Analysis of Machine Learning as a Service

1st Gokay Saldamli
Computer Engineering Department
San Jose State University
San Jose, CA 95192, USA

2nd Nishit A. Doshi
Computer Engineering Department
San Jose State University
San Jose, CA 95192, USA

3rd Vishal Gadapa
Computer Engineering Department
San Jose State University
San Jose, CA 95192, USA

4th Jainish J. Parikh
Computer Engineering Department
San Jose State University
San Jose, CA 95192, USA

5th Mihir M. Patel
Computer Engineering Department
San Jose State University
San Jose, CA 95192, USA

6th Levent Ertaul
Department of Computer Science
California State University East Bay
Hayward, CA 94542, USA

***Abstract*—Machine Learning as a Service is a set of services that allows the Machine Learning models to run on the cloud using ready-made, easily configurable tools. The demand for MLaaS is increasing day by day as Machine Learning revolves around data processing, algorithms, and computational power, which further requires a highly skilled workforce. The market value of the industry is likely to grow from USD 1.0b to USD 8.48b within the next five years according to the study conducted by Modor Intelligence. Machine Learning as a Service offers services for data transformation, predictive analysis, data visualization, and advanced machine learning algorithms for the same. Cost-effectiveness and faster product delivery make MLaaS one of the most in-demand cloud-based services.MLaaS being a relatively new technology requires many factors to be considered such as the accuracy of the ML model, availability and cost of the service before choosing any cloud provider. Users get overwhelmed while selecting one cloud service provider over another because of the high volume of the vendors. Accordingly, there is a need for research to find the most suitable provider based on the requirements.In this paper, we propose a comparison of MLaaS provided by different cloud vendors to identify the most efficacious one. Our approach includes a thorough analysis on Natural language Processing APIs to draw conclusions on cost, time, accuracy, and ease of use of these service providers. The outcome will be better and cost-effective decision making for users in a lesser amount of time.**

***Index Terms*—Cloud computing, Natural language processing, Machine Learning, Machine Learning as a Service, Sentiment Analysis, Key Phrase Detection, Named Entity Recognition, Text Classification**

I. Introduction

In the last few decades, the field of Machine Learning and Artificial Intelligence has seen massive advancements. The rise in the capabilities of computers and the development of numerous Machine Learning Algorithms have driven this movement. Among many application domains of Machine Learning, Natural Language Processing is one of the most widely used domain. The domain of Natural language processing includes context categorization, sentiment analysis, topic discovery and modeling, contextual extraction, document summarization, speech-to-text and text-to-speech conversion, machine translations. With a large influx of textual data and machine learning users, the demand for NLP has seen huge growth as well.

The other novice term that has emerged in the last decade is Cloud Computing. Cloud providers give easy to use, cost-effective, and highly configurable services to users which have drastically changed the way the software industry operates. This movement has led to the convergence of Machine Learning and Cloud Computing which gave birth to Machine Learning as a Service. The term essentially means providing a configurable and cost-effective way of using the power of machine learning without massive investment in labor and capital.

The resources required to independently manage the Machine Learning ecosystem are far from rudimentary and without domain experts, the process of building a functioning Machine Learning system is significantly difficult. However, MLaaS solves these problems by providing ready to use machine learning services. With minimalistic knowledge of the domain and a little capital investment, an organization or an individual can relish the benefits of Machine Learning. The "Big Four" in the race of MLaaS are Google, Amazon, IBM, and Microsoft with each of them providing a variety of services. However, the tools and services of MLaaS are a black-box for many, and choosing the right provider which suits their requirements is still an obstacle for them.

Given the plethora of choices in deciding an MLaaS which includes taking into account costs, accuracy, ease of use, time, and efficacy of the services, a customer with minimalistic knowledge of the domain could face many dilemmas. Some cloud providers allow full customizability of models, while some do not even reveal the models used by them. In addition to this, they also provide API endpoints to use their service, however, the integrations of these endpoints to the user's system is a conundrum. Requirements vary from customer to customer and questions like, what kind of ML models would fit the requirements? What customizations or feature tunings are needed? How much cost would it incur to use MLaaS models? How much knowledge would be needed to operate MLaaS? What are the accuracy, fault tolerance, and availability of these services? Lack of documentation which would answer these questions satisfactorily is enough to stop the customer from leveraging the power of MLaaS.

In this paper, we propose a comparison of Natural Language Processing APIs provided by these Cloud Vendors. The "Big Four" provides API endpoints to use NLP services with a few of them offering options for configurations and customizability with a different set of applications. We have based our analysis on four factors namely Cost, Accuracy, Time, and Ease-of-use.

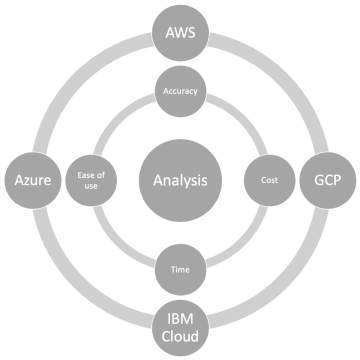

Fig. 1. Comparison matrices

- **Cost** refers to the cost incurred on a set of API calls.
- **Time** indicated time taken for a set API calls to be completed.
- **Accuracy** refers to the accuracy of the ML model used by the provider for a particular domain of data.
- **Ease-of-us** indicates the efforts required in integrating MLaaS in their application.

