VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

WEB DATA ANALYTICS

BEHAVIOR OF CUSTOMER IN AMAZON

(STARBUCK COFFEE)

Lecturer : PHD.Ho Trung Thanh

Class: 221IS9901 – Group 2

2022- 2023

NUM	FULL NAME	STUDENT CODE	MISSION	EVALUATE
1	Ngô Đức Huy	K214131983	Code, Chap 4, Chap 5	75%
2	Nguyễn Gia Phong	K214130899	Chap 3, Research Method	75%
3	Nguyễn Tuấn Anh	K214130877	Chap 1, Power Point	75%
4	Đỗ Trần Duy Khang	K21413	Chap 2, Research Method	75%
5				

Leader’s information: Ngo Duc Huy

Gmail: huynd21413@st.uel.edu.vn

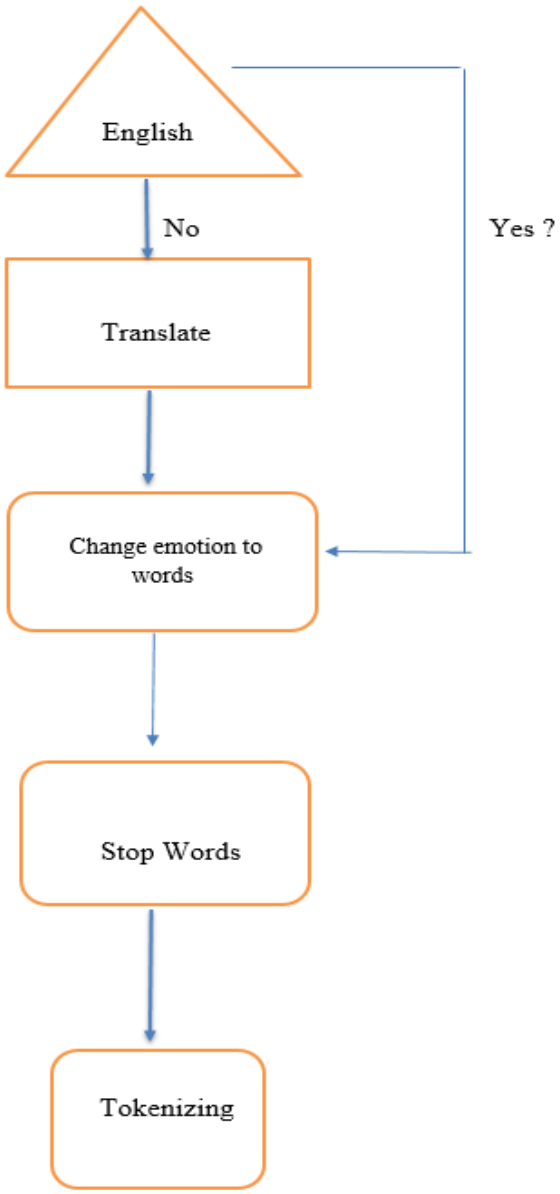
Phone number : 0816189203

Lecturer’s comment :

.....
.....
.....
.....
.....
.....
.....
.....

LECTURER	**LEADER**
(Signature)	(Signature)
PHD.Ho Trung Thanh	Ngo Duc Huy

TABLE OF CONTENT



CONTENT

I . PROJECT OVERVIEW

1. Reasons for choosing this topic

In the global economy, e-commerce and e-business have become an essential element of business strategy and a powerful catalyst for economic development. The integration of information and communication technology (ICT) into business has revolutionized relationships within organizations and between organizations and individuals. In particular, the use of ICT in business has enhanced productivity, encouraged greater customer attendance, and facilitated large-scale customer service, in addition to reducing costs. With the development in Internet technology and Web technology, the difference between the traditional markets and the global network market. Moreover, covid-19 has affected the buying trend from traditional to online shopping. Following that, e-commerce platforms also developed, including the giant Amazon.

Customer opinion is extremely important for e-commerce platforms and their commercial success. The goal of e-commerce platforms is to generate sales and increase customer satisfaction, which is why helping customer reach their goals efficiently and effortlessly is an essential part of the online customer journey. Customer opinion analysis is essential and helps businesses understand more deeply about the characteristics, wants, needs and problems that customers are facing. Thereby, it is possible to predict the buying behavior of consumers in the future and offer more attractive solutions.Every day on the amazon e-commerce platform, there is a large amount of online comment data in the form of text generated daily by customers after making a purchase on the website.Because it is difficult to understand customers through their experience, we wanted to analyze the customer's point of view in this area. For the purpose of understanding customers to focus on developing marketing and business strategies more effectively.

2.Objectives of the study

- Understanding customers through opinion analysis to increase experience

- The model of aspect extraction, opinion mining and experimentation proposed model in Vietnamese is customer feedback in the field of e-commerce (amazon), in which:
    - Technical application
    - Experiment to build an automatic reporting system from model results.

- Of course, in business, mistakes and risks cannot be avoided, no matter which game you join. Amazon is no exception! But at such a risk, many people are still trying to race to sell coffee on Amazon? Because, we see opportunities in it. And when we want to seize opportunities, we must take risks.

- Benefits of selling on Amazon:
    - Amazon's reach and influence: Amazon's global e-commerce sales will reach $416.48 billion in 2020, leading other e-commerce sites with a market share of 38.7%. Amazon owns 18 Websites in major markets such as North America, Europe, Japan, India, etc. Up to now, Amazon has been present in 185 countries and territories. Order fulfillment center (Amazon's Fulfillment Center) currently has more than 175 warehouses and this number is still growing. As of the first quarter of 2019, Amazon has collected more than $ 145.2 million - high in revenue. in the world in the e-commerce industry. The latest financial figures from Amazon show that the company posted $75.5 billion in sales in the first quarter of 2020. Of these sales, net income accounted for $2.5 billion.
    - The benefits of selling on Amazon that make you want to join right away: Huge customer base; Amazon has the ability to change the shopping behavior of customers; Booming revenue; "Leverage" brand, affirming prestige and product quality; Enterprises have the opportunity to quickly expand and develop business activities; Collect information, adjust business strategies; build good relationship with customers;

Amazon helps businesses manage their inventory with the world's #1 warehousing service.

- Situation in recent years: Since the end of 2019, the Covid-19 epidemic began to form, chain stores, also known as F&B, are going downhill. To solve this problem, e-commerce chains started to float and rise faster and faster. In which amazon is a typical example. Thanks to that, the team decided to make a report on an e-commerce platform, especially amazon.

3.Object and scope of the study

a.Study Object

- Method of extracting aspects and mining opinions from customer feedback
    - Customer feedback in e-commerce is an extremely useful and valuable source of information to reflect the quality of products or services, not only helping Customers decide whether to buy or not, but also help businesses understand customer behavior and experience to improve product quality and better service. However, over time, the increasing number of comments and increasing interference make it increasingly difficult to understand customers with manual methods.
    - Extract customer opinions, then ask the question: Why choose amazon, why buy on amazon?

- Customer perspectives and aspects to be found in response to product reviews on AMAZON
    - Customers review products for different reasons and from various perspectives. On the one hand, Amazon Reviews are of a positive nature and other customers want to be sincere in recommending an item, but at the same time, customers are expressing their frustration regarding insufficient product information or a poor product quality, and want to dissuade others from purchasing. Particularly in the latter case, sellers want to enter into direct communication with customers, to gather feedback, remove any doubts, or answer questions. Negative reviews are lessened through this and shoppers are given the positive feeling that they are cared about and that the seller is being transparent. However, from the 15th December 2020, Amazon removed the opportunity to communicate directly with customers – to the dismay of many

sellers. Since then, unwarranted or unintelligible negative reviews can now no longer be commented on. Amazon considers this function superfluous, as it is apparently so rarely utilised, but according to Amazon's own sources, they are working on other ways of facilitating connecting with customers and continuing to support successful selling. Given this, it is even more critical to address as many potential customer questions or problems as possible through an optimised Amazon product description.The question-and-answer section is also ideal for sellers to clarify customer queries in advance.

- A valuable aspect of Amazon Reviews is the possibility to use customer feedback as a form of market research. Sellers should pay attention to useful customer voices, to continually optimise their products and receive input for new listings. Customers often only really know what they want when they have it in their hands. Numerous aspects of a product can subsequently be optimised afterwards if new customer needs are discovered post-purchase:
    - ➔ Any missing product information in the description or on the packaging can be adapted.
    - ➔ Problems with packaging can be addressed, so that customers do not receive damaged or incorrect products.
    - ➔ Accompanying information like instruction manuals can be improved.
    - ➔ Do customers want additional functions? This data can be used for new products.

b.Research Scope

- The analyzed data is a set of online customer comments in Vietnamese, including usernames, comment content, rating points on a scale of 1-5,...
- Data is collected from the most popular e-commerce AMAZON about Starbucks.

