Consumer Analytics using Natural Language Processing and Artificial Intelligence in the Cloud

An UberCloud Case Study

With Support From

IMT Institute of Management Technology Ghaziabad

UberCloud

UberCloud Case Study 221

http://www.TheUberCloud.com

July 01, 2021

Welcome!

The UberCloud* Experiment started in July 2012, with a discussion about cloud adoption in technical computing and a list of technical and cloud computing challenges and potential solutions. We decided to explore these challenges further, hands-on, and the idea of the UberCloud Experiment was born, then also due to the excellent support from INTEL generously sponsoring these experiments since the early days!

We found that especially small and medium enterprises in digital manufacturing would strongly benefit from technical computing in HPC centers and in the cloud. By gaining access on demand from their desktop workstations to additional and more powerful compute resources in the cloud, their major benefits became clear: the **agility** gained by shortening product design cycles through shorter simulation times; the superior **quality** achieved by simulating more sophisticated geometries and physics and by running many more iterations to look for the best product design; and the **cost** benefit by only paying for what is really used. These are benefits that obviously increase a company's **innovation and competitiveness**.

Tangible benefits like these make computing - and more specifically technical computing as a service in the cloud - very attractive. But how far are we from an ideal cloud model for engineers and scientists? At first, we didn't know. We were facing challenges like security, privacy, and trust; traditional software licensing models; slow data transfer; uncertain cost & ROI; lack of standardization, transparency, cloud expertise. However, in the course of these cloud experiments, as we followed each of the 221 teams closely and monitored their challenges and progress, we've got an excellent insight into their roadblocks, how our teams have tackled them, and how we are now able to reduce or even fully resolve them.

This UberCloud Experiment #221 is about "Consumer Analytics using Natural Language Processing and Artificial Intelligence in the Cloud". The enormous popularity of big data in social media allows purchasers to voice their opinions such as expressing their pleasure or displeasure with certain items or services, or to express their satisfaction with their purchases. Such numerous consumer opinions and product reviews contain rich and valuable information and the use of Natural Language Processing (NLP) machine learning techniques is essential to extract data and opinions from the huge amount of information present on the web. Therefore, the scope of this project has been to develop a consumer analytics framework considering e-commerce review data using AI machine learning techniques in the Cloud. This study enables the objective to understand the consumer in real-time and to identify a critical issue affecting the business. As this process is resource-driven, and the data extracted from social media is huge, it requires a higher number of CPU cores and RAM for faster data processing, model building & data visualization. This study explores this relationship through a cloud-based solution infrastructure to speed up computation time and achieve faster turn-around times required in the model building process.

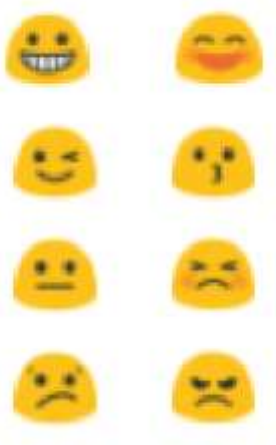
Now, enjoy reading!

Veena Mokal, Wolfgang Gentzsch, and Burak Yenier

**) UberCloud is the online community & marketplace where engineers and scientists discover, try, and buy Computing Power as a Service, on demand. Engineers and scientists can explore and discuss how to use this computing power to solve their demanding problems, and to identify the roadblocks and solutions, with a crowd-sourcing approach, jointly with our engineering and scientific community. Learn more about the UberCloud at: http://www.TheUberCloud.com.*

Team 221

Consumer Analytics using Natural Language Processing and Artificial Intelligence in the Cloud



> "UberCloud's HPC and Cloud computing capabilities have aided in the processing of large volumes of consumer data in order to build AI models for making better decisions and game-changing strategies."

MEET THE TEAM

End-User/Data Science Expert: Veena Mokal, Data Science Expert, MBA in Business Analytics, Institute of Management Technology, INDIA
Software Provider: Anaconda Python distribution platform
Resource Provider: On-premises systems. And next step: UberCloud Engineering Simulation Platform
HPC Expert: Praveen Bhat, HPC/Python Technology Consultant, India, Wolfgang Gentzsch, UberCloud

USE CASE

In recent years, advancements in internet connectivity have brought significant opportunities to customers and shoppers. Because of these advancements in internet connectivity, rapidly rising e-commerce enterprises have yielded true big data. The enormous popularity of big data on social media allows purchasers to voice their opinions and views on a wide range of topics such as the status of the economy, or to express their **displeasure** with certain items or services, or to express their **satisfaction** with their purchases.

