Economic Insights – Trends and Challenges	Vol.VIII(LXXI) No. 3/2019	33 - 43

Using Machine Learning Techniques to Gain Insights from Enterprise Unstructured Data

Cristian Bucur

Faculty of Economic Sciences, Petroleum-Gas University of Ploieşti, Bd. Bucureşti 39, 100680, Ploieşti, Romania
e-mail: cristian.bucur@upg-ploiesti.ro

Abstract

This article addresses the new IT challenges of business environment in context of Big data. As digitalization extended in all areas and aspects of business processes companies gather huge quantities of data. Business environment realize now that data is an important asset of organization but as studies show most of this data is unstructured so gaining insights is a complex and tedious process. There is now a shift in the way data is processed as new technologies become more accessible to companies. The article presents a practical problem related to document understanding and propose a solution concept and methodology for solving it. Using machine learning and natural language processing has now become a must in dealing with unstructured data for companies that need a competitive advantage.

Keywords: *big data; machine learning; natural language processing; unstructured data; information retrieval; information extraction; document understanding; named entity recognition; pos tagging*

JEL Classification: *A12; C63; C88; M15*

Introduction

Nowadays digitalization and informatization is the normality in all business environments. Companies now record data just about anything related to their activities. As interactions and business flows become more and more complex the volume of data produced increase exponentially. Big Data has become the reality the IT department of companies deal with.

In this context companies invested in systems for storing and managing this increasing volume and variety of data. But now as they are aware of the data the focus has changed on using more efficiently Big data, especially as now appeared new ways of doing this.

We are living now in the era of data. If in earlier decades some of the main disruptors in business environment were technical evolutions, introduction of computers and evolution of internet now the shift is driven by introduction on large scale of artificial intelligence, machine learning and natural language processing-based solutions. Companies can now gain insights more accurate and more precise from multiple sources of data. The ones who adopt faster these technologies allowing them to improve their business decisions will become the leaders (Hack, 2014).

Big Data captures data regarding all aspects of a company, regarding internal processes, business flows, customers, partners, suppliers, employers, sales, market, internal and external communications, and becomes one of the strategic assets for the organization. But without a way to utilize the potential of that data this asset is left unexploited. This is where machine learning and advanced analytics technologies come into aid of business environment helping them on leveraging insights from all kind of data (Bucur, 2015).

Machine Learning (Evolution of machine learning: https://www.sas.com/en_us/insights/analytics/machine-learning.html) (ML) (https://en.wikipedia.org/wiki/Machine_learning) is computer science domain, a branch of artificial intelligence, dealing with algorithms and statistical methods to find patterns and make predictions based on data, with minimal human intervention. ML is more efficient on large volumes of variate and fast changing data, than other older methods. It helps companies to create accurate models to predict future actions and identify unseen patterns in their activity. ML systems adopted in businesses are changing focus from studying reports about past activity to future based analysis.

Natural language processing (NLP) (Natural Language Processing (NLP) for Machine Learning https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b) is a machine learning domain dealing with the ability of machines to analyze, process and understand human language. System based on NLP technologies could also help companies acquire more insights from their documents.

We now reached the state where these advanced states of the art technologies are no longer attribute of few scientists, they become accessible to standard business users. This dramatically changes the way information is analyzed in organizations.

Unstructured Data and Document Understanding

We can divide company data into three groups:

- structured,
- semi-structured
- unstructured.

Structured data is data easily accessible by machines in a well-organized format usually as database, structured as rows and columns. This data is easy accessed and used in analysis. Examples of structured data are: numbers, dates, strings, database records.

Semi structured data is referring to non-standard data, without structure that has some form of annotation or semantic tags. (Azati Software, 2019)

Unstructured data has no standard format, so it cannot be easily processed by machines, but represent almost all data that usually human operate with. Examples of unstructured data are: emails, articles, business documents, images, videos, spreadsheets, word processing documents, portable document formats.