4. Research methods

Survey and study secondary data from previous studies relevant to the research objective.

Survey customer comments on the Website and find the right data source for research to solve problem posed.

Survey the theory of customer experience in e-commerce, survey models and mining methods

User opinions, analyze survey results to find gaps from previous studies

Use descriptive system method to survey and evaluate the collected data.
• Apply method partitioning classes in machine learning such as Logistic Regression, Bayes Naïve, Support Vectors and
Decision tree on file architecture client.
• Conduct model testing, compare and evaluate the model's results

5. New Contributions

- Building a dictionary of opinions (negative or positive) in the field of e-commerce in Vietnamese language based on adjectives.
- The proposed research model, methods and experimental results can be applied in practice in the field of behavior analysis and customer experience in e-commerce in general and especially Amazon in particular. .
- The research results can also be applied to building data analysis systems, network listening systems (Social listening). Analysis of the online community's opinion on e-commerce platforms

6. Project Structure

I. Project overview
- Step 1: Identify the topic
- Step 2: Develop a proposal
- Step 3: Create a research plan
- Step 4: Collect and process information
- Step 5: Write a thesis statement
- Step 6 Seek new ideas

II. Review of related studies
- Perspective analysis
- Perspective analysis based on opinion

III. Theoretical basis

IV. Proposed and experimental research model
- CInfluencing factors for online purchases
- Model of the overview research Data
- Collection Data
- Exploratory analysis
- Data preprocessing
- Model application and point of view

V. Evaluation of experimental results and conclusions

II. LITURATURE REVIEW

The articles are pretty thrilling and associated with this system the organization is researching:

- **Behind a cup of coffee: international market structure and competitiveness**
    - ✓ Author : Tafarel Carvalho Gois, Karim Marini Thomé, Jeremiás Máté Balogh ,Competitiveness Review
    - ✓ Publishing year : 2022
    - ✓ Purpose: This study aims to analyse the structure and the competitiveness of the international coffee market.
    - ✓ Design/methodology/approach: To describe the international market structure, this study uses Herfindahl–Hirschman index, net export index (NEI), and to measure export competitiveness revealed symmetric comparative advantage (RSCA). Finally, survival function analyses were developed using the Kaplan–Meier product-limit estimator to characterize the stability and duration of the competitiveness in the international coffee market. The results reveal that the imports and exports market structure are unconcentrated. NEI shows that several countries are stable in their commercial characteristics (imports, exports and re-exports), nevertheless, NEI also revealed countries

transitioning through the commercial characteristics, that the international coffee market structure presents dynamic commercial characteristics. The result for (RSCA shows that Uganda, Ethiopia, Honduras, Brazil, Colombia, Guatemala and Indonesia had the highest values and also resulted in better survival rates along with Italy, India, Mexico and Switzerland. The stability of RSCA indices is investigated by regression analysis, showing a tendency to increase expertise in coffee exports from 2015.

- ✓ Keywords: coffee market, stability,Revealed symmetric comparative advantage, market structure.

IN OUR OPINION:

- The strong point of the article is the results reveal that the imports and exports market structure are unconcentrated. NEI shows that several countries are stable in their commercial characteristics (imports, exports and re-exports), nevertheless, NEI also revealed countries transitioning through the commercial characteristics, that the international coffee market structure presents dynamic commercial characteristics. The result for (RSCA shows that Uganda, Ethiopia, Honduras, Brazil, Colombia, Guatemala and Indonesia had the highest values and also resulted in better survival rates along with Italy, India, Mexico and Switzerland. The stability of RSCA

indices is investigated by regression analysis, showing a tendency to increase expertise in coffee exports from 2015.

- The weak point of the article is NEI also revealed countries transitioning through the commercial characteristics, that the international coffee market structure presents dynamic commercial characteristics.The result for (RSCA shows that Uganda, Ethiopia, Honduras, Brazil, Colombia, Guatemala and Indonesia had the highest values and also resulted in better survival rates along with Italy, India, Mexico and Switzerland.

The group came to 1 conclusion about the paper: This study provides a comprehensive and recent analysis of the international coffee market structure and competitiveness, contributing to the analysis of the international market of the product.

- **Online customer reviews: insights from the coffee shops industry and the moderating effect of business types**
    - ✓ Author: Shuting Tao, Hak-Seon Kim
    - ✓ Publishing year: 2022
    - ✓ Purpose: This study aims to explore the hidden connectivity among words by semantic network analysis, further identify salient factors accounting for customer satisfaction of coffee shops through analysis

of online reviews and, finally, examine the moderating effect of business types of coffee shops on customer satisfaction.

- ✓ Design/methodology/approach: Two typical major procedures of big data analytics in the hospitality industry were adopted in this research: one is data collection and the other is data analysis. In terms of data analysis, frequency analysis with text mining, semantic network analysis, CONCOR analysis for clustering and quantitative analysis with dummy variables were performed to dig new insights from online customer reviews both qualitatively and quantitatively. Different factors were extracted from online customer reviews contributing to customer satisfaction or dissatisfaction, and among these factors, the brand-new factor "Sales event" was examined to be significantly associated with customer satisfaction. In addition, the moderating effect of business types on the relationship between "Value for money" and customer satisfaction was verified, indicating differences between customers from different types of coffee shops.
- ✓ Key word: COFFEE SHOP, ONLINE CUSTOMER REVIEWS , TEST MINING, SEMANTIC NETWORK ANALYSIS, BUSINESS TYPES.

IN OUR OPINION :

- The Strong point of the article : Two typical major procedures of big data analytics in the hospitality industry were adopted in this research: one is data collection and the other is data analysis. In terms of data analysis, frequency analysis with text mining, semantic network analysis, CONCOR analysis for clustering and quantitative analysis with dummy variables were performed to dig new insights from online customer reviews both qualitatively and quantitatively.
- The Weak point of the article : The present study broadened the research directions of coffee shops by adopting online customer reviews through relative analytics. New dimensions such as “Sales event” and detailed categorization of “Coffee quality”, “Interior” and “Physical environment” were revealed, indicating that even new cognition could be generated with new data source and analytical methods. The industry professionals could develop their decision-making based on information from online reviews.

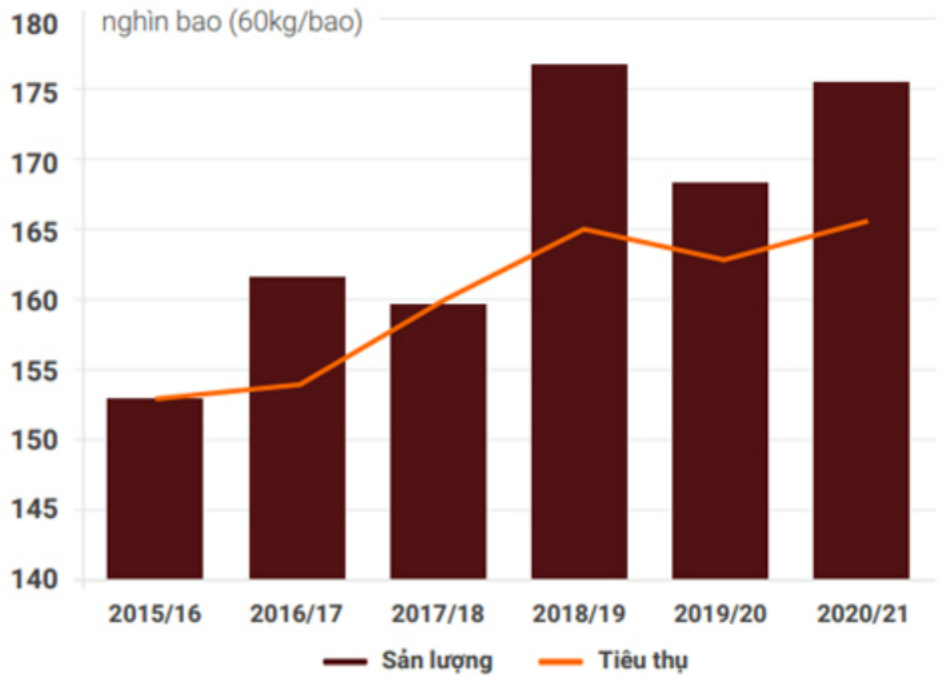
  The group came to 1 conclusion about the paper: The present study used online reviews to understand coffee shop costumer experience and satisfaction through a set of analytical methods. The textual reviews and numeric reviews were concerned simultaneously to unearth qualitative perception and quantitative data information for customers of coffee shops.

  Generally :

Currently, along with certain growth steps in both clean and pure coffee products and coffee drinks, the status of dirty, padded, and cheap coffee is becoming more and more popular in the market, causing confusion for consumers. user.

