COMP 8043 Machine Learning

Research Paper

Global Terrorism Database

Darren Smith

BSc.(Hons.) in Software Development

Department of Computer Science

Student ID: R00117899

Lecturer: Dr. Ted Scully

Date: 6th December 2016

### Declaration of Authorship

I, Darren Smith, declare that this thesis titled, “Research Paper - Global Terrorism Database” and the work presented in it are my own. I confirm that:

* This work was done wholly or mainly while in candidature for an undergraduate degree at Cork Institute of Technology.
* Where any part of this paper has previously been submitted for a degree or any other qualification at Cork Institute of Technology or any other institution, this has been clearly stated.
* Where I have consulted the published work of others, this is always clearly attributed.
* Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
* I have acknowledged all main sources of help.
* Where the paper is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: \_\_\_\_\_\_\_\_\_Darren Smith \_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_6/12/2016 \_\_\_\_\_\_\_\_\_\_

### CORK INSTITUTE OF TECHNOLOGY

### *Abstract*

### Faculty of Engineering and Science

### Department of Computer Science

### Bachelor of Science

### By Darren Smith

The aim of this research paper is to investigate, model and implement a technical solution that will improve the understanding of terrorist attacks. The dataset used in this paper from the Global Terrorism Database. This research in this paper will consider different ways to achieve as high an accuracy as possible from the dataset. The areas covered in this paper are decision tress, k-nearest neighbours, SVMs, random forest and naive bayes. It will then will focus on SVMs and use different approaches of improving them in relation to the dataset. Some of the ways used are one-hot encoding, label encoder, hyper-parameter optimization and feature selection.

Table of Contents

**1 Introduction4**

1.1 Motivation4

1.2 Problem4

1.3 Approach4

**2 Related Research5**

2.1 Current State of the Art5

**3 Algorithm/Model Detail6**

3.1 Overview6

3.2 Algorithms6

**4 Empirical Evaluation7**

4.1 Initial Results7

4.1 Subsequent Results8

**5 Conclusion and Future Work14**

5.1 Conclusion14

5.1 Future Work14

**Bibliography 15**

### Chapter 1

### Introduction

### Motivation

With the increase of high-profile attacks on major cities such as Belgium, France and the United States, the world is on edge. It has become ever more important to understand why these attacks happen and who is responsible. The word “Terrorism” has been used in many different context and depends on the person’s point of view. To better understand the meaning of the word we will look at the definition from the International Terrorism and Security Research organisation:

“*Terrorism has been described variously as both a tactic and strategy; a crime and a holy duty; a justified reaction to oppression and an inexcusable abomination. Obviously, a lot depends on whose point of view is being represented. Terrorism has often been an effective tactic for the weaker side in a conflict. As an asymmetric form of conflict, it confers coercive power with many of the advantages of military force at a fraction of the cost. Due to the secretive nature and small size of terrorist organizations, they often offer opponents no clear organization to defend against or to deter. “* [1].

### 1.2 Problem

The aim of this paper is to solve a multi classification problem, using different types of supervised learning algorithms in a large dataset. This problem will be addressed by trying to arcuately predict what organisation is responsible after an attack has happened.

### 1.3 Approach

The approach taken in this research is to initially perform basic pre-processing on the dataset. This is done by removing features, imputing the mean values for missing values, encoding categorical variables as continuous variables and one-hot encoding. After which a range of classifiers using cross fold validation will be run on the dataset. From these initial results a classifier will be chosen. Different implementations of this classifier and approaches of how to improve it will then be reviewed and tested.

### Chapter 2

### Related Research

### 2.1 Current State of Art

Machine learning is the science of getting computers to act without being explicitly programmed. Many researchers think it is the best way to make progress towards human-level AI. Machine learning tasks are typically classified into three broad categories - supervised learning, unsupervised learning, and reinforcement learning. The focus of this paper will be on SVMS which is under the category of supervised learning. After researching various journals, articles and papers in relation to SVMs, the thesis “Design and Training of Support Vector Machines” by Shilton, A. [2], gives a basic understanding on SVMs and how to design and train them. This thesis reviews a novel form of SVM known as regression with inequalities, in addition to the standard SVM formulations of binary classification and regression. The conclusion from this thesis was there are a few advantages and disadvantages SVMs.