Company legacy systems dealt only with structured data, now in context of increased volumes of data there is a shift toward managing large unstructured data volumes.

In the table below we present some key differentiators between structured and unstructured data. Unstructured data is much larger and requires more storage than structured data (Chiang, 2018). According to Gartner (https://www.gartner.com/en), unstructured data represents most of a company data, approx. 80%.

Also unstructured data has a steeper increasing rate that structured data. According to IDC (https://www.idc.com/), unstructured data has an annually growing rate of 26.8%. The growing rate of structured data is only 19.6%. IDC also predicts that by 2025 almost 80% of worldwide data will be unstructured.

Sources for unstructured data have also increased in latest years. Unstructured data existed for many years in companies as emails, reports, documents produced by employees, but now is also generated by machines used in production process, by software used in business processes, modelling, monitoring, etc. Now there are also external organization sources for unstructured data, like web marketing data, social media, streaming data from devices and sensors (IoT). All this sources contribute to create critical mass of data that become challenging for organization from processing and analyzing point of view as storage solution was partially solved by Big Data systems (King, 2019).

Table 1. Comparation Structured vs. Unstructured Data

Structured Data	Unstructured data
Can be formatted in rows, columns and relational databases	Cannot be formatted in rows, columns and relational databases
Numbers, dates, strings	Images, audio, video, doc, e-mails, xls, pdf
Approx. 20% of companies' data (Gartner)	Estimated as 80% of companies' data (Gartner)
Requires less storage	Requires more storage
Easy to manage and secure with legacy solutions	More difficult to manage and protect with legacy solutions

Source: (Chiang, 2018)

More companies realize what important asset their Big Data systems represent and after implementing storage solutions they focus on gaining insights from data. But as volume and sources of unstructured data is complex, processing of data poses challenges.

In the figure below we see a Gartner study (Heudecker, 2016) regarding the sources of Big Data existing in companies and the percent of companies that use or plan to use that source of unstructured data. With blue is represented the percent of companies using the unstructured data source and with gray the percent of companies that intend to use that source.

We can observe that companies acknowledge the benefits of analyzing unstructured data and even if currently only about 30% of unstructured data is analyzed more are planning in future to address this problem.

In the last years machine learning has become accessible for business environment and become a key factor for evolution. One of the capabilities enabled by these ML systems is Document understanding. Document understanding is combining machine learning and natural language processing techniques to gain insights from unstructured human generated text.

This technology would play an important role in a company as it becoming capable to extract more knowledge from operational unstructured data. As Gartner predicts, document understanding technology would produce another shift in companies as unstructured data analysis moves from reactive searches to proactive insights generation. By 2022 reactive searching would be reduced by 20% (Khan, 2019).

