rigid, abstracted, understandable, and unreadable, this led to an attempt to automatically detect tweets with terrorist content. Facebook recently blocked many accounts of Taliban leaders with the rise of this group in Afghanistan [18]. According to Facebook, the company considers the Taliban a terrorist group, and therefore, the content related to the group should not be allowed on the social platform [18]. However, Twitter desisted from adopting this strategy and clarified that the Taliban spokesman may continue to maintain the account on Twitter as long as the account abides by the terms of service and does not incite violence [19]. Twitter is being criticized for this soft policy, and tweet classification and sentiment analysis are the only solutions to track and report the terror content.

There are several English sentiment analysis datasets including terrorism detection. However, the datasets in the Arabic language are limited. The Arabic language has extended attributes with different characteristics when compared with Western languages. In addition, there are several extended Arabic social features such as culture, race, and education. The Arabic social characteristics make it difficult to be used across language techniques to detect a topic like terrorism in tweets using machine learning techniques [3]. There is some limited literature on the problem of radical discovery on Arab social media [9].

In this study, Arabic radical content on Twitter is detected and Arabic tweet is classified as supporting radicalism on Twitter or no by machine learning. Also, it focuses on classifying the text, using linguistic heuristics (language model) in the Arabic language. The ways on how to trend and implement the techniques to detect the tweets will be stated more clearly in the methodology section of the study.

3. Methodology

The terror orientation process adopted here is divided into three main phases: data analysis, feature extraction, and training model. By applying these three phases in the proposed methodology, the researcher accomplishes the research objectives. As part of the first research objective, different machine learning algorithms were tested for their effectiveness and relevance in detecting Arabic terrorism messages. As part of the second research objective, tweets' classification for Arabic messages was carried out to demonstrate the effectiveness of proposed text parsing algorithms.

3.1. Data Analysis

This section shows how data was collected and divided, as well as the application of statistical methods. It also shows how Cronbach's Alpha, which is a measure of scale reliability, was calculated. The preprocessing process is also explained in this section.

A. Dataset

Data in March 2008 was collected using Twitter API and was further built as a data set. The number of collected tweets was 135,069. The number of tweets was reduced to 346 tweets in the cleaning step. Arabic tweets were retained. Further, tweets that were not related to terrorism, repeated tweets, incomplete tweets, tweets with links only, and tweets of news accounts that broadcast news stories about events that were not considered supportive or against terrorism were discarded. In this work, the dataset collected was divided into two separate datasets: TW-PRO and TW-CON. The datasets used and analyzed during the study are available from the corresponding author on reasonable request.

B. Manual Labeling

The Arabic tweets were collected from Twitter through the number of specific hashtags based on previous studies (Ali, 2016; Omer, 2015; Salmi Abdellatif, 2016). Then the data set was cleaned, and preprocessed (346 tweets). The 346 tweets were presented to four experts (native speakers of Arabic, who are also Arabic language teachers), who classified the tweets based on a questionnaire.

C. Preprocessing

Getting rid of the unwanted elements in the raw data will improve the performance of the classifier. The following preprocessing tasks were accomplished:

- Tokenization of the tweets.
- Cleaning of the data by removing the stop words, punctuation, blank spaces, diacritic marks (pronunciation signs), etc.

3.2. Feature Extraction

Feature extraction was an integral component of the research methodology. The process began with initial datasets of tweets and derived meaningful information concerning the terrorism content by applying machine learning algorithms. The key benefit of feature extraction was that it significantly reduced the amount of data that was required to be processed. It enabled the researcher to focus on tweets relevant to the terrorism content. This technique also enabled the extraction of new features based on the combination of the available features. This learning from the original datasets highlighted new dimensions of Arabic message tweets.

A. Feature Vectors

A feature vector can be thought of as an n-dimensional vector, which is based on numerical features (Omer, 2015). Here, language models of n-grams (unigram, bigram, and trigram) were chosen, as they are cost, time, and size effective. (Choi et al., 2014).

Table 1. Size of features.

