CSE 587 - LAB 3
Predicting Explosions and Remote Violence using Apache Spark

Anupriya Goyal(anupriya,50287108)
Sravanthi Adibhatla(sadibhat,50288587)

USE CASE: Predicting Explosions and remote violence using Apache Spark

ABSTRACT:

Unfortunately, every day we hear news related to violence. So, we want to perform predictive analysis to predict the possibility of explosions and remote violence and find the areas prone to such events**.**

Reason for using Spark: We are using Apache Spark to do it because Apache Spark provides real-time streaming, interactive processing, graph processing, in-memory processing and batch processing with a very fast and simple interface. That is why it has gained a lot of importance to use with ML applications. In comparison to MapReduce and other Apache Hadoop components, the Apache Spark API is very friendly to developers, hiding much of the complexity of a distributed processing engine behind simple method calls.

Methodology: We are using Supervised Machine Learning (SVM) algorithm through Spark ML-Lib library and are building model with training data and validate it with test data.

PROBLEM STATEMENT:

To design a Real Time Explosion and Remote Violence Predictive Model in predicting the upcoming events that might happen in future in the countries which are more prone to remote violence. This will help the government to take necessary steps to avoid such events.

DATASET:

We are collection data related to our topic from ACLED(www.acleddata.com), where we will be collecting data from different countries and regions for the Explosions/ remote violence.

MACHINE LEARNING WITH ML-LIB:

- MLlib stands for Machine Learning Library. Spark MLlib is used to perform machine learning in Apache Spark.
- We are using Apache Spark via Scala API to generate our feature matrix and also use ML-Lib to build and evaluate supervised learning model. We are exploring a predictive model for explosions and remote violence. Our source dataset is ACLED, and it includes details about explosions or remote violence filtered by countries from the past 5 years.
- We will build a supervised learning model to predict explosions and remote violence using the year 2017 and 2018 data to build the model, and test its validity using data from 2019.

METHODOLOGY:

The detailed step-by-step construction of the feature matrix, machine learning model, implementation details as well as the model evaluation steps are demonstrated .The methodology entails the following steps:

- Defining functions for pre-processing and feature generation.
- Features we are considering from our dataset are: **Event_date, Event SubType-(Landmine IED/ Grenade), parties/organization involved and Location(Country, state, city, place).**
- Using Spark and Scala to compute the feature matrix.
- Modeling with Spark and ML-Lib, using Support Vector Machines.
- Constructing a predictive model for explosions and remote violence delays by using ACLED data.

BLOCK DIAGRAM:

- We are collecting data from ACLED and streaming it using Spark Streaming.
- Then using Spark MLLib we are performing training and find the places which are prone to explosions and remote violence.
- This predicted data is visualized, and we are using **lat-lon** pairs of the location to plot it on map.

