Pais *et al. Journal of Big Data* (2022) 9:54
https://doi.org/10.1186/s40537-022-00603-5

Journal of Big Data

SURVEY PAPER Open Access

NLP-based platform as a service: a brief review

Sebastião Pais[1,2,3*], João Cordeiro[1,4] and M. Luqman Jamil[1]

*Correspondence:
sebastiao@di.ubi.pt
[1] Department of Computer Science, University of Beira Interior, Covilha, Portugal
Full list of author information is available at the end of the article

Abstract

Natural language processing (NLP) refers to the field of study that focuses on the interactions between human language and computers. It has recently gained much attention for analyzing human language computationally and has spread its applications for various tasks such as machine translation, information extraction, summarization, question answering, and others. With the rapid growth of cloud computing services, merging NLP in the cloud is a significant benefit. It allows researchers to conduct NLP-related experiments on large amounts of data handled by big data techniques while harnessing the cloud's vast, on-demand computing power. However, it has not sufficiently spread its tools and applications as a service in the cloud and there is little literature available that discusses the scope of interdisciplinary work. NLP, cloud Computing, and big data are vast domains and contain their challenges and potentials. By overcoming those challenges and integrating these fields, great potential for NLP and its applications can be unleashed. This paper presents a survey of NLP in cloud computing with a key focus on the comparison of cloud-based NLP services, challenges of NLP and big data while emphasizing the necessity of viable cloud-based NLP services. In the first part of this paper, an overview of NLP is presented by discussing different levels of NLP and components of natural language generation (NLG), followed by the applications of NLP. In the second part, the concept of cloud computing is discussed that highlights the architectural layers and deployment models of cloud computing and cloud-hosted NLP services. In the third part, the field of big data in the cloud is discussed with an emphasis on NLP. Furthermore, information extraction via NLP techniques within big data is introduced.

Keywords: Natural language processing, Cloud computing, Big data

Introduction

Natural language processing (NLP) is a rapidly developing field of artificial intelligence and data science that deals with speech and text processing technologies. The goal of this direction is the development of methods for automatic analysis and human language presentation [1]. NLP uses a variety of methodologies to interpret the ambiguities in human language, including automatic summarization, part-of-speech tagging, disambiguation, entity and relation extraction, sentiment analysis, natural language understanding, and speech recognition. Many NLP-related software tasks have been successfully solved and integrated that are used on the internet, such as morphological

and syntactic analysis. Therefore, now they can be carried out as software, like Software as a Service (SaaS) in cloud computing [2].

Cloud computing is a model to allow flexible on-demand network access to a shared pool of configurable computing resources that can be accessed and delivered quickly with minimal management effort. This paradigm is based on many existing technologies such as the internet, the resource for virtualization, grid computing, and web services. Hence, Cloud Computing combines software as a service and utility computing. Cloud computing is designed to provide flexible and low-cost on-demand computing infrastructure with good service reliability [3, 4].

Nowadays, cloud services for NLP analysis are becoming very popular among scientists and all users interested in the field. It allows researchers to deploy, share, and use components and resources for language processing, following the paradigms of data as a service and software as a service. However, there are only a few reviews of natural language processing services in the cloud are available.

A few examples of prominent natural language processing APIs and cloud-based services are Amazon Comprehend [5], Microsoft Azure Cognitive Services [6], Google Cloud Natural Language [7], and third-party options.

Amazon Comprehend (AWS) service uses machine learning to extract key phrases and identify the language in a given text. Amazon Comprehend works with any AWS-supported application, and it has features such as sentiment analysis, tokenization, and automated text file organization.

The Microsoft Azure Cognitive Services is a portfolio of natural language processing tools broken out into several different, more targeted services and uses. For example, if developers want to build applications that can analyze the sentiment or identify the language of a given text, Azure Text Analytics API can be used. Alternatively, Azure Language Understanding Intelligent Service can understand things such as user intent. It is precious when developers build chatbots, voice-powered products, and even customer service platforms.

Google Cloud Natural Language also emphasizes entity extraction, sentiment analysis, syntax analysis, and categorization. However, this API differs because it is powered by Google's own in-depth learning modules – the same ones that drive the query comprehension behind Google Search and the language understanding system behind Google Assistant.

Third-party options are many natural language processing APIs and services in the market. For example, companies such as Diffbot [8] offer features via commercial API that let users extract data precisely from websites, while other vendors, like MonkeyLearn [9], provide services to automate workflows based on unstructured data.

Different from the above reviews, we investigate the existing applications and techniques of NLP and cloud-hosted NLP services and discuss methods and technologies related to NLP and big data and information extraction via NLP within big data. The challenges of NLP in big data are introduced including the requirements and methodology. The NLP platforms, layers, and deployment models of cloud computing are also discussed.

This paper aims to provide a survey on NLP in cloud computing and NLP in big data. It highlights the existing NLP services and their main features for the sake of comparison.

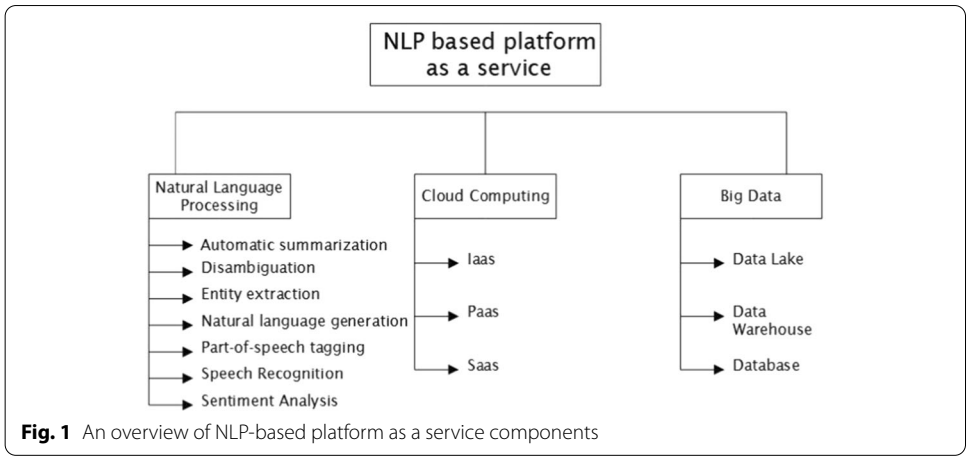

Fig. 1 An overview of NLP-based platform as a service components

Similarly, different big data techniques and cloud computing paradigms are presented. The concept of this review and the main components of NLP, cloud computing, and big data are provided in Fig. 1. The organization of the paper is as follows: "Natural language processing" section presents the various important terminologies of NLP and its applications. "Cloud computing" section discusses cloud computing and presents its architectural layers and taxonomy, followed by presenting cloud services for NLP and towards the NLP platform as a service: requirements and methodology. "Big data in cloud" section introduces big data in the cloud, methods, and technologies related to NLP, and big data are discussed, followed by presenting machine learning and NLP in big data. Moreover, information extraction via NLP techniques within big data is introduced. In addition, challenges of NLP and big data are also discussed. "Conclusion" section provides the conclusion of this survey.

Natural language processing

The field of Artificial Intelligence and linguistics aims to make computers understand statements or words written in human languages. NLP was developed to ease the user's work and satisfy the desire to communicate in natural language with the computer. Since all users may not be well versed in the machine-specific language, NLP supports those users who do not have enough time to learn or perfect new languages. A language can be described as a set of rules or a set of symbols. Symbols are combined and used for conveying information or broadcasting the information. NLP can essentially be divided into two components, natural language understanding (NLU) and natural language generation (NLG), which evolve around the task of understanding and generating the text.

Levels of linguistic analysis for NLP

Linguistics is a science that comprehends the meaning of language, the linguistic context, and various forms of language. Language levels are one of the most analytic approaches to representing the NLP, helping to produce the NLP text by performing the content planning, sentence planning, and surface implementation phases. Various key

terminologies of the NLP are divided into seven levels, from the simplest to the most difficult [10], namely: morphology, lexicon, syntax, semantics, phonology, discourse, and pragmatics.

Morphology, the term itself is a Greek word derived from "morph" and "ology". The term "morph" means to form or structure, while "ology" means to study something. Therefore, morphological analysis can be defined as the scientific study, identification, analysis, and description of the structure or form of words in a language and the relationship of the words to one another. The morphological analysis enables NLP systems to process new words more flexibly [11]. The morphological analysis can be inflection or derivation. Inflectional morphologies are forms of the same stem, for example, the words eyed and eyes come from the root eye, while derived morphologies are new words derived from existing words; for example, the word heartburn is derived from the words heart and burning.

