Machine learning models and methods aspects of processing unstructured data

Oleksandr Bryk[1,†], Ivan Mudryk[1,*,†], Mykhailo Holubovskyi[1,†] and Yurii Stoianov[1,†]

[1] *Ternopil Ivan Puluj National Technical University, 56 Ruska str., Ternopil 46001, Ukraine*

Abstract

The ever-increasing amount of unstructured data, including text, images, audio, and video, poses a serious challenge to traditional data mining techniques. Machine learning (ML) offers powerful tools and techniques to unlock the valuable insights hidden in this vast amount of information. This article explores the role of machine learning models and methods in processing unstructured data.

We delve into key aspects of unstructured data processing, including data cleaning, feature development, and model selection. We explore specific ML techniques developed for different types of data, such as natural language processing (NLP) for text analysis and computer vision for image recognition. The paper also discusses the challenges and considerations involved in building and deploying ML models to handle unstructured data.

By understanding the capabilities of ML on unstructured data, organizations can gain a competitive advantage by deriving valuable insights for various applications. This information can range from understanding customer sentiment in social media posts to detecting anomalies in sensor data for predictive maintenance.

Keywords

Machine Learning, unstructured data, text analysis, image recognition, natural language processing (NLP), computer vision, feature engineering, model selection, predictive maintenance

1. Introduction

Unstructured data has become an important source of information in today's world. They can be obtained from various sources such as social media, web pages, touch devices, medical records, and many others. Information in an unstructured form can be extremely valuable, but processing and analyzing it can be difficult due to a lack of clear organization and format.

Unstructured data is data that does not have a clear structure or format, such as text, images, audio, and video. Their processing requires significant computing resources and using of advanced methods in comparison to structured data, organized in the form of rows and columns within a database.

However, thanks to the development of artificial intelligence, many methods and tools for processing unstructured data have appeared. These techniques allow computers to extract

[1]*BAIT'2024: The 1st International Workshop on "Bioinformatics and applied information technologies", October 02-04, 2024, Zboriv, Ukraine*

* Corresponding author.

† These authors contributed equally.

alex.lenberg@gmail.com (O. Bryk); i1mudryk@ukr.net (I. Mudryk); m.holubovskyi@gmail.com (M. Holubovskyi); yuriy556s@gmail.com (Y. Stoianov)

0009-0005-6564-1102 (O. Bryk); 0000-0002-4305-1911 (I. Mudryk); 0009-0003-9479-8454 (M. Holubovskyi); 0000-0003-1848-2258 (Y. Stoianov)

knowledge and useful information from this data, which can be useful for a variety of tasks, such as:

Sentiment analysis: Determining the overall mood of a text, such as positive, negative, or neutral.

Text Classification: Assigning categories to text, such as the topic of a news article or the type of email.

Information Mining: Identifying and extracting key information from text, such as people's names, dates, or places.

Object recognition: Detect and identify objects in images, such as people, cars or animals.

Speech Recognition: Converting spoken language to text.

Machine translation: Translation of text from one language to another.

IDC estimates that by 2025, more than 80% of business information will consist of unstructured data.

Unstructured data processing unlocks a treasure trove of insights across various industries: analyze social media posts and reviews to understand customer sentiment towards a brand or product; fraud detection to identify suspicious patterns in financial transactions to prevent fraud, in medical diagnosis: analyze medical images like X-rays and CT scans to aid in diagnosis; content recommendation of user behavior and preferences to recommend relevant content on streaming platforms [3].

2. Challenges and Considerations in solutions for processing unstructured data

Unstructured data is vast and comes in many formats. Processing requires robust systems and tools to handle the load and heterogeneity. Unstructured data can be noisy, with errors, inconsistencies, and missing information. Techniques like data cleaning and normalization are crucial for data Quality. As data volume grows, processing needs to scale efficiently. Cloud-based solutions and distributed processing frameworks are often used for scalability.

The rigid structure of traditional data storage options can exacerbate the problem of its pre-defined structure may lack the flexibility and adaptability required for unstructured data. Due to the nature of unstructured data, their processing requires significant computing power and large amounts of storage. This means that operating software that works with such data requires complex IT infrastructure. Such infrastructure can consist of various components to provide a repository for original unprocessed data, relational and non-relational databases to store processing results, artifacts storage, and an environment for running applications with many components and various technologies.