In fact, no one can be sure that they are drinking a cup of clean, quality coffee that does not contain impurities or does not contain any harmful ingredients. Through a number of articles talking about the current dirty coffee crop, it can be seen that most of the ingredients for making coffee are soybeans, the cob (corn) is roasted black and mixed with a little coffee to create the smell. fragrant, then pureed, packaged, labeled beautifully and then released to the market.

Coffee with impurities or using a lot of chemical fertilizers, gas, and coal in the coffee roasting process also makes coffee prices go down. The toad shops on the sidewalk, in the labor alley, the price of drinking an iced coffee is about 0.3USD-0.4USD/1P. The selling cost of the shopkeeper when taking each kilogram of coffee is about 3-4 USD. At that price, the so-called real coffee, clean coffee only has roasted soybeans and flavoring chemicals. The process of processing cheap coffee is as follows: soybeans, foaming chemicals, cocoa flavorings, butter, bitter chemicals, sugar, salt, fish sauce...

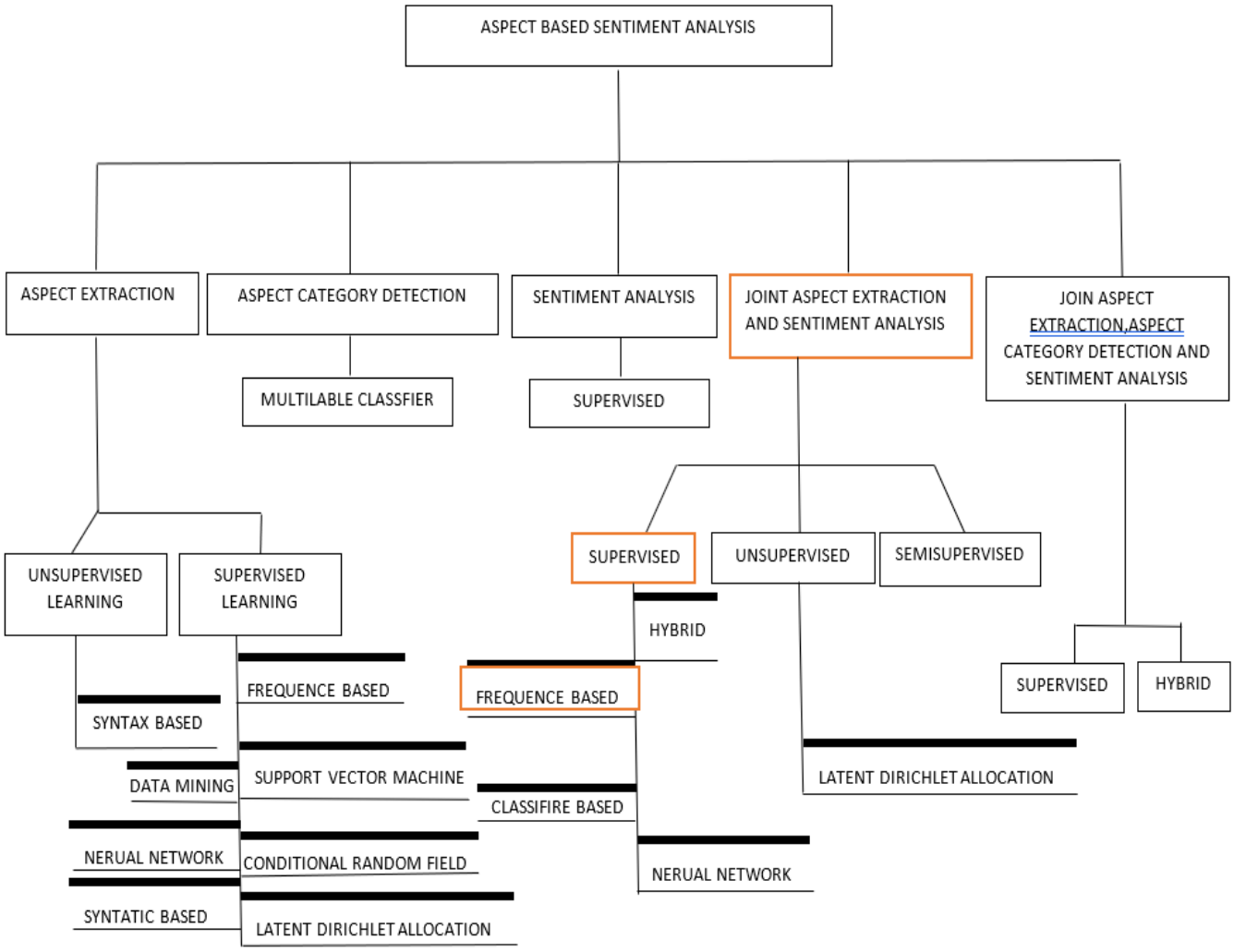


World coffee consumption

Through 15 reports the group has made, the team took representatives of 2/15 of those reports to evaluate the current coffee situation and came to the following conclusions:

- Coffee is good for you. Or it’s not. Maybe it is, then it isn’t, then it is again. If you drink coffee, and follow the news, then perhaps you’ve noticed this pattern.
- A recent study showed that coffee, even sweetened, was associated with health benefits. But other studies have come to more mixed conclusions.

- What's driving these pendulum swings in the health status of coffee? Like a good cup of coffee, the answer is complex, but seems to boil down to human nature and scientific practice
- Is coffee good for you? Yes, in the sense that it will wake you up, brighten your mood, maybe even give you an excuse to get out of the house and chat with friends at a local coffee house.
- Will drinking coffee make you healthier or help you live longer? Probably not. Sure, the antioxidants in our morning cup could actually be helping our bodies, but there are far better ways to boost your antioxidant intake. So, wake up with a strong cup of coffee, but stay healthy with a complex and varied diet.



Choosing Methods to Analytics

III. THEORETICAL BASIS

III.1. Amazon's Orientation

Amazon’s retail ecommerce income was progressively increasing, with a spike visible in 2020 reaching $309 billion bucks way to consumers’ pandemic-brought on demands. As the store who dominates the ecommerce industry, it’s no marvel that they’re making strides to keep—and expand—that position. How? Learn about these Amazon trends unfolding for 2022.

1. Online grocery

- Online grocery is a number of the main Amazon ecommerce trends. The business enterprise has experimented with numerous grocery codecs over the last decade, however its cutting-edge services include:
  - Amazon.com
  - Amazon Fresh
  - Whole Foods Market
  - Amazon Go

- Insider Intelligence estimates that Amazon’s US grocery ecommerce income will develop 12.9% this 12 months to $29.12 billion, amounting to 23.8% of all virtual grocery income. It’s the second-biggest virtual grocer after Walmart, with a view to seize 25.3% of virtual grocery income this 12 months (and 28.9% whilst such as its subsidiary Sam’s Club).

- We forecast that US grocery retail ecommerce income will almost double over the subsequent 4 years, developing from $122.39 billion in 2021 to $243.sixty seven billion in 2025. In the US, virtual income is nonetheless a small portion (9.6%) of normal grocery income, and Amazon’s grocery commercial enterprise isn’t almost as aggressive offline, however given Amazon’s placement as an ecommerce powerhouse, it's miles primed to generate a massive chew of this growth.

2. **Amazon one**
   - Amazon One is a biometric-primarily based totally price terminal that permits clients to authenticate in-man or woman transactions with a palm scan. It was released in September 2020, after which accelerated into Seattle and New York locations, which includes a few Whole Foods Market stores, in overdue wintry weather and early spring 2021.

   - Launch situations had been pretty favourable, as contactless bills surged whilst the pandemic made clients extra health-aware and touch-averse. Increased familiarity with price alternatives like cellular wallets ought to make clients even extra inclined to check different new platforms, mainly in the event that they provide even much less touch than a phone.

   - Amazon hopping into the biometric scanning ring now needs to provide it an early-mover advantage, specifically seeing that there isn't a whole lot specially like Amazon One available in the marketplace yet. But the circulate may sign to opponents that hobby is mounting and boost up industry wide era development. Down the line, the organisation plans to licence the era to different involved parties, together with retailers, workplace buildings, and stadiums, which can convey in a brand new sales stream. And transferring past retail may want to preview a push into identification verification greater broadly.

3. **Amazon advertising.**
   - Amazon netted $26.31 billion international and $20.47 billion withinside the US in 2021 in marketing and marketing revenues, in step with our maximum latest forecast. Amazon commercials had been increasing swiftly for years, and the boom extended currently due to the pandemic and the related step extrade in the percentage of retail transactions performed online.

   - Amazon's advertising business has several facets:

     - Advertisers that sell products on Amazon can buy impressions based on cost per click on Amazon's retail properties. We

estimate that the majority of Amazon's net US ad revenues come from these types of placements.

