International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Unstructured Data Processing using Spark for Topics Modelling

Adjovi Irène Sokegbe, Ayushi Nainwal

Abstract*: Information Technology domain is facing changes day by day. Furthermore, the size of data increases, as well as the demand to process them. There are two types of data: structured and unstructured data. The multiple sources and the variety of data today involve the use of "Big data" instead of data. It is related that 80% of enteUprise's data is unstructured [1]. However, the procedures to handle unstructured data are more complex than those for structured data. Thus, it becomes necessary to have a clear idea about this type of data and to know how to extract useful information from this data set. In this paper we will study how to retrieve useful information from unstructured data in E-commerce area using data analysis tools: Spark. To solve this issue, first an overview on structured and unstructured data and data analysis is provided, then information retrieval algorithm will be implemented using Spark MLlib tool in order to determine for a set of reviews, negative or positive, which subjects are more discussed by the customers. This study is needed in order to improve business based on customer satisfaction reviews. In that case, Unsupervised Machine Learning Latent Dirichlet Allocation (LDA) algorithm constitutes our model. Finally, the evaluation of the model will be given based on some parameters.

***Keywords*: Unstructured data, data analysis, Spark MLlib, LDA.**

I. INTRODUCTION

In December 2019, Amazon visitors were up to 2729 Million [3], so it generates a considerable number of reviews. It becomes necessary to study such data in order to gain knowledge. Reviews data on online platform helps improve Business products. In previous studies, it is used a classification method on reviews dataset to categorize them based on the existing categories. What if the categories are missing? The LDA algorithm solution comes on to solve this issue.The aim of this study is to search for relevant topics that were discussed in a set of Amazon reviews.

A. Background study

- *Unstructured data Vs Structured data*

When combined structured data and unstructured data, it will result in Big data concept. To understand technologies related to Big data, it must be cleared what are structured and unstructured data.Structured data is the type of data that has a pre-defined schema.

Revised Manuscript Received on June 17, 2020.

*** Correspondence Author**

Adjovi Irène SOKEGBE*, Computer science and Engineering, Alakh Prakash Goyal Shimla University, Shimla, India. E-mail: isokegbe9@gmail.com

Ayushi Nainwal, Computer science and Engineering, Alakh Prakash Goyal Shimla University, Shimla, India. E-mail: ayushinainwal@gmail.com

Unlike structured data, the unstructured data gathers the type of data whose format or model is not defined.
The estimation of unstructured data in the world is around 80% [4]. Due to the growth of unstructured data, it becomes primarily to develop an intelligent solution to handle them as they cannot be processed using traditional tools like: Relational Database, Data warehouse, OLAP.
The unstructured data are those data which come from sensors, satellite, blogs, emails, social media, etc. They are raw data with no structure. The techniques or technologies used to process and store (Data Lake) them are also different from structured data one.

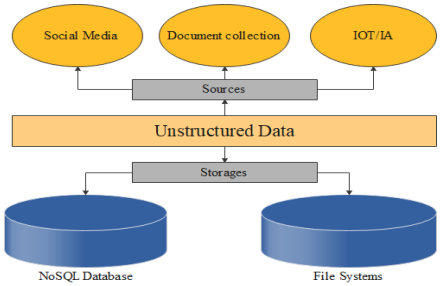

Figure 1: Unstructured Data overview

- *Data analysis*

The term analysis comes from Greek that means 'breaking-up or releasing' [5]. So, data analysis can be defined as a breaking up of data.
Clearly, data analysis is the process of collecting, cleaning, transforming and modelling data in order to get insights of data according to the requirements. There are five (5) types of data analysis that are: Text analysis, Statistical analysis, Predictive analysis, Prescriptive analysis, Diagnostic analysis.

