Fouille de données web

Web mining for the detection of terrorism's hints

OUEDRAOGO José [*] — GUINKO Ferdinand[**] — TRAORE Yaya[***]

[*] Ecole Supérieure d'Informatique
Université Nazi BONI
BURKINA FASO
josearthur.oued@outlook.com

[**] Institut Burkinabè des Arts et Metiers
Université Joseph KI-ZERBO
BURKINA FASO
tonguimferdinand@guinko.net

[***] Institut Burkinabè des Arts et Metiers
Université Joseph KI-ZERBO
BURKINA FASO
yayatra@yahoo.fr

RÉSUMÉ. Depuis les événements du 11 septembre 2001, le terrorisme s'est propagé dans le monde entier, menaçant la sécurité et la tranquillité de l'humanité. En tant que contributeur potentiel au développement des technologies avancées, l'exploration de données trouve de multiples applications pour prédire les attaques terroristes ou toute trace de terrorisme sur le Web.Cet article présente une étude sur les indices de prédiction du terrorisme basée sur les méthodes d'exploration de données pour découvrir un moyen efficace de détecter et prévoir des indices de terrorisme. L'objectif de cet article est de comparer et d'identifier un modèle précis pour prédire ou détecter les indices de terrorisme à travers des commentaires sur les réseaux sociaux.

ABSTRACT. Since the events of September 11th 2001, terrorism has spread throughout the world, threatening the safety and tranquility of humanity. As a potential contributor to the development of advanced technologies, data mining finds multiple application in predicting terrorist attacks or any trace of terrorism on the web.This paper presents a study about terrorism's hints prediction based on data mining methods to discover an effective way to detect and predict terrorism's hints.The objective of this paper is to compare and identify an accurate model to predict or to detect terrorism's hints through comments on social networks.

MOTS-CLÉS : Data mining, Weka, Facebook graph api,terrorism's hints

KEYWORDS : Data mining, Weka, Facebook graph api,terrorism's hints

1. Presentation

Web mining is the application of data mining techniques to discover constants, patterns or models in Internet resources or data about it. Nowadays, web data mining is a powerful software technology used to automatically obtain correlations between computerized data from different sources predict events or establish strategies. Thus, for example, "data mining" allows financial organizations to reduce credit card fraud by analyzing transactions in real time. Adapted to large-scale distribution, these software are able, after having analyzed the behavior of consumers through their purchases, to suggest the outline of a marketing campaign... In the security context, web data mining nowadays appears to be a program that uses queries, searches or analyzes in one or more databases to discover or locate predictive forms or indicator abnormalities in a terrorist database or criminal act. Aware of this reality, we decided to conduct studies on a current topic that is entitled:**«Web mining for the detection of terrorism's hints»**. Terrorism is an act of aggression that uses violence and force against civilians and aims to weaken the enemy's morale by terrorizing civilians by various violent means. Terrorism takes place in multiple places between the enemy and the battlefield that is initiated by the use of violence. The challenge of predictive modeling is to create models that have good performance making predictions upon new unseen data. Consequently, it is critically important to use extended methods to practice and evaluate your models on your available training data. The more reliable the estimation of the performance and be confident it will translate to the operational use of your model. The present paper is organized as follows. The first part presents a motivation and background of our study and the proposed methodology for extracting knowledge from social networks. The last section presents the results and the discussion related to our study.

2. Related Work and Methodology

2.1. Related Work

2.1.1. Similar projects in the literature

In depth research related to terrorism's hints and web mining was carried out. However, we have not been able to find articles that deal directly with the search of web mining for the detection of terrorism. But so we did research on projects similar to the one realized in this memoir.The articles selected for this literature review are those that we believe to be the most relevant. They all follow a general approach of Web mining or data mining techniques by detailing all steps. Through our reading, web mining is a powerful software technology used to automatically obtain correlations between computerized data from different sources in order to predict events or strategies in many domains of daily life. Web mining through the use of data mining techniques is widely used in domain medical for the prevention of certain pathologies including for the prediction of the presence or the absence of a problem of heart type, [3]. The second most important issue according to our research is the prediction of all types of cancer, especially breast cancer [1, 2, 5]. Having seen the impact of data mining in the health field, we also contemplate the influence of data mining on the security aspect. Web mining, through the use of data mining techniques is much in the field of the national security of a country collecting the sensitive data to threaten the security to propose new measures of security. It is in this sense that during our reading articles dealing with the classification of terrorist threats,[5]. Also articles dealing with the comparison of jihad, homicides and crimes were met, the purpose of these readings is to apply data mining techniques to predict terrorist attacks or to find measures to predict terrorism,[6, 7, 8]. In the literature review we are completely seduced by the performance of the SVM showing the high level of comparison with other classifiers. Thus, the SVM shows the concrete results of the files of patients with breast cancer[5]. Therefore, the SVM classifier is suggested for diagnosis of breast cancer classification based on the disease for better results with accuracy, low error rate and performance.We will be inspired by it for the rest of our work. Finally,these techniques prove to be an important tool for countering terrorism, [9].