We use a different collection of datasets from various domains and test models on the basis of these four parameters. This paper aims to help customers make a well-informed decision on choosing the right cloud vendor for their needs and leveraging the power of MLaaS to its full potential.

II. Motivation

Enterprises have started understanding the potential of Machine Learning and Cloud Computing in recent years. The bulk of the businesses have started opting for the "Cloud Way" for running their infrastructure instead of building bare metal services. Machine Learning is becoming the core of numerous ventures and, senior corporations have started incorporating Machine Learning in their infrastructure. However, administrating and handling such an infrastructure is a cumbersome task and involves high domain knowledge.

Machine Learning as a Service integrates ML and Cloud computing to provide an affordable, scalable, and convenient way to use the power of Machine Learning in their stack. It provides remotely managed, secured, and well-integrated Machine Learning tools on demand. Although, there are a plethora of choices to choose from while using these services which can be overwhelming for MLaaS users. One solution does not work for all hence, cloud vendors provide certain customization options on top of fundamental Machine Learning services. Selecting a cloud provider and an MLaaS which suits their needs is a critical choice as these choices define the capital and time investments that would decide the future of organizations. Hence, knowing which MLaaS provider supports and offers services required by them and selecting the right MLaaS is of paramount importance.

Since MLaaS is an amalgamation of a vast set of machine learning services, we aim to provide an in-depth comparison of Natural Language Processing (NLP) services provided by major cloud providers. Cloud Vendors provide API endpoints to access their pre-trained ML models. In addition to that customization options to retrain models are also provided to support varying needs. This comparison would help MLaaS users in understanding the NLP services and customizations available and hence would aid in making a well-informed decision.

A. Goals

Which platform is the best for starting off with your first implementation of AI and Machine Learning solutions? There can't be one universal answer to this open-ended question, but through the research in this study, it is aimed to provide some conclusive analysis to choose the best MLaaS platform for different use cases.

There are myriads of AI solutions available and their numerous implementations. However, the easiest way would be to use the MLaaS solution provided by the cloud service vendors. This study aims to analyze the MLaaS solutions provided by the top four businesses - Google, AWS, IBM, and Azure using the most widely accepted Machine Learning application - Natural Language Processing. With well-defined measurement criteria including parameters like cost, time, accuracy, and ease of use it will be possible to find the best possible answer to the question raised at the start of this chapter.

III. Background and related work

MLaaS is one of the most demanding services in the cloud field due to its out-of-box, ready-to-use, and cost-effective Machine Learning solutions. In simple terms, MLaaS is the way to use cloud providers' computing power for your benefit just like any other cloud computing service. This cuts down the overhead of setting up an infrastructure needed for running heavy ML models which eventually saves a lot of time for the businesses.

In the paper [1], authors propose an efficient and scalable way to create machine learning as a service. This paper proposes a novel approach for machine learning, providing a scalable, flexible, and non-blocking platform as a service based on the service component architecture. By taking advantage of service-oriented architecture, the proposed approach becomes easily scalable and easy to adapt by adding, removing, changing, and linking any component. This also makes the system more flexible for handling multiple data sources and different machine learning algorithms at the same time.

The approach of building Machine Learning as a service can be divided into three phases. First of which is Building, where the user asks to build the model providing the parameters for tuning the model. The next phase is the Training phase

where the tuned model is trained on the data provided by the consumer. This phase consists of training the model over the data, Validating the model, and generating reports about the process. The third and last phase is prediction where the model is provided data and it predicts. Thus, in paper [1] authors provide an architecture to create a Machine Learning as a Service.

Various Big Organizations like Google, Amazon, Microsoft provides Machine Learning as a Service. They provide numerous capabilities to configure and fine-tune the models and use them to make predictions. Although all of them provide Machine Learning as a Service, they differ in many aspects such as the number and types of models (i.e classifiers, clustering algorithms, regression algorithms) costs of API calls, the functionality of pre-processing data, etc. In paper [2] authors provide a detailed comparison of services provide by these vendors in binary classification. Among the myriad of services provided, the focus of the paper is on the service of binary classification provided by them. It discusses how minutely a machine learning model can be configured and try to compare how does a Machine Learning model provided as a service, perform in front of a model made from scratch. They compare the accuracy, feasibility, and effectiveness of the service as these models are a black box when compared to a Machine Learning model made from scratch. Although these models are highly configurable, they would never be able to provide the fine-tuning a model made from scratch can provide. However, these services provide ease of use and cost-effectiveness which is a major issue when it comes to building Machine Learning models. Few of the major questions they tackle are as follows:

1. How does the complexity of ML system correlate with model accuracy?
2. Can increased configuration options lead to higher risks?
3. Which key knobs have biggest impact on the performance?
4. Can we design a generalized technique to optimize the knobs?

Key conclusions from the paper [2] are a tradeoff between ease of use and user-control, a near-optimal result by choosing the model wisely, and a fully automated (black-box) model outperforms the model with default settings but lags when compared to a fully tuned model. The paper [2] also shows that the performance of the model increases with the increase in the user-control and an initial experiment with a small random data can reduce the complexity of choosing a classifier. Fully automated services provided by vendors like Google use internal testing to increase the accuracy of the model. While these models perform better than their default counterparts, but they are far behind in the race when compared to fully configurable models provided by Microsoft, PredictionIO, and local scikit-learn.

