2016 International Conference on Computational Science and Computational Intelligence

Sentiment analysis of ISIS related Tweets using Absolute location

Tarun B. Mirani and Sreela Sasi
Department of Computer and Information Science
Gannon University
Erie, PA-16541, USA
mirani001@knights.gannon.com; sasi001@gannon.edu

Abstract— **Twitter is a free broadcast service for the registered members to the public limited to 140 characters that may include text, photos, videos and hyperlinks. People share news, opinions and information to support or against media. The most petrified topic is the ISIS terrorist attacks taking place around the world. ISIS takes advantage of the social media to continuously communicate using coded words or to establish their indirect presence. Hashtags associated with ISIS can be analyzed and capture the sentiment of the tweets. This paper presents a novel process for sentiment analysis on the ISIS related tweets and to organize the opinions with their geolocations. The Jeffrey Breen algorithm is used for sentiment analysis. The data mining algorithms such as Support Vector Machine, Random Forest, Bagging, Decision Trees and Maximum Entropy are applied for polarity based classification of ISIS related Tweets. The results are compared and presented.**

Keywords—Sentiment Analysis; Random Forest; Linear Support Vector Machine; Logistic Regression; Geolocations;

I. INTRODUCTION

People share most of their emotions, thoughts and the current world information in 140 characters on Twitter. Twitter stated that it has 313 million active users monthly, which in turn causes one billion unique visits monthly in more than 40 languages [1]. Twitter is well known for sharing real time information about sports, political views, terrorist activities, etc. Sentiment analysis of social media such as Twitter helps to understand the public opinion about any situation. The process of determining the emotional tone behind these words helps in understanding the people's attitude and their view. According to Kevin Thau [2], Twitter provides an environment for delivering the opinions, news, crime, facts on ISIS attacks, etc. ISIS has shaken the world with their terror and these groups have been continuously sharing contents through this media. ISIS contents are being removed from Twitter every now and then, and it actively shuts down the extremist's handles [3].

In this research, the tweets related to ISIS are identified from Twitter with their exact location, and sentiment analysis is performed on them. Initially, the Jeffrey Breen algorithm is used to identify the polarity of the tweets [4]. Then the data mining algorithms such as Support Vector Machine, Maximum Entropy, Random Forest, Decision Tree, and Bagging are applied for the classification of the sentiments of the tweets. The results are compared and presented.

The remaining portion of this research paper is organized as follows: Section II provides the background research for the analysis of the Twitter data; Section III provides the architecture for the process of doing sentiment analysis. Section IV discusses the methods for classification of the sentiments and the simulation results. Section V presents the conclusion and future work.

II. RELATED WORK

A. Identifying ISIS Tweets

Govind A Ali analyzed the ISIS related tweets both in English and in Arabic in his thesis [5]. He used the k-means Nearest Neighbor algorithm to find out the frequently appearing words in the tweets. He has identified and verified three user accounts that contained ISIS supporting tweets by using network graph. Enghin Omer used hashtags to identify ISIS related tweets in his thesis [6]. Three different datasets were collected: one that supports ISIS, one that is anti-ISIS, and another one on random tweets that are not related to ISIS. He used Naïve Bayes, Ada Boost and support vector machine (SVM) for the classification. Matthew Rowe et. al. [7] studied the response of people after ISIS attacks. They examined the behavior of people during pre- and post-ISIS attacks from tweets and validated by using lexicon based approach.

B. Sentiment Analysis using machine learning

Agarwal et. al. [8] presented a practical approach for analyzing sentiments of tweets by using polarity where text is classified into positive, negative and neutral. Three models were used to classify the tweets: unigram, tree kernel based and feature based approach. They used Support Vector Machine (SVM), and achieved 71.35% average accuracy on unigram based approach. Alec go et. al. explored a novel approach to classify sentiment of tweets using machine learning algorithms [9]. A polarity based approach was used to find the sentiment of tweets using an emoticon dictionary. They achieved an accuracy of more than 80% by using Naïve Bayes, Support Vector machines and Maximum Entropy for unigram feature extraction.