Lexical level connects people with NLP systems to interpret the meaning of individual words. Several types of processing contribute to word-level comprehension: the first is the assignment to each word of a single part-of-speech label. This processing applies the most likely part of the speech tag to words that can function as more than one part of speech depending on the context in which they occur. Furthermore, at the lexical level, words with only one possible meaning or interpretation can be replaced by a semantic representation of this meaning. The type of representation varies depending on the semantic theory used in the NLP system.

The lexical level may require a lexicon, and the specific focus of an NLP system will determine whether a lexicon is used and the type and amount of information encoded in the lexicon. Lexica can be very simple, containing only words and their part(s) of speech, or perhaps increasingly complex, containing information about the semantic class of the word. What arguments are needed and the semantic constraints on those arguments, the definitions of the meanings in the semantic representation used in the particular system, and even the semantic field in which each meaning of a polysemous word is used.

Syntactic level emphasizes examining the words in a sentence to reveal the grammatical structure of the sentence. At this level, grammar and parsers are required. The result of this level of processing is the representation of the sentence, revealing the structural dependency relationships between the words. There are different grammars that can be used, which in turn influence the choice of a parser. Not all NLP applications require full sentence analysis. Thus, the current challenges of parsing prepositional suffixes and controlling conjunctions no longer stand in the way of defending which clause- and phrase-dependent dependencies are appropriate [10]. In most languages, syntax conveys meaning, while order and dependency create connotations. For example, the following two sentences: "*The cat chased away the mouse.*" and "*The mouse chased the cat.*" differ only in grammar, but they express completely different meanings.

Semantic processing determines possible meanings by activating interactions between meanings at the word level in the sentence. This level of processing may include the semantic uniqueness of words with multiple meanings; in the context of syntactic disambiguation of words that can serve multiple parts of speech, it is useful at the syntactic level. For example, "file" as a noun can mean, among other things, a folder for collecting papers, a tool for shaping nails, or a row of people in a queue. The semantic level

examines words for their dictionary explanation and mid-sentence explanations. The semantic medium where most words have more than one explanation, but the correct one can be found by looking at the rest of the sentence [12].

Phonology is the study of language at the phonetic level, including the study of combinations of sounds in organized units of speech, the formation of syllables, and larger units. Phonetic and phonological information is essential to speech-based systems because they deal with the relationship between words and the sounds they recognize.

Discourse involves syntactical and semantics, working with sentence length units. It does not consider multi-sentence texts as chain sentences that can be explained individually. Instead, the discourse focuses on the properties of the text as a whole, which conveys meaning by connecting the sentences of the [10] component. The most common level is *Anaphora Resolution*. Anaphora resolution replaces words as pronouns that are semantically intertwined with the actual object they refer to. Recognition of discourse structure influences the functions of sentences in the text, which in turn contributes to the meaningful representation of the text.

Pragmatic level deals with the intentional use of language in situations and uses the context and content of the text for understanding. Its purpose is to explain how additional meaning is read into texts without encoding them. Some NLP applications may use knowledge bases and modules for inference. For example, the following two sentences require the resolution of the anaphoric term "they", but this resolution requires some practical knowledge, also known as world knowledge, to resolve the exact meaning of "they" [10].

"The ministry refused to increase the petroleum prices because they feared violence."
"The ministry refused to increase the petroleum prices because they advocated revolution."

Natural language generation

NLG generates phrases, sentences, and paragraphs that are meaningful from an internal representation. It is an important task from natural language processing and happens in four phases: identifying the goals, planning how objectives can be accomplished by assessing the situation, available sources of communication, and realizing the plan as a text. According to Khurana et al. [13], the NLG components consist of speaker and generator; components and levels of representation (content selection, textual organization, linguistic resources, and realization); and application or speaker.

Speaker and generator, To generate a text, we need to have a speaker or an application and a generator or a system that renders the application's intentions into fluent phrases relevant to the situation.

Components and levels of representation—the language generation process involves the following interwoven tasks:

Content Selection, the information should be selected and included in the set. Depending on how this information is parsed into representational units, parts of the units may need to be omitted, while others may be inserted by default.

Textual Organization, the information must be textually arranged according to the grammar; it must be ordered sequentially and in terms of linguistic relations such as modifications.

Linguistic Resources, to support the realization of information, linguistic resources must be chosen. Eventually, these resources will come down to choices of particular words, idioms, syntactic structures.

Realization of the chosen and organized resources must be realized as an actual text or voice output.

Application or speaker is only for maintaining the model of the situation. Here the speaker starts the process, does not engage in the generation of languages. It stores the history, constructs the potentially relevant content, and deploys a representation of what it knows. All these form the situation while selecting a subset of propositions that the speaker has. The only obligation is to make sense of the situation by the speaker.

Applications of NLP

Natural language processing provides both theory and implementations for a range of applications. Any application that utilizes text is a candidate for NLP. The most common applications utilizing NLP include information retrieval and summarization. NLP can be applied in various areas such as machine translation, information extraction, summarization, and question answering among other applications.

Information Retrieval, according to Copestake [14], information retrieval is the process of returning a set of documents in response to a user query. A typical example of information retrieval is search engines like Google to match a user's query to a collection of documents and then return the most similar or relevant documents. The major drawback of information retrieval is that the set of documents returned is not organized, and thus it becomes difficult for the end-users to navigate the documents. Natural language techniques such as tokenization, sentence splitting, and morphological normalization are usually used during information retrieval.

Summarization, the type of text summarization, depends based on the number of documents, and the two major categories are a summarization of single-document and summarization of multi-document [15, 16]. There are also two forms of summaries, generic or query-focused [17, 18]. Summarization tasks can be either supervised or unsupervised [19]. In a supervised system, training data is required to select relevant material from the documents. For learning techniques, a large amount of annotated data is required. There are certain methods as follows:

- Bayesian Sentence based Topic Model (BSTM) uses both term-sentences and term-document associations for summarizing multiple documents [20];
- Factorization Given Bases is a language model where sentence bases are the given bases, and it utilizes document-term and sentence-term matrices. This approach groups and summarizes the documents simultaneously [21];
- Topic Aspect-Oriented summarization is based on topic factors. These topic factors are various features that describe topics such as capital words are used to represent

the entity. Various topics can have various aspects, and various preferences of features are used to represent various aspects [22].

Information Extraction aims to find specific information from unstructured natural language texts [23]. Extracting entities such as name, location, activities, dates, times, and prices for many applications is a powerful way to summarize the information relevant to a user's needs. In the case of a domain-specific search engine, the automatic identification of critical information can increase the accuracy and efficiency of a directed search. Over recent years, the exploration of information has become an important area of research. Knowledge discovery research uses various techniques to retrieve helpful information from source documents, such as part-of-speech tagging, chunking or shadow parsing, stopwords and stemming. The former has good accuracy but a higher cost of implementation, while the latter has a lower cost of implementation but is generally insufficient for information retrieval: compound and statistical phrases index multi-token units instead of single tokens. Word sense disambiguation is the task of understanding the correct sense of a word in context. When used for information retrieval, terms are replaced by their senses in the document vector. Its extracted information can be used for various purposes, such as preparing a summarization, building databases, defining keywords, classifying text objects by specific predefined categories, among other purposes.

Question-Answering is a discipline of computer science in the fields of AI and NLP, focusing on building systems that automatically answer the questions posed by human beings in their natural language. Question-answering returns a set of relevant documents in response to a user's query [24]. A computer system that understands the natural language has the capability of a program system to translate human-written sentences into an internal representation to produce valid answers. The exact answers can be given by doing syntax and semantic analysis of the questions. Lexical gap, ambiguity, and multilingualism are some of the challenges for NLP when it comes to building a good question answering system. The exact answers can be generated by doing syntax and semantic analysis of the questions. Lexical gap, ambiguity, and multilingualism are some of the challenges for NLP in building a good question answering system.

Sentiment analysis is used to identify the sentiments among several posts. It is also used to identify the sentiment where the emotions are not expressed explicitly. According to Prabowo and Thelwall [25] and Saif et al. [26], sentiment can be categorized into two groups, which are negative and positive words.

Sentiment analysis refers to the general method of extracting polarity and subjectivity from semantic orientation, which refers to the strength of words and polarity text or phrases [27]. There are two main approaches for extracting sentiment automatically, which are the lexicon-based approach and the machine-learning-based approach [28]. Companies are using sentiment analysis and application of NLP to recognize the opinion and sentiment of their customers online. Companies can also use sentiment analysis to determine their overall reputation from customer posts. It will assist companies to understand what their clients think about the products and services. In this way, we may conclude that, in addition to assessing essential polarity, sentiment analysis understands sentiments in a given context, helping humans to better understand what lies behind the expressed opinion.