2.1. Application of cloud computing

Building an infrastructure for complex applications that work with unstructured data from scratch requires significant resources and time. To solve these problems, it is advisable to use cloud computing. Instead of maintaining physical computing resources cloud computing allows access to computing services, storage, and databases on an as-needed basis [1]. Thus, the cloud service provider is responsible for maintaining the physical infrastructure. An infrastructure engineer works with the provider’s API to create virtual resources, such as

database clusters, storage, and computing clusters. It is worth emphasizing the advantage of using the Infrastructure as a code (IaC) approach for managing complex cloud infrastructure. IaC and tools that implement it allow management resources through code instead of manual interaction and settings. The code is stored in the version control system to provide versioning, reusability, observability, and consistency. The IaC approach allows the implementation of robust testing and deployment approaches for IT infrastructure [2].

2.2. Unstructured data storage and processing

Due to the large volumes, heterogeneity, and complexity of unstructured data, special approaches to storage and processing are required. The data should be stored in an environment that meets certain requirements:

- Scalability. The solution must easily scale up to accommodate large amounts of data. This is especially important for unstructured data that can grow rapidly in volume, such as images, video, audio, etc.
- Availability and durability. The system must ensure data durability (be configured to provide data backup and automatic recovery in the event of a failure on one of the nodes) and availability (data should be available according to the defined level of performance).
- Speed. The system should offer high-speed data access, high throughput, and low latency.
- Security. The solution must provide a high level of data security, including user authentication, access authorization, and data encryption.

The systems or repositories for storing unstructured data in raw format in the form of files or binary objects are called data lakes. There are several widely used solutions for data lake implementation from popular cloud providers.Amazon Web Services offers the following architecture for implementing the data lake. Amazon S3 is used to store datasets. The service allows various options for configuring data security, durability, and scalability. Amazon DynamoDB is used to manage corresponding metadata for the dataset. Once a dataset is cataloged, its attributes and descriptive tags are available to search on. Amazon OpenSearch Service is offered to perform search and interactive analytics on data. Amazon Cognito is the service used to implement user authentication and authorization. AWS Glue can be used for data transformation, building ETL pipelines, and interactive data exploration. Amazon Athena is a service that can be used for building analytics applications.

Google Cloud Platform offers Cloud Storage (GCS) as a backbone for its data lake architecture. It's an object storage service that can be easily integrated with Google data processing services. Data from GCS can be used in the BigQuery analytics platform. It supports structured data in various formats and unstructured data. Google Dataflow can provide real-time insights from your data with streaming and machine learning. Google Cloud Data Fusion is the service that allows the visual building of the ETL/ELT data pipelines [4].

NoSQL databases such as documents, key-value, wide-column, and graph databases can also be useful for storing unstructured data. One of the most commonly used NoSQL databases are MongoDB and Apache Cassandra. MongoDB offers storage for vast amounts of unstructured data in JSON format with flexible horizontal scaling. Apache Cassandra is

known for scalability and high availability, used to handle enormous amounts of unstructured data.

To solve the challenge of processing and storing unstructured data there are several distributed computing systems like Apache Hadoop. Apache Hadoop is an open-source software framework for building distributed, fault-tolerance computing clusters. The main Hadoop components are HDFS (a distributed filesystem for storing large datasets), YARN (a platform for managing cluster compute resources, and using them for scheduling users' applications), and Apache Spark (a cluster computing framework for large-scale data processing).

2.3. The Complexity of a Machine Learning System

When developing a software product using machine learning models, in addition to model development, other factors affecting the system's overall complexity should also be considered. Machine learning models are relatively easy to develop but maintaining the systems based on them is difficult and expensive as they are complex and tend to accumulate technical debt [6].

Data quality plays a crucial role in the performance of the resulting machine learning system. To be a reliable basis data should be tested, unified, and continually improved during the system life cycle. The machine learning system should implement a data collection process to create and keep datasets updated. It should provide the tools and workflows for data exploration, running of interactive experiments, and model performance evaluation. The system must implement the deployment process, continuous testing, and monitoring of the application based on ML models.

3. Methods of processing unstructured data

Artificial intelligence is a broad field of computer science that enables computer programs to understand natural language, reason, learn, analyze information, and act in a way that resembles human intelligence.

Machine learning is a subset of artificial intelligence that has proved to be especially useful in unstructured data handling. The idea behind machine learning methods is to build programs that can learn and make predictions based on provided data without being explicitly programmed. Such programs are able to find patterns and discover complex relations in unstructured data that traditional analysis methods would miss.

There are many methods for processing unstructured data recognition based on machine learning and artificial intelligence, each with its own advantages and disadvantages. Some of the more common methods include:

Natural Language Processing (NLP): NLP is a field of artificial intelligence that deals with the interaction between computers and human language. NLP techniques are used to analyze and understand text, for example, to determine its meaning, grammar and structure.