- Amazon DSP is a demand-side platform (DSP) that allows advertisers, including non endemic brands, to buy display impressions across Amazon's retail properties; its non retail properties like Twitch and IMDb TV; and the programmatic web.

- Amazon Publisher Services provides header bidding integrations for publishers to access both Amazon demand and demand coming through other supply-side platforms (SSPs).

- In its upward thrust to turn into a member of the virtual marketing and marketing triopoly, Amazon revolutionised the quest advert marketplace, which has lengthy been ruled by way of means of Google. Google nonetheless nets the bulk of US seek advert revenues, however it's Amazon—now no longer every other popular seek engine, inclusive of Bing—that has taken a big percentage of the marketplace at Google's expense.

4. Fulfilment by Amazon

- As Amazon's ecommerce commercial enterprise took off, it found out that achievement—now no longer best transport, however additionally returns, warehousing, and stock management—become a vital part of its retail operations. Amazon achievement facilities are characteristic one-of-a-kind centres for receiving goods, sortable select out and facilities for small items, and regions devoted to big items, to name a few.

- Separate from its logistics prowess is the transportation community Amazon has built. Amazon is the fourth-biggest transportation community withinside the US, in keeping with Bank of America Global Research. It shipped 415 million applications in July 2020 alone, handing over 66% of its very own applications throughout the month, a growth of 12 percent factors YoY, according to ShipMatrix. This

demonstrates Amazon's dedication to developing a cease-to-cease fulfilment/transport arm—one which third-celebration dealers usually experience being obliged to apply to be extra a hit at the platform, notwithstanding its better charges whilst as compared with different suppliers. Nearly 85% of Amazon's largest dealers use its Fulfilment with the aid of using Amazon (FBA) service.

- The greater Amazon has over fulfilment, and frequently forgotten however friction-inducing a part of the patron journey, the much more likely it's far to satisfy consumers' expectancies of comfort and preserve its logo ethos whilst including to its backside line.

5. Gaming

- Amazon's Twitch subsidiary is a pioneer and class chief in game-oriented streams and is increasing right into a broader social video platform. We expected that Twitch might have 31.four million US customers in 2021, developing its base to 36.7 million with the aid of using 2025. Amazon is weaving this commercial enterprise unit into broader leisure reviews that consist of influencer advertising lifestyle and social video.

- Like different elements of Amazon's commercial enterprise, Twitch benefited from pandemic-brought on lockdowns that boosted home-primarily based totally amusement sports inclusive of gaming. This Amazon commercial enterprise unit helped deliver an upward push to esports and could benefit an increasing number of enjoy the developing reputation of sport streams and the influencers who create them. The service's "Just Chatting" feature—a catchall venue in which creators can speak approximately their paintings in informal streams—has become its most-watched class in Q3 2020 and has persevered to develop considering the fact that then, illustrating the engagement that the Twitch network can generate.

III.2 Overview of supervised machine learning methods.

Supervised machine learning is an algorithm that predicts the output of a new data based on pairs called data,label.

❖ Logistic regression

▪ What is logistic regression?

- Logistic regression is a statistical evaluation technique that expects a binary outcome, inclusive of sure or no, primarily based totally on earlier observations of a records set.

- A logistic regression version predicts a based information variable with the aid of studying the connection among one or greater current impartial variables. For example, a logistic regression might be used to predict whether or not a politician will win or lose an election or whether or not an excessive faculty scholar can be admitted or now no longer to a specific college. These binary results permit truthful selections among alternatives.

- A logistic regression version can take into consideration multiple input criteria. In the case of university acceptance, the logistic characteristic may want to keep in mind elements consisting of the student's grade factor average, SAT rating and wide variety of extracurricular activities. Based on historic information approximately in advance effects regarding the identical enter criteria, it then ranks new instances on their chance of falling into one in all final results categories.

- Logistic regression has turned out to be an critical device withinside the subject of gadget studying. It lets in algorithms utilised in gadget studying programs to categorise incoming facts primarily based totally on ancient facts. As extra applicable facts

come in, the algorithms get higher at predicting classifications inside facts sets.

- Logistic regression also can play a function in facts coaching sports via means of permitting facts units to be positioned into especially predefined buckets throughout the extract, transform, load (ETL) system to be able to grade the data for analysis.

- **What is the purpose of logistic regression?**
  - Logistic regression streamlines the arithmetic for measuring the effect of more than one variable (e.g., age, gender, advert placement) with a given outcome (e.g., click-via or ignore). The ensuing fashions can assist tease aside the relative effectiveness of diverse interventions for special classes of people, consisting of young/vintage or male/female.

  - Logistic fashions also can remodel uncooked information streams to create capabilities for different styles of AI and device mastering techniques. In fact, logistic regression is one of the normally used algorithms in device mastering for binary class troubles, which can be troubles with elegance values, such as predictions such as "this or that," "yes or no," and "A or B."

  - Logistic regression also can estimate the chances of events, inclusive of figuring out a courting among functions and the chances of outcomes. That is, it is able to be used for categories with the aid of using a version that correlates the hours studied with the probability the pupil passes or fails. On the turn side, the equal version might be used for predicting whether or not a specific pupil will by skip or fail whilst the wide variety of hours studied is supplied as a function and the variable for the reaction has values: sure and fail.

- **Advantages and disadvantages of logistic regression**
  - The main advantage of logistic regression is that it is much easier to set up and train than other machine learning and AI applications.

  - Another advantage is that it's far one of the maximum green algorithms whilst the distinct effects or differences represented through the statistics are linearly separable. This method that you may draw instantly line keeping apart the effects of a logistic regression calculation.

  - One of the largest points of interest of logistic regression for statisticians is that it is able to assist screen the interrelationships among unique variables and their effect on outcomes. This ought to quickly decide whilst variables are definitely or negatively correlated, inclusive of the locating mentioned above that extra reading has a tendency to be correlated with better check outcomes. But it's essential to be aware that different strategies like causal AI are required to take the plunge from correlation to causation.

- **Mathematical formulation.**

  - **<u>The Logistic Curve</u>**

    The logistic curve relates the independent variable, X, to the rolling mean of the DV, P ($\overline{Y}$). The formula to do so may be written either

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

    Or

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

where P is the probability of a 1 (the proportion of 1s, the mean of Y), *e* is the base of the natural logarithm (about 2.718) and *a* and *b* are the parameters of the model. The value of *a* yields P when X is zero, and *b* adjusts how quickly the probability changes with changing X a single unit (we can have standardized and unstandardized b weights in logistic regression, just as in ordinary linear regression). Because the relation between X and P is nonlinear, *b* does not have a straightforward interpretation in this model as it does in ordinary linear regression.

- **Loss Function**

A loss function is a measure of fit between a mathematical model of data and the actual data. We choose the parameters of our model to minimize the badness-of-fit or to maximize the goodness-of-fit of the model to the data. With least squares (the only loss function we have used thus far), we minimize SS_{res}, the sum of squares residual. This also happens to maximize SS_{reg}, the sum of squares due to regression. With linear or curvilinear models, there is a mathematical solution to the problem that will minimize the sum of squares, that is,

$$\mathbf{b = (X'X)^{-1}X'y}$$

Or

$$\mathbf{b = R^{-1}r}$$

With some models, like the logistic curve, there is no mathematical solution that will produce least squares estimates of the parameters. For many of these models, the loss function chosen is called *maximum likelihood*. A *likelihood* is a conditional probability (e.g., P(Y|X), the probability of Y given X). We can pick the parameters of the model (*a* and *b* of the logistic curve) at random or by trial-and-error and then compute

the likelihood of the data given those parameters (actually, we do better than trail-and-error, but not perfectly). We will choose as our parameters, those that result in the greatest likelihood computed. The estimates are called maximum likelihood because the parameters are chosen to maximize the likelihood (conditional probability of the data given parameter estimates) of the sample data. The techniques actually employed to find the maximum likelihood estimates fall under the general label *numerical analysis*. There are several methods of numerical analysis, but they all follow a similar series of steps. First, the computer picks some initial estimates of the parameters. Then it will compute the likelihood of the data given these parameter estimates. Then it will improve the parameter estimates slightly and recalculate the likelihood of the data. It will do this forever until we tell it to stop, which we usually do when the parameter estimates do not change much (usually a change .01 or .001 is small enough to tell the computer to stop). [Sometimes we tell the computer to stop after a certain number of tries or iterations, e.g., 20 or 250. This usually indicates a problem in estimation.]