In this research, the sentiment analysis is done on more than 5000 tweets collected during three days using multiple validated hashtags related to ISIS data. Data mining algorithms are used for the classification of the sentiments of the tweets and computed the evaluation metrics. The exact location of the user is also identified so that the geolocation of the majority of

978-1-5090-5510-4/16 $31.00 © 2016 IEEE
DOI 10.1109/CSCI.2016.215

ISIS related tweets could be understood for the law enforcement purposes. Additionally, more than 70 misspelled words are added to the existing Opinion Lexicon dictionary by Liu and Hu [10] on purpose to achieve better results. This is done because twitter consists mostly of slang/misspelled words, and if it is removed then it may affect in finding the actual sentiment of the tweets.

III. METHODOLOGY

A novel process, 'Sentiment analysis of ISIS related Tweets using Absolute location (SITA)' for sentiment analysis is presented in this research. It recognizes tweets related to ISIS by using ISIS related hashtags. Also, it will classify and predict the sentiment of the tweets. Every year, there has been an escalation of the number of ISIS attacks; it happens once in every 84 hours in Europe [11] and the coordination of the attack is done with the help of social media also [12].

The architecture for SITA process is shown in Fig. 1. The data is fetched using the Twitter Api, and then it is stored into multiple data tables based on their hashtags. Initial cleaning of the tweets is done using regular expressions, tokenization and the creation of the dictionary. The Opinion Lexicon dictionary is used for comparing the words. Jeffrey Breen algorithm is used for doing sentiment analysis and for calculating the polarities with scores. There are still chances of unorganized data even after the initial cleaning of these tweets. Hence, a secondary cleaning is done and is used for finding the frequently occurring words. This is done by placing these sentences into a matrix and then converting to a corpus. The same matrix is used for applying linear Support Vector Machine (SVM), Maximum Entropy (ME), Random Forest (RF), Decision Trees (DT), and Bagging (BG). All these algorithms were applied for classification and predictions. The evaluation metrics were computed during post-processing to measure the performance. The geolocation is also extracted from user profiles of ISIS related tweets.

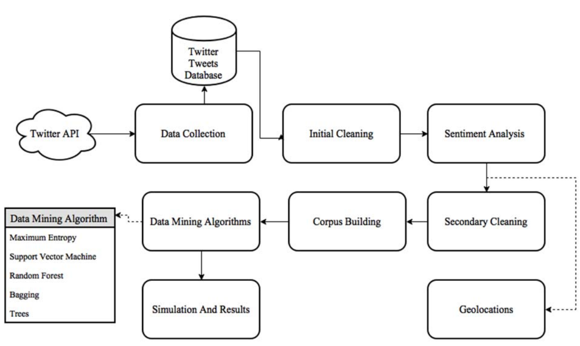

Fig 1. Architecture of SITA process for Sentiment analysis

A. Twitter API

The Twitter API is an interface provided by Twitter that allows to interact globally. Initially, streaming of ISIS related hashtags are done. The current privacy policy of Twitter allows users to retrieve only during the past 7 days of data from the site.

B. Data Collection

ISIS organizes hashtag campaigns for supporting ISIS or its cause. The hashtags such as #DAESH, #ISIS, #ISIL #IS #ISLAMICSTATE are used in ISIS related tweets and these are verified using hashtagify.com [13]. This website provided the related popular hashtags when searched for #ISIS with their popularity percentages and frequency of usage. Twitter data is retrieved using these hashtags. The tweets along with their information were placed into data frames that support tabular data, which can be accessed easily.

