ALFA BK UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGIES
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
International Scientific Conference

Alfa BK Univerzitet

UDK: 004.6-022.59
DOI: 10.46793/ALFATECHproc25.188P
Original scientific paper

Analysis of Security and Intelligence Data Obtained through OSINT Techniques Using the Apache Hadoop Big Data Platform
Instructions for Authors

Nikola Petrović[1], Vojkan Nikolić[2].

Abstract: Apache Hadoop is a platform for storing, processing, and analyzing large amounts of data. Some of the capabilities of this platform include data storage in HDFS (Hadoop Distributed File System) and the execution of complex HiveQL queries. In addition to Apache Hadoop, which is used in this paper for processing and analyzing the collected data, convolutional neural networks were also employed for image analysis. Data collection was carried out using various OSINT (Open-Source Intelligence) techniques, which involve locating, selecting, and gathering information from publicly available sources.
Keywords: Big Data, Apache Hadoop, OSINT

1 INTRODUCTION

The concept of a Smart City aims to improve the standards and quality of life through modern technologies, while also providing a higher level of security for citizens. All of this would not be possible without smart cameras and the application of artificial intelligence, machine learning, big data processing systems, and other modern technologies in use today.

Big data processing systems are now present almost everywhere, precisely because a large amount of heterogeneous data is generated daily, which cannot be processed by traditional systems. This paper will combine the Apache Hadoop platform for big data analysis, artificial intelligence, big data processing systems, and OSINT (Open Source Intelligence) techniques to demonstrate how the capabilities of smart cameras can be used to improve citizens' security.

The paper will present how, with the help of smart cameras and the application of artificial intelligence for real-time facial and object recognition, potential incidents can be detected and prevented. The recognized facial image will be compared using artificial intelligence techniques and OSINT with publicly available content on the internet to identify similarities. All internet content will be downloaded and stored in the big data processing system's storage, after which the data will be further analyzed.

2. APACHE HADOOP PLATFORM AND BIG DATA PROCESSING SYSTEMS

When we mention the term "big data," we most often associate it with analysis within the context of Apache Hadoop. Unlike traditional data storage and processing systems, Hadoop offers the ability to store and process heterogeneous data. For example, relational databases can store large amounts of data and analyze it, but their main limitation is that they can only store one type of data. Hadoop allows us to store heterogeneous data such as: text files, photographs, videos, sensor data, etc. Big data is essentially defined by the three "Vs": Volume, Velocity, and Variety.

For data storage, Hadoop provides its distributed file system – Hadoop Distributed File System (HDFS). To run queries, it uses the HiveQL query language, which is quite similar to SQL. HBase is a distributed Hadoop database that runs on top of HDFS.

A major advantage of big data processing systems is real-time analytics, which provides the ability to analyze data in real time.

MapReduce is a programming model that allows for parallel processing of large amounts of data distributed across multiple computer nodes. The model consists of two main functions: "Map" and "Reduce." The Map function takes input data and divides it into smaller chunks, which are then distributed for processing across different nodes. The Reduce function aggregates the results from all the nodes to produce the final result. This approach enables efficient processing and analysis of vast amounts of data in distributed systems like Apache Hadoop.

3. OSINT AND ARTIFICIAL INTELLIGENCE

OSINT (Open Source Intelligence) refers to the process of collecting, analyzing, and interpreting information that is publicly available through open sources. The concept of OSINT is used in many fields such as journalism, security, and by professionals in various research areas on the internet. OSINT techniques are relatively new and have primarily developed with the growth of the internet and social media.

Publicly available sources refer to any information accessible to the general public. These do not necessarily have to be published on the internet; they can include data published in newspapers, magazines, or on television. Public or open sources are data that can be found on: general internet searches, social media monitoring, website analysis, photo analysis, and other multimedia content, the use of public databases, and publicly available Web GIS.

Information that can be gathered through OSINT includes: personal data, geographical data, social data, business data, political data, technical data, cultural information, etc. The most common application of OSINT is in the military, security agencies, and journalism.

The artificial intelligence techniques used in this paper involve Convolutional Neural Networks (CNN) for facial and object recognition in images by analyzing their characteristic patterns. CNN is trained on large datasets, learning to recognize specific facial features, such as eyes, nose, mouth, and their relative positions. Once trained, the network can recognize the same faces in different photographs, even under varying lighting conditions, poses, or angles. Using the OpenCV library, which contains tools for detecting faces and objects, AI can quickly analyze images and identify objects such as cars, people, or weapons, as is the case in this study.

4. ANALYSIS OF DATA COLLECTED FROM CAMERAS IN A SMART CITY USING BIG DATA PROCESSING SYSTEMS, ARTIFICIAL INTELLIGENCE, AND OSINT

For the purposes of the research, a scenario was designed that represents a possible method for identifying a suspicious person in large gatherings in a smart city.