Image Processing: Image processing is an area of computer vision that deals with the analysis and manipulation of images. Image processing methods are used to detect and identify objects in images, as well as to extract information from them.

Audio Processing: Audio processing is an area of computer science that deals with the analysis and manipulation of audio signals. Audio processing techniques are used for speech recognition, speech synthesis, and music analysis.

Data Analysis: Data analysis is the process of discovering useful information from data. Data mining techniques are used to process and analyze unstructured data such as text, images, and audio.

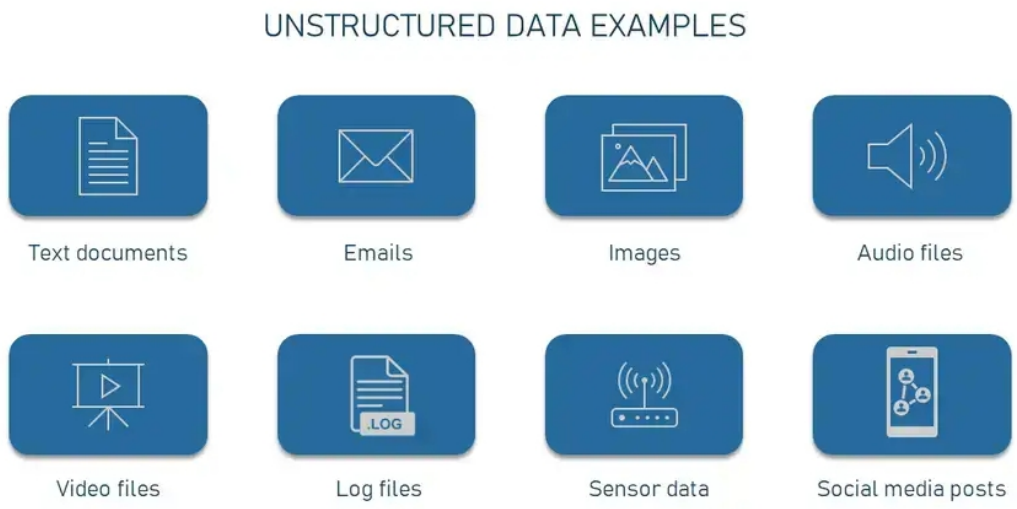

Figure 1: Some examples of unstructured data

The way data is processed depends on whether it is text, image, audio or video. This is how video data is processed: Speech-to-text software transcribes audio to video. The platform extracts and analyzes the subtitles that appear in the video, with the goal that no potentially meaningful entity is missed in the process. The next step recognizes and captures the image and text data using optical character recognition. The intelligent scanner then performs an in-depth scan to identify any logos that appear in the video. Ultimately, the platform recognizes and extracts all the text.

Tools for processing unstructured data, each of which has its own characteristics. Some of the more common tools include:

Scikit-learn - is a Python machine learning library that includes many tools for processing unstructured data such as NLP, image processing, and audio processing.

TensorFlow - is an open-source numerical computing library used for machine learning and deep learning. TensorFlow can be used to develop complex models for processing unstructured data.

NLTK - NLTK is a Python toolkit for NLP. NLTK includes many tools for text processing, such as tokenization, stemming, part-of-speech tagging, and sentiment analysis. Various methods and libraries are available to perform tokenization. NLTK, Gensim, Keras, TextBlob, spaCy are some of the libraries that can be used for the task.

OpenCV is an open-source library for image processing and computer vision. OpenCV includes many tools for image processing, such as object detection, face recognition, and motion tracking.

3.1. Stages of processing unstructured text data using NLTP

Unstructured text data processing is a multi-step process that converts informal text into a format that can be understood and used by computers. This process includes:

1. Data collection means identifying data sources such as websites, social networks, forums, or internal databases. Download and save data in a convenient format, for example, TXT, CSV or JSON.
2. Data cleaning: Removing noise and errors such as misspellings, duplicate entries, special characters, and HTML tags. Text normalization, such as converting all letters to lowercase, removing extra spaces, and converting date and time to a standard format.
3. Tokenization and Breaking text into individual words or phrases called tokens: identifying and removing stop words that have no informative value, such as "the", "a", "an", "this", "that", "is", "etc."
4. Stemming or lemmatization by reducing words to their basic form, for example, "running" to "run", "cities" to "city", "studied" to "study". This helps to reduce the dimensionality of the data and improve the accuracy of the analysis.
5. Frequency analysis: Determining the frequency of occurrence of each word or phrase in the text. This can help identify key themes, concepts, and emotions in the text.
6. Positional analysis: identifying the context in which words or phrases appear. This can help to better understand the meaning of the text and the connections between words.
7. Text classification with automatic assignment of text to categories or labels. This can be used to filter spam, identify the subject of documents, or segment text by genre.
8. Extraction of information, identifying and extracting key facts, essences, and connections from the text. This can be used to create resumes, build knowledge bases, or automatically generate reports
9. Sentiment analysis with determining the general emotional tone of the text, for example, positive, negative, or neutral. This can be used to measure people's opinions about a product, service, or event.
10. Data visualization allows converting text data into visual formats such as graphs, charts, and word maps. This can help better understand data distributions, trends, and relationships between words.