Information about the tweet include the following: ID (identification), Screen name (user chooses the Screen name), Created (time when the tweet was created), Text (content of the tweet), Status source (The HTML location of the tweet), Is Retweet (Boolean value indicating whether the tweet is Retweet or not), Retweet count (the number of times the tweet is Retweeted), Favorite count (the number of times the tweet is being liked), Reply to Screen name (the Screen name of the user who replied to the owner of the tweet), Reply to the user ID (The ID of the user that was replied by owner of the tweet.), and Reply to the Status ID (The ID of the tweet that was replied by owner of the tweet).

The following user information was also extracted based on the screen name of the tweet: ID, Screen Name, Created, Text, Lang (the language set by the user), Status Count (the number of tweets that the user has posted), Followers Count, Friends Count, Favorite Count, Location, URL (the twitter profile of the user), Profile image URL (the twitter profile image set by the user).

C. Initial Cleaning

One thousand tweets per hashtag were collected for this research. The Initial Cleaning is done before applying sentiment analysis on the tweets data. This involves cleaning the data thoroughly as the raw tweets is noisy. The Initial Cleaning consists of Regular Expression, tokenization and, creating dictionaries. Regular Expression is used for the removal of retweets, names of people, punctuations, numbers, and html links. The unnecessary spaces and 'not available' (NAs) data are also removed and the tweets are converted to lower case. An example for the 'initial cleaning' of the tweet using Regular Expression is shown in Table 1.

TABLE 1. Example for the 'initial cleaning' of the tweet

Initial Cleaning	**Actual Tweet Sample: RT @PressTV: #Daesh sympathizer killed by US police after stabbing 8 people https://t.co/T7ahSdbKBg https://t.co/SfDD4khE3E**
Remove retweet entities	"#Daesh sympathizer killed by US police after sta bbing 8 people https://t.co/T7ahSdbKBg https://t.c o/SfDD4khE3E"
Remove people	Remove the people same as removal of retweet

Remove Punctuation	"Daesh sympathizer killed by US police after stab bing 8 people httpstcoT7ahSdbKBghttpstcoSfDD4 khE3E"
Remove numbers	"Daesh sympathizer killed by US police after stab bing people httpstcoTahSdbKBg httpstcoSfDDkh EE"
Remove html links	"Daesh sympathizer killed by US police after stab bing people "
Remove unnecessary spaces	"Daesh sympathizer killed by US police after stab bing people"
Lower Case	"daesh sympathizer killed by us police after stabbi ng people"

Tokenization is generally used in lexical analysis. It is the process that splits up a text into words, phrases, symbols or other components. These components are called tokens for further processing [14]. Tokenization does the vital task that helps in classifying tweets based on unigram. For example, the actual clean tweet used as a sample from #Daesh is "the saudi royal family is supporting terrorism all over the world daesh". After Tokenization, it will be transformed to token: token1: "the", token2: "saudi", token3: "royal", token4: "family", token5: "is", token6: "supporting", token7: "terrorism", token8: "all", token9: "over", token10: "the", token11: "world", token12: "daesh". Creation of dictionaries is explained in the Sentiment Analysis section.

D. Sentiment Analysis

A lot of research has been done on sentiment analysis of tweets by classifying them into positive, negative and neutral; but very less research has focused on classifying tweets related to terrorism and ISIS activities. In this research, a score is assigned to the sentiment of each tweet by using each hashtag mentioned in the Data Collection section. Jeffrey Breen Algorithm is used to classify tweets based on polarity. The opinion lexicon dictionary that contains approximately 6,800 words as positive and negative is used. More than 3000 tweets are physically observed and identified related hashtags, such as terrorism, ISIS, alqaeda. More than 70 such words were added to the dictionary to give better results. This includes misspelled/ slang words because the tweets also contain misspelled /slang words.

Clean tweet sentences are used for sentiment analysis. They are split into words and are used as tokens. These tokens are compared to the dictionary of positive and negative words. The words that match with the dictionary will be used to calculate the score of the sentiment. This score is calculated by subtracting the sum of positive matches from the sum of negative matches for a sentence. The scores and the tweets are placed into a proper tabular format. In order to analyze the sentiment scores, the score distribution is built by plotting histograms. All the sentiment scores of each hashtag are combined together for comparison.