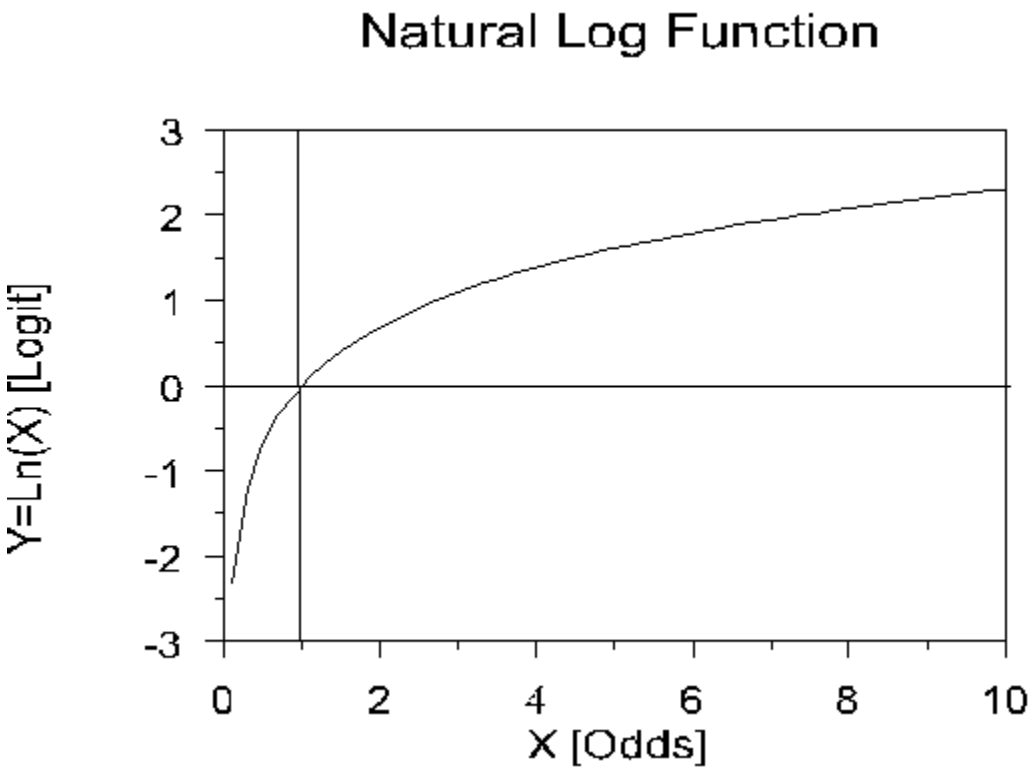
- **Where on Earth Did This Stuff Come From?**

  Suppose we only know a person's height and we want to predict whether that person is male or female. We can talk about the probability of being male or female, or we can talk about the odds of being male or female. Let's say that the probability of being male at a given height is .90. Then the odds of being male would be

$$odds = \frac{P}{1 - P}.$$

  (Odds can also be found by counting the number of people in each group and dividing one number by the other. Clearly, the probability is not the same as the odds.) In our example, the odds would be .90/.10 or 9 to one. Now the odds of being female would be .10/.90 or 1/9 or .11. This asymmetry is unappealing, because the odds of being a male should be the opposite of the

odds of being a female. We can take care of this asymmetry though the natural logarithm, ln. The natural log of 9 is 2.217 (ln(.9/.1)=2.217). The natural log of 1/9 is -2.217 (ln(.1/.9)=-2.217), so the log odds of being male is exactly opposite to the log odds of being female. The natural log function looks like this:



Note that the natural log is zero when X is 1. When X is larger than one, the log curves up slowly. When X is less than one, the natural log is less than zero, and decreases rapidly as X approaches zero. When P = .50, the odds are .50/.50 or 1, and ln(1) =0. If P is greater than .50, ln(P/(1-P) is positive; if P is less than .50, ln(odds) is negative. [A number taken to a negative power is one divided by that number, e.g. $e^{-10} = 1/e^{10.}$ A logarithm is an exponent from a given base, for example $\ln(e^{10}) = 10$.]

Back to logistic regression.

In logistic regression, the dependent variable is a *logit*, which is the natural log of the odds, that is,

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find

logit(P) = a + bX,

Which is assumed to be linear, that is, the log odds (logit) is assumed to be linearly related to X, our IV. So there's an ordinary regression hidden in there. We could in theory do ordinary regression with logits as our DV, but of course, we don't have logits in there, we have 1s and 0s. Then, too, people have a hard time understanding logits. We could talk about odds instead. Of course, people like to talk about probabilities more than odds. To get there (from logits to probabilities), we first have to take the log out of both sides of the equation. Then we have to convert odds to a simple probability:

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

❖ Support Vector Machine – SVM

- Support vector machines (SVMs) are hard and fast supervised mastering strategies used for classification, regression and outlier detection.

- Effective The advantages of support vector machines are:

  - In excessive dimensional spaces.

  - Still powerful in instances wherein quantity of dimensions is more than the quantity of samples.

  - Uses a subset of schooling factors withinside the choice function (known as assist vectors), so it's also reminiscence efficient.

  - Versatile: distinctive Kernel features may be exact for the choice function. Common kernels are provided, however it's also viable to specify custom kernels.

- The disadvantages of support vector machines include:

  - If the range of capabilities is an awful lot more than the range of samples, keeping away from over-becoming in selecting Kernel features and regularisation time period is crucial.

  - SVMs no longer immediately offer possibility estimates, those are calculated using a pricey five-fold cross-validation (see Scores and probabilities, below).

- The guide vector machines in scikit-research guide each dense (numpy.ndarray and convertible to that via way of means of numpy.asarray) and sparse (any scipy.sparse) pattern vectors as input. However, to apply an SVM to make predictions for sparse data, it ought to have been matched on such data. For superior

performance, use C-ordered numpy.ndarray (dense) or scipy.sparse.csr_matrix (sparse) with dtype=float64.

- **Mathematical formulation.**

  - Linear SVC

    The primal problem can be equivalently formulated as:

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \max(0, 1 - y_i(w^T \phi(x_i) + b)),$$

    wherein we employ the hinge loss. This is the shape this is immediately optimized with the aid of using LinearSVC, however not like the twin shape, this one does now no longer contain internal merchandise among samples, so the well-known kernel trick can't be applied. This is why most effective the linear kernel is supported with the aid of using LinearSVC (ø is the identification function).

  - NuSVC

    The ν-SVC formulation is a reparameterization of the C-SVC and therefore mathematically equivalent.
    We introduce a new parameter v (instead of C) which controls the number of support vectors and margin errors: v € (0,1] is an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors. A margin error corresponds to a sample that lies on the wrong side of its margin boundary: it is either misclassified, or it is correctly classified but does not lie beyond the margin.

  - SVE

Given training vectors $x_i \in R^p$, i=1,…, n, and a vector $y \in R^n$ ε-SVR solves the following primal problem:

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*)$$
$$\text{subject to } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i,$$
$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*,$$
$$\zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n$$

Here, we are penalizing samples whose prediction is at least ε away from their true target. These samples penalize the objective by ζi or ζi∗, depending on whether their predictions lie above or below the ε tube.

The dual problem is

$$\min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon e^T (\alpha + \alpha^*) - y^T (\alpha - \alpha^*)$$
$$\text{subject to } e^T (\alpha - \alpha^*) = 0$$
$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, n$$

where e is the vector of all ones, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The prediction is:

$$\sum_{i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

These parameters can be accessed through the attributes dual_coef_ which holds the difference $\alpha_i - \alpha_i^*$, support_vectors_ which holds the support vectors, and intercept_ which holds the independent term ***b***

- LinearSVR
  The primal problem can be equivalently formulated as

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1} \max(0, |y_i - (w^T \phi(x_i) + b)| - \varepsilon),$$

  where we make use of the epsilon-insensitive loss, i.e. errors of less than ε are ignored. This is the form that is directly optimized by **LinearSVR**.

III.3 Feature extraction technique.

- **Bag of words.**

  Bag of words (Zhang, et al., 2010) is one of the most basic methods to transform tokens into features, where each word is used as a feature to train the classifier.

  Bag of words is a Natural Language Processing method of textual content modelling. In technical terms, we will say that it's miles a technique of function extraction with textual content data. This method is an easy and bendy manner of extracting capabilities from documents.

  A bag of words is an illustration of textual content that describes the prevalence of phrases inside a report. We simply hold music of phrase counts and push aside the grammatical information and the phrase order. It is known as a "bag" of phrases due to the fact any data approximately the order or shape of phrases with inside the report is discarded. The version is handiest involved with whether or not recognised phrases arise within side the report, now no longer wherein within side the report.
  So, why bag-of-words, what is incorrect with the easy and clean textual content?

  One of the largest troubles with textual content is that it's far messy and unstructured, and device getting to know algorithms select structured, properly described fixed-period inputs and with the aid of using the usage of the Bag-of-Words approach we will convert variable-length texts right into a fixed-length vector.

