Twitter Sentiment Analysis on Covid19 Tweets

AWS Services Implementation for Sentiment Analysis

Ahmad Tariq Ali Alsaqqaf (2017176)
Mert Burak Burabak (2020699)
Raziye Ataseven (2020676)

June 14, 2021

Introduction

Cloud computing is around quite some time and can collaborate and be utilized with other disciplines such as big data analysis, software development, devops etc. which by handling, storing, visualizing and processing the huge amount of data within the challenging situations. Since institutions that benefit from cloud computing increase the reliability and usability of their technological products, this also directly affects their own reliability and profit. When we consider a case involving big data, there are many more scenarios and requirements that must be handled.

In our study, we are taking advantages of cloud computing such as Comprehend; pre-defined natural language process models, S3; strong file interface and storage power, SageMaker; Python development environment for create, train and building ML models and making big data analytics, DynamoDB; flexibility and high availability, QuickSight; as last and also powerful tool for analysing data for in a minutes. In this study, the sentiment analysis examined tweet datas in Covid19 context. Beside the sentiment analysis, there had been a statistical, linguistic and geographic analysis using the specified AWS tools and frameworks.

1. Group Members and Project Contributions

	Ahmad	Mert	Raziye
Scenario definition	R	R	R
Data collection	S	S	R
Data preprocessing and exploring	R	R	R
AWS Components implementation	S	R	R
TextBlob sentiment analysis with Python	S	S	R
Word cloud and frequency	R	S	S
Data Analysis and visualization	R	R	R

R: Responsible, S: Supportive

2. Scenario Definition

The project is about sentiment analysis for Twitter tweets (English tweets only) in covid19 time, in which we are going to analyze the positivity and negativity of the tweets in (April 2020) and (April 2021) and find the difference between them. Also, we are going to study the emotions of the people while they are sitting at their home during the pandemic and what is the most used words were during that time, and the most used words after one year from that time, to find if it still the same or it is different, based on countries.

Also, it's a good opportunity for government and health institutions to study how people react toward covid19 while they are sitting at their homes doing nothing, and how their reactions when vaccination was found, and the possibility that people accept to take this vaccine. In this research, we will study the sentiments and difference between two time periods, and see if people's emotions, feelings, and mood if it is getting better or still the same since the start of covid19 in 2020 using Twitter tweets.

3. Solution

3.1. Data collection

Detailed tweet data, which includes tweet text, creating date, context and much more information can not be found as a dataset in open source as it violates the data privacy policy. In order to reach full tweet text with specific context, in this case it is Covid19, we must collect data from twitter's api itself. First we create a twitter developer account to fetch tweets. [5] There was some tweet id dataset for Covid19 related [6], and we can dehydrate these records by twitter api. Api requests are made with the Python's Tweepy library and the return value of json type is saved in the file system.

3.2. Data preprocessing and exploring

Create date, full text, id string, location, and language property of the data collected and created a dataset. For each row, It will implement sentiment analysis on text data of tweets on Comprehend and TextBlob, and then It will be analyzed regarding their creation time and place such aspects as sentiments and words used. First of all, we checked if the dataset includes any null value for full text and location. After that, we filtered by language column to eliminate non-English texts. After null or empty value checking, tweet text is cleaned from the stop words and hyperlinks and all characters except alphabets upper and lower case and numbers. It uses pandas, regex, and nltk libraries for this process. Also we did not have any geospatial data for twitter users, instead we have manually added data from users in their biography, such as "Istanbul, Turkey" or non-real locations that made up or null values. For grouping the location data, we splitted the location string and used excel functions to match and aggregate city or district names within the country.

3.3. AWS Components implementation

In order to realize our project on AWS, in the first step, we created the sagemaker authority with the necessary privileges over the IAM Role. Required authorizations are as follows: s3 full access, sagemaker full access, dynamodb full access, comprehend full access. Later, we have uploaded the data prepared on the AWS S3 service, which allows more storage and sharing, to be used in other services. After the bucket we prepared, we created a dynamodb table so that we can print the results. AWS dynamodb is a no-sql database.