The sample tweets collected during the process and their corresponding scores of positivity and negativity are shown in Table 2. The tweets that are collected during the three days didn't have any score of '4' of positivity, and hence this is missing in Table 2.

TABLE 2. Sample Tweets with their Sentiment

Score	Text	Sentiment
0	isil is daesh islamicstatethe reality of isis and their ideology by hh younusalgoharmi mi cia nsa	neutral
1	i have no doubt that isil will be defeated in iraq the challenge is to restore trust between the communities and win.	positive
2	watch iraqi security forces celebrate the liberation of shirqat from daesh	very positive
3	kurdish victory dance kurdish women fighters ypj celebrate their victory over the islamicstate in alshadadi	very very positive
-1	daesh fighters assassinate a political security officer in mansura in adan yemen	negative
-2	islamicstate actively abuses child soldiers to carry out their wicked plansthis kid carried out a vbied suicide	very negative
-3	new horrific isis propaganda video they slaughter people in front of a small kid then make the kid shot a man isil	very very negative
-4	warning graphic pictures iraq islamicstate isis barbaric terror group throw gay man to his death in nineveh	seriously negative

E. Secondary Cleaning and Corpus Building

Secondary cleaning involves complex steps in cleaning the tweets. Sometimes the tweets are not completely cleaned using the regular expressions. Hence, they need to be re-cleaned in order to achieve optimal results. Secondary cleaning is also needed for finding the frequently appearing words using the corpus. It involves removal of stop words, use of stemming and the rest of the removals are similar to the initial cleaning. Example for secondary cleaning is shown in Table 3.

TABLE 3. Sample Tweet for Secondary Cleaning

Secondary Cleaning	**Actual Tweet Sample: RT @PressTV: #Daesh sympathizer killed by US police after stabbing 8 people https://t.co/T7ahSdbKBg https://t.co/SfDD4khE3E**
Remove Html link	RT @PressTV: #Daesh sympathizer killed by US police after stabbing 8 people
Lower case	"rt @presstv: #daesh sympathizer killed by us poli ce after stabbing 8 people"
Remove numbers	rt @presstv: #daesh sympathizer killed by us police after stabbing people"
Remove stop words	"rt @presstv: #daesh sympathizer killed us police stabbing people"
Remove punctuations	"rt presstv daesh sympathizer killed us police stabbing people"
Remove strip white space	"rt presstv daesh sympathizer killed us police stabbing people"
Stemming	"rt presstv daesh sympath kill us polic stab peopl"

Corpus is a collection of documents. Stop words are the frequently used words in a sentence that plays a very small role in sentiment analysis [15]. Hence, the removal of stop words is necessary. "by" and "after" are two examples of stop words and are given in Table 3. Stemming is the process of removing the ending of the words to their root [15]. In this secondary cleaning, Porter stemming algorithm is used to improve the cleaning process. Example of stemming is also given in Table 3. The data is converted into a plain text document after the Secondary Cleaning. Then the word cloud is plotted to visualize the most frequently occurring words and it will help to understand the frequency of ISIS related words in the tweets. Frequently appearing words may provide the related links to ISIS, which may lead to get more and more information on this topic. Sample word cloud of #ISIL is shown in Fig. 2.

Fig 2. Word Cloud for #isil

F. Data Mining Algorithms

The Support Vector Machine, Maximum Entropy, Random Forest, Bagging and Decision Trees are applied for classification of the sentiment of the data. These algorithms are used for training and testing on the ISIS related tweets and to predict their sentiments.