Sample	No of features	Approaches
S1	3317	Benchmark
S2	8424	Proposed
S3	13628	Proposed

Table 1 shows three sample features that were used in this work (S1, S2, S3). "S" is referring to the sample and subscript to the n-gram models that are used. Uni-gram, bi-gram, and tri-gram were represented as S1, S2, and S3 in the language models, respectively.

B. Feature Extraction Vector Representation

One of the most widely applied weighing schemes is tf-idf (term frequency-inverse document frequency). The representation is applied to assess the importance of a word to a given document in a corpus of texts.

TF (t) = (the number of incidences term t is found in a document) / (the total terms).
IDF (t) = log e (Document total number / the number of documents with word t).
Tf-idf weight = Tf (t) * IDF (t)
In this way, the matrix is created with TF-IDF weights by placing the values in their assigned rows.

3.3. Training and Testing Model

The aim is to create an Arabic terrorism model using machine learning algorithms and text classification.

A. Classification Algorithms

The process of classification algorithms utilized in this study is explained below. The process is sourced and mapped from the study of Sharma and Hoque [15]:

- Extract the relevant tweets from the obtained datasets.
- Given a set of tweets, estimate tweets' distribution and classify them into positive, negative, and neutral classes based on terrorism content.
- Develop a machine learning model by comparing the available models of sentiment analysis and text mining. Recommend the best model for predicting future tweets regarding their positive, neutral, or negative connotation.

In the classification setting, a set of training datasets is used to train a classifier. Classification algorithms were implemented to extract terrorism detection, and the results were evaluated through matrix confusion.

B. Evaluation Measures

The cross-validation method splits the data into two parts: testing data and training data. It is an important step for testing the efficiency of the classifier, which performs multiple evaluations on different test sets, and combines the scores from these evaluations. In particular, the k-fold cross-validation method was used, in which the data was divided into k parts. The methods used to evaluate the performance of classification models: precision, recall, F-measure, and accuracy.

4. Result and Evaluation

4.1. Experimental Setup

Table 2. Experimental setup of terror orientation detection.

Name	Description
Data collecting	Twitter
Validity	Statistical approach
Number of tweets	346
Language programing	Python 3.6
NLP tools	NLTK, API python
Classification tools	Sklearn
Classification Methods	-SVM - Naïve Basie - Logistic Regression - AdB_SAMME -AdB_SAMME.R
Statistical Programs	SPSS version 24

Table 2 shows the experimental setup used in this research. The process of classification algorithms was executed in the following steps:

- Step1: The dataset was collected from Twitter using python API techniques.
- Step2: The statistical analysis approaches were applied to evaluate the collected dataset in step (1) using SPSS version 24.
- Step3: The feature extraction approaches of text were vectorized using several approaches discussed in the earlier section.
- Step4: The feature extraction vectors were trained and tested by several classification approaches, as shown in Tables 3 to 17 using the Sklearn library in python.

The 346 tweets were given to experts (three teachers and an educational supervisor in Saudi Arabia) who evaluated and judged the tweets to determine whether they were radical or not radical. As the value of the Cronbach's Alpha obtained was >0.8, it confirms the validity and reliability of the data to conduct further experiments on them.

4.2. Experiment Result Analysis

Two research objectives in this study were focused on evaluating different classifiers and proposing the best classifier for detecting Arabic terrorism messages on Twitter. The analysis below is based on evaluating five renowned classifiers using three features S1 (unigram, one word), S2 (bigram, word pair), and S3 (trigram, word triplet). Based on the analysis, the best classifier has been indicated. Moreover, the viability of other classifiers has been mentioned if time constraints and other factors are considered.

A. The Experiment Analysis using F1-Score

Fifteen experiments were conducted with five classifiers (AdB_SAMME, AdB_SAMME.R, Linear, NB, and LR) on each of the 3 features (S1, S2, and S3) separately, class-wise and the average for Precision, Recall and F-measure were presented. Precision is mathematically defined by $TP/(TP + FP)$ and Recall by $TP/(TP+FN)$.