Cloud computing

Cloud computing is a generic term for anything involving the delivery of hosting services over the Internet. The name cloud computing has been inspired by the cloud icon often used to describe the Internet. Authors of [29] define cloud computing as a style of computing that provides massively scalable IT-enabled capabilities' as a service to external customers using Internet technologies. According to NIST (National Institute of Standards and Technology), cloud computing is on-demand access to a shared pool of computing resources [30]. Cloud computing can also be described as a collection of services presented as a layered cloud computing architecture. Cloud computing typically involves three distinct levels of service combinations: IaaS, PaaS, and SaaS [31–33].

Another study [34] suggests that cloud computing refers to a pool of abstracted, highly scalable, and managed infrastructure capable of hosting end-customer applications and billed for consumption. Cloud computing's broader aim is to provide the masses with supercomputing. These definitions include cloud architecture and deployment strategies. In particular, there are clearly articulated essential elements of cloud computing: (i) On-demand self-service that allows users to consume computing capabilities, for example (applications, server time, network storage) as required; (ii) Resource pooling that allows multiple customers to be served by combining computer resources (hardware, software, processing, network bandwidth) – these resources are dynamically allocated; (iii) Rapid elasticity and scalability allowing fast and automated provision and scaling of functionalities and resources; (iv) Measured arrangements to automate resource allocation and provide metering capabilities to assess billing usage, extension to existing hardware and application resources, thus reducing the cost of additional resource provisioning.

Cloud computing is an enticing paradigm that has many advantages, such as:

Reduced cost Cost containment is a clear advantage of cloud computing regarding both capital and operational expenses. The reduction in capital expenditure is apparent because a company can invest the necessary increase in capacity and does not need to build infrastructure for total or excess capacity. Enterprises can use a cloud service, or they can reduce operational and maintenance costs by implementing cloud paradigms internally;

Improved automation Cloud computing is based on the premise that services are provided in a highly automated manner and distributed. This feature provides companies with significant efficiencies.

Flexibility The advantages of versatility arise from the rapid provisioning of new capacity and rapid relocation or workload transfer. Cloud computing, for example, enables consistency in procurement and development processes and schedules in public sector environments.

Sustainability Cloud computing uses far less electricity and other resources than a conventional DATA center by exploiting economies of scale and the ability to manage assets more effectively. Due to poor architecture or inefficient use of materials, the low energy efficiency of most existing data centers is environmentally and economically unsustainable.

Cloud computing deployment models

NIST summarized the cloud computing characteristics as on-demand self-service, ubiquitous network access, resource pooling, rapid elasticity, and pay-per-use. The rapid transition towards cloud computing has increased the demand for far more deployment models. The cloud computing model has four main deployment models identified by NIST [30]: public cloud, community cloud, private cloud, and hybrid cloud.

The public cloud deployment is the dominant form of the current cloud computing model. The general cloud consumers use the public cloud, and the cloud service provider has full ownership of the public cloud with its policies, values, profit, costing, and charging model. Many popular cloud services are public clouds such as Amazon EC2, S3, Google AppEngine, and Force.com.

A community cloud deployment is implemented and followed by a specific community of users, including institutions and organizations, that share well-defined common goals/interests/missions.

Private cloud deployment is run solely within a single organization and managed by the organization or a third party regardless of whether it is located on the premise or off-premise. The motivation to set up a private cloud within the company has several aspects. First, to maximize and optimize the utilization of existing in-house resources. Second, security concerns, including data privacy and trust, often make private clouds a choice for many businesses. Third, there is still a considerable cost of transferring data from local IT systems to a public cloud. Fourth, companies are always in need of complete control of mission-critical operations behind their firewalls. Last, academic often builds a private cloud for research and teaching purposes.

A hybrid cloud deployment combines two or more clouds (private, community, or public) that remain unique entities but are connected by standardized or proprietary technology that allows data and application portability, for example, (cloud bursting for load-balancing between clouds). Organizations use the hybrid cloud model to maximize their resources by marginalizing peripheral business functions on the cloud while controlling core on-premises operations via a private cloud.

Cloud-hosted NLP services

Many NLP applications demand some necessary linguistic processing (tokenization, part-of-speech tagging, named entity recognition and classification, syntactic parsing, coreference resolution, among other applications) to carry out more complex tasks. Generally, NLP annotation must be as accurate and efficient as possible, and current tools have focused chiefly on efficiency very rightly. However, this generally means that NLP suites and tools usually require researchers to use such tools to perform complex compilation/installation/configuration procedures. At the same time, many small and medium enterprises are currently offering services in the industry that rely on NLP annotations in one way or another. There are many ways to provide cloud-based NLP services by APIs. In this section, we discuss many other NLP frameworks built around cloud services with varying objectives.

SYSTRAN [35] platform is a series of APIs for translation, multilingual dictionary lookups, natural language processing (Entity recognition, Morphological analysis,

Part-of-Speech tagging, Language Identification), and Text Extraction from documents, audio files, or images. SYSTRAN Platform allows the user to use and analyze both structured and unstructured multilingual content, such as user-generated content, social media, Web content, and more. It is simple to use, scalable, and reliable. Moreover, it is free for small volumes and testing purposes; monthly subscriptions for higher volumes are available.

AYLIEN Text API [36] is a package of information retrieval, machine learning, and natural language APIs that make the analysis of text on a scale easier. It offers eight APIs with different functionalities such as article extraction, concept extraction, entity extraction, summarization, classification, semantic labeling, image tagging, sentiment analysis, hash-tag suggestion, language detection, and microformat extraction. These functionalities are available in six languages: English, German, French, Italian, Spanish and Portuguese. One interesting feature of Aylien is a text analysis add-on for Google Spreadsheets that allows users to run the API functionality through a familiar interface.

Text Summarization API [37] provides a professional text summary service, which relies on advanced natural language processing and machine learning technologies. It can be used, to sum up short essential texts from the URLs or documents provided by users.

Twinword Text Analysis Bundle API [38] is an all-purpose API for text analysis, thus including a wide variety of tasks on demand, like sentiment analysis, topic tagging, lemmatization, word associations, among others. Their goal is to gather various NLP tools in just one place, aiming to analyze and understand human sentences.

AlchemyAPI [39] provides cloud and on-premises text processing services for text analysis. It integrates NLP systems with existing tools for processing applications, resources, and data. It provides some of the tasks such as language detection, text extraction, keyword extraction, entity extraction, sentiment analysis, and text categorization.

RxNLP Text Mining [40] provides access to some advanced functionalities of text analytics over the cloud. This API offers a range of functionalities such as word generation and n-gram counts, computes text similarity, extracts topics (keywords) from the text, clusters sentences, extracts text from HTML pages, and summarises opinions.

Stanford Core NLP [41] provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people and between other applications, normalize date, time, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said. Stanford CoreNLP integrates many of Stanford's NLP tools, including part-of-speech (POS) tagger, named entity recognizer (NER), parser, the co-reference resolution system, sentiment analysis, bootstrapped pattern learning, and open information extraction tools. Moreover, an annotator pipeline can include additional custom or third-party annotators. CoreNLP's analysis provides the foundational building blocks for higher-level and domain-specific text understanding applications.

Text-Processing API [42] offers functionalities such as review of emotions, stemming and lemmatization, part of speech tagging and chunking, phrase extraction, and named entity recognition.

The *NLPTools API* is created by Atrilla [43] for a text processing framework to analyze Natural Language by performing operations and tasks on corpus data. Hence, this approach focuses on the statistical track of NLP. It is primarily focused on text classification and sentiment analysis.

Stemmer API [44], this API takes a paragraph and returns the text with each word stemmed using porter stemmer, snowball stemmer, or UEA stemmer.

The *WebKnox Text-Processing API* [45] can process natural languages and detect text language, quality of writing, find entity mentions, tag part of speech, extract dates, locations, and determine the sentiment of the text.

Topics Extraction API [46], created by Meaning Cloud, tags locations, people, companies, dates, and many other elements appearing in a text written in Spanish, English, French, Italian, Portuguese, or Catalan. This detection process is implemented by integrating many complex natural language processing techniques to acquire morphological, syntactic, and semantic analysis of a text and use them to identify various types of significant elements.

Fluxifi NLP API [47] is a cloud-based natural language processing API designed to detect the input text language and its sentiment.

Mountain Fog [48] developed *Cloud NLP API*. It is a set of web service APIs for natural language processing to perform functions such as interpreting feelings.

Linguakit API [49] is a multilingual suite of tools aimed at performing several tasks in linguistic analysis and information extraction. It provides features such as language identifier, keyword extractor, named entity recognizer, part of speech tagger, syntactic analyzer, tokenizer, and a text summarizer.

Semantria [50] provides a variety of NLP services based on the Salience Lexalytics engine but can only be accessed via an API or Microsoft Excel.