- *Data analysis techniques and tools*

Data analysis techniques are a set of operations, methods, tools that are required for the analysis of data. There are three (3) techniques for analysing data: Data Mining techniques, Business Intelligence (BI) analytics tools and Artificial Intelligence (AI). Data mining approaches are used to retrieve knowledge from a huge amount of data. It is a process that allow to extract patterns from unstructured data. Business Intelligence is a set of mechanisms that enhances the business. The approach results in a set of techniques, tools, architectures that provides results on dashboard, chart, etc. for data visualization. It is mostly applied on structured data. Artificial Intelligence for data analysis refers to the computation of complexes algorithms that makes machines to think like a human being or to perform tasks that require human being intelligence.

Machine learning, Natural Language processing (NLP), Neural Network (NN) are used to implement data analysis solutions.

There are several data analysis tools, commercial and free tools. The big challenge is to find out which one is the best for the requirements. To know that, the kind of data that will processed must be known and the characteristics of such data as well as the requirements must be clearly defined.

As commercial tools there are: Tableau Public, RapidMiner, Microsoft Excel etc.

As free tools there are: Apache Spark, Apache Storm, Apache Hadoop, R, Python, etc.

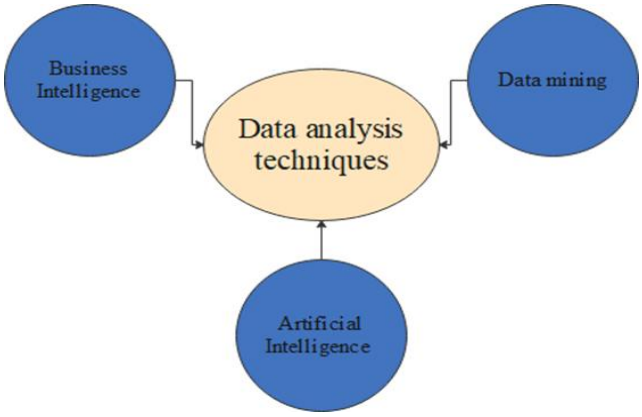

Figure 2: Data analysis techniques

B. Problem statement

The problem to be solved is to determine the parameters or factors that can influence the customer satisfaction rate and the customer feedback.

C. Literature review

Various papers attempt to study unstructured data and explore various techniques to handle them from data mining techniques to Machine learning techniques. During our study, many researchers have used different tools to solve their problems. A. MadhaviLatha [5] in 2016 compares Spark algorithms to be 100 times faster than Map Reduce. K. Aziz, D. Zaidouni and M. Bellafkih [6] in 2018 conclude that, in the case of structured data MapReduce is the suited one. But for analysing unstructured data Spark is the suited one. Spark support real time batch and streaming processing. Lovenika Kushwahain in 2016 [7], "opinion Mining of Customer Reviews based on their Score using Machine Learning Techniques" proposed a solution for the extraction of knowledge from online shopping customer reviews. This solution aimed to provide a recommendation list based on review of the customer. Then a comparison of Naïve Bayes classifier and AdaBoost classifier was provided. The proposed solution implements AdaBoost classifier, and it results in better accuracy and better time execution than Naïve Bayes. As shown above, there are many choices of tools and algorithms that can solve a particular problem. Every researcher may be chosen [8] Hadoop instead of Spark and vice versa, but to provide better insights by the analysis of online reviews using LDA in Pyspark are quietly unknown. The use of Spark and Hadoop promises to give an effective result for the processing of huge amount of data.

II. METHODOLOGY

A. Approach

one major approach used to solve our problem is : relevant topic retrieval along with negative reviews filtration. The solution will be implemented using unsupervised machine learning algorithm "Latent Dirichlet Allocation, LDA".

B. Processing methodology

The unstructured data processing involves the consideration of defining the steps in the problem solving. Steps involving in the algorithm model are:

- *Data collection*

In the first step, the dataset of our building model is Amazon-fashion reviews [9] from year 2014. It counts 883k reviews and the raw data has given in JSON file. The data are about: overall, verified, review time, reviewerID, asin, reviewerName, reviewerText, summary, rating and UnixReviewTime. Our model will focus on reviewerID, reviewText and rating.