One of the biggest concerns while testing the ML models is the accuracy of the trained model. Many factors contribute to the accuracy of any model such as data size, data quality, preprocessing algorithm, train and test data distribution, and parameters fine-tuning. The paper [5] trained the AWS ML model on a Banking dataset to test its accuracy in two ways, first using a single input and then on a batch input. The research concludes that by fine-tuning the parameters, accuracy of over 90% can be achieved.

Though the cloud platforms provide an easy-to-use solution to train ML models, not all the providers take utmost care when it comes to protecting the privacy of the data. The papers [3] and [4] shed some light on how privacy is an important factor that needs to be considered while determining a particular service. The researchers performed encryption on datasets like Crab Dataset, Fertility Dataset, and Climate Dataset to conclude how encrypting the data from the cloud provider's end can help in protecting privacy.

IV. Understanding of MLaaS

MLaaS are a set of Machine Learning Services that leverage the power of Cloud Computing. Clients of MLaaS do not have to worry about building Machine Learning infrastructure from bare metal and pay per usage. Cloud vendors host data centers and create infrastructure on which they provide various services. These services are accessible to clients through a web interface and have options to either directly integrate these Machine Learning services into their application or use cloud vendors' platforms to build personalized Machine Learning services to satisfy their requirements.

Among many cloud vendors which offer MLaaS, the most popular and recognized are Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, and IBM Cloud. These cloud vendors offer a wide range of MLaaS services with varying functionalities and choices to customize their services. These services are offered as a web interface where services can be subscribed to and used on the go. The control over the pipeline from providing Data and customizability varies from provider to provider. Following is a detailed explanation of the features offered by these four mainstream cloud vendors.

A. Amazon Web Services (AWS)

Amazon Web Services (AWS) is a heavily used cloud computing platform because of its wide range of services including On-Demand computing, Internet of things, Media services, and Machine Learning, to name a few. Especially, the ML services hosted by AWS can be leveraged by a novice user who wants to get his hands dirty in this field or wants to integrate some intelligence into his application without having much hassle, to an expert level ML user who knows how to build, train, and tweak ML models but has limited on-prem resources. AWS provides models that are ready to deploy on the cloud just by a single click as well as Integrated Development Environment to build, train, test, and deploy models on the cloud. It has a high-performance, cost-effective, and scalable infrastructure for heavy ML models to run. One of the major components of AWS's ML stack is its Amazon SageMaker. This comes in many flavors such as Studio, a

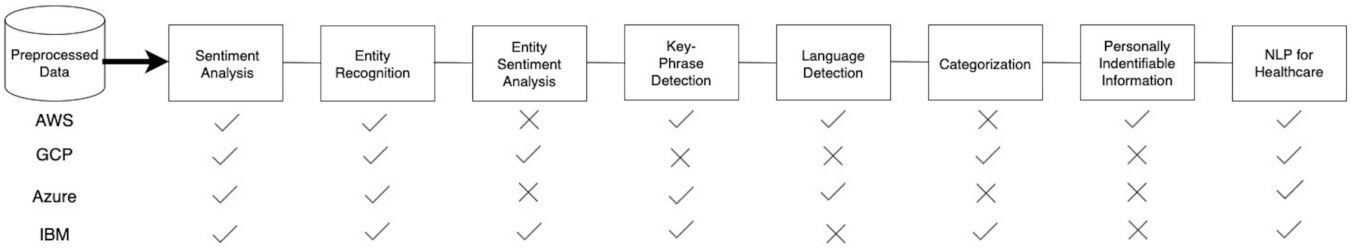

Fig. 2. Platforms and NLP Features

cloud IDE for building and training ML models, Autopilot, an automated ML model which is suited for user's data out of the box, and JumpStart, an easy interface to deploy models in one click from the common library of ML models. Apart from this, AWS also has special services for Computer Vision and Natural Language Processing.

Since NLP has seen huge growth in the past years, AWS has also focused on providing targeted services in this field. Amazon Textract is a service that adds functionalities like text detection and analysis to your application. It can also detect typed and handwritten texts from documents. Amazon Transcribe is an advanced pre-trained ML algorithm that detects speech from audio and video files and generates text from it. Amazon HealthLake is HIPPA enabled data storage, monitor, and analysis service that is capable of handling petabytes of data at once specifically for the Healthcare domain. It has integrated medical NLP services that make it one of the favorite choices for the Healthcare field. All these services are available in the form of APIs but Amazon Studio is all in one solution to build, train, and deploy our own ML models.

One of the widely used AWS NLP services is Amazon Comprehend which is an API based solution for the tasks including Entity Recognition, Key Phrase Detection, Personally Identifiable Information (PII) Detection, Language Detection, Sentiment Analysis, Syntax Generator, Topic Modeling.

Just hitting a single API of Amazon Comprehend with the user data, the above-mentioned analysis can be generated. It is also possible to customize the pre-trained models to adjust them according to the user's need and use them for document clustering. This API also processes the data in different forms such as Single-Document Processing, Multiple Document Synchronous Processing, Asynchronous Batch Processing.