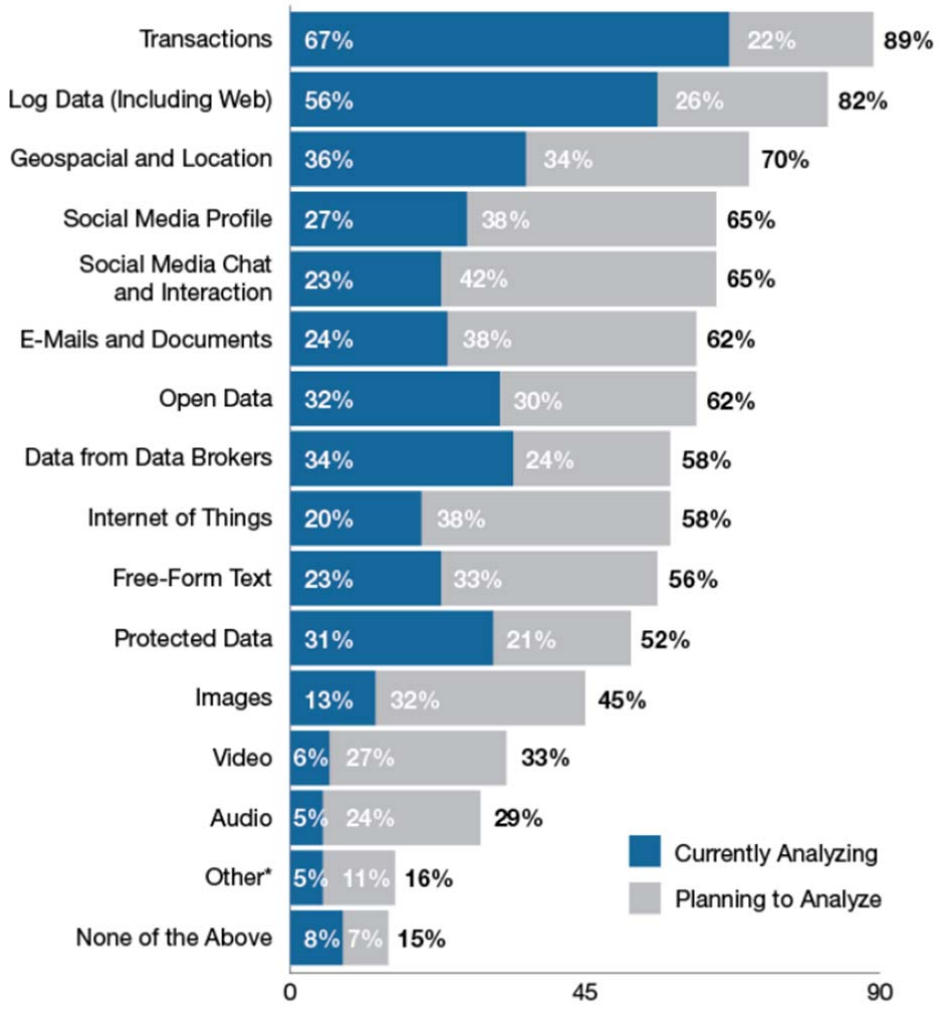

Fig. 1. Data sources for Big Data Analysis and intention of analyzing. Percentage of respondents
Source: https://www.gartner.com/smarterwithgartner/digital-demands-cfos-rethink-how-to-deliver-value/

Extracting knowledges from unstructured data using legacy systems was difficult as it involved complex and long processes. As 80% percent of a company data is unstructured, addressing this issue, and finding efficient methods become more important. Evolution of modern system using machine learning and natural language processing provides companies precious insights and competitive advantages.

Development of document understanding systems could be useful in multiple domains as:

- legal departments – computer analyzing legal documents could easily detect risk clauses or may suggest mitigation of liabilities;
- company correspondence rerouting, government agencies – by analyzing the mails a system could route relevant correspondence to right department, eliminating the effort of human sorting;
- recruiting – matching cv to job posting, detecting patterns for hiring skills;
- banks, finance – cross-analyzing profile of clients with their loans;

- content creation, recommendation system – automatic identification of main theme of analyzed article and finding similar sources;
- storage optimization – using automated business rules determine the management of documents and storage handling and location, detecting duplicate or redundant data (Khan, 2019).

Study of a Practical Scenario

Given the above context, we proposed to apply some of the modern algorithms used in document understanding systems, to a specific study. We started from a concrete problem in a large corporate activating in multiple domains. They operate on a large number of assets managed by multiple entities each with specific operating substructures and procedures. For each asset (there are hundreds of assets and several tens of entities involved) there are a number of working groups (usually between 3 to 5) that participate to multiple meetings (between 6 to 12 meetings per group). Each meeting deals with dozens of documents necessary for group activity and also from each meeting decisions results another set of documents.

The setup described above has a complex process for administrative and business operations. Those processes involve a large number of meetings that require large number of documents. A common scenario would be a number between 10.000 and 40.000 meetings that would produce between 50.000 to 1.000.000 documents per each year. This represent a large amount of unstructured data in non-governed sources. In current condition there is no coordination between sources and no guideline for storage or management procedures.