Such numerous consumer opinions and product reviews contain rich and valuable information and recently became important sources for both consumers and business firms. Consumers commonly seek quality information from online reviews before purchasing a product, while many firms use online reviews as important feedbacks of their products, marketing and consumer relationship management. Therefore, understanding the psychology behind online consumer behaviour became the key to compete in today's markets which are characterized by ever-increasing competition and globalization.

Sentiment analysis & text analysis are the applications of big data analysis, which aim to aggregate and extract emotions and feelings from different kinds of reviews. These big data which is growing exponentially are mainly available in an unstructured format, and they are not machine-processable and interpretable. Therefore, the use of the Natural Language Processing (NLP) machine learning technique is essential which focuses on extracting this data and opinions from the huge amount of information present on the web.

Consumer sentiment analysis is one of the popular techniques of discovering the emotion to understand your customer related to your product or service. The scope of this project is to develop a consumer analytics framework considering e-commerce review data using AI machine learning techniques with a Cloud solution capability. This study enables the objective to understand the consumer in real-time and to identify a critical issue affecting the business. As this process is resource-driven, and the data extracted from social media is huge, it requires a higher number of CPU cores and RAM for faster data processing, model building & data visualization. This study explores this relationship through a cloud-based solution infrastructure to speed up computation time and achieve the faster turn-around time required in the model building activity.

PROCESS OVERVIEW

The following defines the step-by-step approach in setting up the model environment for consumer analytics using NLP and Python.

a. Data Pre-Processing & Feature Extraction

The pre-processing of data involves Text cleaning & Feature Extraction:

Text Cleaning: Text cleaning activity is one of the major steps in data pre-processing. In this step punctuations, stop words, URL Links, HTML tags, and special characters like emoticons and emoji are removed and the text converts into lower case. In the final step, text cleaning performs spell checks and corrects it grammatically.

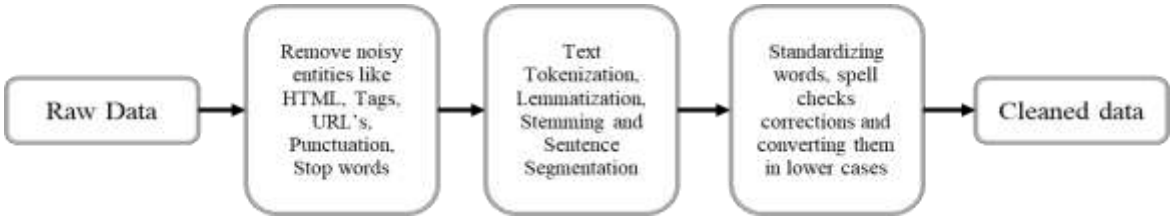


Figure 1: Text cleaning process

Feature Extraction: The cleaned data is then processed to extract features (variables) that help to understand the distribution of the review text. This includes the distribution of characters, numerical values, lowercase words, and average word length. This completes the data pre-processing step.

b. Performing Sentiment Analysis

This section mainly focuses on generating sentiment scores for the review text data. Sentiment score is a function of polarity & subjectivity. Both parameters are extracted from the review text using NLP algorithms to understand the overall sentiment. Typically, the overall sentiment is often inferred as positive, neutral or negative from the sign of the polarity score. Polarity [11] is a floating-point number that lies in the range of [-1,1] where 1 means a positive statement and -1 means a negative statement. Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float that lies in the range of [0,1].

c. Topic Modeling

Topic modelling is the process of identifying topics in a set of documents. It enables search engines on the topics of documents that are important. There are multiple methods of doing this, however, this project includes Latent Dirichlet Allocation (LDA). LDA is a form of unsupervised learning that views documents as bags of words. It works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words. To do this it does the following for each document m:

The following algorithm steps are implemented using Python and LDA for topic modelling:

- Assume there are k topics across the whole document
- Distribute these k topics across document m (this distribution is known as α and can be symmetric or asymmetric, more on this later) by assigning each word a topic
- For each word w in document m, assume its topic is wrong, but every other word is assigned the correct topic
- Probabilistically, assign word w a topic based on two things:
    - what topics are in document m
    - how many times word w has been assigned a particular topic across all the documents?
- Repeat this process several times for each document to get the list of topics