B. Google Cloud Platform (GCP)

Google cloud is one of the top leaders in the cloud market and their MLaaS solutions are gaining popularity and a large acceptance rate day by day. They provide a fully configurable and end-to-end services for the machine learning domain. Google cloud's "AI Platform" provides fully managed ML services for uncomplicated development, faster time to production, and easy maintenance with trained as well as pre-trained models.

As we can see in the diagram above, Google's AI platform services in the below phases of ML workflow.

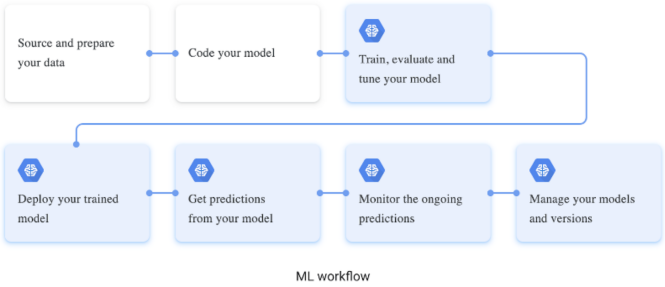

Fig. 3. Google Cloud ML Flow

- Train a machine learning model on custom dataset
- Containerization and deployment of the models
- predict responses from the model
- Continue to keep an eye on the predictions
- Management of the models and version control

Users can leverage different services as per their needs. AI platform also provides REST API endpoints for each of these services.

For the domain of NLP, Google cloud has an MLaaS platform called "Natural Language". Natural Language makes use of machine learning to discover the text's form and context. Users will find out more about individuals, sites, and activities, as well as get a greater understanding of social media opinion and consumer conversations. Customers can use Natural Language to analyze text and align it with their Cloud Storage document storage. The Natural Language platform has been divided into two major domains,

1. AutoML natural language - that allows the user to create a custom machine learning model
2. Natural language API - that allows the user to use a pre-trained machine learning model to reveal important information from the text.

For the scope of this paper, we are more interested in natural language API. Natural language API supports numerous NLP features including Entity recognition, Sentiment analysis, Entity sentiment analysis, Syntax analysis, and Classifier. They provide client libraries in all popular languages for API integration. They have REST, RPC, and command-line references for each of the services mentioned earlier.

C. Microsoft Azure

Azure is a set of cloud computing services provided by Microsoft. Azure AI is a unit of Azure that provides services to build, customize, train, deploy, and use Machine Learning models with minimalistic domain knowledge. Azure AI platform offers a quick and easy way to build, train and deploy machine learning models using Azure Machine Learning, Azure Databricks, and ONNX Runtime. It supports a wide range of frameworks, operating systems, and hardware platforms. Along with options to build custom models, Azure also offers ready-to-use knowledge mining services using Azure Cognitive search to extract data from large information corpus. In addition to pre-trained and custom solutions, Azure supports hybrid machine learning solutions where it offers state-of-the-art Machine Learning models to integrate with user infrastructure and possible options to customize them as per requirements.

Azure Natural Language Processing services provide a set of capabilities with Azure HDInsight with Spark MLlib and Spark NLP, Azure Databricks, and Microsoft Cognitive Services. Azure HDInsight is Microsoft's implementation of Apache Spark which is a parallel processing platform with in-built Machine Learning libraries. Azure HDInsight serves as a platform to build and train custom ML models with a large corpus of text data. It provides a parallel processing infrastructure so that users can design personalized machine learning models. Azure Cognitive Services are a set of REST APIs and SDKs those are ready-to-use and can be directly integrated into client applications.

Among the REST API services provided by Azure, Language APIs are a set of APIs which are used for Natural Language Processing. Language APIs include functionalities such as Language Understanding, QnA maker, Text Analytics and Translator. The most widely used APIs are Text Analytics APIs which provide features for text mining including Key Phrase Extraction, Named Entity Recognition, Opinion Mining, Sentiment Analysis, and Language Detection. Workflow is elementary in which text data is submitted using an API call and results are sent in response which is handled and used by clients as per requirements. The API response is in the form of a JSON document which can then be analyzed and visualized to gather actionable insights.

D. IBM Cloud

IBM Cloud is one of the leading cloud service providers with a wide range of business-ready applications, tools, and solutions that are cost-effective and hurdle-free. To name a few, IBM Cloud provides Cloud Compute, Networking, Internet of Things, Blockchain, and Machine Learning services. Lately, Machine Learning has gained a lot of demand, but building and tuning machine learning models could be a daunting task. IBM Watson provides several Machine Learning services which a novice user could easily use. It also includes complex features to customize these services to domain-specific requirements and integrate these services into enterprise-grade applications. IBM Watson cloud has many flavors, IBM Watson Studio to build and train AI Models, IBM Watson Machine Learning to deploy and run ML models, IBM Watson OpenScale to manage and operate AI models. Other than these, there are several other Natural Language Processing, Speech Recognition, and Computer Vision services.

Natural Language Processing has seen massive growth in the past decade. IBM has focused on developing services in this field; IBM Watson Text to Speech is a service that converts written text into natural-sounding audio; IBM Natural Language Classifier can be used to Interpret and classify natural language with confidence; IBM Watson Speech to Text converts real-time audio into text. IBM Cloud also provides the flexibility to enable features like EU support which is useful when processing personal data for European citizens; Financial Services Validated support could help when dealing with regulated financial services information, and HIPAA support if these services deal with Protected Health Information. All of these features can be easily enabled while using the NLP-related services provided by IBM Cloud.