In this work, we are interested in natural language processing tasks and their solutions using SaaS/PaaS. Analysis of available services is needed to clarify the boundaries of their functionality and the quality of performance of existing options. It is also essential to understand the policy of using these services. Cloud services for NLP provide features such as named entity recognition (NER), part of speech tagging (POS), sentiment analysis, stemming, lemmatization, categorization, among others [51]. Several systems support and provide NLP tools using cloud computing.

GATECLOUD [52] is a cloud-based version of the GATE set of NLP tools with support for 24 languages, and it provides many specialty services. It is an open-source software that comes with a limited free version and *pay-as-you-go* options.

Lexalytics [53] is a text analytics service providing entity extraction, sentiment analysis, document summarization, and thematic extraction;

Amenity Analytics [54] is cloud-based natural language processing. It offers NLP text analytics and sentiment analysis tools;

TEXT2DATA [55] is a text analytics platform as a service SaaS. It offers features such as sentiment analysis, document classification, and entity extraction.

Another cloud computing service supporting machine learning development is *BigML* [56], a SaaS approach to machine learning. Users can set up data sources, create, visualize, share prediction models (only decision trees are supported), and use models to generate predictions. All from a Web interface or programmatically using REST API.

Table 1 Comparison of cloud-based NLP technologies

	NLP Services	# Languages	Subscription
GATECLOUD	Multiple	24	Free, Pay-as-you-go
Lexalytics	Multiple (4)	24	Starts at $999
Amenity Analytics	Multiple	–	Paid
TEXT2DATA	Multiple	4	Free, Paid
BigML	Multiple	22	Free, Paid
Google Cloud Natural Language	Multiple (5)	>10	Free, Paid
Eigen Technologies	Multiple	–	Paid
Linguakit	Multiple	4	Free
NLP Cloud	Multiple	–	Free, Paid
Azure Text analytics	Multiple (12)	–	Free, Paid
Amazon Comprehend	Multiple (13)	–	Free, Paid
Watson Natural Language Understanding	Multiple	–	Free, Paid

It provides a variety of options for text analysis along with other machine learning algorithms.

Google Prediction API [57] is Google's cloud-based machine learning tool that can help analyze different data. It is closely connected to Google Cloud Storage, where training data is stored. It offers its services using a RESTful interface, client libraries allowing programmers to connect with Java, JavaScript, .NET, Ruby, Python, and other programming languages. *Google Natural Language* [7] provides sentiment analysis, entity analysis, entity sentiment analysis, content classification, and syntax analysis.

Eigen tech [58] is a natural language processing/machine learning B2B software platform. Eigen's platform in general and sector agnostic is currently focused on the financial services, legal, and insurance sectors. It can classify, extract, organize, and analyze text data.

Matrix [59] is built using Apache MahoutTM. It can be accessed as PaaS using a RESTful interface. It can incrementally update the model once new data is available. It is organized in two layers serving (open source and free) and computation (Hadoop-based).

A French startup *nlpcloud* [60] specializes in providing NLP focused compute platform as a service. It includes various NLP-related services such as; named entity recognition, classification, summarization, question answering, sentiment analysis, text generation, translation, language detection, part-Of-speech tagging, tokenization, and lemmatization. It provides different tiers based on usage including the free tier. Users have options to choose different plans for usage along with a variety of best algorithms for the job.

A brief comparison of cloud-based NLP platforms is given in Table 1 while the details of provided NLP services are discussed earlier in this section.

It is important to note that for all of the services described in this section, the issues of data privacy and security are not handled clearly and transparently. For example, it is sometimes unclear who provides the cloud computing resources, and a third party is involved in handling billing and in managing access to computing resources. There are two main factors for users regarding cloud-based NLP technology namely costs

and services. Many of these services require an ongoing financial commitment to use their services. Some offer free subscriptions but with a limited amount of processing per day that usually offers low complexity basic NLP services. Similarly, the usability of NLP-based cloud technologies also depends on the type of services provided based on the subscription. Table 1 also highlights the limited number of NLP-related cloud services that are free for users.

Cloud services is a rapidly growing market. Modern technologies like big data analytics, IoT, artificial intelligence, and even web and mobile app hosting need massive computing power. Cloud computing offers enterprises an alternative to building their in-house infrastructure. With cloud computing, anybody using the Internet can enjoy scalable computing power on a plug-and-play basis. Since this saves organizations from investing and maintaining costly infrastructure, it has become a trendy solution. Many companies offer cloud platforms for the development, management, and deployment of applications.

The core of Google's business is all in Cloud Computing. Services delivered over network connections include search, email, online mapping, office productivity (including documents, spreadsheets, presentations, and databases), collaboration, social networking, and voice, video, and data services. Users can subscribe to these services for free or pay for increased levels of service and support.

As the world's largest online retailer, the core of Amazon's business is e-commerce. While e-commerce itself can be considered Cloud Computing, Amazon has also provided capabilities that give IT departments direct access to Amazon compute power. Key examples include S3 and EC2. S3 stands for Simple Storage Services. Any internet user can access storage in S3 and access stored objects from anywhere on the Internet. EC2 is the Elastic Compute Cloud, a virtual computing infrastructure able to run diverse applications ranging from web hosts to simulations or anywhere in between. This is all available for a meager cost per user.

Microsoft's core business has historically been in device operating systems and device office automation software. Microsoft has also offered web hosting, online email, and many other cloud services since the early days of the Internet. Microsoft also offers office automation capabilities via a cloud office live approach referred to as Software Plus Services or Software as a service to allow synchronous and asynchronous integration of online cloud documents with their traditional offline desktop resident versions.

The core mission of `salesforce.com` [61] has been in the delivery of Capabilities centered on customer relationship management. `salesforce.com` has established itself as a thought leader in software as a Service and is delivering an extensive suite of capabilities via the Internet. An essential capability provided is the site Force.com, enabling external developers to create add-on applications that integrate into the main `salesforce.com` application and are available on `salesforce.com`'s infrastructure.

VMware [62] provides several technologies of critical importance to enable cloud computing and has started offering its own cloud computing on-demand capability called VMware Cloud and VMware cloud Universal. This type of capability allows enterprises to leverage virtualized clouds inside their own IT infrastructure or host with external service providers.

Although there are many APIs available in the market for NLP-related tasks, Cloud-based platforms as a service remains uncommon. For some of the platforms such as Eigen technologies and Amenity Analytics, the access is restricted to only paid members. Where other alternatives do provide free options, the restriction of service usage still obstructs the potential full use. There is a need for common interest-based services which are beneficial for both parties such as user and provider.

Towards NLP platform as a service: requirements and methodology

NLP platform as a service provide easier administration, automatic update and patch management, compatibility, more comfortable usage, easier collaboration, and global accessibility by being designed as a SaaS. The motivation for the NLP platform as a service (PaaS) comes from the real word problems of sharing the produced NLP resources by different people from the varying level of computer background starting from students or researchers, up to people from other fields such as linguistics. Anybody from any discipline may easily use a web interface to analyze language data or construct more complicated NLP systems. In a cloud computing context, developing an NLP PaaS requires consideration of the following requirements: Straightforward deployment and sharing of NLP pipelines—How can we achieve this transparently for the NLP developer, i.e., NLP applications developed on a desktop machine can run without any adaptation on the PaaS. Furthermore, developers need to be able to share their NLP pipelines easily as SaaS, with on-demand scalability and robustness ensured by the underlying NLP PaaS; Efficient upload, storage, and sharing of large corpora—An NLP PaaS needs to provide users with a safe and efficient way to bulk upload, analyze, and download large text corpora, i.e., batch processing over large datasets. Moreover, users need to share their big text corpora between different NLP pipelines, running on the PaaS both for services bundled within the NLP PaaS and for services generated by the developers. Algorithm-agnostic parallelization—It is best to parallelize the execution of complex NLP pipelines that may contain arbitrary algorithms that are not all implemented or suitable for MapReduce and Hadoop. Load balancing—Determine the optimal number of virtual machines for running a given NLP application within the PaaS, taking into account the size of the collection of documents to be processed and the significant overhead of starting up new virtual machines on demand. Security and fault tolerance—As with any Web application, the NLP PaaS needs to ensure secure data exchange, processing, storage, and be robust in the face of hardware failures and processing errors.

The cloud architecture consists of two segments; frontend and backend, which refer to the separation of concerns between the presentation layer(frontend) and data access layer (backend). The backend contains additional components of infrastructure, data storage, cloud computing environment, and service interface/API while the frontend provides a medium for clients to connect with services. The infrastructure is made up of hardware components, data storage is normally databases for handling and storing data, cloud environment defines cloud deployment structure, service interface/API handles operations between client and cloud. The client accesses the API using the frontend over the internet. The visual representation of cloud architecture is given in Fig. 2.