### Advantages:

1. Produce very accurate classifiers.
2. Less overfitting, robust to noise.

### Disadvantages:

1. SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes).
2. Computationally expensive, thus runs slow.

To try counter act these disadvantages other sources were reviewed. One such source is “Learning in Extreme Conditions: Online and Active Learning with Massive, Imbalanced and Noisy Data” by Ertekin, S. [3]. This thesis addresses improving the performance of machine learning algorithms with a particular focus on classification tasks with large, imbalanced and noisy datasets. It looks into great detail about class imbalance and noise in the data which can degrade the prediction accuracy of standard machine learning algorithms.

Another thesis is “Improved Learning of Structural Support Vector Machines: Training with Latent Variables and Nonlinear Kernels” by Nam Yu, C. [4]. This thesis explores improving the learning of structured prediction rules with structural SVMs in two main areas: incorporating latent variables to extend their scope of application and speeding up the training of structural SVMs with nonlinear kernels.

### Chapter 3

### Algorithm/Model Detail

### 3.1 Overview

The data set used is the Global Terrorism Database [5]. The GTD includes information on terrorist events around the world from 1970 through 2015. It also includes systematic data on domestic as well as transnational and international terrorist incidents that have occurred during this time period and with over 150,000 cases. For each GTD incident, information is available on the date and location of the incident, the weapons used and nature of the target, the number of casualties, and when identifiable the group or individual responsible.

### Characteristics of the GTD

* Contains information on over 150,000 terrorist attacks.
* Includes information on more than 75,000 bombings, 17,000 assassinations, and 9,000 kidnappings since 1970.
* Includes information on at least 45 variables for each case, with more recent incidents including information on more than 120 variables.
* Over 4,000,000 news articles and 25,000 news sources were reviewed to collect incident data from 1998 to 2015 alone.

### 3.2 Algorithms

The initial algorithms used on the dataset were:

* Decision tree
* K-nearest neighbours
* SVM
* Random Forest
* Naïve Bayes

### Chapter 4

### Empirical Evaluation

### 4.1 Initial Results

The first step taken in this research was to run a preliminary test on the dataset. This was achieved by first reducing the features in the dataset from 120 to 13. This was done because most of the features were irrelevant in predicting the organisation responsible for an attack. Following this the mean value was imputed for all missing values. The next step was to identify the top organisations. After this a spread of different classifiers were run on the top 2 and top 5 organisations.

### *Top Organisations*

1. Unknown 71922 (ignored)
2. Taliban 5502
3. Shining Path (SL) 4548
4. Farabundo Marti National Liberation Front (FMLN) 3351
5. Islamic State of Iraq and the Levant (ISIL) 2833
6. Irish Republican Army (IRA) 2670

### *Spread of Classifiers Top 2 (runtime 20mins)*

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Before Standardization | After Standardization | Difference |
| Tree | 99.98% | 97.83% | -2.15 |
| SVM | 99.36% | 92.38% | -6.98 |
| NNeighbour | 98.65% | 93.47% | -5.18 |
| RForest | 99.99% | 99.73% | -0.26 |
| Naive Bayes | 99.98% | 99.80% | -0.18 |

### *Spread of Classifiers Top 5 (runtime 2hrs 31mins)*

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Before Standardization | After Standardization | Difference |
| Tree | 99.93% | 99.90% | -0.03 |
| SVM | 99.35% | 90.28% | -9.07 |
| NNeighbour | 96.02% | 93.57% | -2.45 |
| RForest | 99.94% | 99.96% | +0.02 |
| Naive Bayes | 99.92% | 99.41% | -0.51 |

### *Key Information Top 5*

Rows 18904

Features (before one hot encoding) 13

Features (after one hot encoding) 4271

Values (after one hot encoding) 80738984

Runtime 2hrs 31mins

### *Initial Conclusion*

From these initial results, the accuracy is very high. This may be due to the low number of organisations used in the initial test and that most organisations are usually active in a small geographic area. It should also be noted that the runtime for the top 5 organisations is 2hrs 31mins. This is due to after one hot encoding the number of features went from 13 to 4271. This can be translated to 80738984 values in the dataset. I have decided that research will now focus on SVMs. The reason for this it will give me a more accurate result as I feel that the other classifiers are to overfitted.