The scenario is as follows:

In a large crowd at Republic Square during a concert taking place there, an object detection system on camera footage sends an alert that a weapon has been detected in the possession of a person. The concert security notices the alert, and the goal is to identify the person as quickly as possible.

1. The security officer immediately reviews all available cameras. The camera search program, which has the ability to search both in real-time and retroactively, and which uses software to search for specific objects in the images, including weapons, sets the criteria for the search. The criteria are to separate a middle-aged man who is carrying a weapon as an object.

Search criteria: Man, between 40-50 years old, wearing a green cap on his head, the wanted item is a

FIGURE 1.0: Camera review with search criteria before the search.

The program, with the help of smart cameras, almost in real time, separates a few photos that meet the criteria, on which the sought-after man can clearly be seen. However, the live camera feed shows that the last time the criteria were met was ten minutes ago, meaning that the sought-after person is currently not visible on any of the cameras.

FIGURE 1.1: Display of the person recognized by the system.

3. The search of existing data reveals that the system has not been able to find a match or identify the sought-after person in the current databases. Therefore, in this step, techniques for searching and comparing photos on the internet will be used. The goal now is to search the internet and social media to find a match for the person from the camera. We once again use the Python OpenCV library, which is now used for comparing the same facial images, and alongside it, we use OSINT, which helps us search available sources on the internet with predefined functions for optimal searching.

4. A match has been found, and through visual inspection, it can be confirmed that it is indeed the same person. An account was found on the social media platform Instagram, which contains multiple posts with photos featuring the sought-after person.

FIGURE 1.2: Display of the match between the search sample photo and the photo from the Instagram social media account.

5. It has been established that this is the profile of the hunting association *_hunterssrb_*.

6. In this step, the entire list of accounts, as well as all photos from the account that follow the *_hunterssrb_* profile and all the accounts that *_hunterssrb_* follows, are downloaded. Specialized OSINT techniques are used to download the content from the Instagram social media platform.

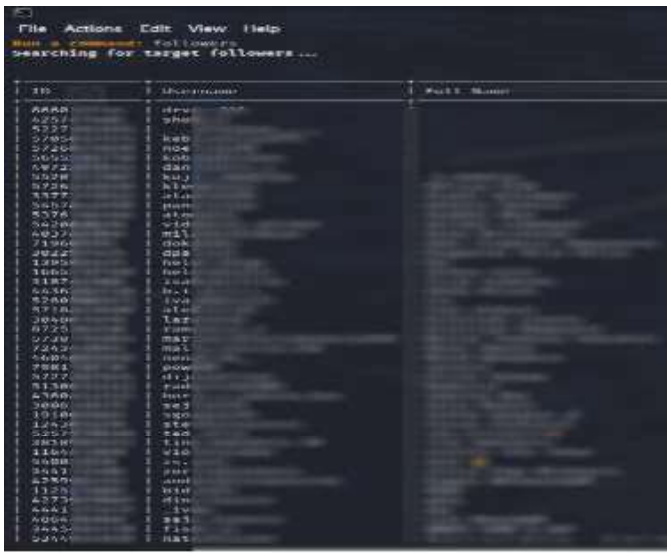

FIGURE 1.3: Display of the command for downloading the list of all followers of the profile.

7. The downloaded material is stored in the big data processing system's storage. A total of 1,221 photos, 213 videos, and 3,000 text files have been downloaded.
8. All photos from the *_hunterssrb_* profile that feature the person we are looking for are downloaded.
9. The downloaded photos from the *_hunterssrb_* profile are now compared using artificial intelligence techniques with the photos and videos downloaded from the profiles that *_hunterssrb_* follows and those who follow him. The assumption is that the sought-after person will be identified in these images, as the likelihood of following a profile in which the person is tagged is high.
10. After the download, no match was found, so now all comments on the posts from the *_hunterssrb_* profile featuring the person to be identified are being downloaded.

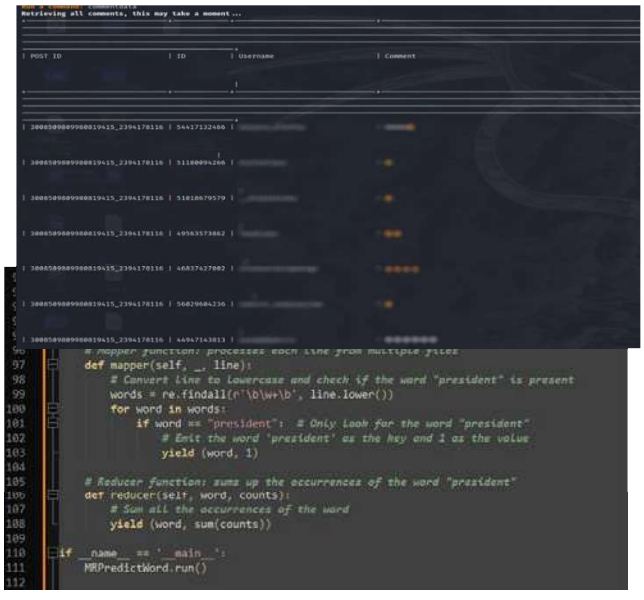

FIGURE 1.4: Display of the command for downloading comments from the profile on the Instagram social media platform.