In addition to these technical requirements, an NLP PaaS needs to offer comprehensive methodological support to underpin the life cycle of NLP applications: Build an

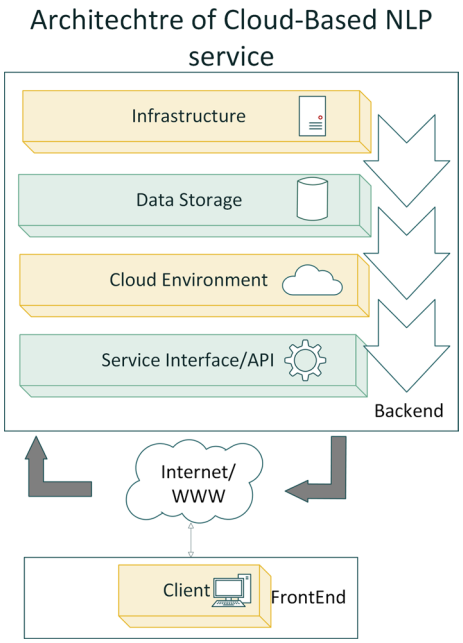

Fig. 2 The architecture of cloud-based NLP platform as a service

initial NLP pipeline prototype, test a small collection of documents, use an NLP application development environment, run on a regular desktop, or a local server computer; Crowd-source a gold-standard corpus for assessment and/or training, using a Web-based collaborative corpus annotation tool, deployed as a service on the PaaS; Evaluate the performance of the automatic pipeline on the gold standard (either locally within the desktop development environment or through the manual annotation environment on the cloud). Return to step 1 as required for further development and evaluation cycles; upload large datasets and deploy the NLP pipeline on the PaaS; run the large-scale text-processing experiment and download the results as XML, JSON, RDF, or `schema.org` formats. Optionally, an NLP PaaS could also offer scalable semantic indexing and search over the linguistic annotations and the document content; Lastly, analyze any errors and, if required, iterate again over the critical system development stages, either on a local machine or on the NLP PaaS. The depiction of the NLP platform as a service is given in Fig. 3.

Big data in cloud

Storing and processing big loads of data requires scalability, easiness-to-growth, fault tolerance, and availability. Cloud computing can provide all these features by virtualization hardware. Big data and cloud computing are two complementary concepts: Cloud can make big data available, scalable, and fault-tolerant.

Data lakes and data warehouses are significant parts of the big data sphere. According to the authors [63], a data lake is a flexible, scalable data storage and management system, which intakes and stores raw data from wide-ranging sources in their original format, and enables query processing and data analytics in an on-the-go format. While data

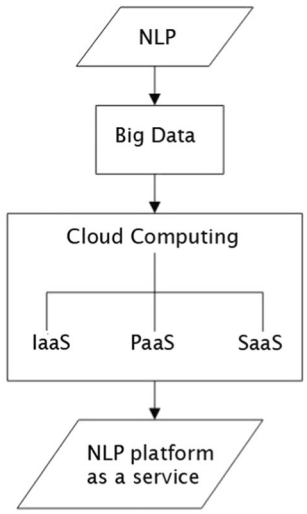

Fig. 3 Stages of NLP platform as a service

warehouse manages, integrates, and aggregates data from various sources, and supplies data analytics for decision making using multi-dimensional data cubes in data marts/data warehouses.

A new business sector has arisen with the rapid increase of data and the demand for new ways to store, process, and analyze data. Companies started to view the cloud and big data as valuable business opportunities. Several new companies have recently focused on providing big data as a Service (BDaaS) or DataBase as a Service (DBaaS). Companies such as Google, IBM, Amazon, and Microsoft also supply customers with ways to access big data on demand.

Big data not just address the ability to store large amounts of data, but also the means to extract and process knowledge from the stored data [64]. In practice, a big data database can contain structured and unstructured data that can overlap, vary, and have various volumes at different speeds. Big data features differ from other data in five aspects: volume, velocity, variety, value, and complexity.

Volume refers to the high load that big data typically handles with processing, and storing large amounts of data is challenging because it involves (among other things): scalability (vertical, horizontal, or both) to accommodate storage growth and drive the processing flow; flexibility that guarantees access to data and the means to perform data-related operations; and bandwidth and efficiency that ensure access to data at the right time.

Velocity refers to the different speeds of each data source. For example, an enterprise data warehouse (EDW) is typically updated once a day, while information is continuously updated from wireless sensor systems. To aggregate data from multiple data sources, Big Data must manage data that arrives at different speeds.

Variety relates to the different types of data from different sources that big data frameworks handle (typically, different sources produce different types of data). Big data is a way to bridge these differences and unify data. The Internet of Things is a topic related to big data that analyzes the data of individual objects in daily life that can be very varied:

Internet traffic, smartphones, wearable technologies, and others. To handle different types of data, Big Data must provide data type abstraction frameworks.

Value refers to the actual value of the data (i.e., the potential value of the data relative to the information it contains). Large amounts of data are worthless unless they add value to the user who discovers them.

Veracity refers to the reliability of the data (i.e., confidentiality, integrity, and availability of the data). The data is meaningless if the source is not reliable. Therefore, organizations must ensure that the data is correct and that the analysis performed on the data are correct.

Methods and technologies related to NLP and big data

Some common methods of NLP include lexical acquisition, word sense disambiguation, part-of-speech tagging, and probabilistic context-free grammar. Several NLP-based techniques have been applied to text mining, including information extraction, topic models, text summarization, classification, clustering, question answering, and opinion mining [65].

In linguistics, the term annotation most commonly refers to metadata used to describe words, sentences, or other metadata. The annotation process is often described as tagging, the automatic assignment of descriptors to input tokens. One prominent example is parts-of-speech tagging, which maps a natural language, like English, to a meta-language composed of word classes, like nouns, verbs, adverbs, or other metadata. Text segmentation is an essential step in any NLP process. Electronic text, in its raw form, is essentially just a sequence of characters. Therefore, it has to be divided into linguistic units. Such units include words, punctuation, numbers, alphanumeric, or other representations. This process is also referred to as tokenization. NLP tasks require linguistic entities (tokens) to be allocated, classes. Parts-of-speech is an essential linguistic concept in NLP, and parts-of-speech taggers are used to assign each word to syntactic categories (e.g., noun, verb, adjective, adverb, or other categories.). Software focusing on text mining and text analysis, such as the Apache UIMA project or General Architecture for Text Engineering (GATE), are designed to facilitate unstructured content analysis [66].

Syntax-centred NLP can be generally classified into keyword spotting, lexical affinity, and statistical methods. Keyword spotting is the most naïve approach and probably the most common because of its accessibility and economy. Lexical affinity is slightly more sophisticated than keyword spotting. It assigns to arbitrary words a probabilistic 'affinity' for a specific category instead of merely detecting apparent words. Since the late 1990s, Statistical NLP has been the dominant course of NLP science. It is based on language models based on popular machine learning algorithms such as maximum-likelihood, expectation-maximization, conditional random fields, and Support Vector Machines. Generally, statistical methods are semantically weak [1].

Web data analysis aims to automatically retrieve, extract, and evaluate information from cloud documents and services to discover useful knowledge. Web analysis is related to database, information retrieval, NLP, and text mining. Web data analysis can be divided into three related fields: mining of web content, mining of web structures, and mining of web use. Web content mining is the process of finding practical knowledge

from web pages that usually involves several types of data, such as text, image, audio, video, code, metadata, and hyperlinks. Web structure mining includes models for discovering web link structures. We use mining, and it is intended to explore auxiliary data generated through web dialogue or activities. The web search problem is divided into three components: web content gathering (crawling), inverted index building (indexing), and ranking documents are given a query (retrieval). Gathering web content and building inverted indexes are, for the most part, offline issues. Both need to be flexible and reliable, but they do not have to work in real-time. On the other hand, recovery is an online issue requiring a sub-second response time [67].

Sentiment analysis of social networking data such as Twitter involves much-preprocessing data. Because of the enormous volume, the highly unstructured nature of the data, and the tremendous rate at which it is produced, parallel implementations of preprocessing algorithms become necessary. Preprocessing involves the conversion of string to vector, removing irrelevant words and symbols if any, frequency mapping, or other preprocessing procedures. Preprocessing helps to reduce the data dimensionality, hence removing many unnecessary features and, in some cases, allowing data processing in memory, resulting in a significant reduction of I/O overhead [68].