TN / True Negative: label was negative, and the model predicted negative
TP / True Positive: label was positive, and the model predicted positive
FN / False Negative: label was positive, but the model predicted negative
FP / False Positive: label was negative, but the model predicted positive]

In the tables below, the "Support" column represents the total number of samples from each class. In "avg/total" row, "avg" refers to the first three columns, i.e. "precision", "recall" and "f1-score", whereas "total" refers only to "Support". "Avg" is calculated by summing up the weighted value of each class. For example, how the "avg/total" row is calculated in Table 3, is presented as follows:

"avg" for precision = .72*(146/(146+200))+.92*(200/(146/200)) ≈ .84
"avg" for recall = .91*(146/(146+200))+.74*(200/(146/200)) ≈ .82
"avg" for f1-score = .81*(146/(146+200))+.82*(200/(146/200)) ≈ .82
"total" for support = (146+200) = 346

F-measure is also called the F-Score. In other words, the F1 score conveys the balance between precision and recall.

Table 3. Feature extraction of unigram sample by AdB_SAMME classification model.

	precision	recall	f1-score	Support
support	0.72	0.91	0.81	146
not support	0.92	0.74	0.82	200
avg / total	0.84	0.82	0.82	346

Table 4. Feature extraction of S1 by AdB_SAMME.R classification model

	precision	recall	f1-score	Support
support	0.72	0.91	0.81	146
not support	0.92	0.74	0.82	200
avg / total	0.84	0.82	0.82	346

Table 5. Feature extraction of S1 by Linear classification model

	precision	recall	f1-score	Support
support	1.00	0.99	0.99	146
not support	0.99	1.00	1.00	200
avg / total	0.99	0.99	0.99	346

Table 6. Feature extraction of S1 by NB classification model

	precision	recall	f1-score	Support
support	0.99	1.00	0.99	146
not support	1.00	0.99	0.99	200
avg / total	0.99	0.99	0.99	346

Table 7. Feature extraction of S1 by LR classification model

	precision	recall	f1-score	Support
support	1.00	0.95	0.98	146
not support	0.97	1.00	0.98	200
avg / total	0.98	0.98	0.98	346

Table 8. Feature extraction of: S2 by AdB_SAMME classification model

	Precision	recall	f1-score	Support
support	0.72	0.91	0.81	146
not support	0.92	0.74	0.82	200
avg / total	0.84	0.82	0.82	346

Table 9. Feature extraction of S2 by AdB_SAMME.R classification model

	precision	recall	f1-score	Support
support	0.72	0.91	0.81	146
not support	0.92	0.74	0.82	200
avg / total	0.84	0.82	0.82	346

Table 10. Feature extraction of S2 by Linear classification model

	Precision	recall	f1-score	Support
support	1.00	0.99	1.00	146
not support	1.00	1.00	1.00	200
avg / total	1.00	1.00	1.00	346

Table 11. Feature extraction of S2 by NB classification model

	precision	recall	f1-score	Support
support	0.99	1.00	0.99	146
not support	1.00	0.99	0.99	200
avg / total	0.99	0.99	0.99	346

Table 12. Feature extraction of S2 by LR classification model

	precision	recall	f1-score	Support
support	1.00	0.98	0.99	146
not support	0.99	1.00	0.99	200
avg / total	0.99	0.99	0.99	346

Table 13. Feature extraction of S3 by AdB_SAMME classification model

	precision	recall	f1-score	Support
support	0.73	0.91	0.81	146
not support	0.92	0.75	0.83	200
avg / total	0.84	0.82	0.82	346

Table 14. Feature extraction of S3 by AdB_SAMME.R classification model

	precision	recall	f1-score	Support
support	0.88	0.73	0.79	146
not support	0.82	0.93	0.87	200
avg / total	0.84	0.84	0.84	346

Table 15. Feature extraction of S3 by Linear classification model

	precision	recall	f1-score	Support
support	1.00	0.99	1.00	146
not support	1.00	1.00	1.00	200
avg / total	1.00	1.00	1.00	346

Table 16. Feature extraction of S3 by NB classification model

	precision	recall	f1-score	Support
support	0.99	1.00	0.99	146
not support	1.00	0.99	0.99	200
avg / total	0.99	0.99	0.99	346

Table 17. Feature extraction of S3 by LR classification model

	precision	recall	f1-score	Support
Support	1.00	0.98	0.99	146
not support	0.99	1.00	0.99	200
avg / total	0.99	0.99	0.99	346

Tables 3 to 17 depict the details of the precision, recall, and F-score. It is evident from the difference in the values of classification based on the quality of the classification technique that there is an enhancement in reading the aptness of the tweets. Besides, it is obvious from the tables that the linear classification technique achieved the best value of the F-score owing to the reason that the polarity is bi-directional (terrorism or not-terrorism).