IBM Watson Natural Language Understanding provides several Natural Language Processing services useful in analyzing text and extract metadata content. We can extract the following metadata content using the IBM Watson Natural Language Understanding service.

Entities	Keywords	Sentiment	Categories
Syntax	Concepts	Emotion	Metadata
Relations	Semantic roles	Classifications	

The features listed above can be easily accessed through an API endpoint provided by IBM Watson. It is also possible to customize the services to understand the linguistic nuances of any industry using IBM Watson Knowledge Studios. It provides features to customize these services using either of the rule-based or the ML models.

V. Dataset

We have used over 25 different labeled datasets spanning different application domains such as Social Media, Entertainment, News, Product Reviews, etc. The majority of the datasets used are popular Kaggle datasets, remaining datasets are from AI Stanford, UCI Machine Learning Repository, and Scikit-Learn. The datasets selected vary widely in terms of the number of records, domain, and the number of features, which would result in an unbiased comparison of the MLaaS providers. The datasets vary from a sample size of 2377 to 774362, we limit the size of datasets keeping in mind the computational complexity involved when processing the data using the cloud services. We performed data pre-processing for all the datasets before uploading the data to the MLaaS providers, all the service providers lack the feature to pre-process the data. We removed the records which have missing fields, removed all the stop words, converted all the letters to lower case, and also performed normalization, and stemming. Finally, we split each of the labeled datasets into three categories - train, test, and validate datasets for comparing

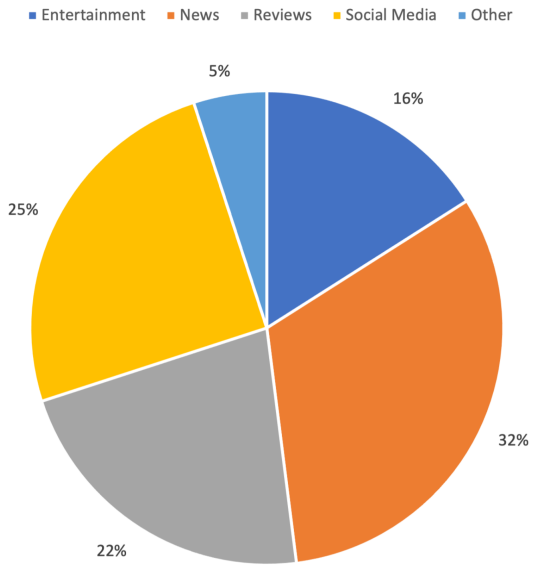

Fig. 4. Dataset Domain Breakdown

services that require domain-specific customizations or build custom Machine Learning Models.

VI. Measurement Criteria

A. Cost

One of the critical decision-making criteria that decide the use of MLaaS is Costs. Organizations have a limited allocation of the finances which they can spend on their Machine Learning infrastructure. Minimizing that expenditure is important while deciding on a cloud provider for MLaaS. Costs in MLaaS are the charges incurred upon making a set of API calls to use Machine Learning models used by Cloud Vendors. These charges vary from provider to provider and are subjective to the plans used by the clients.

Each cloud provider offers different tiers of cost plans where charges vary depending on the plan. Relatively low-priced plans allow a limited number of API calls in a definitive time frame, on the other hand, the number of successful API calls increases with expensive plans. Here we have provided a detailed comparison table of prices of various cloud providers for different batches of API calls.

TABLE I
Amazon Web Services

Units	Sentiment	Key Phrase Detection	Entity Recognition
0-5K	$1	$1	$1
5K-250K	$1	$1	$1
250K-500K	$1	$1	$1
500K-1M	$1	$1	$1
1M-2.5M	$1	$1	$1
2.5M-5M	$1	$1	$1
5M-10M	$1	$1	$1
10M-20M	$0.5	$0.5	$0.5
20M-50M	$0.5	$0.5	$0.5
50M+	$0.25	$0.25	$0.25
Prices here are per 1000 units where 1 unit = 1000 characters			

TABLE II
Google Cloud Platform

Units	Sentiment	Key Phrase Detection	Entity Recognition
0-5K	Free	Free	Free
5K-250K	$1	$1	$1
250K-500K	$1	$1	$1
500K-1M	$1	$1	$1
1M-2.5M	$1	$1	$1
2.5M-5M	$1	$1	$1
5M-10M	$0.25	$0.25	$0.25
10M-20M	$0.25	$0.25	$0.25
20M-50M	$0.25	$0.25	$0.25
50M+	$0.25	$0.25	$0.25
Prices here are per 1000 units where 1 unit = 1000 characters			

TABLE III
Microsoft Azure

Units	Sentiment	Key Phrase Detection	Entity Recognition
0-5K	$1	$1	$1
5K-250K	$1	$1	$1
250K-500K	$1	$1	$1
500K-1M	$0.75	$0.75	$0.75
1M-2.5M	$0.75	$0.75	$0.75
2.5M-5M	$0.3	$0.3	$0.3
5M-10M	$0.3	$0.3	$0.3
10M-20M	$0.25	$0.25	$0.25
20M-50M	$0.25	$0.25	$0.25
50M+	$0.25	$0.25	$0.25
Prices here are per 1000 units where 1 unit = 1000 characters			