Sentiment analysis aims to detect the attitude of text. A basic sub-task of sentiment analysis is defining text's polarity: positive, negative, or neutral. This can be seen as a classification task, and the Naïve Bayes algorithm is ideal for this classification because it produces good accuracy. The supervised learning methods have mainly used algorithms based on statistical Learning. Naïve Bayes, Support Vector Machine and Maximum Entropy are widely used algorithms. Such algorithms rely on tokenizing and quantifying patterns that are essential in the context of opinion mining. The Naïve Bayes algorithm employs probabilities to decide which class matches the best for a given input text. The Naïve Bayes method can provide good results for sentiment analysis than other supervised learning algorithms. In sentiment analysis, extraction techniques are widely used and include term presence and their frequency, part-of-speech information, negations, and opinion words and phrases [69].

In the big data environment, a key advantage of construction-based parsing is that only small parts of the text are required to extract meaning; word category information and the generally small size of constructions mean that the parser can still use error-filled or conventionally unseen text.

Massive parallel processing (MPP) structures such as Hadoop, NoSQL, or MPP databases, have been employed to support big data. Table 1, [70] compares some of the big data technologies in their strengths and weaknesses.

Hadoop / MapReduce techniques and clusters of Hadoop are very successful in processing large quantities of textual data [71]. S4 is an open-source, distributed, flexible, and partially tolerant platform for developing and running distributed programs for the continuous processing of data streams. Due to an unsuccessful attempt to adopt Hadoop, S4 was created to deal with applications consuming large data streams in real-time. S4 lacks a cluster balancing system, which unbalances the system over time. In-stream computing, the storm is an alternative in S4. It was designed to meet the needs of a distributed and scalable platform for real-time computing. Storm, like S4, handles

connectivity within the cluster using *ZooKeeper*. It automatically re-balances to compensate for the processing load between the nodes [72].

Natural language processing in big data

NLP has attracted increasing interest in the big data community, especially in designing systems to facilitate decision-making. Decision modelers often use textual sources to identify relevant information when creating suitable models. However, this is very time-consuming, as only a limited amount of information can be processed manually. Consequently, various NLP approaches have been introduced to bridge efficiency and massive data processing [73].

More specifically, it consists of computational techniques to evaluate and extract knowledge from textual sources for a range of tasks or applications through linguistic analysis. Its goal is to expand its methods to include any language, mode, or genre that humans use to understand better the patterns of information that emerge in human communication. NLP was initially referred to as natural language understanding (NLU), and although the NLP's ultimate goal is "true" NLU, more research is still needed to achieve that. The ability to logically infer conclusions from textual sources is yet being developed and improved to incorporate the richness of language in terms of imprecise knowledge, causality, and ambiguous meaning.

In an NLP system, the most basic level is based on lexical analysis, which deals with the words considered the atomic structure of text documents. In particular, it is the process that occurs when the essential components of a text are analyzed and grouped into tokens, which are sequences of characters with collective meaning. In other words, lexical analysis promotes the understanding of single words depending on the context in which they appear, which can relate to more than one concept. Consequently, the use of simplified lexical representations unifies the meaning across words to generate complex interpretations at a higher meta-level. The lexical analysis may need a lexicon, which usually consists of the specific approach used in a properly defined NLP system and the nature and extent of information inherent to the lexicon. Mainly, lexicons vary in terms of their sophistication as they may contain information about word-related semantic information. Furthermore, accurate and comprehensive sub-categorization lexicons are essential for the development of parsing technology and any NLP application that relies on the structure of information related to a predicate-argument structure. More research efforts have been tried aiming to provide better tools for analyzing words in semantic contexts [74].

In particular, the lexical analysis consists of different tasks, including (i) Lemmatisation, which gathers inflected forms of a word into a single item that corresponds to its lemma (or dictionary form); (ii) Part-of-speech tagging aimed at defining each word's syntactic role; and (iii) Parsing, which is the process to grammatically analyze a sentence, where each word's contribution is regarded as a whole, with the corresponding hierarchy.

The semantic analysis deals with a higher meta-level concerning objects associated with a lexicon. Semantic processing determines the possible meanings by examining the interactions between word-level meanings in the sentence. This approach may also include the semantic disambiguation of words with multiple senses, which

defines the meaning of ambiguous words to be included in the sentence's correct semantic representation. This is particularly relevant in any information retrieval and processing system based on ambiguous and partially known knowledge. Practical applications of NLP can be seen in inferencing techniques where additional information obtained from a broader context effectively addresses statistical properties between concepts within textual sources.

Information extraction via NLP techniques within big data

One of the most investigated data types in big data involves those without a well-defined structure, unstructured. Textual data fall within this category, and NLP techniques allow relevant information to be identified and evaluated between concepts embedded in textual data sources [75]. The primary objective in defining relationships between concepts, which is crucial in any decision-making and information extraction process, is to select the correct concepts and the type of links that connect them. For instance, similar concepts (e.g., "disease" and "illness"), or those at different lexical levels, need to be appropriately analyzed and selected to provide the appropriate amount of information. Also, contradictory information is often found in large textual datasets, as it is common to find opposite relationships in the same context [76].

Different approaches can be used, including statistically dependent methods focused on frequency, co-occurrence, and other indicators to evaluate the overall behavior of the information embedded in texts. Another commonly used approach focuses on the grammatical and syntactic roles of each textual fragment's different components. In [77], the authors use a grammar-based technique to identify influence relationships between concepts by considering the triples (NP1, VB, NP2) where: NP1 and NP2 are noun phrases, for example, phrases with a noun as their headword, which have one or more concepts to contain; VB is the linking verb, which needs to be associated with an influence type of relation.

One component of the NLP, which has been investigated extensively, focuses on sentiment analysis aimed at detecting "opinions" or polarity from textual data sources [78]. It can be instrumental in supporting the extracted specific information. If the overall opinion related to a particular context is "positive," it may imply that the corresponding information is addressed in positive terms [79].

Machine learning-based decision modeling in big data

The driving force behind machine learning in big data is its ability to perform intelligent automation. Moreover, it allows the discovery of secret patterns, industry dynamics, and consumer preferences in a fraction of time, with greater precision over a human counterpart or conventional data analytics model. To apply machine learning techniques in big data processing is crucial to understand the strengths and weaknesses of the different methods to ensure that the most suitable approach is used to solve the specific problem. There are three principal methods, broadly speaking: supervised learning, unsupervised learning, and reinforcement learning. Each group has its unique strengths and weaknesses, creating no ideal single solution. Machine

learning techniques can be used to improve big data analytics in conjunction with traditional methods.

Many real-world problems fall into this field as it can be expensive to use experts in a specific area to mark an entire dataset. The semi-supervised Learning methodology combines both Supervised and Unsupervised Learning that addresses issues that include both labeled and unlabelled data. Unsupervised learning discovers the data structure, while supervised learning creates best guess predictions for the unlabelled data [80]. An example of the implementation of Semi-Supervised Learning is the Python VADER sentiment analysis tool, which assesses the sentiment polarity of each word on social media platforms [81].

The reinforcement learning methodology is also widely used, using sophisticated algorithms to take action based on its current state. Popular reinforcement Learning methods include the Markov decision process (MDP) and a neural network-based NEAT (NeuroEvolution of Augmenting Topologies) [82]. The ability to produce intelligent analytics makes machine learning well placed to address many big data challenges. Machine learning is not limited to one type of data, and its highly versatile analytical process can lead to rapid decision-making assessments and processes.

Challenges of big data and NLP

Big data technologies provide diverse solutions for storing data. Since the process to make the data usable is a pain-staking process and requires additional setup, this extra task makes the NLP application more complicated. Such as, in the case of data lakes, data is stored without any prior processing which can be difficult to integrate into the NLP case. While data warehouse includes extract, transform and load (ETL) stage, which makes it easier to access, and utilize, the limitation of data access and data storage options also restricts the possibilities of various applications. In regards to NLP-related situations, a few problems are discussed below. Keyword search is a classic approach for handling text data on a computer and a primary method in text mining. Other search tools include ontologies and taxonomies, Bayesian classifiers, clustering methods for documents. Bayesian classifiers use probability algorithms to determine the significance of a document by weighing particular words or phrases based on their frequency [83]. Searching for keywords includes developing a list of potentially relevant terms and phrases and then searching for text data for occurrences of those words. The reverse is also true: it is not guaranteed the importance of a document to contain a keyword. Keyword searches always match only exact strings; it is impossible to return words with spelling errors or inflected words. The use of boolean operators and fuzzy search techniques enables flexibility; it is still a search for keywords. Although searching for keywords or key phrases is useful, this approach is far from perfect. If the search terms are too narrow, vital information may be overlooked; if it is too broad, the resulting set of 'hits' could include large numbers of completely irrelevant 'false positives [84].

Automatic Parts-of-Speech taggers have to tackle many challenges, including the ambiguity of word types in their parts-of-speech, and classification problems due to the ambiguity of periods (.), which can either be interpreted as a part of a token (for instance, abbreviation), punctuation (full stop), or both [66].