The Support Vector Machines is used for non-linear classification, regression and density estimation in [16]. They used SVM for classification and for predicting sentiments of the tweets. The Maximum Entropy, which is a multinomial logistic regression model works better for text classification [17]. One of the salient methods under the Classification and Regression Tree (CART) is the Random Forest Classifier [18]. Breiman's random forest algorithm for CART is used in this research. Random Forest uses supervised learning for classification and will build trees similar to the decision trees. Then it builds multiple trees and does average of all trees for the final model. Bagging is also used for classification and regression. Bagging enhances the prediction results as it minimizes the variance related with it. In this research, Bagging is used for classification. Also, Decision Trees are used for the same purpose. The performance metrics are determined for all these algorithms in Post-Processing. Also, the statistics is done on the result obtained from the sentiment analysis.

G. Geolocation

The absolute location of the user is extracted with the help of the Twitter API using their user id and user name. Geocode is the 'google maps geocoding API' [19], which gives the latitude and longitude from the given user address. Revgeocode is also one of the processes of 'google maps geocoding API'. Revgeocode provides the complete physical address of the user including the street address, city, zip code, state and country. According to the google maps policy, the geolocation cannot be found if the user has not set his /her location on Twitter. In this research, the sentiments of tweets are combined with the geolocations and is displayed using Tableau [20], which provides the visualization of the results.

IV. SIMULATION AND RESULTS

In this research, the open source R Language [21] is used, as it is cross platform compatible, provides a large number of packages, better graphs and statistical techniques. Initially, Tweets based on the ISIS related hashtags are retrieved using Twitter API. Then, the initial cleaning of these tweets is done using Regular Expression, tokenizing and creating dictionaries. Regular Expression is done using an inbuilt function in R. Tokenization is done using the "stringr" package in R. Then the polarity based sentiments are determined. Secondary cleaning of the tweets helps in building corpus and finding the frequently occurring words. R provides the text mining package (tm), which provides functions that involves loading the text into corpus and for cleaning the tweets data using the function named "tm_map". The wordcloud is made by keeping the frequency of words as 25. The sentiment classification is done by applying linear SVM, Maximum Entropy, Random Forest, Bagging, and Decision Trees on the sentiment of the data. The performance metrics are computed and are compared. Also, the absolute location is extracted using the geocoding API.

A. Comparison of Hashtags

Initially, 1000 tweets per hashtags were extracted during September 29th 2016 to October 1st 2016. After applying sentiment analysis, the tweets were classified into positive, negative and neutral tweets with their scores. The overall percentage of the applied hashtag with their count is shown in Table 4. The neutral content with their scores are removed and used only positive and negative data.

TABLE 4. Count and Percentage of each Hashtag

Hashtags	Positive Count	Negative Count	Total Count	Negative Percent	Positive Percent
#ISLAMIC STATE	80	279	359	78%	22%
#ISIS	93	241	334	72%	28%
#ISIL	135	308	443	70%	30%
#DAESH	206	256	462	55%	45%
#IS	182	163	345	47%	53%

There is a huge difference between the positive and negative counts based on the Table 4. People have posted more about negative in every hashtag except #IS. The maximum negative Tweets as well as negative percentage is from #ISLAMICSTATE. Neutral sentiments are usually the tweets, whose content contains words other than the dictionary words. The dictionary did not recognize some words because there are plenty of misspelled words that are being used in the tweets. In this 21st century, people have reached a new level of texting or posting tweets that every day new words are invented for the sake of quick texting.

Comparisons of each hashtag with their positive, negative and neutral sentiment scores is done using ggplot2 in R and is shown in Fig. 3. The score for #Daesh is '-4', which means there is someone whose tweet contains extreme negativity. All the hashtags show the scores of negativity in their tweets. #ISIL had more negativity on twitter.