	Word 1	Word 2	Word 3	Word 4	--	--	--	--	Word n
Topic -1	--	--	--	--	--	--	--	--	--
Topic -2	--	--	--	--	--	--	--	--	--
Topic -3	--	--	--	--	--	--	--	--	--
Topic -4	--	--	--	--	--	--	--	--	--
. . .	--	--	--	--	--	--	--	--	--
Topic -k	--	--	--	--	--	--	--	--	--

Figure 2: Details on topic modelling structure

The above figure shows the representation of topic modelling structure and how the individual topics are related to each of the words in the documents.

d. Predictive modeling

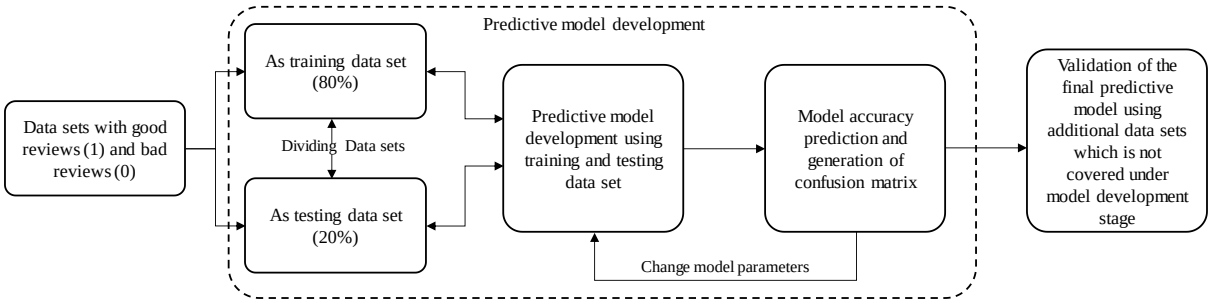


Figure 3: Predictive model development framework

The objective of this phase is to develop a modelling methodology that classifies new input review text into good or bad reviews. The classification accuracy and validation of the model become key criteria for the selection. The predictive model can be developed using both supervised and

unsupervised learning methods. This study covers the following predictive modelling techniques to predict the type of reviews (good or bad reviews):

- Naïve Bayes with
    - Gaussian method
    - Multinomial method
    - Bernoulli method
- Logistic Regression model fitting

RESULTS & DISCUSSION

This section details out the results extracted by running the Python scripts developed for the methodology explained. The following provides details on the results derived and insights gained from the analysis.

Make all text lower case

The first pre-processing step which we will do is transform our reviews into lower case. This avoids having multiple copies of the same words. For example, while calculating the word count, 'Analytics' and 'analytics' will be taken as different words.

```
In [17]: df['reviewText'] = df['reviewText'].apply(lambda x: " ".join(x.lower() for x in x.split()))
         df['reviewText'].head()

Out[17]: 0    we got this gps for my husband who is an (otr)...
         1    i'm a professional otr truck driver, and i bou...
         2    well, what can i say. i've had this unit in my...
         3    not going to write a long review, even thought...
         4    i've had mine for a year and here's what we go...
         Name: reviewText, dtype: object
```

Figure 4: Example showing texts converted to lower case

The first step shows all the review texts converted into lower case followed by removing the punctuation marks and stop word removal, unwanted texts like HTML tags, emoticons, white spaces etc. The data set is further subjected to multiple pre-processing steps involved which will further help to cleanse and structure the data sets for further analysis. This includes - extracting the number of words, characters and average word length.