5. Discussion

It can be observed from the results that linear SVM performs very well on S3 (trigram) features with 99.7 % classification accuracy. It can be seen that S3 (trigram) features tend to perform slightly better than unigram and bigram features.

As depicted in Table 18, it is obvious that the accuracy with the Linear classifier is (99.7%) eliminating the use of complex algorithms like Adaboost here.

Figure 1 shows the results for five different classifiers using all features on all the datasets.

Table 18. The results for technique and the features (S1+S2+S3).

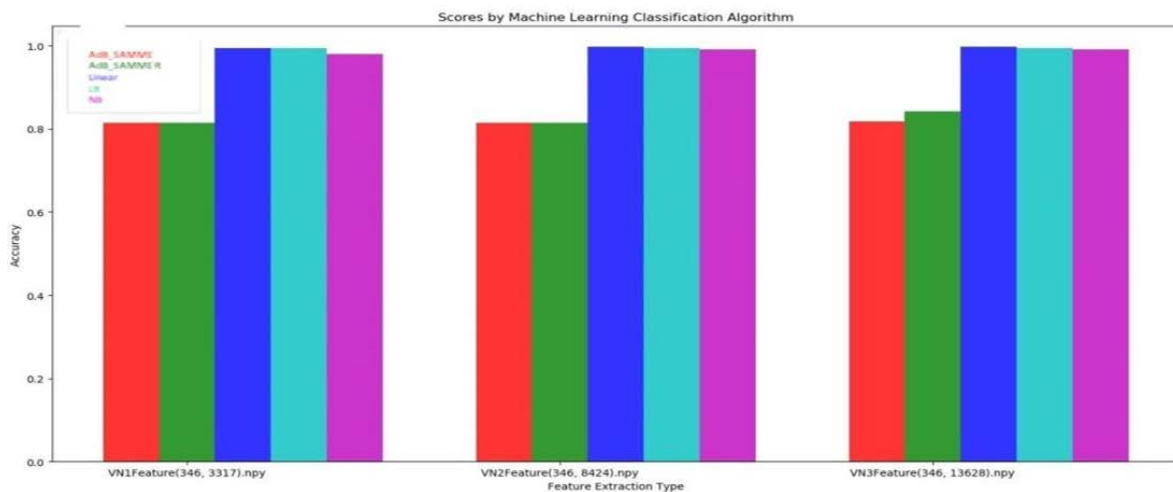| Sample	Technique	Accuracy	Accuracy
S1	AdB_SAMME	0.815028901734104	0.815
S1	AdB_SAMME.R	0.815028901734104	0.815
S1	Linear (SVM)	0.9942196531791907	0.994
S1	NB	0.9942196531791907	0.994
S1	LR	0.9797687861271677	0.980
S2	AdB_SAMME	0.815028901734104	0.815
S2	AdB_SAMME.R	0.815028901734104	0.815
S2	Linear (SVM)	0.9971098265895953	0.997
S2	NB	0.9942196531791907	0.994
S2	LR	0.9913294797687862	0.991
S3	AdB_SAMME	0.8179190751445087	0.820
S3	AdB_SAMME.R	0.8410404624277457	0.841
S3	Linear (SVM)	0.9971098265895953	0.997
S3	NB	0.9942196531791907	0.994
S3	LR	0.9913294797687862	0.991

Fig.1. Scores by machine learning classification algorithms.