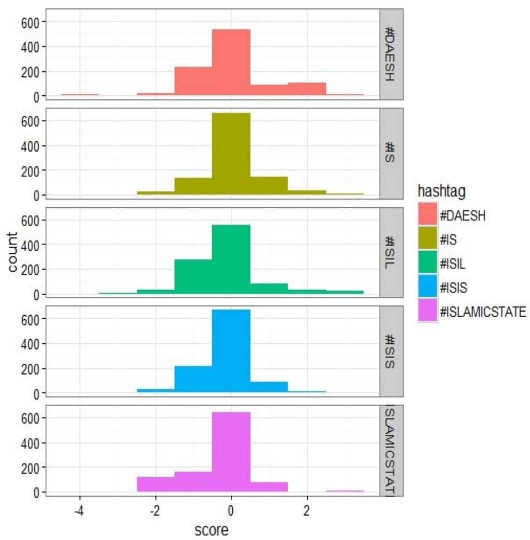

Fig 3. Comparison of hashtag with their Sentiment Scores

B. Geolocation in Tableau

Geolocations of the tweets were extracted and combined with sentiment of the tweets. The visualization is done using Tableau. The sample tweet with -1 sentiment is shown in Fig. 4. The tweet was tweeted from Texas, USA in the Frisco city with the address of 2428-2448 Internet Blvd and zip-code of 75034. In order to verify whether the address actually exist, it was checked on Google maps. Yes, it does exist!

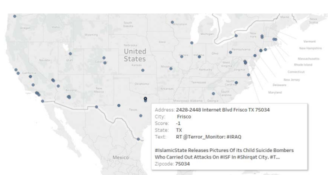

Fig 4. Actual location of the tweet

C. Performance Evaluation

The performance of the Support Vector Machine (SVM), Maximum Entropy (ME), Random Forest (RF), Decision Trees (DT) and Bagging (BG) is evaluated using accuracy, F-measure, recall, Precision. The metrics for evaluating the performance of classifiers are accuracy, F-measure, recall, precision and kappa and are shown in Table 5.

TABLE 5. Accuracy, Precision, Recall, F-Score and Kappa Results for Each Classifier for Each Hashtag

Metrics		#ISIL	#ISIS	#IS	#ISLAMIC STATE	#DAESH
Accuracy	SVM	**0.96**	0.83	0.63	0.94	0.94
	ME	**0.96**	0.86	0.67	0.92	0.94
	RF	**0.96**	0.83	0.56	0.92	0.95
	DT	**0.96**	0.80	0.56	0.91	0.91
	BG	**0.98**	0.82	0.59	0.92	0.97
Precision	SVM	**0.98**	0.70	0.70	0.93	0.95
	ME	**0.98**	0.85	0.72	0.96	0.95
	RF	**0.98**	0.91	0.76	0.96	0.96
	DT	**0.98**	0.90	0.76	0.95	0.92
	BG	**0.99**	0.80	0.77	0.96	0.97
Recall	SVM	0.93	0.72	0.64	0.85	**0.95**
	ME	0.93	0.72	0.67	0.76	**0.95**
	RF	0.93	0.63	0.57	0.76	**0.96**
	DT	**0.93**	0.56	0.57	0.73	0.92
	BG	0.96	0.63	0.60	0.76	**0.97**
F-Score	SVM	**0.95**	0.74	0.60	0.88	0.94
	ME	**0.95**	0.76	0.65	0.82	0.94
	RF	0.93	0.65	0.47	0.82	**0.96**
	DT	**0.95**	0.56	0.47	0.79	0.91
	BG	**0.97**	0.66	0.53	0.82	**0.97**
Kappa	SVM	**0.90**	0.48	0.27	0.76	0.89
	ME	**0.90**	0.53	0.35	0.65	0.89
	RF	0.90	0.35	0.14	0.65	**0.91**
	DT	**0.90**	0.18	0.14	0.59	0.82
	BG	**0.95**	0.35	0.20	0.64	0.94

D. Ensemble Method

The Ensemble Method is used to boost the predictive performance of the single model by combing multiple models. In this research, it is used to compute ensemble coverage and ensemble recall accuracy. Ensemble coverage is the number of algorithms that agree on results among the five datamining algorithms with each other. This means that the ensemble coverage is the percentage of algorithms that match the recall accuracy threshold divided by the total number of algorithms [22]. The ensemble coverage (C) and ensemble recall accuracy (R) are shown in Table 6 and Table 7.