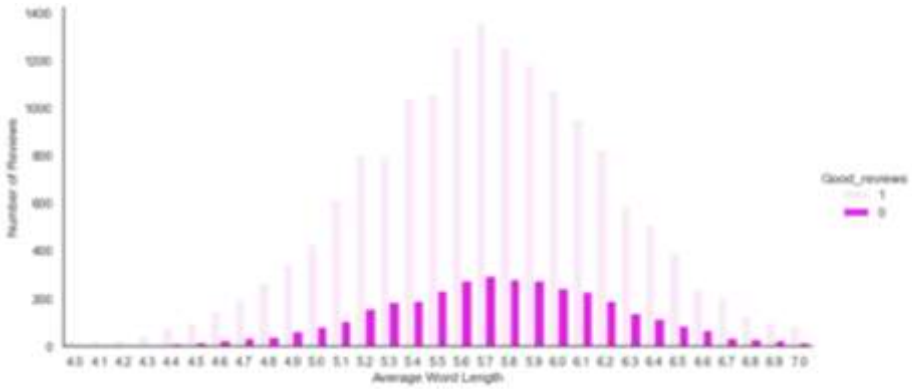


Figure 5: Distribution showing the average word length for good and bad reviews

The above figure shows the average word length for good and bad reviews. From the distribution, the word length for a good review is longer than the bad review and hence the processing time for the good review data is higher when compared to the bad review data.

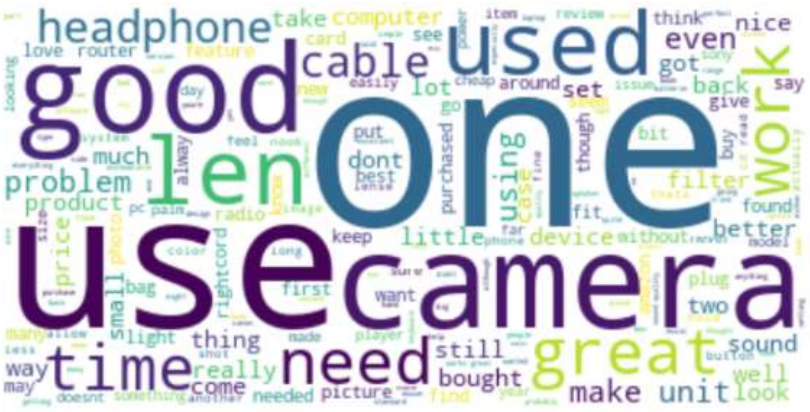


Figure 6: Word Cloud extracted from the processed review data

Figure 6 shows the word cloud which is a visual representation of the word frequency. The more commonly the term appears within the text being analysed, the larger the word appears in the image generated. Word clouds are increasingly being employed as a simple tool to identify the focus of written material.

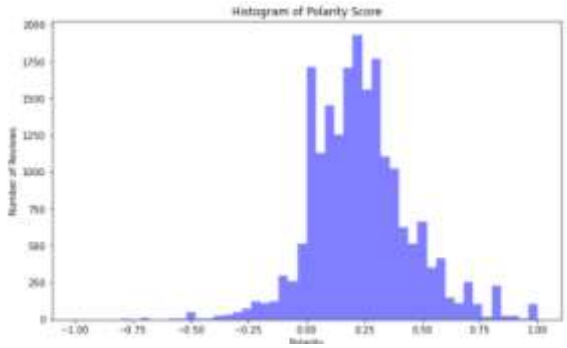


Figure 7: Sentiment scores for the first 2000 reviews

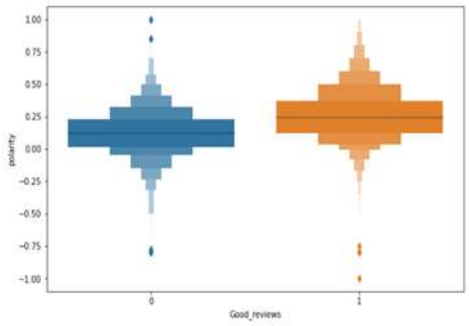


Figure 8: Polarity values for bad (0) & good reviews (1)

Figure 7 shows the sentiment score distribution for the first 2000 reviews where the major portion of the review sentiments shows positive polarity for good reviews. *Figure 8* shows a good review distribution on polarity. Good reviews that have low polarity are categorized in negative sentiments. Bad reviews which have high polarity are categorized in positive sentiments.

