Post Classification in the Social Networks using the Map-reduce Algorithm

Abdoulaye SERE[1], José Arthur OUEDRAOGO[2]
Network of Computer Science Teachers
and Scientists in Faso
Bobo-Dioulasso, Burkina Faso

Boureima ZERBO[3], Oumarou SIE[4]
Network of Computer Science Teachers
and Scientists in Faso
Ouagadougou, Burkina Faso

Abstract—**Wrongdoing is increasing through social media. Detecting them requires highlighting the most interesting topics in the posts. This essential part in the characterization of social network users could be done by a classification of posts. For this, we use a tuple of keywords and the Map-reduce algorithm for data collection and extraction. The main purpose is to achieve success on software realization which will establish a network between social networks to extract data and to speed up the classification of posts. The proposed method consists of verifying a sequence of keywords in the posts, following a grammar in order to determine classes. It allows the categorization of posts and monitoring of social networks. The categorization facilitates research of a particular post containing specific words. Thus, we contribute to increase capacity for wrongdoing prevention and strengthening cyber-security.**

Keywords—*Big Data; map-reduce; social network; cyber-security; classification*

I. INTRODUCTION

Now-a-days, data is playing essential roles in analysis for the enterprises, in taking decisions. Data has become the basic resources for all the applications. Particularly, the applications based on the techniques of deep Learning and Data mining need more data.

Inversely, the internet of Things, the social networks produce enormous quantity of data, named Big Data. Due to the number of posts, the social networks propose a diversity of Big Data with several types such as texts, images and video.

Big Data involves several levels of problems such as the problem of data collection, data processing and data visualization. The problem of Big Data visualisation has been analyzed by Alexandre Perrot in his thesis in [1].

Data extraction is essential to build the collections of data on infrastructures for analysis. Frederic and others in [3] used Netvizz to extract data from Facebook to realize surveys on the 2015 presidential election in Burkina Faso.

Thus, Big Data processing needs the parallelization of data and tasks in order to reduce processing time. Since 2004, Dean and others in [2] proposed the map-reduce framework to count words in enormous documents. The techniques of processing big data have increased more and more with several implementations such as Hadoop , Spark in [5] and in [7]. The map-reduce has been also used in [6] for data-mining and in [4] for parallel algorithmic . A survey on the map-reduce framework has been proposed by N.Alamelu Menaka and others in [10].

Both the map-reduce algorithm and classification are the subjects of several works in [8], [9]. For example, in [8] Ouatik and others have tried a classification of student into four classes, scientific, literary, technical and original, in using the map-reduce algorithm.

Hadoop has been improved to take image processing into account. For example, Hadoop Image Processing Image (HIPI) has been proposed by the University of Virginia Computer Graphics Lab, in 2016 : it increases Hadoop capacity with new functions, giving the processing of distributed images and in taking the techniques of image processing into account. SERE and others in [11] introduced an application of the Hough Transform based on the map-reduce algorithm to improve processing time in straight line detection in distributed big images. The works of SERE and others in [11] have been extended by Mateus Coelho and others in [12], to circle detection in using the map-reduce algorithm.

In others ways, in recent years, scientists have worked on multi-label classification to obtain web page categorization. In [13], Yaya and others have studied the multi label classification in using an ontology in order to classify the web pages.

Every day, users post regularly relevant informations on social networks that can be used to characterize them to determine their behaviour. The posts give a way to obtain information on the users, to predict their future behaviour, to profiling them and to know what they are doing or what they are planning to do.

A selection of keywords allows generally the detection of posts related to specific subjects such as wrongdoing, fake news, terrorism, covid 19. For example, a list of keywords in the posts leads to track the user behaviour.

A Hub between social networks implementing the map-reduce algorithm, using a set of keywords, to extract speedly data from social networks does not exist.

But, there are available interface layers such as Application Programming Interface (API) to extract posts from social networks using keywords. For instance, Twitter proposes an API in python named tweepy that allows to get all the posts extracted into a file.

In this paper, we address the issue of post categorization in social networks. The objective is to facilitate profiling by classifying the most discussed topics in social networks. To do them, the map-reduce algorithm to establish data parallelism is used. Also, in big data, to speed up data processing, we

use parallelism in data extraction. The purpose focuses on monitoring of social networks through a software connected simultaneously on several networks.

The paper is organized as follows: the Section 2 concerns preliminaries followed by the method description in the Section 3. At the end, the Section 4 deals with the simulation of both parallelism using the library thread and parallelism through the map-reduce framework.