### 4.2 Subsequent Results

#### Step 1: Improve Runtime

All figures from here on will be Standardised. The next step in this research was to try an improve the runtime. To do this one hot encoding was removed and label encoder was used instead.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With One Hot Encoding | Without One Hot Encoding | Difference |
| Tree | 99.90% | 99.90% | 0 |
| SVM | 90.28% | 99.53% | +9.25 |
| NNeighbour | 93.57% | 98.75% | +5.18 |
| RForest | 99.96% | 99.43% | -0.53 |
| Naive Bayes | 99.41% | 99.86% | +0.45 |
| Continue | NO | YES |  |

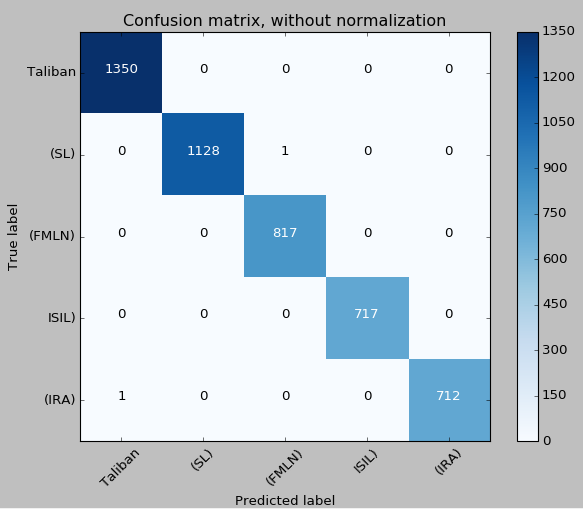
The result from removing one hot encoding improved runtime from 2hrs and 31mins to 90.57 seconds. It also improved all the classifiers except for RForest. The research will now focus on SVMs.

#### Step 2: Hyper-Parameter Optimization

The next step was to run hyper-parameter optimization.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Without Hyper-Parameter Optimization | With Hyper-Parameter Optimization | Difference |
| SVM | 99.53% | 99.93% | +0.40 |
| Continue | NO | YES |  |

### *Confusion matrix Top 5*



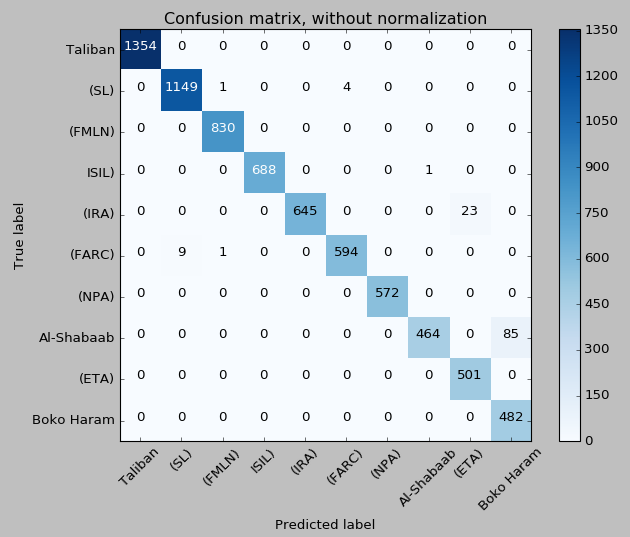
The result from hyper-parameter optimization was kernel = liner and C = 1. This improved the accuracy by 0.40%. From the confusion matrix it shows only 2 instances were wrongly classified.