TABLE 6. Ensemble Coverage and Ensemble Recall Accuracy for #ISLAMICSTATE and #ISIS

	Hashtag			
	#ISLAMIC STATE		#ISIS	
	C	R	C	R
n>=1	1	0.93	1	0.84
n>=2	1	0.93	1	0.84
n>=3	1	0.93	1	0.84
n>=4	1	0.93	0.95	0.88
n>=5	0.95	**0.95**	0.81	**0.97**

TABLE 7. Ensemble Coverage and Ensemble Recall Accuracy for #ISIL, #DAESH and #IS

	Hashtag					
	#ISIL		#DAESH		#IS	
	C	R	C	R	C	R
n>=1	1	0.97	1	0.96	1	0.62
n>=2	1	0.97	1	0.96	1	0.62
n>=3	1	0.97	1	0.96	1	0.62
n>=4	0.99	0.98	0.96	0.98	0.87	0.62
n>=5	0.98	**0.98**	0.90	**0.98**	0.76	**0.63**

E. N-fold Cross Validation

10-Folds Cross Validation (CV) is used to estimate the accuracy of each algorithm based on hashtags. It requires a lot of computation and time; therefore, the CV was calculated for each Hashtag for each algorithm. They were computed using 70% as training set, and the remaining 30% as test set. The average result of these 10 computations for each hashtag was taken as the best performance of each algorithm.

The average accuracy of each algorithm from the 10-Folds CV of #ISIL is shown as example in Table 8. Similar results are obtained for other hashtags also. Using R, the N-fold cross validation function is modified appropriately for this data.

TABLE 8. The Average of 10-Folds CV of #ISIL

#ISIL	**MACHINE LEARNING ALGORITHMS**				
10 Folds	**SVM**	**MAXENT**	**RF**	**Tree**	**Bagging**
1	0.97	0.97	1	1	1
2	0.94	1	0.98	1	0.97
3	0.97	1	0.97	0.98	1
4	0.97	1	0.97	0.95	0.95
5	0.98	0.97	0.97	0.93	0.96
6	0.96	1	0.94	0.92	0.97
7	0.97	0.98	0.98	0.92	1
8	1	1	0.97	0.90	0.97
9	1	1	0.95	0.90	0.96
10	0.94	0.97	0.97	0.98	0.96
Avg. Acc.	0.97	**0.99**	0.97	0.95	0.97

V. CONCLUSION AND FUTURE WORK

In this research, classification of the ISIS related data from Twitter and the prediction of its sentiment are done. A two-step cleaning process, sentiment analysis, frequently occurring words, datamining algorithms, and geolocation are combined. The data was collected for a period of approximately 3 days. There are more negative tweets found than the positive ones. It means that there are more successful attacks done by the ISIS than the number of defeats they received. The most frequently used words among the ISIS related tweets were isil, daesh, Younus AlGohar, Libya, isi and Iraqi. All algorithms performed with an average accuracy of more than 90% after 10-fold cross validation. The Maximum Entropy gave the best result of 99% of average accuracy after validation using #isil.

As a future work, collecting more data from different languages and applying sentiment analysis on n-grams with more accurate geolocations API could be done. Combining more dictionaries could help in providing better results for sentiment analysis. Moreover, analyzing more tweets could lead to the ISIS user accounts and can provide other confidential and illegal data.

REFERENCES

[1] Twitter, "About Twitter Company," Twitter, 30 September 2015. [Online]. Available: https://about.twitter.com/company. [Accessed 01 January 2015].