II. PRELIMINARIES

This section is focusing on the concepts related to the theories of words, languages and explains the meaning of the map-reduce algorithm. That allows forward the best understanding of the following sections.

In the theories of words and languages, a word is a sequence of elements or symbols in the alphabet while a language is a finite set of words. A language can be defined by a regular expression or by a grammar. For instance, let A be an alphabet {a, b, c}. "aa", "abc" are the words of the alphabet A. A^* is the universal language. It is the set of all the words defined by the alphabet A. Any word based on the alphabet A is a member of A^*. Any language based on the alphabet A is a subset of A^*.

In this paper, the keywords could be the words with any alphabet for any language such as French or English used to write posts.

The map-reduce algorithm is proposed initially by Dean and others, in 2004 [2], when working at Google. It is made up of three phases: the mapping phase, the shuffle phase, the reducing phase. The map function defines a transformation that accepts as input a single key value (k, v) pair and produces as output a set of intermediate (key, value) pairs (k_i, v_i). At the end of all the map functions, several pairs (k_i, v_i) are produced for the shuffle phase. For instance the map function transforms the pair (k, v) as input and produces the set of pairs (k_i, v_i) such as (k_1, v_1), (k_2, v_2).

The shuffle phase defines the groups of pairs that have the same k_i. It plays the role of creating groups. Suppose that (k_1, v_1), (k_1, v_2), (k_1, v_3), (k_2, v_4), (k_2, v_5), (k_3, v_3) the pairs produced by all the map functions. The shuffle creates three groups where each group along with its key k_i such as summarized in the Table I:

- the group 1 is $(k_1, < v_1, v_2, v_3 >)$
- the group 2 corresponds to $(k_2, < v_4, v_5 >)$
- the group 3 is (k_3, v_3).

TABLE I. THREE GROUPS CREATED BY THE SHUFFLE PHASE

Groups	Data
Group 1	$(k_1, < v_1, v_2, v_3 >)$
Group 2	$(k_2, < v_4, v_5 >)$
Group 3	(k_3, v_3)

Forward in the method description, the value v_i will correspond to a post. k_i will mean the common criteria for all the posts in a class.

The reduce phase concerns mainly the processing of each group, created by the shuffle phase. That means an operation is introduced by the reduce function to perform the value v_i of each group. For example, considering the operation +, the values of previous groups become:

- the group 1 $(k_1, < v_1, v_2, v_3 >) \Longrightarrow (k_1, < v_1 + v_2 + v_3 >)$
- the group 2 $(k_2, < v_4, v_5 >) \Longrightarrow (k_2, < v_4 + v_5 >)$
- the group 3 $(k_3, v_3) \Longrightarrow (k_3, v_3)$

Finally, the results of the map-reduce algorithm give $(k_1, < v_1 + v_2 + v_3 >)$, $(k_2, < v_4 + v_5 >)$ and (k_3, v_3)

There is a similarity between the wordcount (Dean and others in 2004 [2]) and classification proposed in our method.

III. METHOD DESCRIPTION

This section describes the technique of data extraction used through API layer connected on social networks and the classification of posts based on the map-reduce algorithm.

In the social networks, a post is made of texts, images and video. Often, a text comments an image or a video. That means, the text analysis is useful for the descriptions of an image or a video. A text is a set of sentences or a sequence of words. For instance, the following message posted by the "Ministère du Développement de l'Economie Numérique et des Postes" in French in Facebook presents two sentences such as the sentence 1 and the sentence 2:

- the post is: "Les panélistes du webinaire de dissémination des stratégies de développement du numérique et des postes au Burkina Faso des 25 et 26 juin 2020. Le lien pour l'inscription: https://bit.ly/3gRDsy0";
- the sentence 1 is: "Les panélistes du webinaire de dissémination des stratégies de développement du numérique et des postes au Burkina Faso des 25 et 26 juin 2020." ;
- the sentence 2 is: "Le lien pour l'inscription: https://bit.ly/3gRDsy0".

A sentence is a sequence of words. Inversely, a sequence of words is not always a right sentence. Formally, a sequence of words is represented by a tuple of keywords defined by $(w_1, w_2,..., w_{n-1}, w_n)$ where w_i is a keyword. In the previous post, a tuple $(w_1, w_2,..., w_{n-1}, w_n)$ can be (panélistes , stratégie, développement).