Mining opinions and sentiments from natural language is a challenge because it requires a detailed understanding of the rules of language that are clear and implicit, regular, irregular, syntactic, and semantic. Sentiment analysis researchers struggle with unresolved problems of NLP: coreference resolution, negation handling, anaphora resolution, named-entity recognition, and word-sense disambiguation. Opinion mining is a very restricted NLP problem because the system needs only to understand the positive or negative feelings of each sentence and the target entities or topics [85].

Significant progress has been made in delivering NLP technologies such as extracting information from big unstructured data on the web, sentiment analysis in social networks, or grammatical analysis for grading essays. However, efficiently extracting potentially helpful information and understanding unstructured clinical notes (text) in the proper context remain challenges [86].

Syntactic parsing of natural language sentences is a central task in NLP due to its significance in mediating between linguistic expression and meaning. There are two common shortcomings in standard approaches [87].

Simplifying language assumptions, people often develop an algorithm in NLP and machine learning and then push the data into a compatible format with this algorithm. For example, a common first step in text classification or clustering is representing texts in terms of unordered lists of words (ignoring word order and grammatical structure), a so-called bag of words. This leads to obvious problems when trying to understand a sentence.

In feature representation, most learning systems' efficiency depends crucially on the input's feature representation. Each of these features took a long time to develop and integrate for each new task, slowing down the final algorithm's development and run-time.

Stream processing is complicated in MapReduce, although Hadoop/Mapreduce techniques and Hadoop clusters are instrumental in processing large quantities of textual data [71]. The big data problem influences NLP and the main challenge is not only the collection and scale of documents but also the heterogeneous nature of language [88]. There are many open research challenges related to developing linguistic resources for different languages and covering multilingual and cross-lingual settings. Another important direction of research is guaranteeing the scalability of methods as it is becoming common to deal with big data [89].

Conclusion

Cloud services for Natural Language Processing have complete attention among both academia and industry. It enables researchers to deploy, share, and use language processing components and resources, following the data as a service and software as a service paradigms. In this paper, we performed a survey on NLP in cloud computing. We presented a study on NLP in cloud computing and introduced an overview of NLP by discussing different levels of NLP and components of NLG followed by applications of NLP. This paper also discusses cloud computing with an emphasis on cloud services for NLP. Furthermore, we have highlighted cloud architectural layers and their deployment models. Moreover, we have presented a motivation for the need for NLP as a service in the cloud. A set of technical requirements are identified for developing a platform

as a service. It is aimed at helping researchers to carry out data-intensive text-processing experiments in the cloud. This study also focused on big data in the cloud. We have outlined the methods and technologies related to NLP and big data. Following that, we discussed NLP and machine learning in big data, and information extraction via NLP techniques within big data are discussed. At last, challenges of NLP and big data are presented. Our aim through this paper is to be a source of inspiration for researchers interested in using NLP and building a cloud platform for NLP services.

As for future work, we aim to build an NLP Platform as a service providing easier administration, automatic update and patch management, reliability, user-friendliness, easier collaboration, and global accessibility in the form of SaaS. The research is still in its early stages, and the evaluation scenarios will be elaborated on further soon.

Acknowledgements
This work was supported by National Founding from the FCT Fundação para a Ciência e a Tecnologia, through the MOVES Project—PTDC/EEI-AUT/28918/2017, and by Operação Centro-01-0145-FEDER-000019—C4—Centro de Competências em Cloud Computing, co-financed by the Programa Operacional Regional do Centro (CENTRO 2020), through the Sistema de Apoio á Investigação Científica e Tecnológica—Programas Integrados de IC&DT.

Author contributions
All authors had the same contribution. All authors read and approved the final manuscript.

Authors' information
Sebastião Pais is currently a Professor at the Computer Science Department, the University of Beira Interior (UBI). He holds a Ph.D. from MINES ParisTech, Paris. He is the responsible researcher of MOVES and HULTIG-C scientific research projects, University of Beira Interior. His research and teaching interests are artificial intelligence, statistical natural language processing, lexical semantics, machine learning, and unsupervised and language-independent methodologies.
João Cordeiro is currently a Professor at the Computer Science Department of the University of Beira Interior (UBI). He holds a Ph.D. from the University of Beira Interior, Portugal. He is ahead of the Centre for Human Language Technology and Bioinformatics (HULTIG), University of Beira Interior, and the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto. His main research interests are directed toward the areas of natural language processing, automatic text summarization, information extraction, information retrieval, text mining, sentiment analysis, and automatic machine translation.
Muhammad Luqman Jamil is studying master's in informatics engineering at Universidade Beira Interior (UBI). He is a research fellow and collaborates with two research projects hosted by UBI, MOVES, and HULTIG-C. His thesis focuses on natural language processing, deep learning, and big data.

Funding
This work was supported by National Founding from the FCT Fundacão para a Ciência e a Tecnologia, through the MOVES Project- PTDC/EEI-AUT/28918/2017, and by Operação Centro-01-0145-FEDER-000019—C4—Centro de Competências em Cloud Computing, co-financed by the Programa Operacional Regional do Centro (CENTRO 2020), through the Sistema de Apoio á Investigação Científica e Tecnológica—Programas Integrados de IC&DT.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
The authors declare that they have no known ethics issue that could have appeared to influence the work reported in this paper.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no known competing financial interests or personal relationships or conflicts of interest that could have appeared to influence the work reported in this paper.

Author details
[1]Department of Computer Science, University of Beira Interior, Covilha, Portugal. [2]NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Costa da Caparica, Portugal. [3]GREYC, Groupe de Recherche en Informatique, Image et Instrumentation de University of Caen Normandie, Caen, France. [4]INESC-TEC, Instituto de Engenharia de Sistemas e Computadores (INESC), Porto, Portugal.

Received: 30 October 2021 Accepted: 6 April 2022
Published online: 28 April 2022

References

1. Cambria E, White B. Jumping NLP curves: a review of natural language processing research. IEEE Comput Intell Mag. 2014;9(2):48–57.
2. Dale R. Nlp meets the cloud. Nat Lang Eng. 2015;21(4):653–9.
3. Lamba HS, Singh G. Cloud computing future framework for e-management of ngo's. arXiv:1107.3217 [Preprint]. 2011.
4. Singh G, Sood S, Sharma A. Cm-measurement facets for cloud performance. Int J Comput Appl. 2011;23(3):37–42.
5. Amazon: Amazon Comprehend. 2022. https://aws.amazon.com/comprehend/.
6. Microsoft: Azure Cognitive Services. 2022. https://azure.microsoft.com/en-us/services/cognitive-services/.
7. Cloud G. Natural Language AI. 2022. https://cloud.google.com/natural-language.
8. diffbot: Structure and Understand Natural Language. 2022. https://www.diffbot.com/products/natural-language/.
9. monkeylearn: No-code text analytics. 2022. https://monkeylearn.com/.
10. Liddy ED. Natural language processing. 2001.
11. Friedman C, Johnson SB. Natural language and text processing in biomedicine. In: Springer (ed.) Biomedical Informatics, 2006;312–343.
12. Feldman S. Nlp meets the jabberwocky: natural language processing in information retrieval. ONLINE-WESTON THEN WILTON. 1999;23:62–73.
13. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. arXiv preprint arXiv:1708.05148 2017.
14. Copestake A. Natural language processing: part 1 of lecture notes. Cambridge: Ann Copestake Lecture Note Series; 2003.
15. Zajic DM, Dorr BJ, Lin J. Single-document and multi-document summarization techniques for email threads using sentence compression. Inf Process Manag. 2008;44(4):1600–10.
16. Fattah MA, Ren F. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. Comput Speech Lang. 2009;23(1):126–44.
17. Wan X. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. Inf Retr. 2008;11(1):25–49.
18. Ouyang Y, Li W, Li S, Lu Q. Applying regression models to query-focused multi-document summarization. Inf Process Manag. 2011;47(2):227–37.
19. Riedhammer K, Favre B, Hakkani-Tür D. Long story short-global unsupervised models for keyphrase based meeting summarization. Speech Commun. 2010;52(10):801–15.
20. Wang D, Zhu S, Li T, Gong Y. Multi-document summarization using sentence-based topic models. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009;297–300. Association for Computational Linguistics
21. Wang D, Zhu S, Li T, Chi Y, Gong Y. Integrating document clustering and multidocument summarization. ACM Trans Knowl Discov Data (TKDD). 2011;5(3):14.
22. Fang H, Lu W, Wu F, Zhang Y, Shang X, Shao J, Zhuang Y. Topic aspect-oriented summarization via group selection. Neurocomputing. 2015;149:1613–9.
23. Iman K, Mohammad S. A metric-based approach for web-based question answering. Int J Inf Technol Comput Sci. 2014;9:39–45.
24. Moschitti A, Vergata T. Natural language processing and automated text categorization: a study on the reciprocal beneficial interactions. 2003.
25. Prabowo R, Thelwall M. Sentiment analysis: a combined approach. J Inf. 2009;3(2):143–57.
26. Saif H, He Y, Alani H. Semantic sentiment analysis of twitter. In: International Semantic Web Conference, 2012;508–524. Springer
27. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Comput Linguist. 2011;37(2):267–307.
28. Sharma S. Application of support vector machines for damage detection in structures. Diss. Worcester Polytechnic Institute. 2008.
29. Cearley DW. Cloud computing: key initiative overview. Gartner Report, 2010.
30. Mell P, Grance T. The NIST definition of cloud computing. 2011.
31. Foster I, Zhao Y, Raicu I, Lu S. Cloud computing and grid computing 360-degree compared. arXiv preprint arXiv:0901.0131 2008.
32. Cheng D. Paas-onomics: A cio's guide to using platform-as-a-service to lower costs of application initiatives while improving the business value of it. Technical report: Tech. rep., LongJump; 2008.
33. Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I. Above the clouds: A Berkeley view of cloud computing. Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS 2009;28(13), 2009.
34. Rothon J. Cloud computing explained: implementation handbook for enterprises (2 Kindle ed.). London: Recursive Press; 2009.
35. systran: SYSTRAN.io - Translation and NLP API Documentation (systran)—RapidAPI. 2020. https://rapidapi.com/systran/api/systran-io-translation-and-nlp.
36. aylien: AYLIEN®Text Analysis API—Natural Language Processing API. 2020. https://rapidapi.com/aylien/api/text-analysis.
37. text analysis: Text Summarization API Documentation (textanalysis)—RapidAPI. 2020. https://rapidapi.com/textanalysis/api/text-summarization.
38. twinword: Twinword Text Analysis Bundle API Documentation (twinword)—RapidAPI. 2020. https://rapidapi.com/twinword/api/twinword-text-analysis-bundle.
39. Turian J. Using alchemyapi for enterprise-grade text analysis. AlchemyAPI: Denver, CO, USA; 2020.
40. RxNLP: Text Mining and NLP API. 2020. https://rapidapi.com/RxNLP/api/text-mining-and-nlp/details.

41. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014;55–60
42. text processing: Natural Language Processing APIs and Python NLTK Demos. 2020. http://text-processing.com/.
43. atrilla: nlpTools—Natural Language Processing Toolkit for PHP. 2020. http://www.nlptools.atrilla.net/web/.
44. enclout: Stemmer API: how to use the API. 2020. https://rapidapi.com/collection/natural-language-processing-api.
45. Urbansky D, Thom JA, Feldmann M. Webknox: Web knowledge extraction. In: Proceedings of the Thirteenth Australasian Document Computing Symposium, 2008;27–34. Citeseer
46. MeaningCloud: Text Analytics—MeaningCloud text mining solutions, 2020. https://www.meaningcloud.com/.
47. API, F.: Fluxifi API—ProgrammableWeb. 2020. https://www.programmableweb.com/api/fluxifi.
48. Fog M. Cloud NLP API. 2020. https://www.programmableweb.com/api/fluxifi.
49. Gamallo P, et al. Linguakit: a big data-based multilingual tool for linguistic analysis and information extraction. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018. IEEE.
50. Lexalytics: Semantria Cloud API Text & Sentiment Analysis—Lexalytics. 2020. https://www.lexalytics.com/semantria
51. Dale R. Text analytics apis, part 2: the smaller players. Nat Lang Eng. 2018;24(5):797–803.
52. Tablan V, Roberts I, Cunningham H, Bontcheva K. Gatecloud. net: a platform for large-scale, open-source text processing on the cloud. Philos Trans R Soc A Math Phys Eng Sci. 2013;371(1983):20120071.
53. Lexalytics: Data analytics with NLP and text analytics. 2020. https://www.lexalytics.com/.
54. Analytics A. Amenity analytics—NLP Text Analytics & Mining Software for Finance. 2020. https://www.amenityanalytics.com/.
55. TEXT2DATA: Introducing sentiment analysis and text analytics add-in for excel. 2020. https://text2data.com/Excel.
56. bigml: BigML. 2020. https://bigml.com/.
57. Cloud G. Cloud prediction API is deprecated. 2019. https://cloud.google.com/prediction/.
58. Technologies E. natural language processing/machine learning B2B software platform. 2022. https://eigentech.com/.
59. myrrix: myrrix API. 2019. http://myrrix.com.
60. nlpcloud: NLPCloud.io, 2022. https://nlpcloud.io/.
61. salesforce: Salesforce cloud services. 2020. https://www.salesforce.com.
62. VMware: AYLIEN®Text Analysis API | Natural Language Processing API. 2020. https://www.vmware.com/.
63. Hai R, Quix C, Jarke M. Data lake concept and systems: a survey. CoRR **abs/2106.09592** 2021. arxiv:2106.09592.
64. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU. The rise of "big data" on cloud computing: review and open research issues. Inf Syst. 2015;47:98–115.
65. Chen M, Mao S, Liu Y. Big data: a survey. Mobile Netw Appl. 2014;19(2):171–209.
66. Holzinger A, Stocker C, Ofner B, Prohaska G, Brabenetz A, Hofmann-Wellenhof R. Combining hci, natural language processing, and knowledge discovery-potential of ibm content analytics as an assistive technology in the biomedical field. In: International Workshop on Human–Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, 2013;13–24. Springer.
67. Lin J, Dyer C. Data-intensive text processing with mapreduce. Synth Lect Hum Lang Technol. 2010;3(1):1–177.
68. Nirmal VJ, Amalarethinam DG. Parallel implementation of big data pre-processing algorithms for sentiment analysis of social networking data. Int J Fuzzy Math Arch. 2015;6(2):149–59.
69. Jaswant U, Kumar P. Big data analytics: a supervised approach for sentiment classification using mahout: an illustration. Int J Appl Eng Res. 2015;10(5):13447–57.
70. Dean J. Big data, data mining, and machine learning: value creation for business leaders and practitioners. US: Wiley; 2014.
71. van Banerveld M, Le-Khac N-A, Kechadi M-T. Performance evaluation of a natural language processing approach applied in white collar crime investigation. In: International conference on future data and security engineering, 2014;29–43. Springer.
72. Artola X, Beloki Z, Soroa A. A stream computing approach towards scalable nlp. In: LREC, 2014;8–13.
73. Sanchez-Graillet O, Poesio M. Acquiring bayesian networks from text. In: LREC 2004.
74. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press; 2007.
75. Manning C. Generating typed dependency parses from phrase structure parses 2008.
76. Trovati M, Hayes J, Palmieri F, Bessis N. Automated extraction of fragments of bayesian networks from textual sources. Appl Soft Comput. 2017;60:508–19.
77. Trovati M, Bessis N, Huber A, Zelenkauskaite A, Asimakopoulou E. Extraction, identification, and ranking of network structures from data sets. In: 2014 Eighth international conference on complex, intelligent and software intensive systems, 2014;331–337. IEEE.
78. Liu B. Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol. 2012;5(1):1–167.
79. Ray J, Trovati M. A survey of topological data analysis (tda) methods implemented in python. In: International conference on intelligent networking and collaborative systems, 2017;594–600. Springer.
80. Inoubli W, Aridhi S, Mezni H, Maddouri M, Nguifo EM. An experimental survey on big data frameworks. Fut Gener Comput Syst. 2018;86:546–64.
81. Hutto CJ, Gilbert E. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media. 2014.
82. Stanley KO, Miikkulainen R. Evolving neural networks through augmenting topologies. Evol Comput. 2002;10(2):99–127.
83. Crabb ES. "Time for some traffic problems": enhancing e-discovery and big data processing tools with linguistic methods for deception detection. J Digit Forens Secur Law. 2014;9(2):14.
84. Khan E. Addressing big data problems using semantics and natural language understanding. In: 12th Wseas International Conference on Telecommunications and Informatics (Tele-Info '13), Baltimore. 2013.

85. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst. 2013;28(2):15–21.
86. Priyanka K, Kulennavar N. A survey on big data analytics in health care. Int J Comput Sci Inf Technol. 2014;5(4):5865–8.
87. Socher R. Recursive deep learning for natural language processing and computer vision. PhD thesis, Citeseer. 2014.
88. Cheptsov A, Tenschert A, Schmidt P, Glimm B, Matthesius M, Liebig T. Introducing a new scalable data-as-a-service cloud platform for enriching traditional text mining techniques by integrating ontology modelling and natural language processing. In: International Conference on Web Information Systems Engineering, 2013;62–74. Springer.
89. Mladenić D, Grobelnik M. Automatic text analysis by artificial intelligence. Informatica, 2013;37(1).

Publisher's Note