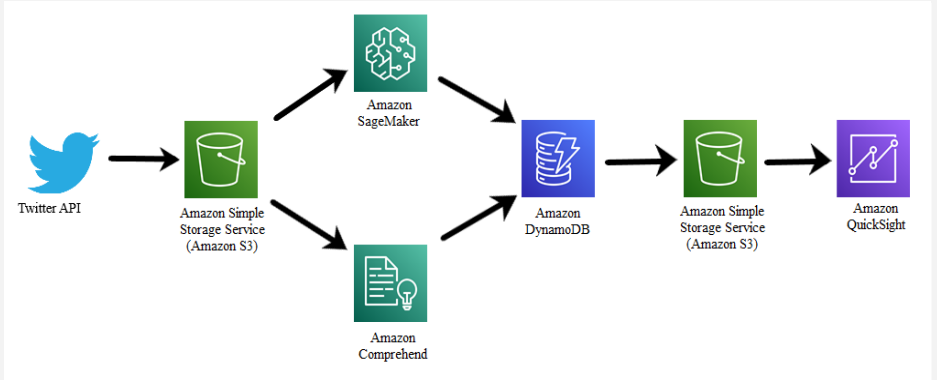
Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It is a fully managed, multi-region, multi-active, durable database with built-in security, backup, restore, and in-memory caching for Internet-scale applications[1]. After creating the table, we created a notebook on sagemaker with the help of AWS sagemaker studio. Amazon SageMaker brings together a suite of built-in ML capabilities to help data scientists and developers prepare, build, serve for education, and deploy high-quality machine learning (ML) models[2] .

In Sagemaker, we first called our tweets that we uploaded to the s3 bucket. We have called the comprehend service and the dynamodb table that we have authorized. Then we performed sentiment analysis with the help of comprehend service for all texts. Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to uncover information in unstructured data. Instead of scanning documents, the process is simplified and invisible information is easier to understand[3]. It is also used for sentiment analysis tasks. Finally, we printed the text itself and the sentiment analysis results to the dynamodb table we created. The pseudocode of the script we created in Sagemaker is as follows:

CALL-> s3_bucket
CALL -> comprehend
CALL-> dynamodb_table
DECLARE texts from s3_bucket

For text in texts:
    Sentiment = comprehend(text)
    Put text_id, text,location,text_date,sentiment into dynamodb_table

We restored the dynamodb table that we created with Sentiment analysis to the s3 bucket for public use and made it public. We used Amazon Quicksight for the final stage. Amazon QuickSight is a scalable, serverless, scalable, machine-learning business intelligence (BI) service built for the cloud. QuickSight lets you easily create and publish interactive dashboards with Machine Learning-enhanced insights[4]. We interpreted our results by analyzing and visualizing the results we produced on quicksight. Below you can find the framework of the process we carried out on Amazon Web Service.



(Figure 2. AWS Framework)

3.4.TextBlob sentiment analysis with Python

3.4.1. Textblob library

Natural Language Processing (NLP) is a popular area with increasing attention and need for chatbots, machine translation etc. Being able to interact with and understand humans is one of the most popular, evolving and challenging topics among many intelligent machine disciplines. TextBlob is a Python library for NLP, which supports analysing, clustering and operating on complex textual data successfully. It is used with Natural Language ToolKit (NLTK) to access some lexical resources. In this way, it plays a very important and facilitating role in solving tasks such as categorization and classification. In lexicon-based approaches like TextBlob, there is a pre-defined dictionary that each word is classified by its negativity and positivity. While analyzing the sentiment of a paragraph or sentence, each word is scored according to its semantic orientation and intensity and averaged in the final.

In this study, sentiment analysis is performed on tweets that are published in certain periods and talk about covid19, as explained before. Each tweet will be scored by sentiment with TextBlob's sentiment.polarity function. Output is between -1 and 1 and its classification is; for 0 value it is labeled as "Neutral", between 0 and -1 "Negative" and between 1 and 0 "Positive". After implementing sentiment analysis to tweet data, we get the sentiment results as 19.8 % Negative, 38.9 % Neutral and 41.3 % Positive regarding the population of the entire dataset.

3.4.2. Comparing the Comprehend and TextBlob sentiment results

We obtain 4 different sentiment analysis categorical sentiment values for each tweet we analyze with the AWS Comprehend API; Neutral, Mixed, Positive and Negative results. In Figure 3, we can say easily that, dominant sentiment is Neutral at 77.7 %. Negative (11.5) represents relatively more than others, but a relatively small percentage of the population in the Mixed (6.8). Positivity remains at 4 percent.