Data extraction is based on the definition of a criteria.

A. Data Extraction

Data extraction concerns extraction of posts regularly in the social networks in following the keywords and to store them in a nosql database, in the perspectives to prepare the research phase on the posts. The problem is also how to determine the manner to realize Data extraction and collection, speedly.

Social networks propose facilities through API to access to the posts with certain constraints. The problems of Data

extraction depends on the permissions of social networks and API layer availability to extract posts. The technique used for extraction of posts depends on each social network.

There are no connections between social networks through a common hub to facilitate extraction and the collection of posts in using keywords. That leads to establish a layer to realize data extractions.

Due to the large volume of posts from social networks, it appears necessary to use appropriate techniques in the layer such as the parallelism of data or tasks to process all the posts. For instance, in Twitter an Application Programming Interface (API) allows extraction of posts in using keywords. A solution is to establish the parallelization of available functions in this API to accelerate data extraction.

Thus, the proposed method introduces a layer named the map-reduce algorithm that will play the role of data extraction and will transfer data to a nosql database for data collection. It consists of injecting codes into the function map and reduce to define criteria for data extraction. The layer will communicate with the API layer in order to access to the social network database of posts. For instance, the Fig. 1 illustrates the global architecture of data extraction and collection.
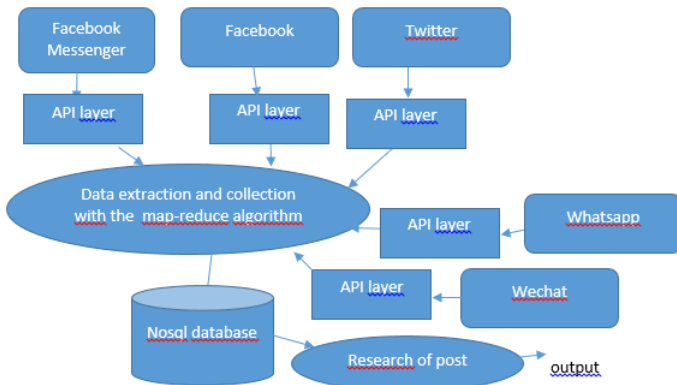

Fig. 1. The Architecture of Data Extraction, Collection and Research

B. Data Classification

In classification, a criteria establishes conditional instructions that describe how to put together data in the same group : It gives the common characteristics or the properties of data or the conditions that data must respect to be considered in the same class. The criteria are based on the tuple (w_1, w_2,..., w_{n-1}, w_n) and the grammar, defined by:

$$A1 : -w_1 \| w_1 A2 \| A2$$
$$A2 : -w_2 \| w_2 A3 \| A3$$
$$\overline{\qquad\qquad\qquad\qquad}$$
$$An - 1 : -w_{n-1} \| w_{n-1} An \| An$$
$$An : -w_n \| \epsilon$$

the grammar leads to respect the ordering of keywords, to align keywords syntactically. But ϵ is not considered as a keyword in this context. The method consists of the verification of a sequence of keywords appeared in the posts. The sequence of keywords produced by the criteria based on the tuple (w_1, w_2,..., w_{n-1}, w_n). The grammar follows the forms (w_i,... w_j) where i < j. The tuple defines the keywords and establishes the order of keywords that is confirmed by the grammar. The grammar also generates the groups of keywords. For instance, four groups of keywords w_1, (w_1, w_4, w_5), w_2, (w_4, w_{10}) will correspond respectively to four classes such as summarized in the Table II, which presents a similarity with the wordcount [2], corresponding to the Table I in preliminaries:

- the class 1 is the set P1 of posts that match the keyword w_1;

- the class 2 concerns the set P2 of posts that match the sequence (w_1, w_4, w_5) in the order;

- the class 3 is the set P3 of posts that match the keyword w_2;

- the class 4 concerns the set P4 of posts that match the sequence(w_4, w_{10}) in the order.

TABLE II. TABLE OF FOUR CLASSES

Post classes	Keywords	Posts	Data
class 1	w_1	P1	(w_1, P1)
class 2	(w_1, w_4, w_5)	P2	((w_1, w_4, w_5), P2)
class 3	w_2	P3	(w_2, P3)
class 4	(w_4, w_{10})	P4	((w_4, w_{10}), P4)

If two posts contains the same sequence of words, they will belong to the same class. Thus, the order of words in the tuples influes on the result of classification. The order of keywords (w_i,... w_j) where i > j, is not considered in this analysis and could be studied in the perspectives.