[2] S. Perez, "Twitter is NOT a Social Network, Says Twitter Exec", ReadWrite, 2010. [Online]. Available: http://readwrite.com/2010/09/14/twitter_is_not_a_social_network_says_twitter_exec/. [Accessed: 09-Oct-2016].

[3] Fox News, "Twitter Shuts Down 235,000 More Extremist Accounts", Fox News, 18 August 2016. [Online]. Available: http://www.foxnews.com/tech/2016/08/18/twitter-shuts-down-235000-more-extremist-accounts.html. [Accessed: 09- Oct- 2016].

[4] J. Breen, "slides from my R tutorial on Twitter text mining #rstats", Things I tend to forget, 2011. [Online]. Available: https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/. [Accessed: 09- Oct- 2016].

[5] Ali, Govand A., "Identifying Terrorist Affiliations through Social Network Analysis Using Data Mining Techniques,". M.S Theses, Dept. Information Technology, Valparaiso Univ., Indiana, USA, 2016

[6] Enghin Omer, "Using Machine Learning to Identify Jihadist Messages on Twitter,". M.S Theses, Dept. Information Technology, Uppsala Univ., Sweden, 2015.

[7] M. Rowe and H. Saif, "Mining pro-ISIS radicalisation signals from social media users," in Proceedings of the 10th International Conference on Web and Social Media, 2016.

[8] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R, "Sentiment Analysis of Twitter Data," in Proc of ACL HLT Conf, 2011.

[9] Go, A., Bhayani, R., and Huang, L, "Twitter Sentiment Classification using Distant Supervision," Technical Report, Stanford Digital Library Technologies Project, 2009.

[10] Liu, B., Hu, M., and Cheng, J., "Opinion Obeserver: Analyzing and Comparing Opinions on the Web," Proceeding of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[11] T. Lister, "'One ISIS attack every 84 hours'", *CNN*, 2016. [Online]. Available: http://www.cnn.com/2016/07/31/europe/isis-attacks-escalating-europe/. [Accessed: 09- Oct- 2016].

[12] J. Berger, "The Evolution of Terrorist Propaganda: The Paris Attack and Social Media | Brookings Institution", *Brookings*, 2015. [Online]. Available: https://www.brookings.edu/testimonies/the-evolution-of-terrorist-propaganda-the-paris-attack-and-social-media/. [Accessed: 09- Oct- 2016].

[13] Hashtagify, "Hashtagify Find, Analyze, Amplify," Hashtagify, [Online]. Available: https://hashtagify.me. [Accessed: 09 Oct 2016].

[14] Jackson, P. and Moulinier, I. Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. John Benjamins Publishing Co., 2002.

[15] C. Manning, P. Raghavan and H. Schütze, *Introduction to information retrieval*. New York: Cambridge University Press, 2008,

[16] J. Moguerza and A. Muñoz, "Support Vector Machines with Applications", *Statistical Science*, vol. 21, no. 3, pp. 322-336, 2006.

[17] Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61–67, 1999.

[18] D. Meyer, E. Dimitriadou, K. Hornik, A. weingessel and F. Leisch. e1071 Package. (2015) [Online]. Available: https://CRAN.R-project.org/package=e1071.

[19] Google, "Getting Started | Google Maps Geocoding API | Google Developers", Google Developers, 2016. [Online]. Available: https://developers.google.com/maps/documentation/geocoding/start. [Accessed: 10- Oct- 2016].

[20] Tableau, "Tableau Software", Tableau Software. [Online]. Available: http://tableau.com. [Accessed: 10- Oct- 2016].

[21] R Core Team. R: A Language and Environment for Statistical Computing. (2016) [Online]. Available: http://www.R-project.org/

[22] T. P. Jurka, L. Collingwood, A. E. Boydstun, E. Grossman and W. Atteveldt. "RTextTools: A Supervised Learning Package for Text Classification," 2013, Vol. 5(1), pp. 6–12.