#### Step 2: Increase number of classes

The next step was to increase the classes 100% from 5 to 10.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Top 5 | Top 10 | Difference |
| SVM | 99.93% | 98.43% | -1.50 |
| Continue | NO | YES |  |

### *Confusion matrix Top 10*



The result from increasing the number of classes from 5 to 10 increased the runtime from 90.57 seconds to 30mins. This was due to running hyper-parameter optimization on the top 10 organisations. The result was the same was kernel = liner and C = 1. As these results are the same, these values will now be used. There is also a decrease in accuracy of 1.50% due the increased number of classes. It should also be noted that 85 of Al-Shabaab class were predicted wrong and were classed as Boko-Haram. 23 of IRA class were predicted wrong and were classed as ETA.

#### Step 3: Feature selection

The next step was to see what features impacted the classification and which didn’t.

***Univariate Feature Selection***

F Score 74.3801674686 for feature year

F Score 2606.66997374 for feature country

F Score 11696.2253815 for feature region

F Score 127.598679478 for feature success

F Score 7.29801287347 for feature attacktype

F Score 18.4375502573 for feature targtype

F Score 22.6547944447 for feature targsubtype

F Score 22.9073753037 for feature weaptype

F Score 7.50251134712 for feature nkill

F Score 0.0405487327375 for feature nwound

F Score 3.1006443857 for feature ishostkid

***Tree-based Feature Selection***

F Score 0.109392808258 for feature year

F Score 0.411641406854 for feature country

F Score 0.411507500609 for feature region

F Score 0.00239769675985 for feature success

F Score 0.0109334456825 for feature attacktype

F Score 0.0162039430854 for feature targtype

F Score 0.017771662302 for feature targsubtype

F Score 0.00918414387603 for feature weaptype

F Score 0.0047654009256 for feature nkill

F Score 0.00340947396293 for feature nwound

F Score 0.00279251768496 for feature ishostkid

The result from feature selection shows that the country feature has the biggest impact on the classification, whilst nwound has the least impact on the classification. Features will now start to be removed.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With nwound + ishostkid | Without nwound + ishostkid | Difference |
| SVM | 98.43% | 98.43% | 0 |
| Continue | NO | YES |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With nkill + attacktype | Without nkill + attacktype | Difference |
| SVM | 98.43% | 98.43% | 0 |
| Continue | NO | YES |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With targsubtype + weaptype + targettype | Without targsubtype + weaptype + targettype | Difference |
| SVM | 98.43% | 98.43% | 0 |
| Continue | NO | YES |  |

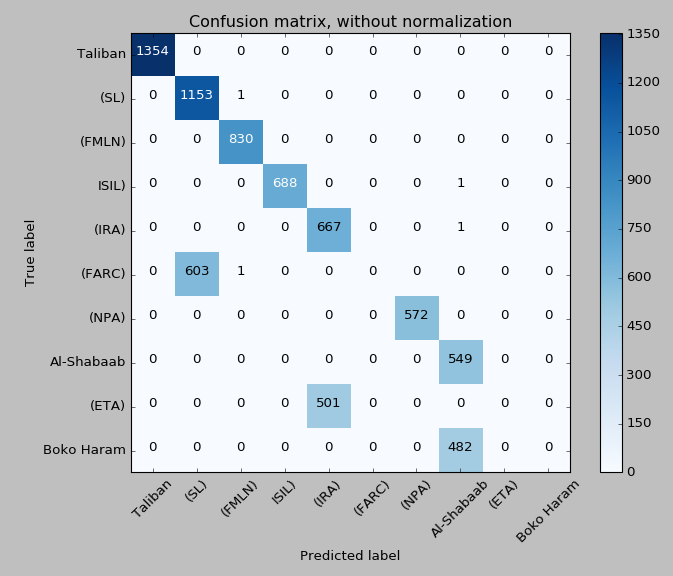
|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With year | Without year | Difference |
| SVM | 98.434% | 98.431% | -0.003 |
| Continue | NO | YES |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With success | Without success | Difference |
| SVM | 98.431% | 98.431% | 0 |
| Continue | NO | YES |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | With country | Without country | Difference |
| SVM | 98.43% | 78.55% | -15.88 |
| Continue | YES | NO |  |

When features start to be removed it is only when the 8th lowest impacting feature is removed that it affects the accuracy rating. When the 9th lowest impacting feature is removed, it stays the same as the 8th lowest impacting feature. When the 10th lowest impacting feature is removed the accuracy, rating drops 15.88%. This means that the region by itself is giving a 78.55% accuracy.