2.2. Methodology and Proposed Solution

2.2.1. Data Set Collector

The «Terrorist Attacks Dataset» where taken from the Global Terrorism Database. The GTD data set is an open source database, most comprehensive and world's most significant data available on terrorism incidents used for the experiment, taken from the National Consortium for the study of terrorism and Responses of Terrorism (START) initiative at the University of Maryland, which broadcasts the terrorism incidents report about the world from 1970 to 2016, and includes information about more than 170,000 terrorist events as well as the vast information more than 120 variables, and include information on more than 83,000 bombings, 18,000 assassinations, and 11,000 kidnappings since 1970. Also, includes information on at least 45 variables for each case.But in the case of our study we used Facebook's Graph API to collect data that is related to the hype of terrorism in posts as well as in communication groups [46].

2.2.2. Data Set collector with using Facebook's GRAPH API

Facebook as a data source has gained a lot of importance in recent years. It is ranked among the 5 most popular websites, with around 2.5 billion registered users and over 3 billion posts generated each day. Additionally, information about APIs, RSS feed or web scraping. Sentiment analysis, stock market, public health, general public mood and finding political alignments are some of the areas where Facebook data has been used.Facebook has several APIs to facilitate the development of solution interacting with its data. These APIs including Graph API. The Graph API is the primary way to get data into and out of the Facebook platform. It's an HTTP-based API that apps can use to programmatically query data, post new stories, manage ads, upload photos, and perform a wide variety of other tasks. The Graph API is named after the idea of a "social graph" — a representation of the information on Facebook. It's composed of :

– **nodes** — basically individual objects, such as a User, a Photo, a Page, or a Comment;

– **edges** — connections between a collection of objects and a single object, such as Photos on a Page or Comments on a Photo;

– **fields** — data about an object, such as a User's birthday, or a Page's name.

Typically you use nodes to get data about a specific object, use edges to get collections of objects on a single object, and use fields to get data about a single object or each object in a collection.

Note : Thanks to the many features offered by the *Graph API* we will use it for data collection because its implementation is very easy. Once this data is collected we will move on to the next step of processing the data.

2.2.3. Data Pre-Processing

Before applying techniques (algorithms) usually some pre- processing is performed on the dataset. It is necessary to improve the data quality to accomplish data processing. There are a few number of techniques used for data processing as data aggregation, data sampling, data discretization, variable transformation, and dealing with missing values.It is in this sense that for the data cleaning, we'll leverage Python's Pandas and NumPy libraries to clean data. For the preparation of the data we will base ourselves on the techniques of the Natural Language Processing (NLP) is a part of computer science and artificial intelligence which deals with human languages. In order to achieve this goal we will use the **NLTK** library of python.

Natural language toolkit (NLTK) is the most popular library for natural language processing (NLP) which was written in Python and has a big community behind it. NLTK also is very easy to learn, actually, it's the easiest natural language processing (NLP) library that you'll use.

2.2.4. Using techniques with Weka Tool

The algorithms in this research are implemented based on WEKA. WEKA is an open source tool created in Java, a collection of machine learning algorithms allows the researcher to mine his own data for trends and patterns and the algorithms can have applied directly to a dataset. In this paper, you will discover the various ways that you can estimate the performance of your machine learning model in Weka tool. How to Evaluating your model using the training dataset, evaluating your model using a random train and test split, assess your model using k-fold cross-validation.There are some model evaluation techniques that you can choose from, and the Weka machine learning workbench others four of them: Training Dataset: prepare your model within the entire training dataset, after that, evaluate the model on the same dataset. Supplied Test Set: divided your dataset is manually utilizing another program. Percentage Split: randomly split your dataset into training and a testing partitions each time you evaluate a model. Cross-Validation: split the data into k-partitions or folds. Train a model on all of the partitions except one that is usually held away as test set after that calculate the average performance of all k models. And you can see these techniques in the Weka tool explorer on the classify tab after you have loaded a dataset. Also, each test option has a time. Evaluation options are concerned with determining the performance of a model on unseen data. Predictive modeling aims to build a model that performs best in a situation that we do not entirely understand, the future with new unknown data. We must use these types of robust statistical techniques to best estimate the performance of the model in this situation and the performance summary is provided in Weka when you evaluate a model. Whenever evaluating a machine learning algorithm on a classification issue, you are given a massive total of performance information to digest because of classification could be the most analyzed type of predictive modeling issue, and there are a wide variety of ways to consider the performance of classification algorithms. Therefore, the first thing to note in the performance for classification algorithms is classification accuracy the ratio of the number of correct predictions away from all forecasts made frequently presented as a percentage where 100% is the foremost an algorithm can perform. The second one is accuracy by class take note of the real positive and false positive rates to get the predictions for each class which may be instructive from the class break down to get the problem is uneven, or you will find more than two classes. As well as the last one is confusion matrix a table showing the number of predictions for each class in comparison to the number of instances that actually participate in each class. The terrorism data set of world wide is splitting into two main sets: Training dataset with 66% percentage and Testing dataset with 34% percentage from the whole data set, and that is applied by using the default setting of Weka tool [47, 48, 49, **?**].

