BUSINESS INTELLIGENCE APPROACH TO MARKET RESEARCH ON FOOD COMMODITY BY USING BIG DATA ANALYSIS, CASE STUDY: FORUM JUAL BELI KASKUS

Andi Wibowo[1), 2)], Widhyawan Prawiraatmadja[2)] , Manahan Siallagan[2)], Jeffri Lingo[3)]

[1)]Badan Pengawas Obat dan Makanan Republik Indonesia, Jakarta, Indonesia
[2)]School of Business and Management, Institut Teknologi Bandung, Jakarta, Indonesia
[3)]PT Mustika Manis Utama, Tangerang, Indonesia

Corresponding author: andi.wibowo@pom.go.id

Abstract

The emerging Covid-19 pandemic and the increasing use of internet access trend in Indonesia have successfully changed most consumer behaviour, shifting into the online market (e-Commerce) instead of the conventional physical market. It also presents vast new business opportunities, particularly in the food and beverage industries. However, the authority must transform the decision support system based on data-driven consideration to ensure consumer protection. This research aimed to present the Big Data Analysis power to find insight from the current condition of e-commerce by conducting market research on Forum Jual Beli Kaskus, which was chosen due to less restriction while implementing the self-programming web scrape engine. The collected data is then analyzed through Rapidminer software for text preprocessing and predictive models shown in the Business intelligence tool. Authors find that the web scraping engine successfully collected the whole population of products listed in the food and beverages category as well as text preprocessing resulted in several keywords which represent the product trend. The prediction model achieved 99.85% accuracy and a minimum 80% precision class while the test dataset was introduced to confirm and test the model. By utilizing The Big Data Analysis and Business Intelligence Tools, the government authority could catch enormous insight based on the real-time market research process to formulate a better policy approach.

Keywords: Business Intelligence, Big Data Analysis, e-Commerce, Market Research, Food Control

Introduction

The rapid changing of consumer behaviour due to the COVID-19 pandemic has benefited e-commerce across Southeast Asia, particularly in Indonesia. The demand for physical distancing during the pandemic prepared the way for a more home-centred lifestyle. Furthermore, the situation has led many micro, small, and medium-sized businesses (MSMEs) to migrate to online platforms. Referring to the Indonesian Minister of Cooperatives and SMEs, during the Covid-19 Pandemic, the number of SMEs utilizing internet platforms increased to 10.2 million. According to PricewaterhouseCoopers (PwC) Global Consumer Insights Survey (2020), 69% of Indonesian respondents said they were purchasing food online due to the Covid-19 Policy imposing social restrictions measures to control the COVID-19 pandemic. Over half of respondents (57%) said they were likely to continue buying online even after the government lifted the mobility restrictions. When questioned about their buying frequency for daily and weekly purchases, half of the Indonesian respondents said they often utilized online mobile channels, up from 37% in 2019.

On the other side, Customers Survey showed that they shopped in-store dropped to 42% from 52% in 2019, suggesting that in-store shopping is becoming less favourable for daily to weekly purchases. Many retail store closures have been consistently occurring across Indonesia over the last five years as a collective impact. Numerous shops were reported as going out of business before the pandemic for several reasons. According to the Indonesian Retail Entrepreneurs Association (APRINDO), around 400 minimarkets went bankrupt or closed their doors during the Covid-19 Pandemic. Between March and December 2020, an average of five to six supermarkets will be forced to shut down.

In comparison, 1-2 shops had the same issue between January and March 2021. Thus, the growth of the e-commerce marketplace has significantly affected traditional business. E-commerce transactions have surpassed in-store sales as the most preferred channel for regular purchases. The Indonesia Statistic Bureau (BPS, 2020) explains that the emerging e-commerce growth revenue, particularly in the Food & Personal Care segment, is expected to reach $5,941 million in 2021. The market's largest segment is Personal Care, with a projected market volume of US$3,470 million in 2021. Otherwise, the food and beverage sector

brought the highest growth since measured from 2018 to 2020. According to DataReportal (2021), the Food Commodity had been the least affected by the Covid-19 crisis.