TABLE IV
IBM CLOUD

Units	Sentiment	Key Phrase Detection	Entity Recognition
0-5K	$0.3	$0.3	$0.3
5K-250K	$0.3	$0.3	$0.3
250K-500K	$0.1	$0.1	$0.1
500K-1M	$0.1	$0.1	$0.1
1M-2.5M	$0.1	$0.1	$0.1
2.5M-5M	$0.1	$0.1	$0.1
5M-10M	$0.02	$0.02	$0.02
10M-20M	$0.02	$0.02	$0.02
20M-50M	$0.02	$0.02	$0.02
50M+	$0.02	$0.02	$0.02
Prices here are per 1000 units where 1 unit = 1000 characters			

B. Time

The importance of time monitoring in machine learning applications is second to none. The response time of the machine learning model is as important as the accuracy of the model. In some cases, it is even more important than accuracy. The mainstream acceptance of high-speed Internet connectivity and its use for routine activities has resulted in significant shifts in consumer preferences for Web site efficiency and reliability. A slower response time at the peak traffic time can cause the loss of thousands of customers in today's competitive environment. For web applications like Grammarly that provide a real-time response through their NLP machine learning models, each fraction of a second is important. With this view, "Time" has been kept as one of the four measurement criteria.

To measure the response time of the API endpoints provided by four cloud providers, we picked the datasets of majorly four different domains - Social media, Entertainment, Product reviews, and news articles. These datasets contained millions of rows. To find the average response time for each domain, we further divided the datasets into batches of 1000. the average of the average response time of all batches would lead to the desired answer. In this manner, the average response time of each domain can be calculated and it would be really helpful in finding the standard deviations and drawing conclusions on which platform is providing better results in which domain. Since the scope of this paper is limited to the performance of MLaaS APIs, we are not considering the time taken by each platform to process and train the model.

TABLE V
RESPONSE TIME MEASURES

Platforms	Mean (S)	Standard Deviation (S)
AWS	0.31	0.12
GCP	0.47	0.05
IBM	0.45	0.24
Azure	0.54	0.10

The graph(fig.5) shows an overall average Response Time for each batch of inputs on different cloud platforms. All the cloud providers performed well in terms of Response Time as the average response time never exceeded 1 second. From all these cloud providers, GCP is the most consistent with a Standard Deviation of 0.05 seconds. Azure, AWS, and IBM follow GCP in order with Standard Deviations of 0.1, 0.12, and 0.24 seconds. Although GCP is more consistent than others, AWS has the lowest overall average time of 0.31 seconds. So if the consistency in the responses is the highest priority then GCP is better but, for the actual lower response time, AWS is better.

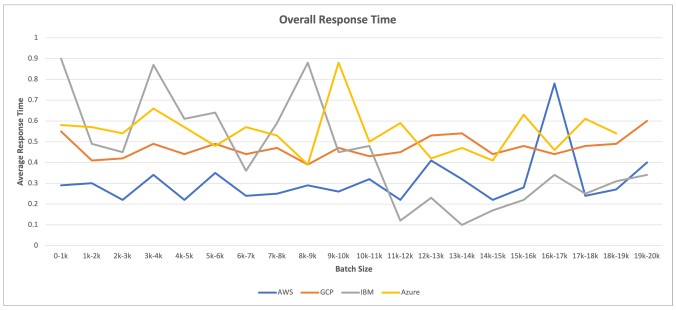

Fig. 5. Response Time

C. Accuracy

Accuracy is one of the crucial metrics for Machine Learning model evaluation. Before choosing between several cloud providers, it is imperative to know the models' accuracy that each of these has. This metric can answer the following questions regarding the ML model.

- How well is the model doing?
- Does the model require more features to be included?
- Is the model over-fitted or under-fitted?
- Will further training of the model increase the accuracy?

Many cloud providers include a Confidence score along with accuracy. A confidence score is a number between 0 and 1 which shows how confident the model is for its output. A score of 0.7 means the model is 70% confident that the output that it has predicted is correct. The following section will elaborate on how we measured the accuracy of each cloud provider's NLP API.

We chose a set of datasets and tested them in terms of accuracy across each cloud vendor in the same environment. The steps to measure this metric are as follows,

1. Collect the dataset
2. Hit APIs of a cloud provider's NLP service and save the response
3. Process the response and pass it through the accuracy measurement algorithm
4. Iterate through steps 2 through 3 for each cloud provider
5. Collect all results and compare

To measure the accuracy, we have developed an algorithm. Each dataset is partitioned into number of batches where each batch consists of 1000 records. There are two measurements here. First, we calculate the average accuracy of all records in the batch and its standard deviation.