### *Confusion matrix Top 10 after with only 1 feature*

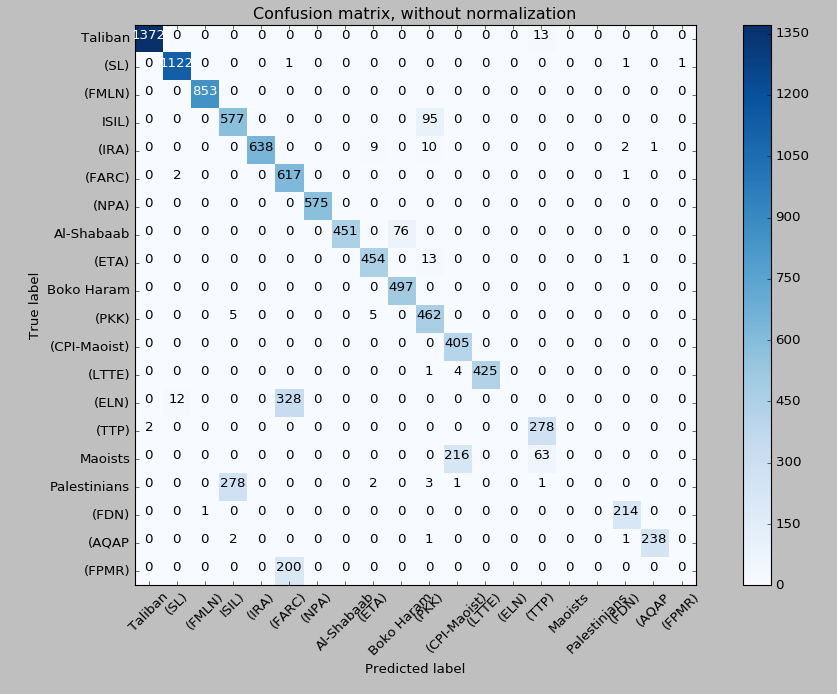


#### Step 4: Increase number of classes with results from previous steps (only 2 features, hyper-parameter optimization, label encoding)

The next step was to increase the number of classes from 10 to 20 using results gathered in pervious steps.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Top 10 | Top 20 | Difference |
| SVM | 98.43% | 87.06% | -11.37 |

### *Confusion matrix Top 20 after with only 2 features*



The result from increasing the number off classes by 100% from 10 to 20, lowered the accuracy rating by 11.37%. This is due to the number classes. It should also be noted as the number of instances lessens the number of false positives increases.