An algorithm with less training time may be preferred even if it is less accurate, especially when time is the constraint. Tables 19 to 21 show the training times for each classifier. In the following tables, "tweet_no" means "total number of tweets in the dataset", "FE_no" means "number of features". For unigram, the "FE_no" signifies the number of unique words in the entire dataset. For bi-gram, the "FE_no" signifies the number of unique consecutive word pairs. For tri-gram, the "FE_no" signifies the number of unique consecutive word triplets. "ML_Name" signifies the classifier used for the classification task, "accuracy", "f1_score" and "time" are self-explanatory.

Ideally, accuracy is needed, and f1_score needs to be high, and time needs to be as low as possible. So if both classification performance and time are considered, it can be seen that the 4th experiment (uni-gram features with NB Classifier) ranks higher than Linear Classifier.

Table 19. The results for the performance time (tf-idf uni-gram)

no	tweet_no	FE_no	ML_name	Accuracy	f1_score	time
1	346	3317	AdB_SAM	81.50289	81.59703	0.207
2	346	3317	AdB_SAM	81.50289	81.59703	0.139
3	346	3317	Linear	99.42197	99.42141	0.808
4	346	3317	NB	99.42197	99.42248	0.05
5	346	3317	LR	97.97688	97.96943	0.004

Table 20. The results for the performance time (tf-idf bi-gram)

no	tweet_no	FE_no	ML_name	Accuracy	f1_score	time
1	346	8424	AdB_SAM	81.50289	81.59703	0.295
2	346	8424	AdB_SAM	81.50289	81.59703	0.561
3	346	8424	Linear	99.71098	99.71085	2.113
4	346	8424	NB	99.42197	99.42248	0.13
5	346	8424	LR	99.13295	99.13167	0.015

Table 21. The results for the performance time (tf-idf tri-gram)

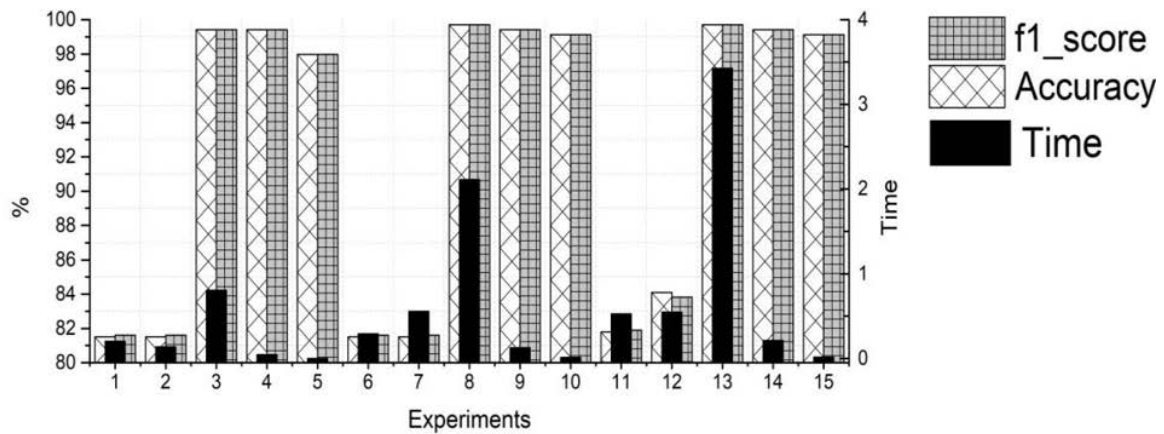| no	tweet_no	FE_no	ML_name	Accuracy	f1_score	time
1	346	13628	AdB_SAM	81.79191	81.88781	0.531
2	346	13628	AdB_SAM	84.10405	83.82738	0.549
3	346	13628	Linear	99.71098	99.71085	3.428
4	346	13628	NB	99.42197	99.42248	0.214
5	346	13628	LR	99.13295	99.13167	0.018

Fig.2.The time and accuracy of classification performance.