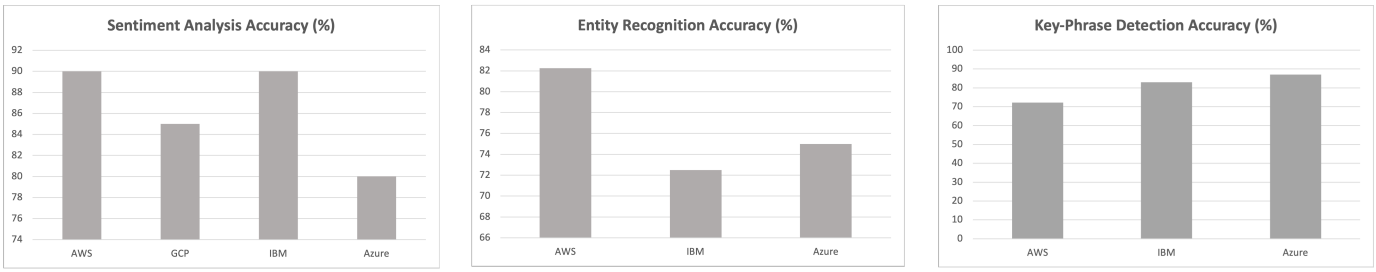

Fig. 6. Accuracy comparison

$$Average\ Accuracy\,(\mu) = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

where N = Batch size, x_i = Each record's accuracy

$$Standard\ Deviation\,(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2}$$

where N = Batch size, x_i = Each record's accuracy, μ = average accuracy of a batch

Once we have the average accuracy and standard deviation of each batch, we calculate the overall average accuracy and standard deviation of the whole dataset using the results from the previous steps.

The pool of datasets we have, includes data from several domains such as social media, news, entertainment, etc. This diversity in datasets helps us determine the performance of ML models on a wide range of data. The performance is measured for three major subdomains of NLP namely, Sentiment Analysis, Named Entity Recognition, and Key Phrase analysis.

AWS is recommended above the other three cloud providers for Sentiment Analysis. For a set of diverse data, the responses provided by AWS are accurate and reliable. IBM, GCP, and Azure follow the respective order in performance. However, the outcome changes if data is specific to a particular domain. For instance, for datasets belonging to News categories GCP and IBM defeats AWS and Azure by a fair margin while Azure shows high accuracy in Social Media datasets when compared to the other three.

Named Entity Recognition is a very specific subdomain that requires custom models to be trained and used as per use case. Since NER services provided by all cloud providers are generic and only limited to pre-defined categories the performance of all cloud providers is subpar. GCP does not provide NER services in their NLP APIs and it is not recommended to use general models for NER services. Custom model training functionalities are supported by all of these cloud vendors which is more suitable for NER.

Azure is advised above the other cloud vendors in Key Phrase detection in textual data. IBM is not far behind Azure in this category as well. Irrespective of the domain, Azure performs well in Key Phrase extraction and is consistent across multiple NLP fields.

Our analysis suggests that ready-to-use models used by cloud providers are reliable and stable for general purpose and widely used data. Although, in sub-domains like Named entity recognition addition of more features to the pre-trained model would result in a significant increase in performance. The results suggest that further training would not make a noticeable difference in accuracy however, custom models are advised if data is highly concentrated to a particular domain.

D. Ease of Use

The Main Selling Point of MLaaS is that it takes away the hassle of building, securing, and maintaining Machine Learning infrastructure. For a client who is unfamiliar with the Machine Learning domain, it can be overwhelming to have an in-house Machine Learning infrastructure. Hence Ease Of Use is a critical aspect while considering MLaaS. The Ease-Of-Use encapsulates how convenient is it for a novice in the field of Machine Learning to harness the potential of Machine Learning through MLaaS.

Cloud Vendors provide a User Interface to navigate ML services on the web. To integrate their Machine learning services directly into client applications they also provide API endpoints and SDKs which makes it easy to use their state-of-the-art Machine Learning models. Each cloud provider offers SDKs in a set of programming languages. Depending on the language and provider's SDKs number of steps to install and integrate MLaaS varies which is the Ease-Of-Use of an MLaaS. Following are the major criteria that are a part of the Ease-Of-Use measurement.

1. Steps to navigate through Cloud Vendors' UI to use MLaaS
2. Steps to get API endpoints
3. Response format of API calls
4. The convenience of SDK installation and integration
5. Documentation of services and issues

VII. Conclusion

A. Key Findings

Our research produced some key takeaways. While pre-trained NLP models provided by all these cloud providers perform well for generic data, the performance declines when

used for specific domains such as Medical or Healthcare data. Results have shown that responses provided by these models for categories like NER are limited and hence are not pragmatic to use them for real-time services. By comparing the gathered responses we can conclude that, unlike its peers, GCP provides a wide variety of information like categorization, sentiment analysis for entities, etc. However, AWS and Azure lead in terms of accuracy in Sentiment Analysis and Key Phrase Detection respectively. IBM Cloud leads the way in terms of costs as it is the most economical among all cloud providers for API calls. All the cloud providers stand the test of high workloads as response times are consistent irrespective of a large number of API calls. Although a particular cloud provider may not fit all the requirements, the set of services provided at this price point with above par accuracy is sufficient for small-scale businesses to integrate MLaaS.