In combinations of words, the number of combination for the keywords in the tuple (w_1, w_2,..., w_{n-1}, w_n) is established by $C_n^1 + C_n^2 + ... + C_n^{n-1} + C_n^n$ and corresponds to the number of post classes.

The ordering of the keywords, in following the tuple, determines a semantic meaning associated to a post that matches these keywords. For instance, a post can be a member of the class w_1 and the class w_2, if it matches the pair (w_2, w_1), not ordered in following the tuple. In this manner, a post can become a member of two different classes.

The classification is realized by the map-reduce algorithm, introduced here as a template to implement the parallelization of data and tasks and to accelerate data processing. By analogy to the works of Dean and others on the wordcount, to count the number of words in a document [2], the proposed model performs the tuple of keywords to verify if they appears in the order in the posts, to output the classes of posts. For instance, the algorithms 1, 2 show respectively the content of the map function and the reduce function that deal with the parallelization of data and tasks. The grammar is implemented in the content of the map function. The operation of classification is automatically done by the shuffle phase. The reduce function creates as outputs the classes.

In the Algorithm 1, tab[1..N] is a vector of keywords where each element corresponds to a class: The keywords generated by the grammar are inserted into the vector tab[1..N]. That means N= $C_n^1 + C_n^2 + ... + C_n^{n-1} + C_n^n$

Algorithm 1: function map(k:PostID, v:Post)

Result: the pairs in the form (k, v)
tab[1..N]: vector $< Keyword >$;
p: Keyword ;
i: Integer ;
for *(i=1; i ≤ n; i++)* **do**
    p=tab[i] ;
    if *match(v, p)* **then**
        emit(p, v);
    end
end

Suppose that the shuffle transforms the set of pairs in the form (k, v) to the pairs $(k, < v_1, v_2, ...v_{k-1}, v_k >)$ where $< v_1, v_2, ...v_{k-1}, v_k >$, having a type vector $< Post >$, in order to prepare the reducing phase. As illustrated in the algorithm 2, the reducing phase removes duplicate posts in a vector of posts.

Algorithm 2: function reduce(p:pattern, list: vector $< Post >$)

Result: the pairs in the form (k, v)
cl: vector $< Post >$;
index :vector $< int >$;
temp: Post ;
b:boolean ;
i, j : Integer ;
b=false;
for *(i=1; i ≤ n; i++)* **do**
    temp=list[i] ;
    for *(j=i+1; j ≤ n; j++)* **do**
        if *same(list[j],temp)* **then**
            b=true;
        end
    end
    if *(b==false)* **then**
        cl.add(temp);
    end
    b=false;
end
emit (p, cl)

Simulations on the proposed algorithms will be studied in the Section IV.

IV. NUMERICAL SIMULATION AND TIME EVALUATION

This section evaluates the time according to the number of posts extracted on social networks.

Consider the tuple $(w_1, w_2, ..., w_{n-1}, w_n)$ where n=2. That corresponds to the tuple (w_1, w_2). Suppose that (w_1, w_2)=(terrorism, covid19). The table tab[1..N] of keywords in the Algorithm 1 become tab[1..3], represented by the Table III.

TABLE III. THREE KEYWORDS

Keyword 1	Keyword 2	Keyword 3
terrorism	covid19	(terrorism, covid19)

TABLE IV. THREE CLASSES

Classes	Keywords
Class 1	terrorism
Class 2	covid19
Class 3	(terrorism,covid19)

The number of classes is then N= $C_2^1 + C_2^2$. That leads to three classes such as summarized in the Table IV.

The classification of posts is established by the shuffle phase in considering the keywords in the vector tab[1..3].

Global time processing depends on the configuration of both Local Area Network (LAN) and Wide Area Network (WAN). For LAN, data extraction uses a computer, connected on the Internet directly or through a proxy in a LAN. The characteristics of this computer are defined by :

- Memory : 3,7 GB

- Processor : Intel® Celeron(R) CPU B830 @ 1.80GHz x 2

- Graphic card : Intel® HD Graphics 2000 (SNB GT1)

- Operating System : Ubuntu 18,04 LTS 64 bits

In considering the pair of keywords (terrorism,covid19), a layer, written in python, is used to extract data from Twitter and Facebook, implementing data parallelism: The layer has been connected to Twitter server and Facebook server in using API, respectively tweepy and Facebook.