The following table shows the similarities and differences in accuracy compared to previous studies, taking into consideration the different languages (English, Arabic):

Table 22. Similarities and differences in accuracy

	Omer (2015)	FE	My approach	FE	Delta Δ
AdB_SAMME.R	100	(Text, Time)	84.1	(text)	-0.159
NB	89.0	(Text, Time)	99.4	(text)	0.117
SVM	97.9	(Text, Time)	99.7	(text)	0.018
LR	-		99.1	(text)	-
AdB_SAMME	-		81.8	(text)	-

Table 23 shows the comparison of results in Twitter according to the accuracy of sentiment analysis.

Table 23. Accuracy comparison between approaches

Research	Domain	ML	Accuracy
Deng(2014)	Opinion	SVM	75.90%
Rui et al.(2014)	Social comment	SVM	76.78%
Li and Li (2017)	Social news	SVM	83.80%
Farra et al. (2017)	Reviews	SVM	87.43%
Ali, F., Khan (2018)	Terrorism English	SVM	87.90%
My Proposal	Terrorism	SVM	89.24%

The accuracy in Table 23 draws a comparison between the Arabic datasets used for terrorism detection in this research to the others used in sentiment analysis. In addition, the Ali dataset that was used in terrorism in the English

language achieved 87.90% in terrorism detection using the same feature extraction and classification approaches. This difference in accuracy can be attributed to more positive terrorism tweets in the dataset collected for this research than the English tweets and also the location chosen for the current study.

6. Conclusions

Two research objectives are restated as follows: a) To identify different classification algorithms that are available in the contemporary context for tweets' classification; b) To evaluate how the tweets' classification algorithms can be optimized and mapped to detecting Arabic terrorism messages on Twitter. The study identified five classification algorithms that are highly relevant and useful in the contemporary context. These included AdB_SAMME, AdB_SAMME.R, Linear SVM, NB, and LR. These classifiers were tested using the language models of unigram (S1), bigram (S2), and trigram (S3). The study found that Linear SVM performed exceptionally well on S3 with 99.7% classification accuracy. When both classification accuracy and training time were considered, the NB classifier demonstrated the performance on S1 with 99.4% accuracy, which was comparable with Linear SVM. A viable way to detect radical texts in Arabic was proposed. The proposed dataset with an improvement in the initial database extracted by API from Twitter was evaluated for reliability by using statistical tools. The results yielded from the experiments proved that linear classifiers achieved the highest accuracy.

These are remarkable results in the field of data mining and sentiment analysis. The higher level of accuracy demonstrated by the two classifiers was also validated by the Arabic experts by a questionnaire. The promising results provide a call to action to implement the proposed classifiers, i.e. Linear SVM and NB for detecting Arabic terrorism messages. Terrorism is a global threat and no country and region can handle multiple and diverse avenues of terrorism on its own. There is a need for combined and collective efforts, and the proposed techniques of tweets' classification can help in identifying terrorist activities and threat alerts. Different countries in the Middle Eastern sector are making extensive investments in arms and ammunition to protect their borders from terrorist attacks. The state authorities and private sector entities should also invest in advanced technologies and combat the emerging threats in the cyber world and social networking sites.

6.1. Future Studies

Excluding the unwanted features and including the most relevant features can simplify the training model. An embedded approach for extraction of new few features related to the terrorism field in the Arabic language is suggested. The promising results obtained in this study emphasize the scope to expand the area of collection to include different tweets around the world.

References

[1] Ali, G. A. (2016). Identifying Terrorist Affiliations through Social Network Analysis Using Data Mining Techniques.

[2] Alsaedi, N. (2017). Event Identification in Social Media using Classification-Clustering Framework (Doctoral dissertation, Cardiff University). Retrieved from https://orca.cf.ac.uk/100998/1/2017alsaedinphd.pdf.

[3] Alshari, E. M., Azman, A., Doraisamy, S., Mustapha, N., & Alkeshr, M. (2017, August). Improvement of Sentiment Analysis Based on Clustering of Word2Vec Features. In 28th International Workshop on Database and Expert Systems Applications (DEXA), held on 29 August 2017 (pp. 123-126). IEEE. Retrieved from https://www.uni-weimar.de/medien/webis/events/tir-17/tir17-talks/Eissa2017_improvement-of-sentiment-analysis-based-on-clustering-of-word2vec-featuers_presentation.pdf.