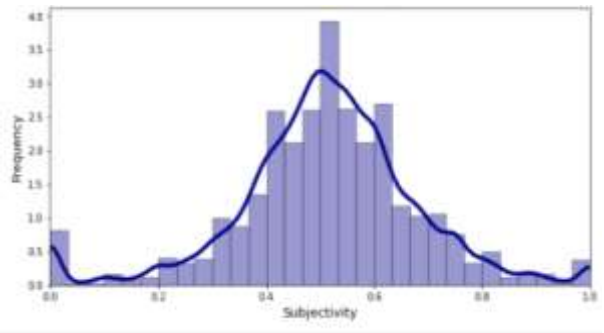


Figure 9: Distribution of subjectivity

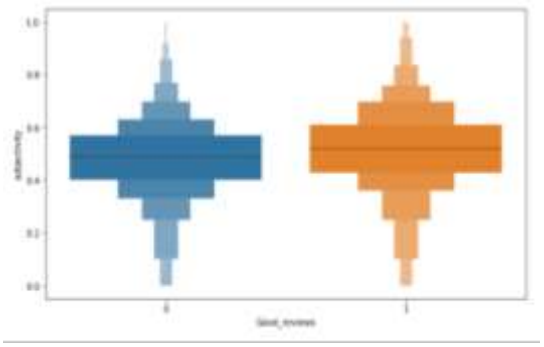


Figure 10: Subjectivity values for good (1) and bad (0) reviews

Figure 9 shows the subjectivity score distribution. *Figure 10* shows a good review distribution on subjectivity. *Figure 11* shows the distribution of subjectivity and polarity for good and bad reviews.

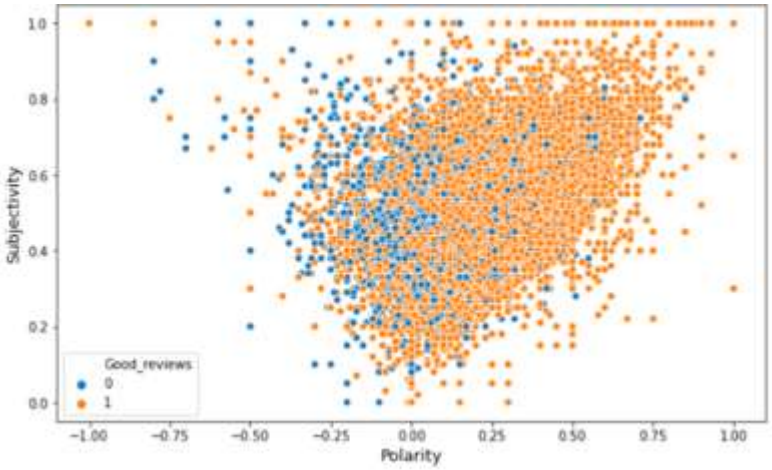


Figure 11: Distribution of subjectivity and polarity for bad and good reviews

The topic modelling activities focus on defining the list of topics from the review data and creating a matrix of topics. The LDA algorithm is used to analyse the topics and generate the probability of occur-ence of topics in a document based on the words. The LDA works with an assumption that each docu-ment is a collection of words, "bag of words", thus the order of the words and their grammatical role are not considered in the model. **Fehler! Verweisquelle konnte nicht gefunden werden.** provides the LDA equation considering 4 topics.

```
# Let's try 4 topics
ldan = models.LdaModel(corpus=corpusn, num_topics=4, id2word=id2wordn, passes=10)
ldan.print_topics()
```

```
[(0,
  '0.044*"headphones" + 0.026*"sound" + 0.018*"price" + 0.017*"quality" + 0.014*"radio" + 0.013*"bass" + 0.012*"pair" + 0.012
*"music" + 0.011*"use" + 0.010*"volume"'),
 (1,
  '0.030*"cable" + 0.013*"product" + 0.011*"mouse" + 0.011*"use" + 0.010*"power" + 0.010*"tv" + 0.010*"router" + 0.010*"cables"
+ 0.010*"computer" + 0.010*"price"'),
 (2,
  '0.015*"palm" + 0.013*"use" + 0.009*"player" + 0.009*"unit" + 0.007*"device" + 0.007*"case" + 0.007*"software" + 0.007*"tape"
+ 0.006*"cd" + 0.006*"battery"'),
 (3,
  '0.051*"camera" + 0.018*"bag" + 0.017*"lens" + 0.016*"use" + 0.014*"canon" + 0.012*"pictures" + 0.009*"quality" + 0.008*"batt
eries" + 0.008*"flash" + 0.008*"card"')]
```

Figure 12: LDA models with 4 topics considered

A similar approach was followed to extract the topics using both nouns and adjectives and then use the LDA algorithm to analyse the topics. The above exercise is performed for both good and bad reviews separately to generate a list of topics and the associated bag of words for further analysis.