The system should index the documents and classify them into several classes associated to working group they are related and should allow retrieving documents based on searching criteria.

The data we used in study is represented by approx. 4000 documents collected from multiple meetings related to 3 working groups. We used word processing documents, pdfs, 3d files, jpg, gif, files related to cad software.

Proposed Solution

We propose a flow for processing the documents and extract relevant information, which is described in Figure 2 below.

By processing the unstructured dataset with the proposed flow, we want to be able to associate metadata of each document with a tag structure. This would allow us to classify documents, relate them to assets they describe and to categorize them into working groups.

By starting with a set of already associated tags based on automated extracted metadata from documents and by manual tagging of a part of data we want to create an initial training set. Using a ML engine and starting from initial set we want to be able to classify the rest of unrelated documents.

Usually document related to a specific working group share some common knowledge, or common terms so by processing the content with a ML algorithm we hope to found the patterns or information related to classes they fit into.

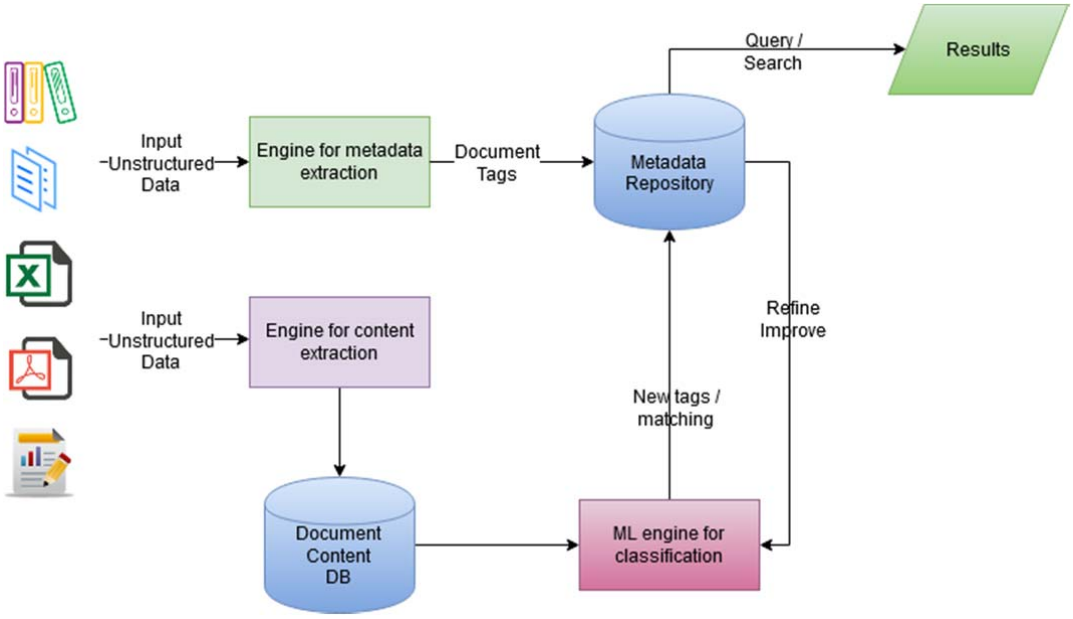

Fig. 2. Flow diagram for document processing and metadata generation

The tag structure we proposed is presented in table below:

Table 2. Proposed tag structure for documents

Entity	Attributes
Time	Year; Quarter; Month; Week; Day; Time
Location	Country; State; County; City; Office Building
Process	Management System Process; Function; Meeting
Document	Document Group; Document Type; Owner; Author; Publisher; Creator; Approver; Revision; Coding;
Source	Folder; Fileserver
Organization	Company; Discipline; Role; Person; External; Partners; Authorities; Vendors; Associations
File	Path; Filename; Filetype; File size; Character count; Page count