- *Data pre-processing*

this step is about the pre-processing of data to fit the inputs required for our model. It results in: Tokenization, null value remover, stop Word treatment and vectorization.

- *Data Processing*

Once the data is collected then prepared, the processing model based on LDA will be built.

- *Data visualization and the model evaluation*

Once the data are processed, the result will be shown in Pycharm output window and in graphs.

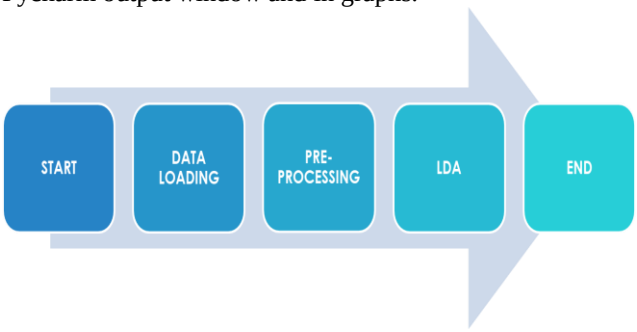

Figure 3: Methodology

C. The proposed solution

LDA whitch stands for "Latent Dirichlet Allocation" is based on Drichlet distribution studied by Peter Gustav Lejeune Dirichlet. The LDA algorithm was developed in machine learning by David Blei, Andrew Ng and Michael I. Jordan in 2003[11].

LDA is an algorithm of Unsupervised Machine Learning for topic modelling.

The LDA model built in this study is described with the below diagram:

The model built has four (4) stages:

• First stage: Data Loading which consist to load the dataset into data frame in the coding environment,

• Second stage: the polarity searching has permitted to separate the dataset that is a set of reviews into negative data frame,

• Third stage: the pre-processing of the data is going here, it is a set of tokenization, stopword remover, null value treatment and vectorization.
• Fourth stage: LDA algorithm was implemented along with the likelihood and perplexity parameters. Those parameters will help further to evaluate the model.

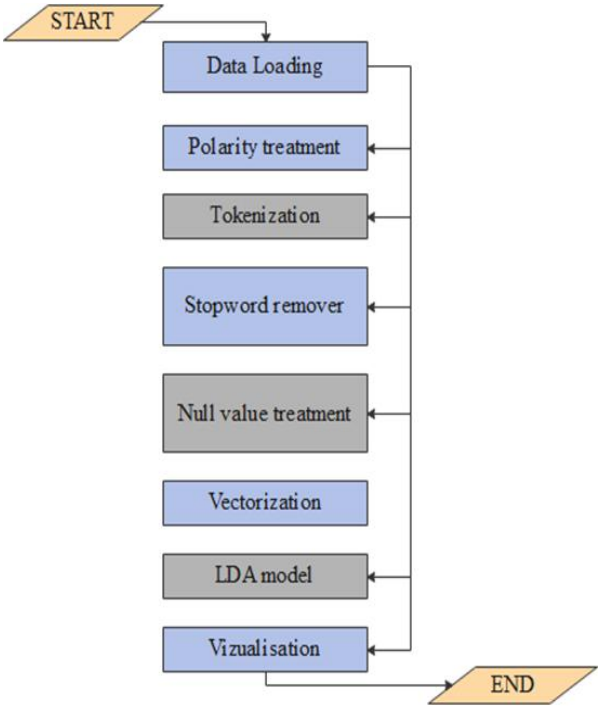

Figure 4: Topic distribution

D. Result and discussion

The model after the stage of pre-processing comes in this result: five (5) rows are showing below because of the size of the document.