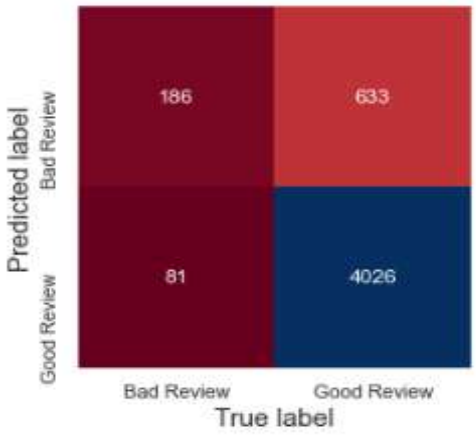


Figure 12: Confusion matrix for logistic regression

The processed data which contains good and bad reviews groups are used for training the predictive models. The review dataset is divided into an 80% training and a 20% testing data group which is used for building the model. The following modelling output is evaluated for model accuracy through a confusion matrix and an accuracy percentage. The Gaussian predictive model shows 71% of prediction accuracy, where around 271 data points are misclassified for the good review and 1142 data points are misclassified for the bad review.

The Multinomial predictive model shows 84% of prediction accuracy where around 749 data points are misclassified as the good review and 21 data points are misclassified as the bad review. The Bernoulli predictive model shows 80% of prediction accuracy where 450 data points are misclassified as the good review and 559 data points are misclassified as the bad review. The Logistic regression predictive model shows 86% of prediction accuracy where 633 data points are misclassified as the good review and 81 data points are misclassified as bad review.
Out of all predictive models, the logistic regression model has a better prediction in terms of accuracy with a lesser misclassification percentage. *Figure 13* shows the comparison done on different predictive modelling methodology.

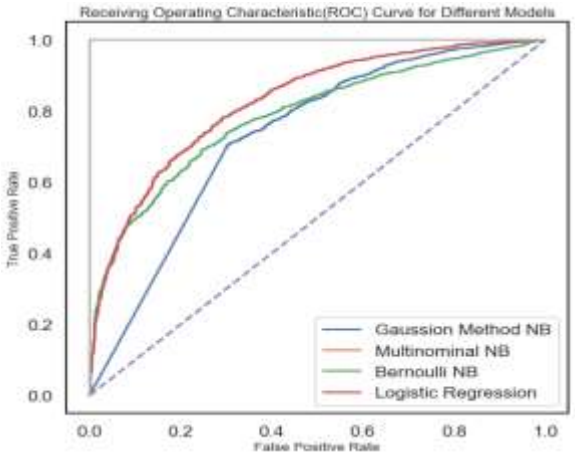


Figure 13: Comparison of different modelling techniques

MODEL VALIDATION

The logistic regression model is validated using the latest review data from the e-com website. The data is then pre-processed using the Python script and analysed using the logistic regression model.

ReviewText	Actual Label	Predicted Label
very secure, we did have to use concrete bolts thought since we were putting it on an outsid	1	1
We bought this for a 55" Vizio and it worked perfectly. I would buy it again and recomm	1	1
Mixed up the review to show how the Nook changed with Updates within 12 hours of openi	0	1
It worked perfectly. Nothing more that I could ask from a home charger. If you need one it w	1	1
I uploaded All of United States and Canada via Garmin and have all the space I need for my C	1	1
Hey, I am a music professional. I make many recordings and service equipment so I know wh	0	0
I have used Maxwell for some time. I liked the price and having used these in the past, you	1	1
It is a good cord. Simply packaged. Nicely priced.There is not much more to say about a sim	1	1
I'm so sick of most local stores overcharging for cables. I always try to buy my cables online.	1	1
The clear tape is great as a way to replace engraving. It does have a gloss like scotch tape an	0	1
I now own three pair of these, They are a bargain, as I have dozens of earbuds (mostly in ea	1	1
Sure these headphones look a bit wierd. A matter of taste at best.However the sound is gre	1	1
I gave it 5 stars because for the price, these are about the best headphones you can get sans	1	1
Gave these as a stocking stuffer and well received. They were worn during workout, and we	1	1
I have the PortaPro, which I bought when it was selling for $50. The sound is incredible for s	1	1
A good one of these can last 20 years. A bad one doesn't last 20 weeks. Either way, getting th	0	1
I have the DYMO Letra Tag Label maker and I love it. I have one at work and one at home. Ex	1	1
Your standard power strip, with surge protection. Does the job, outlets aren't inaccessible a	1	1
I bought this to replace an orange cord I had in my front yard for Christmas lights. The green	1	1
This cord seems as if it will suffice. It is long enough to reach from the outlet on the outside	1	1

Figure 14: Latest data extracted (Jan-Feb 2021) from e-com portal where actual and predicted labels are compared

Prediction accuracy for good review data has 79 misclassified data out of 4000 good review data points.