The documents are first processed with an engine to automate extraction of metadata from files. For metadata extraction we used Apache Tika (https://tika.apache.org/) (Apache, 2019). It is a tool that allows extraction of metadata from text in different file formats (Li, 2019). The resulted metadata from processing a document with Apache Tika is presented in JSON format:

```
"metadata": {
    "date": "2015-03-06T14:55:00Z",
    "Author": "Martin Scott",
    "creator": "Martin Scott",
    "Template": "33877036",
    "modified": "2015-03-06T14:55:00Z",
    "publisher": "Company X Ltd",
    "Line-Count": "739",
    "Page-Count": "38",
    "Word-Count": "15562",
    "dc:creator": "Person 1",
    "Last-Author": " Person 2",
    "X-Parsed-By":
"org.apache.tika.parser.DefaultParser",
    "cp:revision": "2",
    "meta:author": " Person 1",
    "Content-Type":
"application/vnd.openxmlformats-
officedocument.wordprocessingml.docume
nt",
    "Last-Printed": "2015-02-
17T10:23:00Z",
    "dc:publisher": "Company X Ltd",
    "Creation-Date": "2015-03-
06T14:55:00Z",
    "Last-Modified": "2015-03-
06T14:55:00Z",
    "xmpTPg:NPages": "38",
    "Last-Save-Date": "2015-03-
06T14:55:00Z",
    "meta:save-date": "2015-03-
06T14:55:00Z",
    "Character Count": "88706",
    "Paragraph-Count": "208",
    "Revision-Number": "2",
    "dcterms:created": "2015-03-
06T14:55:00Z",
    "meta:line-count": "739",
    "meta:page-count": "38",
    "meta:print-date": "2015-02-
17T10:23:00Z",
    "meta:word-count": "15562",
    "Application-Name": "Microsoft Office
Word",
    "dcterms:modified": "2015-03-
06T14:55:00Z",
    "meta:last-author": " Person 2",
    "meta:creation-date": "2015-03-
06T14:55:00Z",
    "Application-Version": "14.0000",
    "meta:character-count": "88706",
    "meta:paragraph-count": "208",
    "Character-Count-With-Spaces":
"104060",
    "extended-properties:Company": "
Company X Ltd",
    "extended-properties:Template":
"33877036",
    "extended-properties:AppVersion":
"14.0000",
    "extended-properties:Application":
"Microsoft Office Word",
    "meta:character-count-with-spaces":
"104060"
  }
```


The metadata extracted from unstructured data is stored into a metadata repository database. Later we want to improve the metadata from the repository with the Machine Learning engine.

Apart from metadata extracted from unstructured data with Apache Tika we also process the content of documents. The content is stored in a separate database. This is the content on which we apply machine learning and natural language processing algorithms.

By using described flow, we obtain a sql database with full text content from 3643 documents. Of these we have the following classification of data:

- Working group A: 639 documents;
- Working group B: 1050 documents;
- Working group C: 147 documents;
- Not classified: 1807 documents.

Named-Entity Recognition

Named-entity recognition (NER), known as entity identification, entity chunking or entity extraction is an NLP domain related to extraction of information from text documents. It deals with locating and classifying named entities in predefined classes as person names, organizations, locations, expression of times, quantities, monetary values, percentages (Gupta, 2018).

To use NER technology on our data we need to implement a standard flow for information extraction system (see below figure) using NLTK (https://www.nltk.org/book/ch07.html) (Natural Language Toolkit) library in Python. NLTK library has implemented various tools for text analysis and part of speech recognition.