TextBlob contains 3 types of categorical sentiment values such as Neutral, Positive and Negative. Analysing the same tweet data, which we investigated before with Comprehend, and analysing again using another strong lexicon-based approach, Textblob. Its result could be shown in Figure 4. When we compare it with the sentiment value analyzed with the Comprehend API, it is noticed that the percentages of neutral and positive values are close to each other. When we create a heatmap (in Figure 5.), we can see the relation, -correlation between two different sentiment analysis.

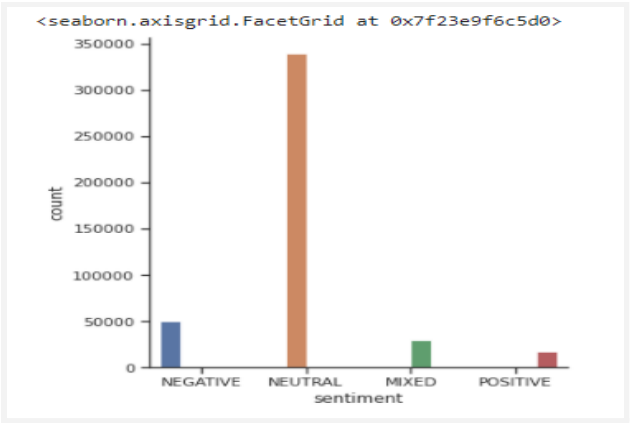


Figure 3. Comprehend sentiment result

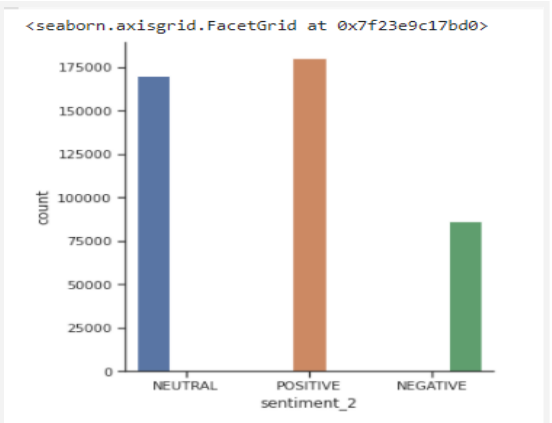
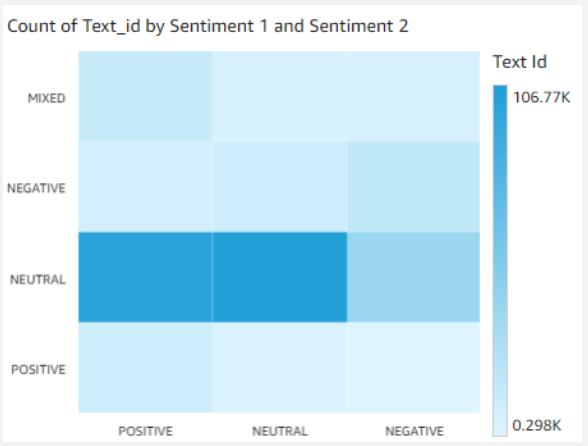


Figure 4. TextBlob sentiment result



When we look at the heatmap, we can see the dark areas as strong relation, brighter colors represent weak relation. There is strong correlation between Comprehend's (rows) Neutral and TextBlob's (columns) Positive and Neutral, and there is weak relation between TextBlob's Negative.

Figure 5. Sentiment heatmap

Word cloud and frequency

Word cloud had been created for this project to see the most frequent words in the tweets in different countries and different times, in order to compare between two times and see the difference. As shown in Figure 6, a sample of word cloud for the United States in April 2021, we can see that there are many "Covid19 vaccine" and "coronavirus vaccine" words used in that time in tweets.