11. After the download, similar comments are separated using HiveQL queries in the big data processing systems. It is established that the word "president" is mentioned in almost all posts.
12. Now, using MapReduce code to search for similar words in all downloaded files, the mentioned word is searched for and counted.

FIGURE 1.5: Display of MapReduce code for searching similar words.

13. The assumption is that the person to be identified is the president of this hunting association, so a search is now conducted in the Agency for Business Registers to determine the name of the association's president and his identification number.

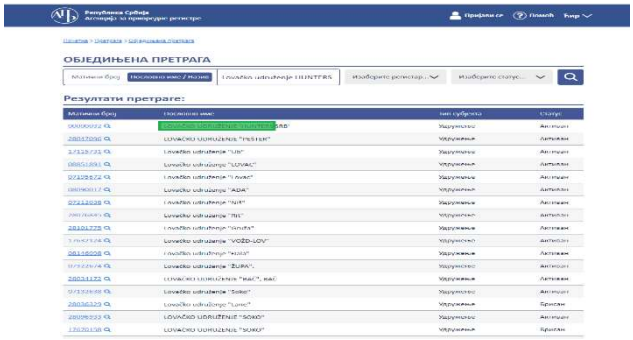

FIGURE 1.6: Display of the search in the Agency for Business Registers.

FIGURE 1.7: Display of representative data from the Agency for Business Registers.

14. The data obtained from the Agency for Business Registers, after further verification, confirmed that it is indeed the person that needed to be identified. Security personnel on the field, after receiving the information, can proceed with their work.

CONCLUSION

The goal of this paper was to demonstrate how a combination of artificial intelligence techniques for object detection, big data processing systems, OSINT, and Smart City technology can contribute to preventing security incidents at large gatherings in smart cities. The application is not limited to this specific case; this method can be applied at any time and can accelerate the process of identifying any individual whose face has been clearly captured by a smart camera. If the case from the example were resolved in a traditional manner, it would take much more time to identify the person seen at the gathering and subsequently carry out identification. However,

with the help of the aforementioned techniques for object detection in images, human face detection, and big data processing systems, identification is completed within minutes, almost in real time. Publicly available data on the internet significantly contribute to this problem-solving approach, as it is rare today to find a person whose photograph has not been uploaded somewhere online.

References

[1] Koops, BP., Hoepman, JP. & Leenes, R. 2013. Open Source Intelligence and Privacy By Design. Journal of Computer Law & Security Review, Elsevier, Tilburg University, The Netherlands

[2] Wiil, U. K. 2011 Counter Terrorism and Open Source Intelligence: Lecture Notes In Social Networks. Vol 2. 15 – 28. Denmark.

[3] Pradeepa, A., &Thanamani, A. S. (2013). Hadoop file system and fundamental concept of MapReduce interior and closure rough set
approximations. International Journal of Advanced Research in Computer and Communication Engineering, 2(10), 5865-5868

[4] J. Ekanayake, S. Pallickara, G. Fox, MapReduce for data intensive scientific analyses, in: Proceedings of Fourth IEEE International Conference on eScience, Indianapolis, Indiana, USA, 2008, pp. 277–284

[5] Verma, A., Mansuri, A. H., & Jain, N. (2016, March). Big data
management processing with HadoopMapReduce and spark technology: A comparison. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-4). IEEE.

[6] Grinev. M., Grineva. M., et ai, "Analytics for the real-time web", In PVLDB, 4(12):1391-1394, September 2011.

[7] Helmi, R., Yusuf, S., & Jamal, A. (2019). Face recognition automatic class attendance system (FRACAS). In *IEEE international conference on automatic control and intelligent systems (I2CACIS 2019)*, Selangor, Malaysia, June 29, 2019.

[8] Zhao, Z.-Q., Zheng, P., Xu, S., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems,30*(11), 3212–3232.

Contact information:

Nikola Petrović[1], Ministry of Internal Affairs, Republic of Serbia,11000 Belgrade,Serbia
Mail
nikola.spetrovic@mup.gov.rs

Vojkan Nikolić[2],
Department for Informatics and Computing, University of Criminal Investigation and Police Studies, 11000 Belgrade, Serbia
Mail
Vojkan.nikolic@kpu.edu.rs