Here, parallelism implementations have been performed by two alternatives, such as data parallelism based on the library Thread in python and data parallelism implementation through the map-reduce algorithm with spark streaming.

A. Data Parallelism using the Library Thread (in Python)

In this way, three threads such as task 1, task 2, task 3 have been created. The task 1 is connected on Twitter to search the posts that correspond to the keyword "Covid19". The task 2 also connected on Twitter, searches the keyword "terrorism" while the tack 3 performs the posts of Facebook. Thus, a common platform is created between Twitter and Facebook, to extract the posts. The Fig. 2 shows step by step according to the number of posts, processing time between the serial case and the parallel case. That also shows clearly, according to the number of posts extracted increasing, an improvement of processing time in the parallel case.

B. Data Parallelism using Spark Streaming

There exist many tools that could be used for data parallelism on real time. For instance, Apache storm [14], Apache Flink [15] and spark streaming [16] implement the framework map-reduce and can be connected directly to Twitter, as a data source, to obtain the posts. But here, the map-reduce in spark streaming, is connected on Twitter that has been implemented,
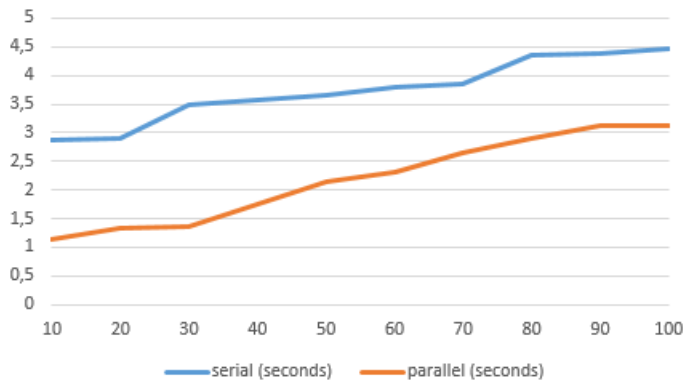
Fig. 2. Curves (Serial, Parallel)

in following the previous Algorithms 1 and 2: the keywords used in the function mapreduce are "terrorism", "Covid19". Processing time in the three cases (serial, parallel and spark streaming) are illustrated in the Fig. 3. That shows:

- from 0 to 70 tweets, tweet collection time in spark streaming is inferior to the serial case one and superior to the parallel case one.

- from 70 tweets to more, collection time in spark streaming is decreasing on the curve

- from 80 collected tweets to 100 collected tweets, the curve for spark streaming is going decreasing and stays under both the curve of the serial case and the curve of the parallel case.

Finally, the experimental results shows that data parallelism using a map-reduce tool in streaming, gives more posts in a short time.
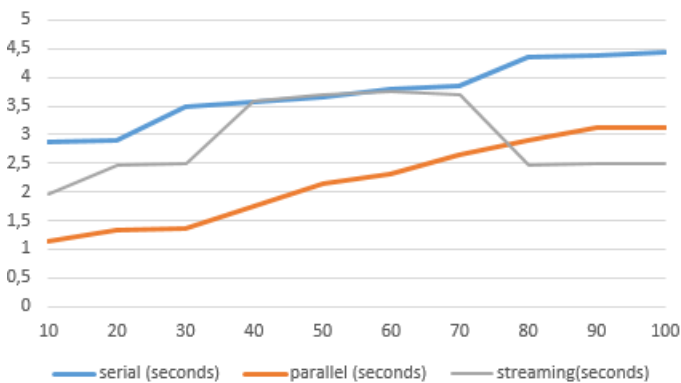

Fig. 3. All the Curves (Serial, Parallel, Spark Streaming)

Thus, the map-reduce tool, spark streaming is better than the serial case and the parallel case, in post extraction in the term of processing time, in considering the number of posts increasing up to 80.

The future research will focus on Deep analysis of network frames produced by spark streaming. We'll study in details the content and the functionalities of spark streaming, noticeably the number of dynamic threads generated for scalability control according to the number of post increasing. It contributes to the best understanding of advantages with spark streaming case than in using directly the library thread.

Moreover spark streaming extension or others tools to take others social networks will lead easily to a software for monitoring social networks about post contents. Because, many social networks don't accept unfortunately connection to the posts in streaming status through a map-reduce tool.