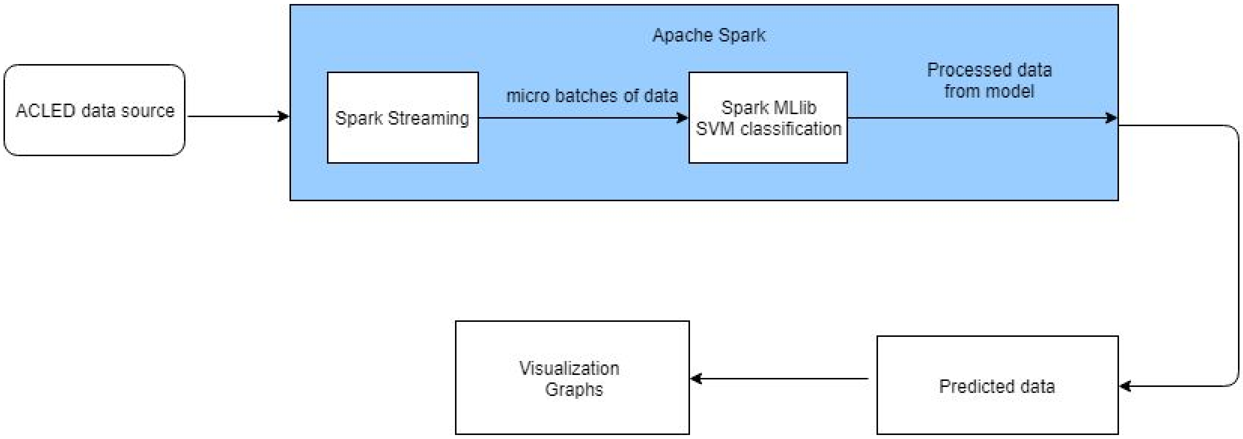
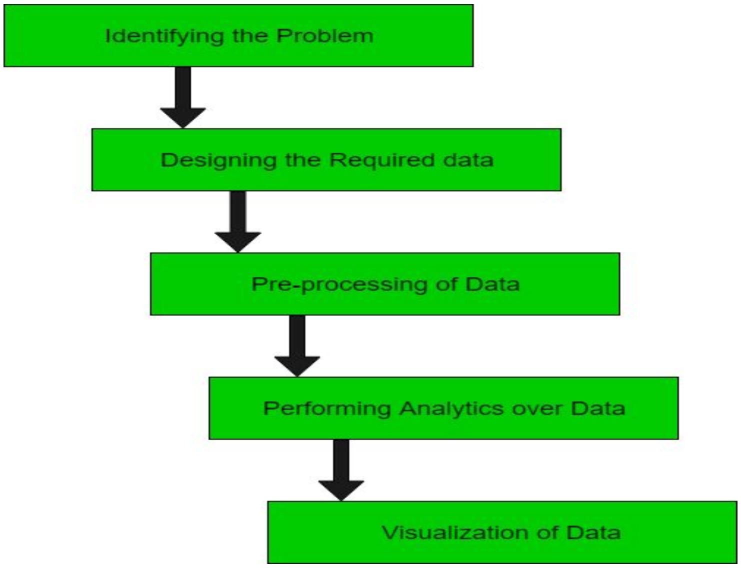


Fig 1: Architecture pipeline block diagram

Flow chart for performing predictive analysis:



PSEUDO CODE:

```
import org.apache.spark._
...
//Creating an Object explosions
object explosions {
 def main(args: Array[String]) {
```

```
//Creating a Spark Configuration and Spark Context
val sparkConf = new SparkConf().setAppName("explosions").setMaster("local[2]")
val sc = new SparkContext(sparkConf)
//Loading the explosions ACLED Dataset file as a LibSVM file
val data = MLUtils.loadLibSVMFile(sc, *Path to the explosions File* )
//Training the data for Machine Learning
val splits = data.randomSplit( *Splitting 60% to 40%* , seed = 11L)
val training = splits(0).cache()
val test = splits(1)
//Creating a model of the trained data
val numIterations = 100
val model = *Creating SVM Model with SGD* (  *Training Data* , *Number of Iterations* )
//Using map transformation of model RDD
val scoreAndLabels = *Map the model to predict features*
//Using Binary Classification Metrics on scoreAndLabels
val metrics = * Use Binary Classification Metrics on scoreAndLabels *(scoreAndLabels)
val accuracy = * Use Test data Classification / (Training data + Testing data)
//Displaying the accuracy on testing and validation dataset
println("Accuracy = " + accuracy)
 }
}
```

EXPECTED OUTCOMES:

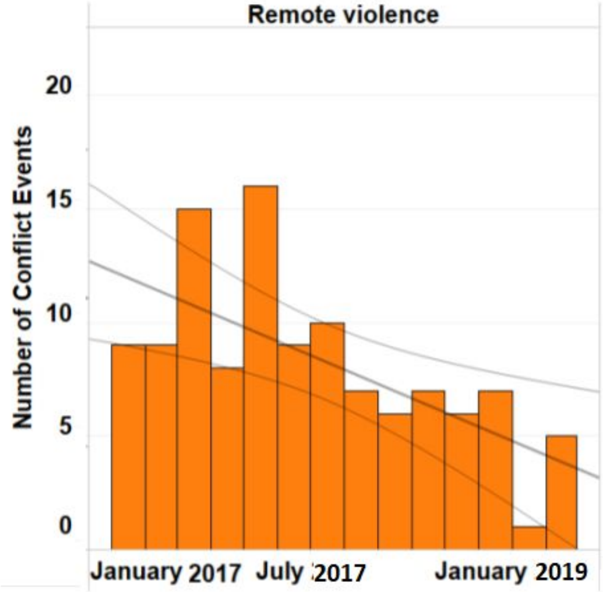
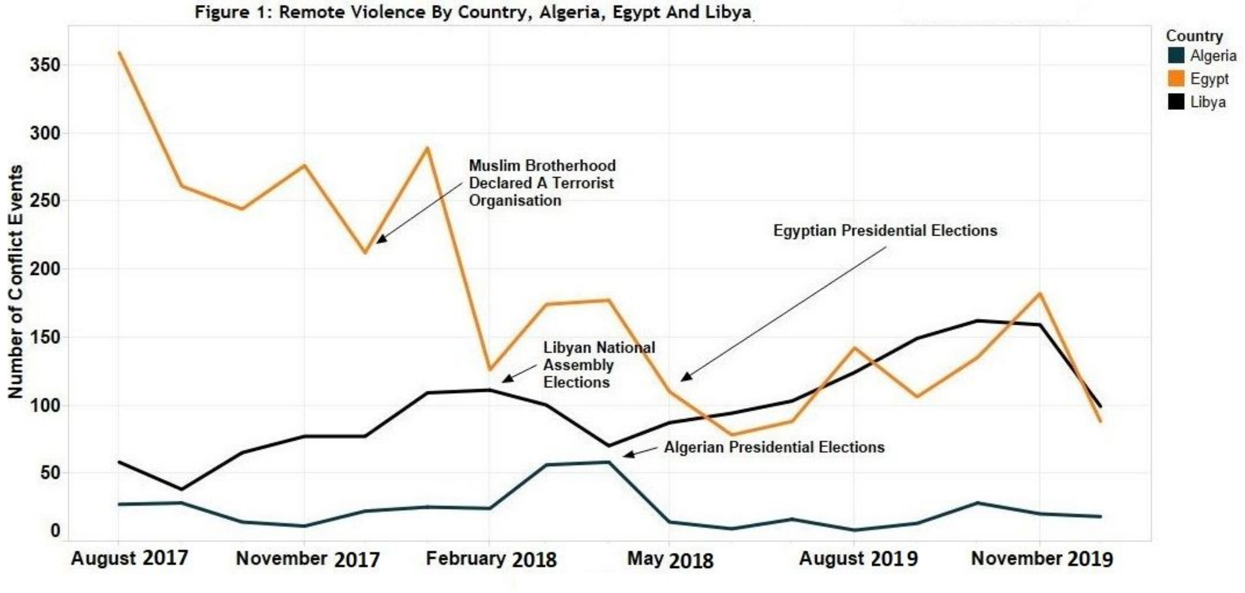


Figure 1: Remote Violence By Country, Algeria, Egypt And Libya



SUMMARY:

- We learned how to use Apache Spark and perform data analysis.
- We understood how to use Spark MLLib and how to implement machine learning algorithm for performing predictive analysis.
- We explored the data flow operations in our Model Pipeline and Spark Architecture.
- We installed Spark environment, we have run the code snippet for SVM algorithm on a small data set similar to our feature data set.

REFERENCES:

https://hortonworks.com/blog/data-science-hadoop-spark-scala-part-2/
https://www.analyticsindiamag.com/how-apache-spark-became-essential-for-machine-learning/
https://blog.cloudera.com/blog/2016/05/how-to-build-a-prediction-engine-using-spark-kudu-and-impala/
https://www.infoq.com/articles/apache-spark-machine-learning