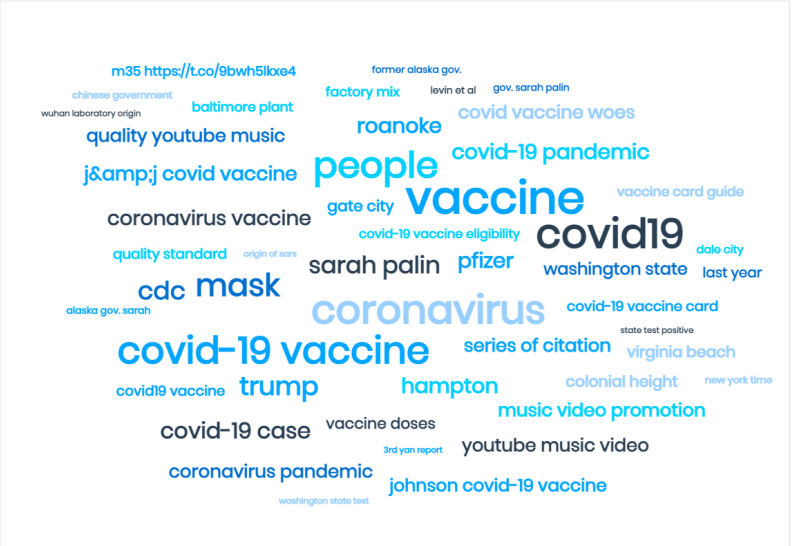


Figure 6. Word cloud for united states tweets for April 2021

3.5.Data Analysis and visualization

After finishing all parts we analyzed the data to get some insights and visualize the data in a good manner to make it clear for the reader. First, we had to collect all data in an excel sheet and start analyzing everything to get these insights.

Second, we drew some charts to compare and show the results more clearly for the tweets for the top countries that tweeted at that time as a sample to show the idea of the project. So, next will show some charts to make the idea more clear, as shown in Figure 7 and Figure 8. It shows the top used words in the United States at April 2020 at the beginning of covid19 and after one year at April 2021 when the vaccination was found. From these figures, we can see how the people react to what happens in the world and each country from the words that are being used and the emotional state for those people while they are writing their tweets, maybe it's not accurate one hundred percent, but at least it gives some indications.

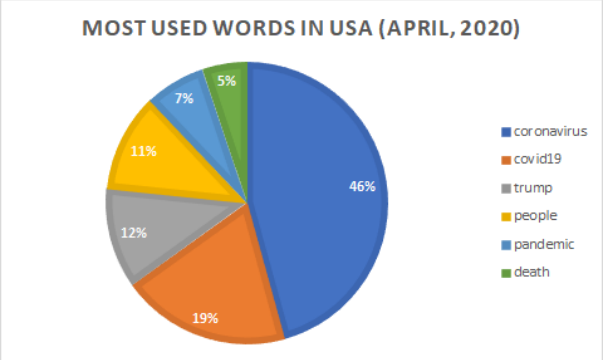


Figure 7. USA (April, 2020)

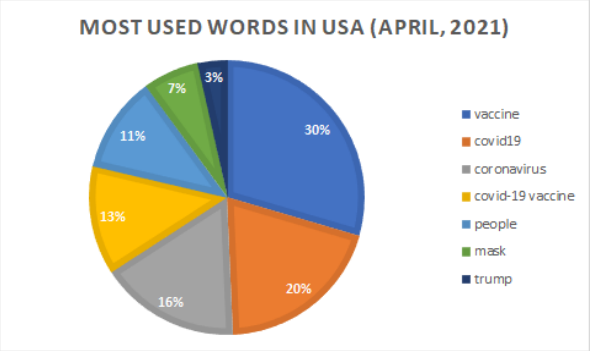


Figure 8. USA (April, 2021)

4. Conclusion

Cloud computing technology surely will not stop here, and everyday more companies will depend on this technology to run their business. In this project we used some services of this technology to conduct sentiment analysis for english twitter tweets from all over the world. In this project we had collected data from Twitter API as we mentioned earlier in this report, and we made some operations and analysis on this data to make them ready for use. Next, we used Comprehend and TextBlob service in Amazon Web Service (AWS) to implement sentiment analysis on the tweets and make it ready for the next step, also we added tweets creation time and location to make more accurate analysis based on these fields.

Moreover, we created sagemaker authority to some privilege over the IAM Role. Also, we used a python library named "Textblob library" in order to help us in analyzing, clustering and operating the complex textual data set in a good way.

Last but not the least, we compared comprehend results with textblob results, then started analysing due to location, time and used words using word cloud tool, and the last step was to draw charts and visuals of the data in a good and understandable manner.

5. References

[0] Dataset: https://group-project-bucket-bau.s3.amazonaws.com/AWSDynamoDB/01622401634167-f5b72026/data/crxroakjsq5mpokvoq7rlbvujm.json.gz

[1] - https://aws.amazon.com/en/dynamodb/

[2] - https://aws.amazon.com/en/sagemaker/

[3] - https://aws.amazon.com/tr/comprehend/

[4] - https://aws.amazon.com/en/quicksight/

[5] - https://developer.twitter.com/en/docs/twitter-api

[6] - https://github.com/thepanacealab/covid19_twitter