3.2. The role of MLOps in managing and using unstructured data

MLOps is the workflow for deployment and maintaining the production machine learning system reliably and efficiently [7]. According to MLOps machine learning is the software engineering discipline and models are reusable software artifacts that can be deployed via deployment pipelines. The MLOps adoption goal is to provide a collaborative development environment for the engineering team with capabilities for experiment tracking, feature engineering, and model versioning and management. The framework can be considered as an adaptation of the DevOps principles for machine learning. In addition to the Continuous

Integration of the model code, it also assumes testing and validating models and data. Continuous delivery involves deploying a multi-step pipeline to automatically retrain and deploy the model. The full workflow for MLOps can be described in the following steps:

- Model building. After building models are stored in version control repositories for future reuse.
- Evaluation. Models' performance is evaluated and measured.
- Testing. Models are continuously tested to confirm they are suitable for deployment, and the performance is better than some baseline.
- Deployment. The validated model is deployed to the target environment.
- Monitoring. The model performance metrics are continuously monitored. If performance is unsatisfactory the new MLOps iteration should be invoked.

Using MLOps in processing unstructured data can bring significant benefits, as it helps you to create, optimize, and maintain complex models that work with these types of data.By harnessing the power of unstructured data processing, businesses and organizations can gain a significant competitive advantage in today's data-driven world. Unstructured data processing (UDP) solutions transform unstructured data into useful data to automate business processes [4].

The MLOps platforms such as MLFlow and Kubeflow can automate various tasks, such as entering data through a streaming API, scheduling a training session, deploying the latest trained models, or sending notifications to stakeholders about an item that needs immediate attention. Additionally, the platform can generate regular reports for stakeholder consumption and provide a baseline for future models.

The next generation of automation is capable of receiving, extracting and processing data from a variety of unstructured formats including images, documents, audio, video and text. Unstructured data processing breaks down the extraction process into smaller, manageable tasks and intelligently directs information to software, artificial intelligence and label trainers/developers to extract useful data with assured quality. It learns from people to continuously increase the level of automation and reduce costs. It is platform and language agnostic. It allows users to use machine learning (ML) models and pre-configured programs from a rich market, create their own, or build their own programs to solve even the most complex extraction tasks.

4. Text analysis matter

4.1. Understanding of textual data with NLP

Text data processing methods include thematic modeling, text classification, detection of emotional tone, etc. Techniques such as NLP (Natural Language Processing), LSTM (Long Short-Term Memory), and BERT (Bidirectional Encoder Representations from Transformers) allow efficient analysis and understanding of textual data. For image processing, convolutional neural networks (CNN - Convolutional Neural Networks) are used, which allow you to effectively perform the tasks of object recognition, image classification, face detection, and pattern recognition in large sets of images.

To predict using streaming data, trained models are further deployed on the MLOps workflow as web services. The streaming data trains the model if the prediction is accepted or rejected. Finally, the trained model can be redeployed as a web service. Deployment frequency can vary from a few minutes to several days. Common techniques used in processing structured data can be applied to unstructured data to simplify operations later. Units of unstructured data are marked with findings for use with subsequent models. NoSQL databases such as MongoDB, Hadoop, and other popular databases can help store data in JSON format.

4.2. Tokenization using NLTK

There are different tokenization techniques that can be applied depending on the language and the purpose of the simulation. Below are a few tokenization techniques used in NLP.

NLTK (natural language toolkit) is a Python library developed by company Microsoft to help with using NLP.

Tokenization can be done to separate words or sentences. If the text is divided into words using some division technique, it is called word tokenization, and the same division done for sentences is called sentence tokenization.