  Also, at a miles granular level, the device getting to know fashions paintings with numerical information in preference to textual information. So to be extra specific, with the aid of using the usage of the bag-of-words (BoW) approach, we convert a textual content into its equal vector of numbers.

- **TF-IDF (Term Frequency – Inverse Document Frequency).**

  TF-IDF (Term Frequency - Inverse Document Frequency) (Sammut & Webb, 2010) is a method of calculating the weight of a word in a document obtained through statistical methods showing the importance of a word.in a document under consideration compared to a set of documents.

TF-IDF (term frequency-inverse document frequency) is a statistical degree that evaluates how applicable a phrase is to a file in a group of files.

This is completed through multiplying metrics: how commonly a phrase seems in a documents, and the inverse file frequency of the phrase throughout a hard and fast of documents.
It has many uses, most significantly in automatic textual content analysis, and may be very beneficial for scoring phrases in system studying algorithms for Natural Language Processing (NLP).

TF-IDF became invented for file seek and records retrieval. It works through growing proportionally to the quantity of instances a phrase seems in a file, however is offset through the quantity of files that comprise the phrase. So, phrases which can be not unusual place in each file, which includes this, what, and if, rank low despite the fact that they'll seem commonly, considering that they don't suggest a lot to that file in particular.

However, if the word Bug seems commonly in a file, at the same time as now no longer acting commonly in others, it possibly approaches that it's very applicable. For example, if what we're doing is looking for out which subjects a few NPS responses belong to, the phrase Bug might possibly emerge as being tied to the subject Reliability, considering that maximum responses containing that phrase might be approximately that topic.

- **Part-of-speech tagging (POS tagging).**

Part-of-speech tagging (L. Márquez, 1999) is a popular natural language processing to the classification of words in a text that correspond to a particular part of speech, depending on the definition of the word and the context of the text. A simplified form of this can be understood customer scores through recognizing words such as nouns, verbs, adjectives, ...

Tag	Description	Example
A	Tính từ (Adjective)	tốt, nhiều, nhanh, hơn, rẻ, cao, ...
C	Liên từ (Coordinating conjunction)	thì, và, nhưng, hay, như, ...
E	Giới từ (Preposition)	vì, của, đến, từ, ...
I	Từ cảm thán (Interjection)	ôi, thay, biết bao, ...
L	Từ hạn định (Determiner)	các, những, mọi, mấy, vài, ...
M	Số từ (Numeral)	một, hai, trăm, nghìn, đôi, triệu, ...
N	Danh từ thường (Common noun)	hàng, mình, sản phẩm, sao, shop, mặt, ...
Nc	Danh từ chỉ loại (loại từ) (classifier noun)	cái, con, quả, củ, tấm, bức, sợi, ...
Ny	Danh từ viết tắt (Noun abbreviation)	t (tao), m (mày), a (anh), e (em), …
Np	Danh từ riêng (Proper noun)	Phú Quốc, Hà Nội, Lê Thánh Tông, ...
Nu	Danh từ chỉ đơn vị đo lường (Unit noun)	tấn, tạ, yến, kg, lít, ...
P	Đại từ (Pronoun)	tôi, chúng ta, nó, ai, mày, ...
R	Phụ từ (Adverb)	không, rất, quá, lại, cũng, còn, ...
S	Subordinating conjunction (Liên từ phụ thuộc)	trong khi, mỗi khi, trước khi, sau khi, ...
T	Trợ từ (Auxiliary)	cả, những, cái, thì, mà, là, …
V	Động từ (Verb)	muốn, đi, chơi, ăn, uống, …
X	Tổ hợp từ không thể xác định (Undetermined group)	
F	Dấu câu (Filtered out - punctuation)	.,:!?

Danh sách mô tả các ký hiệu của phương pháp POS tagging

❖ Model Evaluation Indicators: The Confusion Matrix.

The confusion matrix helps to see how the data points are classified as true/false in the classification method.

	Dự báo: Negative	**Dự báo: Positive**
Thực tế: Negative	True Negative (TN)	False Positive (FP)
Thực tế: Positive	False Negative (FN)	True Positive (TP)

+ True Negative (TN): the number of points of the negative class that are properly classified as negative.

+ False Negative (FN): the number of points in the positive class that were mistakenly classified as negative.

+ True Positive (TP): the number of points of the positive class that are properly classified as positive.

+ False Positive (FP): the number of points of the negative class that were mistakenly classified as positive.

❖ Model evaluation metrics: Confusion Matrix

- Accuracy: là tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

- Recall: là tỉ lệ số điểm true positive (TP) trong số những điểm thực sự là positive (TP+FN).

$$Recall = \frac{TP}{TP + FN}$$

- Precision: là tỉ lệ số điểm true positive (TP) trong số những điểm được phân loại là positive (TP+FP)

$$Precision = \frac{TP}{TP + FP}$$

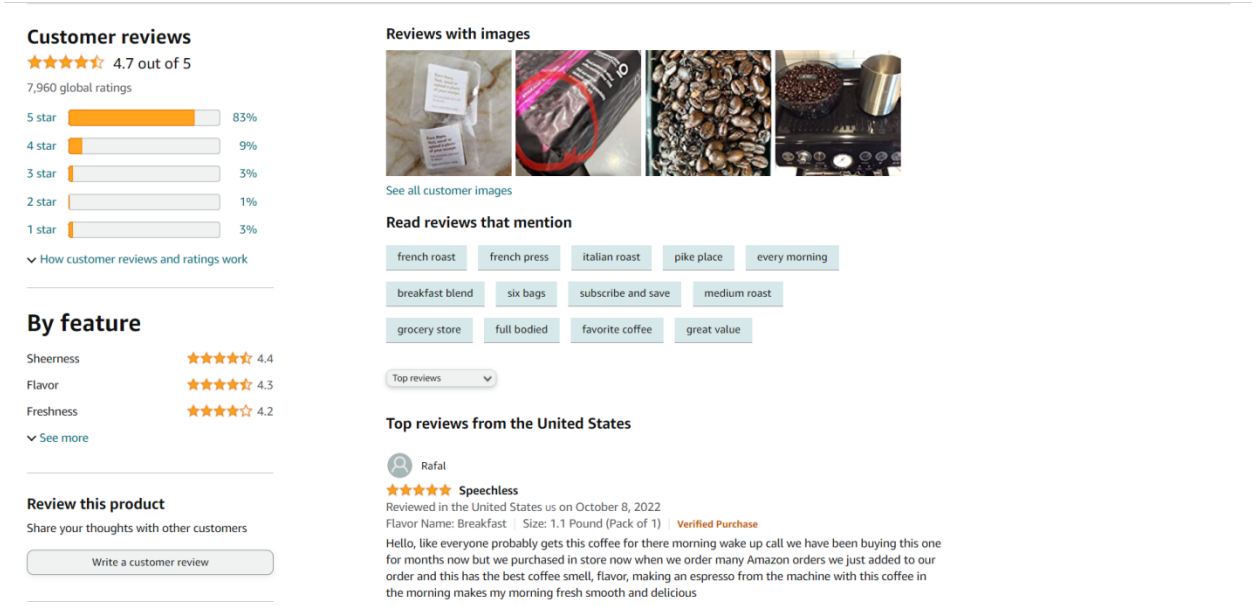
- F-score: là trung bình điều hòa (harmonic mean) của Precision và Recall

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

VI. RESEARCH METHOD AND RESULTS

Collecting data

- Data in our research is public in Amazon.com
- Data collected include 11.530 comment, feedback of customer from 2019-now.