[4] Ashcroft, M., Fisher , A., Kaati, L., & Omer, E. (2015, september). Detecting Jihadists messages on Twitter. Retrieved August 15, 2020, from researchgate.net: http://www.researchgate.net.

[5] Beninati, J. A. (2016). Examining the cyber operations of ISIS (Doctoral dissertation, Utica College).

[6] Choi, D., Ko (Fraiwan, 2020), B., Kim, H., & Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. Journal of Network and Computer Applications, 38, 16-21.

[7] Fraiwan, M. (2020, April 10). identification of markers in twitter data foe data extraction. Retrieved August 15, 2020, from sciencedirect.com: https://www.sciencedirect.com.

[8] Gates, S., & Podder, S. (2015). Social media, recruitment, allegiance and the Islamic State. Perspectives on Terrorism, 9(4), 107-116.

[9] Magdy, W., Darwish, K., & Weber, I. (2015). # FailedRevolutions: Using Twitter to study the antecedents of ISIS support. arXiv preprint arXiv:1503.02401.

[10] Omer, E. (2015). Using machine learning to identify jihadist messages on Twitter. Retrieved from https://www.diva-portal.org/smash/get/diva2:846343/FULLTEXT01.pdf.

[11] Oh, O., Agrawal, M., & Rao, H. R. (2011). Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. Information Systems Frontiers, 13(1), 33-43.

[12] Salmi, Abdul Latif (2016). The violence of language in the speech of extremist organizations "ISIS" a model of research in semantic and rhetorical mechanisms. Journal of Semiconductors (pp. 76-95).

[13] Witmer, E. W. (2016). Terror on Twitter: A Comparative Analysis of Gender and the Involvement in Pro-Jihadist Communities on Twitter. Retrieved from https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1008&context=sociology_masrp.

[14] Donia Gamal, Marco Alfonse, El-Sayed M. El-Horbaty, Abdel-Badeeh M.Salem, "Twitter Benchmark Dataset for Arabic Sentiment Analysis", International Journal of Modern Education and Computer Science, Vol.11, No.1, pp. 33-38,

2019.
[15] Sudhir Kumar Sharma, Ximi Hoque,"Sentiment Predictions using Support Vector Machines for Odd-Even Formula in Delhi", International Journal of Intelligent Systems and Applications, Vol.9, No.7, pp.61-69, 2017.
[16] Shah Zaib, Muhammad Asif, Maha Arooj, "Development of Aggression Detection Technique in Social Media", International Journal of Information Technology and Computer Science, Vol.11, No.5, pp.40-46, 2019.
[17] Arnisha Akhter, Uzzal K. Acharjee, Md Masbaul A. Polash," Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic", International Journal of Mathematical Sciences and Computing, Vol.5, No.4, pp.1-12, 2019.
[18] S. Shead, "Facebook, Tiktok won't Lift Ban on Posts that Promote Taliban after the Fall of Afghanistan," 2021. Retrieved from https://www.cnbc.com/2021/08/17/taliban-content-banned-on-facebook-instagram-whatsapp.html
[19] P. Hall, "Why Twitter Allows the Taliban to Maintain Accounts on its platform," 2021. Retrieved from https://finance.yahoo.com/news/why-twitter-allows-taliban-maintain-142546810.html

Authors' Profiles

Norah Al-Harbi has received her Master of Computer Science from Faculity of Computer Science and information Technology at University of Malaya(MU), Kuala Lumpur, Malysia. She received her B.Sc. Computer Science and IT from Computer Science and information Technology Faculty at Taif University, Taif, Saudi Arabia. Her research interests include Human Computer Interaction (HCI), Machine Learning, AI, Data Mining.

Amirrudin Bin Kamsin has received his Phd in Human Computer Interaction, University College London, Uk. He received his master of Animasi Komputer, Ncca, Bournemouth University, Uk. He received his B.Sc. Information Teqnology, Universiti Malaya (UM). His research interests include Human Computer Interaction (HCI),Virtual Reality, Multimedia.