HPC PERFORMANCE BENCHMARKING

The NLP – Machine Learning algorithm for e-com reviews is a very compute intensive technique, therefore, to complete the study, we have run a performance analysis using a high-performance desktop machine that has 16 CPU Cores and 32 GB RAM. The performance analysis was conducted to study the computing system requirement to run millions of review data.

CHALLENGES

The challenges faced in the project were related to processing massive volumes of data from different social media and e-commerce websites. The project requirement was to build an AI-driven model to process the reviews from different web sources and analyse the sentiments behind these reviews. Processing review data with local computers was a challenge to handle the data size. This resulted in the need to use a high-performance compute node, where we could load the big volume of data to perform data analysis and develop required steps to pre-process the data.

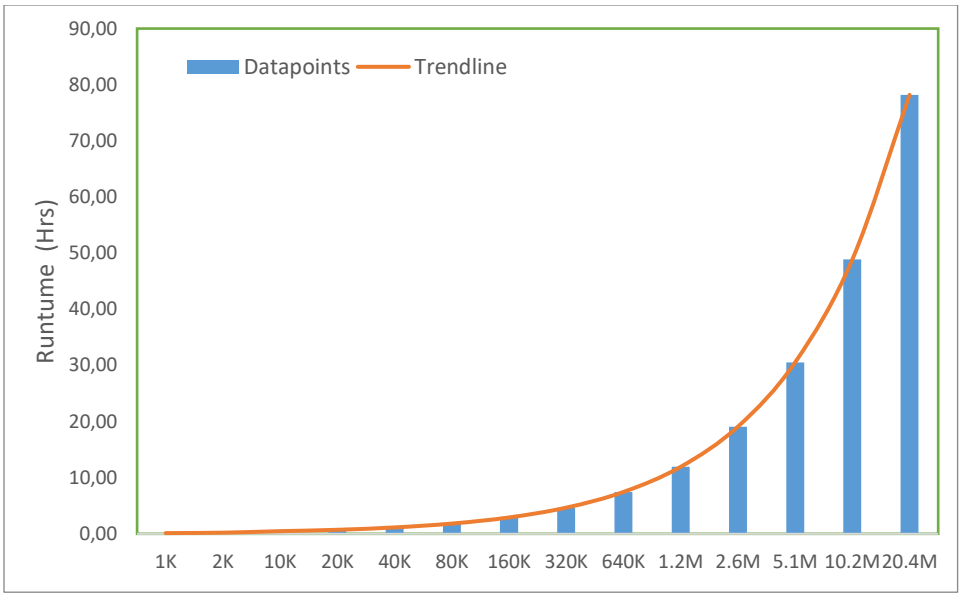


Figure 15: Compute runtime captured for different sizes of review data sets

BENEFITS

1. The HPC cloud computing environment features the Python-based Anaconda platform that aided in data analysis and the construction of predictive models. Dealing with a large volume of data and pre-processing formatting activities was difficult in this project. These tasks demanded a significant amount of computing power. The handling and processing of such massive amounts of data was made possible by cloud HPC.
2. Experiments conducted in the HPC Cloud environment demonstrated the ability to remotely set up and run Big Data analysis as well as build AI models in the cloud. The AI - Machine learning model setup requirements were pre-installed in the HPC container, allowing the user to access the tools without installing any kind of prior set up.

CONCLUSION & RECOMMENDATIONS

- Advanced machine learning technology, such as NLP, is a field of study that examines people's sentiments, attitudes, or emotions toward specific entities. This study addresses the fundamental problem of consumer behavior by using sentiment analysis, sentiment polarity categorization, and high-performance computing containers to speed up the process.
- All of these applications show how sentiment analysis can be a useful resource for analyzing affective information in social platforms, relying not only on domain-specific keywords but also on commonsense knowledge bases that allow for the extrapolation of cognitive and affective information associated with natural language text.
- UberCloud's HPC resource was a good fit for performing NLP analytics and building AL-ML models involving social media and e-commerce big data that could not be executed on a standard workstation. UberCloud's environment aided in speeding up model development activities within a set timeline and completing the project successfully.