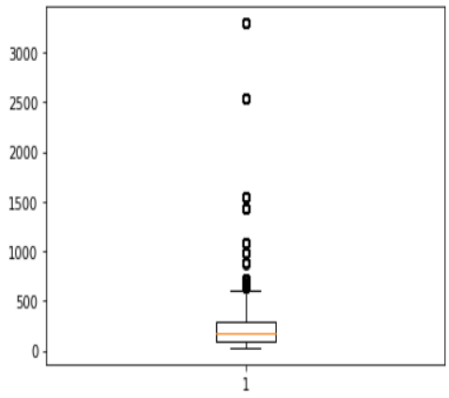


index	review_text	review_text_processed	review_text_lenmatize	review_text_word_token
0	I'd say this is worth the money if you like Pumpkin Spice, considering what Starbucks charges for cups of coffee at their restaurants... Making regular cups of coffee with these is good, but for the full Pumpkin Spice Latte experience I'd get one of those fancy Keurig machines.If you're simple like me, a decent amount of cream and sugar makes a good standard pumpkin-flavored cup right at home. Read more	i d say this is worth the money if you like pumpkin spice considering what starbucks charges for cups of coffee at their restaurants making regular cups of coffee with these is good but for the full pumpkin spice latte experience i d get one of those fancy keurig machines if you re simple like me a decent amount of cream and sugar makes a good standard pumpkin flavored cup right at home read more	i d say this is worth the money if you like pumpkin spice considering what starbucks charge for cup of coffee at their restaurant making regular cup of coffee with these is good but for the full pumpkin spice latte experience i d get one of those fancy keurig machine if you re simple like me a decent amount of cream and sugar make a good standard pumpkin flavored cup right at home read more	say worth money like pumpkin spice considering starbucks charge cup coffee restaurant making regular cup coffee good full pumpkin spice latte experience get one fancy keurig machine simple like decent amount cream sugar make good standard pumpkin flavored cup right home read
1	A coffee pod will never compare favorably to freshly ground beans but Starbucks Italian Roast K-cups are the only brand and roast I have found, after trying several varieties, that provide a decently strong cup of single-serve coffee. Read more	a coffee pod will never compare favorably to freshly ground beans but starbucks italian roast k cups are the only brand and roast i have found after trying several varieties that provide a decently strong cup of single serve coffee read more	a coffee pod will never compare favorably to freshly ground bean but starbucks italian roast k cup are the only brand and roast i have found after trying several variety that provide a decently strong cup of single serve coffee read more	coffee pod never compare favorably freshly ground bean starbucks italian roast k cup brand roast found trying several variety provide decently strong cup single serve coffee read
2	In under 24 hours I was sent Cinnamon Dulce Latte pods in a Seattle's Best coffee box. This is a Starbucks product, it's not a Latte, it's normal, coffee pods. I love Toasted Graham by Starbucks. I've ordered it many, many times, usually from Keurig directly. They were OOS, so I used Amazon who graciously sent ir within 4 hours the first time. It was stolen, having been left in the wrong place. I called, got a promis of a refund, advised to Reorder.I reordered, it was delivered overnight, thank you, but it's not even close to what I ordered and although can't Return it, I was issued a full refund. I'm left holding the ball about disposal.And need to waste more time, today, trying to find a Seller who knows the difference between a Latte and nit Latte, who can read "Toasted Graham" vs.	in under hours i was sent cinnamon dulce latte pods in a seattle s best coffee box this is a starbucks product it s not a latte it s normal coffee pods i love toasted graham by starbucks i ve ordered it many many times usually from keurig directly they were oos so i used amazon who graciously sent ir within hours the first time it was stolen having been left in the wrong place i called got a promis of a refund advised to reorder i reordered it was delivered overnight thank you but it s not even close to what i ordered and although can t return it i was issued a full refund i m left holding the ball about disposal and need to waste more time today trying to find a seller who knows the difference between a latte and nit latte who can read toasted graham vs cinnamon dulce who has clean starbucks box not an obviously recycled seattle s best box and i ll have to get a	in under hour i wa sent cinnamon dulce latte pod in a seattle s best coffee box this is a starbucks product it s not a latte it s normal coffee pod i love toasted graham by starbucks i ve ordered it many many time usually from keurig directly they were oos so i used amazon who graciously sent ir within hour the first time it wa stolen having been left in the wrong place i called got a promis of a refund advised to reorder i reordered it wa delivered overnight thank you but it s not even close to what i ordered and although can t return it i wa issued a full refund i m left holding the ball about disposal and need to waste more time today trying to find a seller who know the difference between a latte and nit latte who can read toasted graham v cinnamon dulce who ha clean starbucks box not an obviously recycled seattle s best box and i ll have to get a	hour wa sent cinnamon dulce latte pod seattle best coffee box starbucks product latte normal coffee pod love toasted graham starbucks ordered many many time usually keurig directly oos used amazon graciously sent ir within hour first time wa stolen left wrong place called got promis refund advised reorder reordered wa delivered overnight thank even close ordered although return wa issued full refund left holding ball disposal need waste time today trying find seller know difference latte nit latte read toasted graham v cinnamon dulce ha clean starbucks box obviously recycled seattle best box get neighbor shop locally grocery super store read

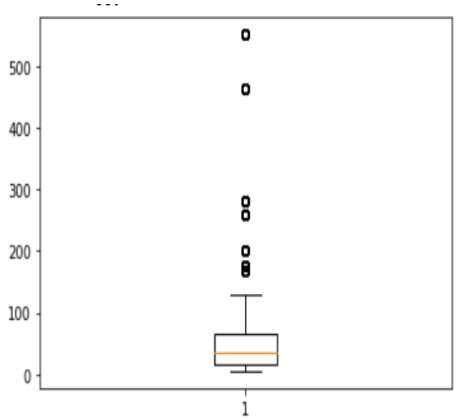
Comment data on Amazon.com

EDA (Exploratory Data Analysis)

- EDA is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods (Wikipedia)
- Num_char, num_word, avg_word_lenght
- Draw boxplot from data :



Num_char



num_word

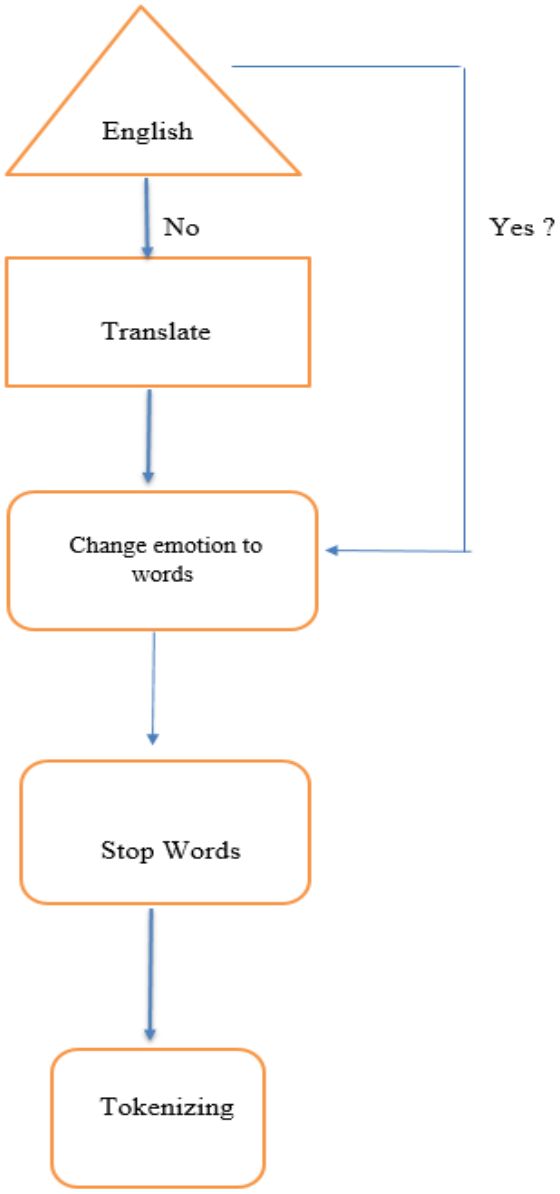
index	review_text
0	I'd say this is worth the money if you like Pumpkin Spice, considering what Starbucks charges for cups of coffee at their restaurants... Making regular cups of coffee with these is good, but for the full Pumpkin Spice Latte experience I'd get one of those fancy Keurig machines.If you're simple like me, a decent amount of cream and sugar makes a good standard pumpkin-flavored cup right at home. Read more
1	A coffee pod will never compare favorably to freshly ground beans but Starbucks Italian Roast K-cups are the only brand and roast I have found, after trying several varieties, that provide a decently strong cup of single-serve coffee. Read more
2	In under 24 hours I was sent Cinnamon Dulce Latte pods in a Seattle's Best coffee box. This is a Starbucks product, it's not a Latte, it's normal, coffee pods. I love Toasted Graham by Starbucks. I've ordered it many, many times, usually from Keurig directly. They were OOS, so I used Amazon who graciously sent ir within 4 hours the first time. I was stolen, having been left in the wrong place. I called, got a promis of a refund, advised to Reorder.I reordered, it was delivered overnight, thank you, but it's not even close to what I ordered and although can't Return it, I was issued a full refund. I'm left holding the ball about disposal.And need to waste more time, today, trying to find a Seller who knows the difference between a Latte and nit Latte, who can read "Toasted Graham" vs. Cinnamon Dulce, who has clean, Starbucks Box, not an obviously recycled Seattle's Best box.And I'll have to get a neighbor to shop for it locally in a grocery super store. Read more

Num_char > 86

11528	The package arrived flat and crushed - but that was not important - the important part is that the price is way too high and the taste horrible. I experienced with different amounts of coffee in my coffee maker and it was all bad - terrible aftertaste. I had to discard it. Read more
11529	It will expire in 2 months. Are you serious? I was expected to drink it for at least half of the year. And the package looks old... Read more