reviewText	words	tokens	filtered	stpW	features
There are two hol...	[there, are, two,...	9	[two, holes, wash...	5	(4314,[4,26,253,3...
These shoes look ...	[these, shoes, lo...	27	[shoes, look, gre...	12	(4314,[13,41,50,7...
In Amazon's adver...	[in, amazon, s, a...	65	[amazon, advertis...	37	(4314,[0,13,14,17...
The costume is ok...	[the, costume, is...	327	[costume, ok, bes...	145	(4314,[3,5,6,12,1...
The reviews said ...	[the, reviews, sa...	87	[reviews, said, r...	38	(4314,[10,11,12,2...

only showing top 5 rows

Figure 5: pre-processing result

In the figure (5), it is observed that the number of words contains in each array of the column "reviewText" decreases after the end of the pre-processing stage.
"tokens" counts the number of words of each array of the column "words" into the column "tokens". After transforming each row of sentences in "reviewText" into tokens refers column "words" on which the stopword remover was applied, the column "stpW" counts the number of words in "filtered". The last column is showing the arrays of words transformed into vectors.

filtered	features	topicDistribution
[two, holes, wash...	(4314,[4,27,260,3...	[0.95252587705514...
[shoes, look, gre...	(4314,[13,41,49,7...	[0.97787537282972...
[amazon, advertis...	(4314,[0,13,14,17...	[0.70399323820409...
[costume, ok, bes...	(4314,[3,5,6,12,1...	[0.99801493113809...
[reviews, said, r...	(4314,[10,11,12,2...	[0.99258396089502...

Figure 6: Topic distribution

The figure (6) is showing the distribution of topics over the set of reviews. As the figure above describes it, four (4) subjects have been deducted from customer review as well as the indices corresponding to the terms in the corpus, their weight and their corresponding terms.
The results below, figure (7) was obtained by choosing four (4) topics and four (4) words to describe each topic. This permit to have:
Lower Likelihood = -203324.64874793414
Upper Perplexity = 7.625150899978779

topic	termIndices	termWeights	topics_words
0	[0, 1, 2, 7]	[0.011429475619295255, 0.009851348176123466, 0.007713931717284412, 0.006842615887982506]	[like, one, picture, quality]
1	[686, 642, 539, 466]	[0.011336234277271822, 0.011211212132935838, 0.008814212530026471, 0.008302410485139515]	[wax, milk, pendant, breast]
2	[86, 72, 228, 34]	[0.011375076792560672, 0.011118578040784812, 0.009394038445541474, 0.007266244265151885]	[days, shoes, sole, return]
3	[250, 670, 857, 427]	[0.005716037149828402, 0.005597642236161679, 0.004320736784589461, 0.004054727997314912]	[socks, stretcher, footwear, model]
4	[561, 355, 1297, 693]	[0.005030711738510333, 0.004273924962154505, 0.003971610768811158, 0.003801420682874639]	[piercing, jewelry, bar, wrist]

Figure 7: topics retrieved

To evaluate the model, we will variate the number of topics to observe what will be the behaviour of the likelihood and perplexity parameters. As shown below, with the increase in number of topics the likelihood decreases and the perplexity increases. The model with higher likelihood and lower perplexity is considering being good.

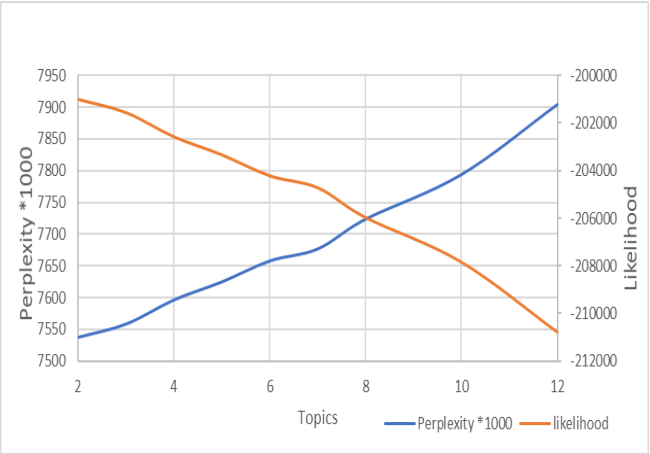

Figure 7: graph of evaluation

The figure (7) shows the behavior of the variation of the topic's numbers over the "likelihood" and "perplexity" parameter. To solve the problem defined in this paper, we have using the free data analysis tool Spark.