We will use *Word_tokenize* and *sent_tokenize* - these are very simple tokenizers available in NLTK:

```
[ ]  import nltk
     nltk.download('punkt')
     sent_detector = nltk.data.load('tokenizers/punkt/english.pickle')
     print('\n-----\n'.join(sent_detector.tokenize(data1.strip())))
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
[11] from nltk.tokenize import word_tokenize
     print(word_tokenize(data1))
```

Figure 2: Using Word_tokenize and sent_tokenize tokenizers

Sent_tokenize splits a string into multiple sentences. sent_tokenizer derives from the PunktSentenceTokenizer class. sent_tokenize uses a pre-trained model from tokenizers/punkt/english.pickle. There are pre-trained models for different languages to choose from. PunktSentenceTokenizer can be trained on our own data to create our own sentence tokenizer.

custom_sent_tokenizer = PunktSentenceTokenizer(train_data)

There are some other special tokenizers like Multi Word Expression tokenizer (MWETokenizer), Tweet Tokenizer. The MWETokenizer takes a string that is already tokenized and re-tokenizes it, concatenating multi-word expressions into a single token using the MWE lexicon. TweetTokenizer handles specific things for tweets, such as emoji handling.

```
import nltk
nltk.download('punkt')
from textblob import TextBlob
# create a TextBlob object
blob_object = TextBlob(text)
# tokenize paragraph into words.
print(" Word Tokenize :\n", blob_object.words)
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
 Word Tokenize :
```

Figure 3: Using special tokenizers NLTK in Python

Tokenization with Textblob. Textblob is used to process text data and is a Python library. Like other packages, it provides APIs for sentiment analysis, tagging parts of speech, classification, translation, and more. Below is a snippet of code to tokenize into sentences and words, and you can notice that in the output, the emojis are removed from the punctuation marks [10].

5. Appendices

In the field of data analysis, unstructured data presents both obstacles and opportunities due to its diverse and dynamic nature. Although it defies traditional patterns and can appear disorganized, the use of modern techniques such as machine learning and artificial intelligence can reveal key insights.

Despite the successes, there are some challenges associated with processing unstructured data using machine learning. These include problems related to large volumes of data, heterogeneity of data, as well as the need for effective data management [15].

Methods and models for processing unstructured data based on machine learning open new perspectives for the analysis and use of this important category of data. Despite the challenges faced by researchers and practitioners, the development of technologies that facilitate the efficient processing and analysis of unstructured data continues, making it even more accessible and useful in various fields.

References

[1] Richard Szeliski. "Computer Vision Algorithms and Applications." Springer Cham 978-3-030-34371-2 January 2022, https://doi.org/10.1007/978-3-030-34372-9.

[2] Bird, Klein, & Loper. Natural Language Processing with Python: https://www.nltk.org/book/ (A practical guide using Python libraries).

[3] M. Petryk, M. Bachynskyi, V. Brevus, I. Mudryk, D. Mykhalyk. Analysis technology of neurological movements considering cognitive feedback influences of cerebral cortex signals. ITTAP CEUR Workshop Proceedings 3309 (2022): 45–54.

[4] Silge, Julian & Robinson, David. (2017). Text Mining with R [Online]. Available from: https://dl.acm.org/doi/10.5555/3165010 -1.
[5] D. Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” NIPS, pp. 2494–2502, Jan. 2015.
[6] C. Breuel, “ML Ops: Machine Learning as an Engineering Discipline,” Medium. Accessed: May 27, 2024. [Online]. Available: https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f
[7] Medium. Unstructured Data Analytics with MLOps: Available from: https://xpressoai.medium.com/unstructured-data-analytics-with-mlops-b6ac2672430
[8] Computerweekly. Unstructured data and the storage it needs; Available from: https://www.computerweekly.com/feature/Unstructured-data-and-the-storage-it-needs
[9] NLTK documentation: https://www.nltk.org/ Course: "Natural Language Processing" Available from: https://web.stanford.edu/class/cs224n/
[10] MR Petryk, A Khimich, MM Petryk, J Fraissard Experimental and computer simulation studies of dehydration on microporous adsorbent of natural gas used as motor fuel. Fuel 239, 1324-1330. 2019
[11] University of Washington. (n.d.). Natural Language Processing Specialization. [Online]. Available from: https://www.coursera.org/specializations/natural-language-processing
[12] Stanford University. (n.d.). Natural Language Processing. [Online]. Available from: https://online.stanford.edu/courses/xcs224n-natural-language-processing-deep-learning
[13] Altexsoft/Unstructured Data: Examples, Tools, Techniques, and Best Practices; Available from: https://www.altexsoft.com/blog/unstructured-data/
[14] Amazon Web Services, Inc. What is Cloud Computing? - Cloud Computing Services, Benefits, and Types – AWS. Available from: https://aws.amazon.com/what-is-cloud-computing/
[15] Petryk, M.R., Boyko, I.V., Khimich, O.M. et al. High-Performance Supercomputer Technologies of Simulation and Identification of Nanoporous Systems with Feedback for n-Component Competitive Adsorption. Cybern Syst Anal 57, 316–328 (2021).