Num_count > 21

Data Pre_Processing



Step by step to processing data

MACHINE LEARNING TO ANALYSIS

- Attributes selected to be included in the model include feature variable "content" and label “score”

- Based on the research of (Liu, 2017), the study will apply the evaluation score (score) according to two negative and positive groups with 8.269 comments as negative (with scores from 1-3 labeled 0) and 3.261 comments are positive (with a score of 4-5 labeled 1)

- Convert "content" variable to TF-IDF vector using Sklearn . library

(sklearn.feature_extraction.text), the TfidfVectorizer function as a variable named vectorizer and then call fit_transform() function

	abernook	able	absolute	absolutely	abundance	abused	access	accustomed	acidic	actually	...	world	worn	worried	worse	worth	would	wrong	year	yes	yet
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.14379	0.000000	0.000000	0.000000	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.000000	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.082806	0.000000	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.000000	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11526	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.000000	0.0	0.0
11527	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.000000	0.0	0.0
11528	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.000000	0.0	0.0
11529	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.000000	0.213791	0.0	0.0
11530	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.304719	0.000000	0.000000	0.0	0.0

11531 rows × 1022 columns

TF-IDF Table

Training model :

- The training process is conducted according to the Hold-Out method, meaning the entire data set
- will be divided into 2 subsets of training data and testing data
- intersect.
- - Use the train_test_split function in sklearn to scale 80% of the corresponding training set
- with 633,562 comments, and 20% of the test set corresponds to 158,391 comments.
- - Four classification algorithms belonging to the group of supervised machine learning are applied in this study are Logistic Regression, Support Vector Machine

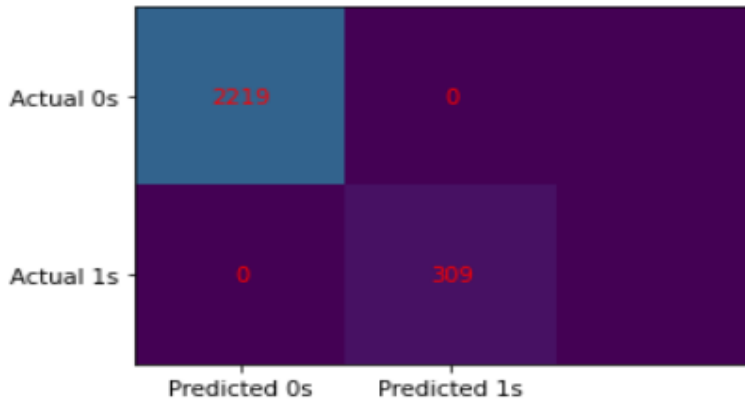
V. EVALUATE RESULTS AND DISCUSSION

5.1 Evaluate results

Logistic Regression

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2219
2	1.00	1.00	1.00	309
3	1.00	1.00	1.00	734
4	1.00	1.00	1.00	1256
5	1.00	1.00	1.00	7013
accuracy			1.00	11531
macro avg	1.00	1.00	1.00	11531
weighted avg	1.00	1.00	1.00	11531

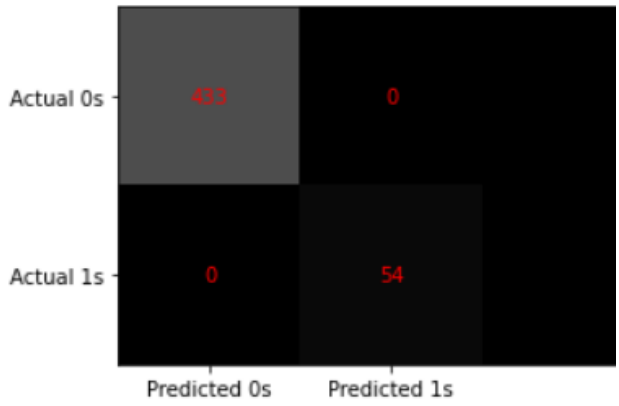
Confusion matrix (LR)



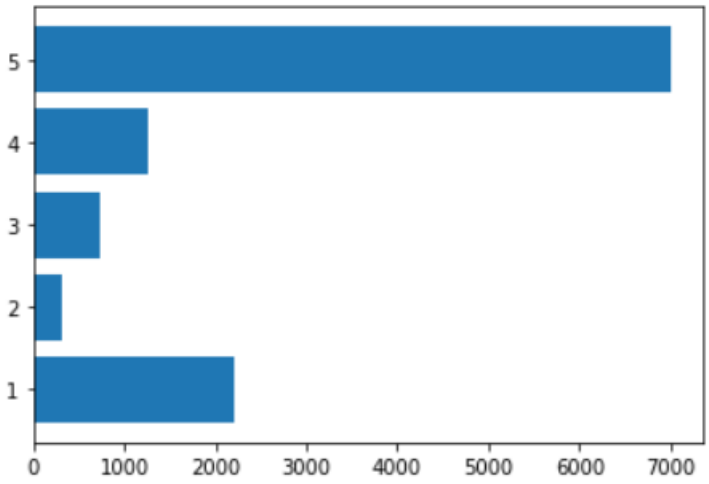
SVM – Support Vector Machine

	precision	recall	f1-score	support
1	1.00	1.00	1.00	433
2	1.00	1.00	1.00	54
3	1.00	1.00	1.00	143
4	1.00	1.00	1.00	249
5	1.00	1.00	1.00	1428
accuracy			1.00	2307
macro avg	1.00	1.00	1.00	2307
weighted avg	1.00	1.00	1.00	2307

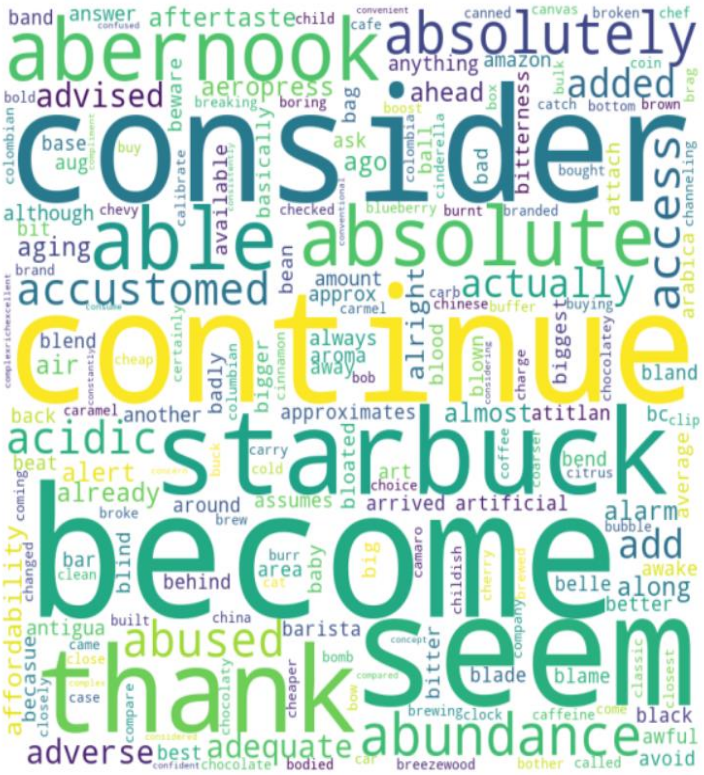
Confusion matrix (SVM)



Results :



Line chart about rating of customer in Amazon



Word cloud comment in Amazon

v

Discussion


Delivery is good
Exam : convenient, fast,.
Contact to delivery company to keep going
Quality of products
Exam : Weak, worse, badlys
Improve products or introduce product which adapt with each of customer
Customer will come back
Exam : continue, good
Discount to customer who will buy again

CONCLUSION:

Summary of the research process and contributors

(1) Building a dictionaries specifically for the field of Ecommerce;

(2) Propose a hybrid model between clearing and classification of opinions;

(3) The results of testing the method of learning on the file computer with the accuracy in turn Logistic regression: 100%, SVM: 100%,. In which, logistic regression the highest standard is selected to test for the model proposal on the analysis score of the word and phrase.

With the results achieved, it is possible to apply to the built-in dashboard to track customer experience directly

Over time, to assist managers in making time decisions, to improve products and services, enhance customer satisfaction.

RESTRICT :

- Data collection sill small and haven't collect from another Ecommerce
- Research focus on model with possitive comment and negative comment.

IMPROVE IN FUTURE:

In the future, research will expand to another Ecommerce, and neutral comment will be collected. And we will improve to collect image, video of customer's comment.

REFERENCES

(1) https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

(2) http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

(3) https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB

(4) https://scikit-learn.org/stable/modules/svm.html

LINK SOURCE CODE :
https://colab.research.google.com/drive/1yHQX2pTLP5Cp_pKJBOh5UJgFnpKQ6XQ7?usp=sharing