V. CONCLUSION

This paper has proposed the map-reduce algorithm, to extract speedly posts from social networks such as Twitter and Facebook, in order to realize a classification of posts, based on a tuple of keywords and a grammar. The tuples has defined the order of keywords while a grammar has generated the groups of keywords, corresponding to the classes. The keywords also play the role of class indexation.

The experimental resuls reveal that spark streaming is better than the parallelism using the library thread (in python) in post extraction in the term of processing time, in considering the number of posts to extract increasing, for instance superior to 80 posts in our study.

In perspectives, data extraction will be extended to more social network to have more posts, according to API layer availability to connect on social network. Through spark streaming for instance, it will be interesting to introduce more social network accessibility to the posts, not only with Twitter, but with proposed API Layer for each social network in order to strengthen the common platform as the main purpose.

The future works will also focus on content analysis of spark streaming functionalities and the study of class index-ation, to facilitate research on stored posts to highlight data characteristics.

The tuples of keywords (w_1, w_2,..., w_{n-1}, w_n) could be the parameters for the tuples of ontologies in a specific domain. That lead to others deep analysis with the map-reduce algorithm, in perspectives.

REFERENCES

[1] A. Perrot, *La visualisation d'information à l'ère du Big Data : résoudre les problèmes de scalabilité par l'abstraction multi-échelle*, thèse de doctorat, Université de Bordeaux. Bordeaux, France, 2017.

[2] J. Dean, S. Ghemawat, *Mapreduce: Simplified data processing on large clusters*, In 6th Symposium on Operating Systems Design and Implementation, USENIX Association, pages 137-150, 2004.

[3] Frédéric T.Ouédraogo, Abdoulaye Séré, Evariste Rouamba, Safiatou Soré, *Analysis of the 2015 Presidential Campaign of Burkina Faso Expressed on Facebook*,In 8th EAI International Conference on e-Infrastructure and e-Services for Developing Countries, AFRICOMM 2016. Ouagadougou, Burkina Faso, 2016

[4] Younghoon Kim, Kyuseok Shim, *Parallel top-k similarity join algorithms using MapReduce*, In IEEE 28th International Conference on Data Engineering, DOI:10.1109/ICDE.2012.87; 2012.

[5] Jeffrey Shafer, Scott Rixner, and Alan L. Cox. *The Hadoop Distributed Filesystem: Balancing Portability and Performance*, In 2010 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), DOI: 10.1109/ISPASS.2010.5452045, 2010.

[6] Moturi, Maiyo. *Use of MapReduce for Data Mining and Data Optimization on a Web*, International Journal of Computer and Applications (0975-8887) Volume 56– No.7, 2012.

[7] Apache : Apache Hadoop.http://hadoop.apache.org, 2010.

[8] Ouatik, Fahd and Erritali, Mohammed and Jourhmane, M., *Comparative study of MapReduce classification algorithms for students orientation*, Procedia Computer Science, page 1192-1197,volume 170, doi: 10.1016/j.procs.2020.03.030, 2020.

[9] Bagui, Sikha and Devulapalli, Keerthi and John, Sharon, *MapReduce Implementation of a Multinomial and Mixed Naive Bayes Classifier*,

[10] N.Alamelu Menaka, Dr.Jabasheela, *Survey on Big Data processing using Hadoop, Map Reduce*, International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O) ISSN: 2349-7009(P) Volume 1 Issue 3, 2014.

[11] Abdoulaye SERE, Dario COLAZZO, Oumarou SIE, *A Hough Transform Based On a Map-Reduce Algorithm*,International. Journal of Engineering Research and Application ISSN : 2248-9622, Vol. 6, Issue 8 (part 2), 2016.

[12] Mateus Coelho ; Dylan Sugimoto ; Gabriel Melo ; Vitor Curtis and Juliana Bezerra *A MapReduce based Approach for Circle Detection*, In Proceedings of the 14th International Conference on Software Technologies - Volume 1: ICSOFT, 454-459, Prague, Czech Republic, 2019.

[13] Yaya TRAORE, Malo SADOUANOUAN, Didier BASSOLE, Abdoulaye SERE, *Multi-Label Classification using an Ontology*, International Journal of Advanced Computer Science and Applications (IJACSA), ISSN : 2156-5570 (Online) ISSN : 2158-107X (Print) Vol. 10, No. 12, DOI : 10.14569/issn.2156-5570, 2019.

[14] Apache : http://storm.apache.org/, 2020.

[15] Apache : https://flink.apache.org/, 2020.

[16] Apache : https://spark.apache.org/streaming/, 2020.