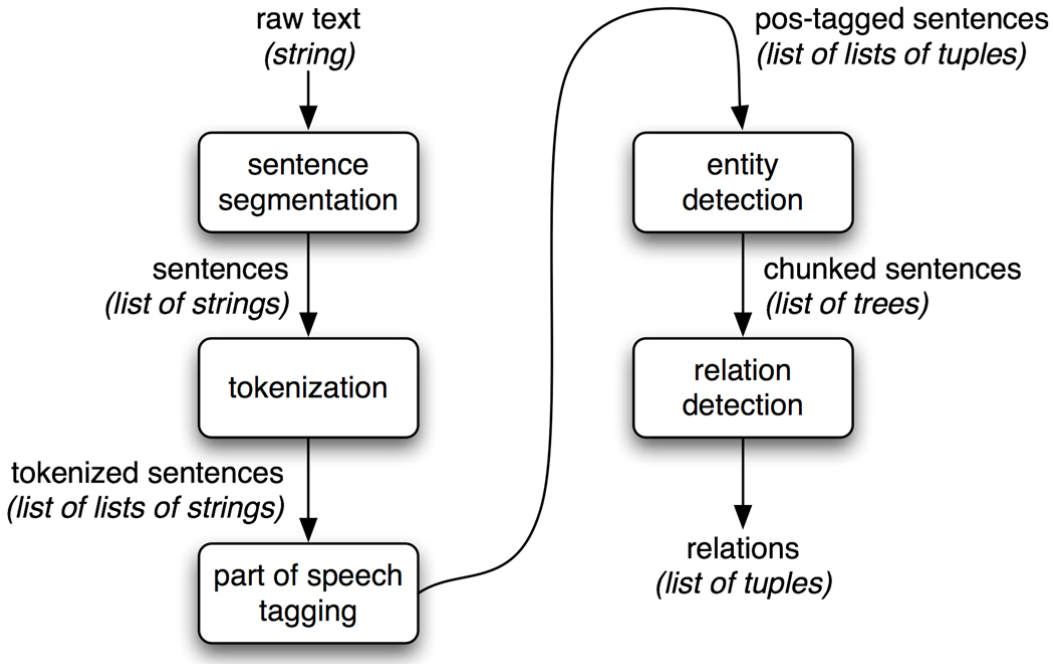

Fig. 3. Flow for an Information Extraction System

Source: Bird, Klein and Loper, 2019, https://www.nltk.org/book/ch07.html

An important part of this process is part-of-speech tagging (POS tagging) also named grammatical tagging or word category disambiguation (Bucur, 2014). This process matches a word from a text with corresponding grammatical part of speech based on context (see Table 3). This process is in our case using probabilistic tagging. NLTK tagging process is classifier based, trained on PENN Treebank corpus (Bird, Klein and Loper, 2019).

Table 3. Parts of speech

Part of Speech Tag	Signification
NN	Noun, singular or mass
CC	Coordinating conjunction
VBG	Verb, gerund or present participle
JJ	Adjective
NNS	Noun, plural
NNP	Proper noun, singular
DT	Determiner
CD	Cardinal number

For identifying in our context, the entities (NER process), we would extract proper names NNP from POS tagged text. Below we present in Figure 4 an example or POS tagging process.

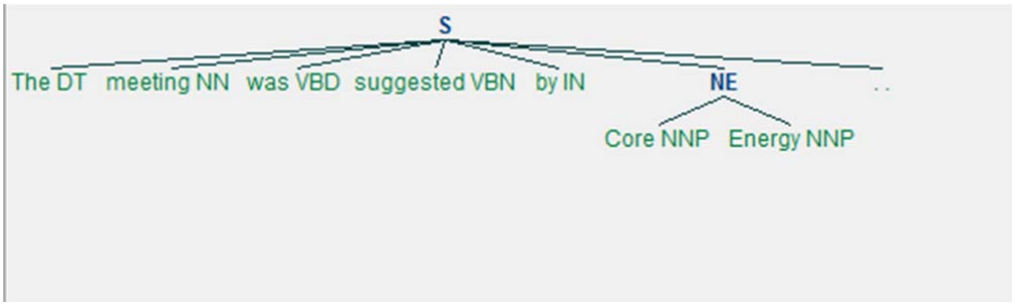

Fig. 4. POS tagging example on a sentence from our extracted full text content

Using this methodology, we can extract for each document the following entities:

Label	Signification
NE	Named Entity
GPE	Geo-Political Entity
LOCATION	
PERSON	
ORGANISATION	
FACILITY	