It matches well for the type of data that are processed. Spark execute the algorithm in few second with complete result.

III. CONCLUSION END FUTURE SCOPE

Unstructured data processing is a crucial methodology being used in current technology due to its simplicity in implementing the solutions and giving high performance.
In the light of everything that has been done, it is concluded that using Spark in the case of unstructured data analysis is very determinant in problem solving. As a proposed solution the unsupervised method LDA with the review's polarity was developed. Such method gives a satisfying result with a low execution time. Further, various data pre-processing methods can be added for more relevant results when the dataset becomes very important in terms of volume. Also, Some NLP algorithm such as NER can be implementing to get insight in topics interpretation.

REFERENCES

1. Schneider, C., 2020. The Biggest Data Challenges That You Might Not Even Know You Have - Watson. [online] Watson. Available at: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.
2. Schneider, C., 2020. The Biggest Data Challenges That You Might Not Even Know You Have - Watson. [online] Watson. Available at: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.
3. Statista. 2020. Web Visitor Traffic To Amazon.Com 2020 | Statista. [online] Available at: https://www.statista.com/statistics/623566/web-visits-to-amazoncom/
4. TechRepublic. 2020. Unstructured Data: A Cheat Sheet. [online] Available at: https://www.techrepublic.com/article/unstructured-data-the-smart-persons-guide/.
5. Etymonline.com. 2020. Analysis | Origin And Meaning Of Analysis By Online Etymology Dictionary. [online] Available at: https://www.etymonline.com/word/analysis.
6. Madhavilatha and G. V. Kumar, "Streaming Data Analysis using Apache Cassandra and Zeppelin," vol. 3, no. 10, pp. 8–15, 2016.
7. L. Kushwaha and S. D. Rathod, "Opinion Mining of Customer Reviews based on their Score using Machine Learning Techniques," pp. 2198–2203, 2016.
8. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Unstructured data analysis on big data using map reduce," Procedia Comput. Sci., vol. 50, pp. 456–465, 2015.
9. McAuley, J., 2020. Amazon Review Data. [online] Jmcauley.ucsd.edu. Available at: http://jmcauley.ucsd.edu/data/amazon/.
10. "MLlib: Main Guide - Spark 2.4.5 Documentation", Spark.apache.org, 2020. [Online]. Available: https://spark.apache.org/docs/latest/ml-guide.html..
11. "10.1162/jmlr.2003.3.4-5.993", CrossRef Listing of Deleted DOIs, vol. 1, 2000. Available: 10.1162/jmlr.2003.3.4-5.993.

AUTHORS PROFILE

Adjovi Irène SOKEGBE is completing her master studies in Computer Science Engineering at APG Shimla University after obtaining a professionnal Bachelor in maintenance and computer network at the University of Lomé (TOGO). She is deeply interrested in the field of the Data Analysis and Information System. Previous works:Technology, crimes and its changing patterns (http://www.jetir.org/papers/JETIR1907V74.pdf) Process Unstructured Data using Data analysis: Apache Spark and Hadoop(http://www.i2or-ijrece.com/vol.-8-issue-1.html)

Ayushi Nainwal is an Assistant Professor in the Department of Computer Science at Alakh Prakash Goyal Shimla University, where she has been since 2017. She received her bachelor degree in Information Technology from Baddi University in 2013, and her masters in Computer Science and Technology from Himachal Pradesh University in 2016. She qualified her State Eligibility Test in 2019. Her research interests span both computer networking and network science. In the networking area, she has worked on identifying node categorization algorithm in Wireless Sensor network. She has around 6 publications and has also guided student under her in Masters of Technology in Computer Science and Engineering.