REFERENCES

1. Collobert, R. Weston, J. Bottou, L. Karlen, M. Kavukcuoglu, K. Kuksa, (2011). Natural Language Processing (almost) from scratch. Journal of Machine Learning Research 12, pp.2493-2537.
2. Hai, Z. Chang, K. Kim, J. and Yang, C (2014). Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance. IEEE transactions on knowledge and data engineering, Vol. 26, No. 3.
3. Hatzivassiloglou, V. & Wiebe, J.M. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. Proc. 18th Conf. Computational Linguistics, pp. 299-305.
4. Pang, B. & Lee, L (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics.
5. Mcdonald, R. Hannan, K. Neylon, T. Wells, M.& Reynar, J (2007). Structured Models for Fine-to-Coarse Sentiment Analysis. Proc. 45th Ann. Meeting of the Assoc. of Computational Linguistics, pp. 432- 439.
6. Qu, L. Ifrim, G. & Weikum, G (2010). The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. Proc. 23rd Int'l Conf. Computational Linguistics, pp. 913-921.
7. Yessenalina, A. & Cardie, C (2011). Compositional Matrix-Space Models for Sentiment Analysis. Proc. Conf. Empirical Methods in Natural Language Processing, pp. 172-182.
8. Jin, W. & Ho, H.H (2009). A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining. Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 465-472.
9. Qiu, G. Liu, B. Bu, J. & Chen, C (2011). Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics, vol. 37, pp. 9-27.
10. Pang, F. B. & Lee, L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics.
11. https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

APPENDIX: Summary of the UberCloud Software Containers Suitable for NLP

UberCloud High-Performance Computing Containers are ready-to-execute packages of software. These packages are designed to deliver the tools that an engineer needs to complete his task. The ISV or Open-Source tools are pre-installed, configured, and tested, and are running on bare metal, without loss of performance. They are ready to execute, literally in an instant with no need to install software, deal with complex OS commands, or configure. The UberCloud Container technology allows wide variety and selection for the engineers because they are portable from server to server, Cloud to Cloud. The Cloud operators or IT departments no longer need to limit the variety, since they no longer have to install, tune and maintain the underlying software. They can rely on the UberCloud Containers to cut through this complexity. This technology also provides hardware abstraction, where the container is not tightly coupled with the server (the container and the software inside isn't installed on the server in the traditional sense). Abstraction between the hardware and software stacks provides the ease of use and agility that bare-metal environments lack. Finally, these containers run on top of the UberCloud Engineering Simulation Platform, which is fully automated, self-service, and are easily migrated from any cloud to any other cloud.

Join the UberCloud Experiment or Contact Us for Your Proof of Concept

If you, as an **end-user**, would like to participate in this Experiment to explore hands-on the end-to-end process of on-demand Technical Computing as a Service, in the Cloud, for your business then please register at: http://www.theubercloud.com/hpc-experiment/

If you, as a **service provider**, are interested in promoting your services through the UberCloud then please send us a message at https://www.theubercloud.com/help/

UberCloud Library of Case Study Compendiums at https://www.theubercloud.com/ubercloud-compendiums

HPCwire Readers Choice Award 2013: http://www.hpcwire.com/off-the-wire/ubercloud-receives-top-honors-2013-hpcwire-readers-choice-awards/
HPCwire Readers Choice Award 2014: https://www.theubercloud.com/ubercloud-receives-top-honors-2014-hpcwire-readers-choice-award/
Gartner Names UberCloud a 2015 Cool Vendor in Oil & Gas: https://www.hpcwire.com/off-the-wire/gartner-names-ubercloud-a-cool-vendor-in-oil-gas/
HPCwire Editors' Choice Awards 2017 & 2018: https://www.hpcwire.com/2017-hpcwire-awards-readers-editors-choice/
IDC/Hyperion Innovation Excellence Awards 2017 & 2018: https://www.hpcwire.com/off-the-wire/hyperion-research-announces-hpc-innovation-excellence-award-winners-2/

If you wish to be informed about the latest developments in technical computing in the cloud, then please **subscribe to our monthly newsletter http://info.theubercloud.com/subscribe-to-newsletter**