We applied the above method on documents. The results are provided below (for privacy purpose all the names and entities in entire paper are replaced by some placeholders):

Table 4. Results for doc id 18

TYPE	NAMES
PERSON	Person1
PERSON	Person2
ORGANIZATION	Org1
ORGANIZATION	Org2
PERSON	Person3
PERSON	Person4
ORGANIZATION	Production Operations
ORGANIZATION	Operations Center Manager
PERSON	Person5
ORGANIZATION	Org3
PERSON	Person6
PERSON	Person7
PERSON	Person8
GPE	Org2
ORGANIZATION	Org4
PERSON	Technology Manager1
PERSON	Superintendant1
GPE	Country1
GPE	Corporate 1
ORGANIZATION	Org5

Conclusions

Machine learning and Natural language processing technologies become more and more daily processes in system implemented today in business environment. Using this kind of technologies, that now are becoming accessible to standard business users, not to a restricted group of scientists, offers companies that adapt quickly to new technologies a competitive advantage.

In this paper we propose a methodology to apply ML and NLP to a specific practical problem found in many corporations. They generate a large number of unstructured data, and even if they have system for storing, they cannot access easily the data or extract knowledge from. By implementing this document understanding system they could have a primary step toward gaining insights from unstructured data.

The proposed solution represents only a primary step toward full unstructured data processing, but is easy to implement and offer good results. The processing should continue with more NLP algorithms to identify more features from documents, and correlate them already classified ones in categories. Based on this training data we could implement advanced ML algorithms to automatically classify the entire unstructured data.

References

1. Apache, 2019. *Apache Tika - a content analysis toolkit.* [Online]. Available at: https://tika.apache.org/ [Accessed Sept 2019].
2. Azati Software, 2019. *Unstructured Data Analysis with Machine Learning.* [Online]. Available at: https://azati.ai/unstructured-data-analysis-with-machine-learning/ [Accessed Sept 2019].
3. Bird, S. Klein, E., Loper., E., 2019. *Natural Language Processing with Python.* s.l.:O'Reilly Media.
4. Bucur, C., 2014. Opinion Mining platform for Intelligence in business. *Economic Insights-Trends and Challenges,* 3(3), pp. 99-108.
5. Bucur, C., 2015. *Using big data for intelligent businesses.* Brasov, Proceedings of the Scientific Conference AFASES.
6. Chiang, C., 2018. *Defining the Terms: Structured Data vs. Unstructured Data.* [Online]. Available at: https://www.igneous.io/blog/structured-data-vs-unstructured-data. [Accessed Sept 2019].
7. Gupta, M., 2018. *A Review of Named Entity Recognition (NER) Using Automatic Summarization of Resumes.* [Online]. Available at: https://towardsdatascience.com/a-review-of-named-entity-recognition-ner-using-automatic-summarization-of-resumes-5248a75de175 [Accessed Sept 2019].
8. Hack, M., 2014. *Use Data to Tell the Future: Understanding Machine Learning.* [Online]. Available at: https://www.wired.com/insights/2014/03/use-data-tell-future-understanding-machine-learning/ [Accessed Sept 2019].
9. Heudecker, N., 2016. *Best Practices for Designing Your Data Lake.* [Online].Available at: https://www.gartner.com/en/documents/3483017/best-practices-for-designing-your-data-lake [Accessed Sept 2019].
10. Khan, K., 2019. *Unlocking value from unstructured data.* [Online] Available at: https://www.accenture.com/ro-en/insights/digital/unlocking-value-unstructured-data [Accessed Sept 2019].
11. King, T., 2019. *80 Percent of Your Data Will Be Unstructured in Five Years.* [Online]. Available at: https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/ [Accessed Sept 2019].
12. Li, S., 2019. *Apache Tika: What is it and why should I use it?.* [Online] Available at: https://medium.com/@simonli_18826/apache-tika-what-is-it-and-why-should-i-use-it-f4d74d7350b6 [Accessed Sept 2019].