3. Results and Discussions

In our experience, we have applied different algorithms to our database dealing with terrorism's hints. The pre-processed dataset consists of 286 instances of data is converted to *. ARFF which is a file to use by the Weka tool.These data were collected by the api graph of Facebook and then cleaned by datacleaning algorithm in R before being submitted to the analysis under weka. Each test option has a time. Therefore, the algorithms result obtained according to two test options which are:

– Evaluation on Test split that divides the input data set into 66% for the training data and 34% for the test set.

– 10 Fold Cross-Validation.

The results from the applied classification algorithms in the two approaches will be evaluated according to four performance measures which are the classification Accuracy, Recall, F-measure also called F-Score. In case of dividing the input data set into 66% for the training data and the remaining 34% for testing, the results are shown in **Table 1** which provide a clear comparison among the selected classifiers according to accuracy, precision, recall and F-measure which shows that:

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Precision	Recall	F-Measure
Decision Tree	209	77	0.777	0.866	0.819
KNN	203	83	0.748	0.886	0.811
Naive Bayes	204	82	0.769	0.846	0.819
J48	216	70	0.757	0.96	0.846
SVM	196	90	0.742	0.846	0.791

Tableau 1. *Test split and accuracy results*

From the accuracy point of view, J48 correctly classified about 75.7% of the data it means 118 items out of 184 in the 34% test split of the data SVM is outperformed Bayes Net which correctly approximately 61.413% of the data. It is evident that the accuracy of KNN achieved the lower accuracy 51.087% among the other classifiers although it has the lowest precision, recall, and f-measure from other classifiers. Bayes Net classifier produces higher precision, f-measure values than other classifiers, and J48 classifier has the higher recall from different classifiers. The overall performance of NB is near from SVM classifier. It is obvious that a comparison is applied on our five classifiers due to precision, recall, and F-measure which shows us that J48 has the highest accuracy than other classifiers, it does not mean that it performs well in other results. Bayes Net classifier performs also well in all results. The overall performance of NB is near from SVM results. KNN has lower accuracy than all other classifiers although it performs well in other measures.

These pictures above shows the performance measures in case of using classification based on 10-fold cross-validation where J48 has lower precision, F-measure values than SVM, but it could not consider more accurate than J48. NB classifier performs well and very near from Bayes Net especially in recall and F- measure results. KNN has the lowest classification accuracy also it performs well in other measures.

Conclusion : Weka software is very powerful software for study or research needs in the field of prediction and automatic data learning. He has a plurality of classification filtering and learning algorithms that vary from one others according to their use and performance. This architecture available and accessible open source allows to add to this powerful tool other algorithms. Indeed, some classifying and learning algorithms written in Python could be introduced in the software by python communications interfaces. The translation of many mathematical functions in the Python language especially for series geometric and arithmetic is very advanced thanks to the development of numerous libraries and its proximity to computing units. This project allows us to have a broad initiation to data analysis with artificial intelligence algorithms by increasing knowledge about the different classifiers and their implementation in the language python.

During this thesis, we presented several models of prediction, some of regressive type and other classification. We have been able to detect and predict terrorism's hints from many explanatory variables. So we responded to our objective which was to detect terrorism's hints.

In the case of test split of the input data with divisions, 75% for training data and 25% for testing data showed that J48 is more accurate than another classifier especially Naive Bayes, SVM and Bayes Net, the overall performance of NB and SVM is very near.

KNN has the lowest accuracy, but it performs well in other measures. In 10-fold cross validation case, NB classifier is very near to the Bayes Net accuracy, recall, and f-measure. NB classifier performs as Bayes Net in most measures, and J48 performs worse than other classifiers in precision and F- measure.