B. Limitations

We have identified some limitations of this research due to the scope. First, we have only compared and analyzed four major cloud providers like AWS, IBM, GCP, and Azure but, as the growth of MLaaS increases, so many new cloud providers are emerging. Some of them focus on specific fields such as Computer Vision, Audio/Video Analysis. All of these platforms are not covered in this paper. Second, there are so many subfields in Machine Learning like Computer Vision, Data Analysis, Predictive Analysis, Service Personalization, to name a few. But, our concentration was solely on one of the subfields, Natural Language Processing. And third, we only used some datasets in several domains but this can be extended to many domains and their related datasets. We leave all these limitations as future scope.

REFERENCES

[1] Mauro Ribeiro, Katarina Grodinger, Miriam A.M. Capretz, MLaas: Machine Learing as a Service 2015 IEEE 14th International Conference on Machine Learning and applications Available: https://ieeexplore.ieee.org/abstract/document/7424435

[2] Yuanshun Yao, Zhunjun Xiao, Bolun Wang, Biman Vishwanath, Haitao Zheng, Ben Y. Zhao, Complexity vs Performance: Empirical Analysis of Machine Learning as a Service IMC '17: Proceedings of the 2017 Internet Measurement Conference, November 2017 Available: https://doi.org/10.1145/3131365.3131372

[3] Lucjan Hanzlik, Yang Zhang, Kathrin Grosse, Ahmed Salem, Max Augustin, Michael Backes, Mario Fritz, MLCapsule: Guarded offline Deployment of Machine Learning as a Service February 6, 2019 Cornell University Available: https://arxiv.org/abs/1808.00590

[4] Ehsan Hesamifard, Hassan Takabi, Mehdi Ghasemi, Rebecca N. Wright, Privacy- preserving Machine Learning as a Service Proceeding on Privacy Enhancing Technologies: 2018(3): 123-142 Available: https://doi.org/10.1515/popets-2018-0024

[5] Ranjith Ramesh , Predictive analytics for banking user data using AWS machine learning as a service 2017 2nd International Conference on Computing and Communications Technologies Available: https://ieeexplore.ieee.org/abstract/document/7972282

[6] G. Sun, T. Cui, S. Chen, W. Guo and J. Shen, "MLaaS: A Cloud System for Mobile Micro Learning in MOOC," 2015 IEEE International Conference on Mobile Services, New York, NY, 2015, pp. 120-127, doi: 10.1109/MobServ.2015.26. Available: https://ieeexplore.ieee.org/abstract/document/7226680

[7] J. Yi, C. Zhang, W. Wang, C. Li and F. Yan, "Not All Explorations Are Equal: Harnessing Heterogeneous Profiling Cost for Efficient MLaaS Training," 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS), New Orleans, LA, USA, 2020, pp. 419-428, doi: 10.1109/IPDPS47924.2020.00051. Available: https://ieeexplore.ieee.org/abstract/document/9139864

[8] Li Erran Li, Eric Chen, Jeremy Hermann, Pusheng Zhang, Luming Wang; "Scaling Machine Learning as a Service", Proceedings of The 3rd International Conference on Predictive Applications and APIs, PMLR 67:14-29, 2017

[9] Alejandro Baldominos, Esperanza Albacete, Yago Saez, Pedro Isasi, "A scalable machine learning online service for big-data real time analysis"[Online] 2014 IEEE Symposium and Computational Intelligence in Big data Available: https://ieeexplore.ieee.org/abstract/document/7011537

[10] Tyler Hunt, Congzheng Song, Reza Shokri, Vitlay Shamatikov, Emmett Witchel, "Chiron: Privacy-preserving Machine Learning as a Service"[Online] March 15, 2018 Cornell University Available: https://arxiv.org/abs/1803.05961

[11] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, Wenqi Wei , "Demsifying Membership Inference Attacks in Machine Learning as a Service"[Online] February 5, 2019 IEEE Transaction on Services Computing Available: https://ieeexplore.ieee.org/document/8634878

[12] Haytham Assem, Lei Xu, Teodara Sandra Buda, and Declan O'Sullivan, "Machine Learning as a Service for enabling Internet of Things and People", Pers Ubiquit Comput 20, 899–914 (2016), Available: https://doi.org/10.1007/s00779-016-0963-3

[13] Tafti A.P., LaRose E., Badger J.C., Kleiman R., Peissig P. (2017) Machine Learning-as-a-Service and Its Application to Medical Informatics. In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2017. Lecture Notes in Computer Science, vol 10358. Springer, Cham. https://doi.org/10.1007/978-3-319-62416-7-15

[14] Sokolova, Marina and Lapalme, Guy. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management. 45. 427-437. 10.1016/j.ipm.2009.03.002.

[15] Candido de Oliveira, Diulhio and Wehrmeister, Marco. (2016). Towards Real-Time People Recognition on Aerial Imagery Using Convolutional Neural Networks. 27-34. 10.1109/ISORC.2016.14

[16] Poggi, N., Carrera, D., Gavaldà, R. et al. A methodology for the evaluation of high response time on E-commerce users and sales. Inf Syst Front 16, 867–885 (2014). https://doi-org.libaccess.sjlibrary.org/10.1007/s10796-012-9387-4

[17] Lim, TS., Loh, WY. and Shih, YS. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning 40, 203–228 (2000). https://doi-org.libaccess.sjlibrary.org/10.1023/A:1007608224229

[18] Merz, C. J. and Murphy, P. M. (1996). UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California, Irvine, CA (http://www.ics.uci.edu/ mlearn/MLRepository.html)