### *Key Information*

Rows 42117

Features 2

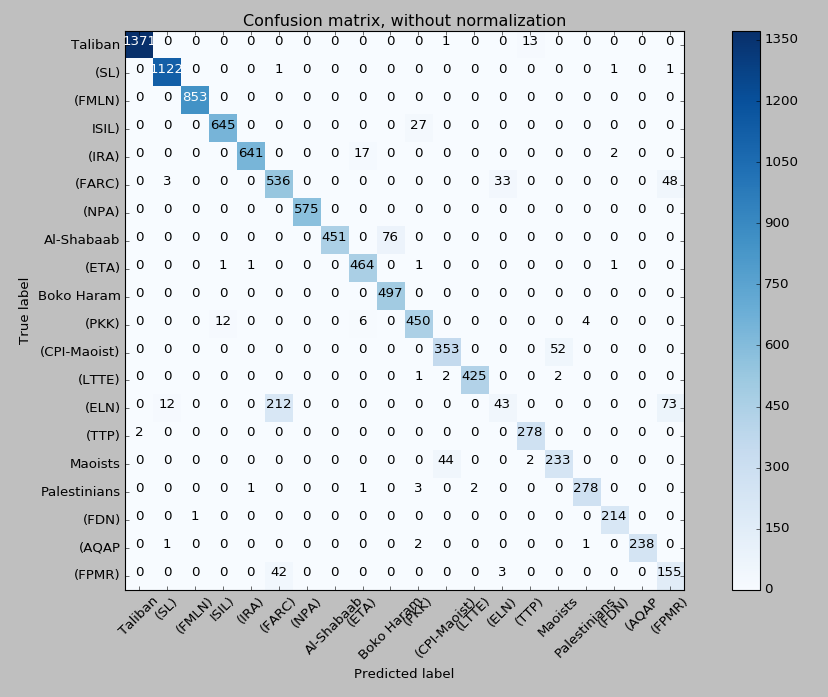
Runtime 142.23 seconds

#### Step 5: Increase number of classes with results from previous steps (12 features, hyper-parameter optimization, label encoding)

The next step was to increase the number of classes from 10 to 20 using results gathered in pervious steps but without removing features.

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Top 10 | Top 20 | Difference |
| SVM | 98.43% | 91.94% | -6.49 |

### *Confusion matrix Top 20 after with 12 features*



The result from increasing the number off classes by 100% from 10 to 20, but not removing features lowered the accuracy rating by 6.49%. This shows that removing features, whilst may seem like a good approach when testing on a lower number of classes does not always work when the number of classes is increased. It should also be noted as the number of instances lessens the number of false positives increases.

### *Key Information*

Rows 42117

Features 12

Runtime 196.56 seconds

### Chapter 5

### Conclusion and Future Work

### 5.1 Conclusion

The results gathered in this research were very intriguing. The first conclusion that was made was that whilst one-hot encoding is great in theory it is not always the best approach when dealing with large datasets with many features and those features having many different values. Label encoder seemed to work very well with this dataset improving runtime and accuracy. Following this hyper-parameter optimization was used, which increased the accuracy even further. The number of features were then increased, lowing the accuracy but helped understand the dataset better.

The last approach used was feature selection and this gave the most enlightening and surprising results. After running univariate feature selection and tree-based feature selection it showed that one feature predicted accurately 78.55% of the instances. When only two features were kept it only dropped the accuracy by 0.003%. The number of classes was then increased by 100% to 20. The classifier was then run on the data twice, with 2 features and 12 features. Whilst it was predicted that the accuracy would drop this was not the most surprising result. The results showed that the pervious conclusion that only 2 features were needed to predict the class was wrong. The accuracy of 2 features on 20 classes was 87.06% and the accuracy on 12 features on 20 classes was 91.94%. This shows that removing features may initially look positive but when more classes are added this is not the case. The conclusion to be taken from this is to always back track and test the model with and without changes made.

### 5.2 Future Work

Further work may be done by increasing the classes even further and seeing how it effects the predicted accuracy. Feature selection should be then run again on this new dataset. Also as the classes increase it is predicted that the accuracy will lower. This should give more false positives. It may also be useful to try adding weights to the classes, as more classes that are added may by be over shadowed by the larger classes.

### Bibliography

[1] International Terrorism and Security Research (no date) Terrorism research - what is terrorism? Available at: http://www.terrorism-research.com/ (Accessed: 19 November 2016).

[2] Shilton, A. (2006) Design and Training of Support Vector Machines. Available at: http://people.eng.unimelb.edu.au/shiltona/publications/Thesis.pdf (Accessed: 19 November 2016).

[3] Ertekin, S. (2009) Learning in Extreme Conditions: Online and Active Learning with Massive, Imbalanced and Noisy Data. Available at: http://web.mit.edu/seyda/www/SeydaErtekin\_PhDThesis.pdf (Accessed: 19 November 2016).

[4] Nam Yu, C. (2011) Improved Learning of Structural Support Vector Machines: Training with Latent Variables and Nonlinear Kernels. Available at: http://www.cs.cornell.edu/~cnyu/papers/thesis\_cnyu.pdf (Accessed: 19 November 2016).

[5] Kaggle(2016) Global Terrorism Database. Available at: https://www.kaggle.com/rishisankineni/global-terrorism-database-gtd (Accessed: 19 November 2016).