Finally, some researchers could perform a modification of this research by using different methods. Others could use different test options to test the performance of the classification algorithms.

4. Bibliographie

[1] Vivek Kumar, Brojo Kishore Mishra, Manuel Mazzara, Dang N. H. Thanh and Abhishek Verma. Feb 2019, *«Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications»*, p 8.

[2] Delen, D., G. Walker and A. Kadam. 2005, *«Predicting breast cancer survivability : a comparison of three data mining methods»*,Artificial intelligence in medicine, vol. 34,no2, p. 113–127.

[3] Dangare, C. S. and S. S. Apte. 2012, *«Improved study of heart disease prediction system using data mining classification techniques»*, International Jo urnal of Computer Applications, vol. 47, no10, p. 44–48.

[4] Lee, S.-M., J.-O. Kang and Y.-M. Suh. 2004, *«Comparison of hospital charge predictionmodels for colorectal cancer patients : neural network vs. decision tree models»*, Journal of Korean medical science, vol. 19, no5, p. 677–681.

[5] Bing Bu, Zhenyang Pi and Lei Wang. 2019, *«Support Vector Machine for Classification of Terrorist Attacks Based on Intelligent Tuned Harmony Search»*, article, p. 12.

[6] Allemar Jhone P. Delima. August 2019, *«Applying Data Mining techniques in Predicting Index and non-Index Crimes»*,International Journal of Machine Learning and Computing, Vol. 9, No. 4, p. 6.

[7] Jeff Gruenewald, Brent R. Klein, Joshua D. Freilich and Steven Chermak. October 2016, *«American jihadi terrorism: A comparison of homicides and unsuccessful plots»*, Terrorism and Political Violence, p. 22.

[8] Bart Schuurman. 2019, *«Topics in terrorism research: reviewing trends and gaps, 2007-2016»*, article, p. 12.

[9] Bhavani Thuraisingham, *«Data Mining for Counter-Terrorism»*, The MITRE Corporation Burlington Road, Bedford, MA, On leave at the National Science Foundation, Arlington, VA,

p. 28.

[40] Alexander Pak and Patrick Paroubek, *T«witter as a corpus for sentiment analysis and opinion mining»*,LREC (2010), pp. 1320-1326.

[41] Johan Bollen, Huina Mao and Xiaojun Zeng, *«Twitter mood predicts the stock market J Comput Sci»*, (2011), pp. 1-8.

[42] Michael J. Paul and Mark. Dredze, *«You are what you tweet: analyzing twitter for public health»*, Proceedings of the fifth international AAAI conference on weblogs and social media (2011), pp. 265-272.

[43] Johan Bollen, Huina Mao and Alberto Pepe, *«Modeling public mood and emotion: twitter sentiment and socio-economic phenomena»*,Proceedings of the fifth international AAAI conference on weblogs and social media (2011), pp. 450-453.

[44] Conover Michael D, Gonçalves Bruno, Ratkiewicz Jacob, Flammini Alessandro and Menczer Filippo, *«Predicting the political alignment of twitter users»*, In: 3rd IEEE international conference on privacy, security, risk and trust (Passat); 2011. p. 192–9.

[45] Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti and Basabi Chakraborty, 2018, *«A review on application of data mining techniques to combat natural disasters»*, Ain Shams Engineering Journal.

[46] Dilkhaz Yaseen Mohammed and Murat KARABATAK, 2018, *«Terrorist Attacks In TURKEY»*,6th International Symposium on Digital Forensic and Security (ISDFS),p.3.

[47] Dilrukshi, Inoshika, Kasun De Zoysa, and Amitha Caldera, 2013, *«Twitternews classification using SVM.»*, In Computer Science and Education (ICCSE), 2013 8th International Conference on, pp. 287-291.

[48] Patil, Pritam H., Suvarna Thube, Bhakti Ratnaparkhi, and K. Rajeswari, 2014, *«Analysis of Different Data Mining Tools using Classification,Clustering and Association Rule Mining.»*, International Journal of Computer Applications 93, no. 8.

[49] Nijhawan, Vani Kapoor, Mamta Madan, and Meenu Dave, 2017, *«The Analytical Comparison of ID3 and C4. 5 using WEKA.»* International Journal of Computer Applications 167, no. 11 .

[50] Thankachan, Tulips Angel, and Kumudha Raimond, 2017, *«A Survey on Classification and Rule Extraction Techniques for Data mining.»*, IOSR Journal of Computer Engineering 8, no. 5.

[51] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, October 2006, *«Introduction to Data Mining»*, International Edition,770 p

[52] *«Techopedia explains Web Mining»*,