This condition shown above should be alarming the authorities. According to Kotsanopoulos et al. (2017), the law enforcement mechanism on food commodities to ensure food safety and quality prior to receiving by the customer relies heavily on inspection and monitoring programs. For a long time, the emphasis of the food-control inspection was on end-product sampling and testing to establish whether or not they met regulatory requirements. However, the global agreement exists on the change from a "reactive" to a "preventive" strategy to mitigate the food safety risks. Rather than just verifying compliance with prescriptive laws, this new strategy involves inspectors acquiring new skills, such as the ability to analyze processes and evaluate the relevance and efficiency of several routes to a food standards conclusion. In contrast to traditional inspections, preventive programs attempt to collect information on the overall state of food commodities in a particular area, providing a comprehensive picture of a given scenario.

Along with the authority's effort to shift strategy for ensuring food safety risk from "reactive" to a "preventive," many food distribution chains also shift their strategy from "Offline" to Online "market." This rapid change is also determined by consumer behaviour, as discussed. The rapid digitalization of social life and commerce has resulted in digital footprints of consumer and business conduct. This massive data in terms of volume, diversified in structure, and often widely available on the internet represents a potential goldmine for marketing academics seeking to acquire insights on seller behaviour that would be impossible or incredibly expensive to observe in other ways. Many e-commerce shows the necessary data on their website, which is incredibly valuable to market research insight. Big Data Analysis offers the latest, fast and reliable data science as the modern way of market research. Looking forward to this specific information could help the organization focus on particular problems and allocate their limited resources for a better decision.

Research Purpose and Question

The present study is dedicated to identifying the most up-to-date and attractive food product topics in e-commerce by finding the keywords for further analysis. This research utilized data collection techniques for web scrapping identifying the emerging trends of food products, then building a model to establish a predictive measure for product classification in terms of focused consideration by employing The Big Data Analysis method and then interpreting the result from the Business Intelligence output.

This research is trying to seek the answer to the following questions:

Q1: What insight from the descriptive result after conducting Big Data Analytics and Business Intelligence in the food and beverages category on Forum Jual Beli Kaskus?

Q2: How could the result of interpretation benefit the Indonesian Food Control Authority?

Literature Review.

According to Ho (2020), The four basic processes of traditional market research are as follows: establishing goals and challenges; manually collecting data; manually analyzing data, and visually visualizing and interacting with data. Due to the enormous quantity of data on the internet, the data collecting challenge is the most repetitive and time-consuming process. Numerous participants in the Ho (2020) research acknowledged the difficulties of finding and collecting meaningful data. Additionally, inconsistent manual data gathering is another factor contributing to the inefficiency of traditional market research. The Big Data Analysis offered a solution for eliminating those constrain on market research. A study by Chen et al. (2012) shows that since 2008, the US government has promoted Big Data Analytics to generate Business Intelligence. This finding was confirmed by Parks et al. (2018), who examines masters in business analytics programs at primary US business schools. There is strong demand for data scientists, but in Indonesia, the high demand is confronted with a skill scarcity (JakartaPost, 2018). Big data analysis is reshaping the commercial sector. Marketing is fundamentally about people, product, place, price, and promotion. Then, data mining has a significant influence on these sectors, and big data analysis is a critical tool for making better judgments (Fan, Lau, & Zhao, 2015).

Table 1. Comparison of Common Data Collection Methods in Online Market Research.

Key Consideration	Scraped Data	Commercial Web Scraping Service	API	Survey
Cost	Low	Medium	Low/Medium	High
Sample frame	Website users	Website users	Website users	Flexible
Customizability of variables	Medium	Low	Low	High
Ease of frequent collection	Easy	Moderate	Easy	Hard
Data type	Behavioural	Behavioural	Behavioural	Attitudinal

Limitations	Time and programming skills	Data may not be suitable to the	Limited availability	Time

Source: (Han & Anderson, 2021), Note. (API) Application Programming Interface

Speed is one of the primary benefits of employing a web scraper for data collection to build The Big Data sets for analysis. It eliminates repeated activities and increases the process's efficiency. Users merely need to pick the necessary information and submit the order to the service after deploying the web scraper (Milev, 2017). Moreover, using a web scraper for collecting market data improves the accuracy of the data acquired. As seen in Table 1, food market researchers may investigate numerous distinct ways to automate the collecting of online data. Han and Anderson (2021) explain that one intriguing possibility is to use an Application Programming Interface (API) given by the data hosting company. However, even when researchers could access an API, it is sometimes difficult to use, only available for a short period, or requires upfront costs. Additionally, APIs often do not disclose all essential variables, and then Web scraping is the most effective and scalable method of collecting such data.

Web scraping is critical for assembling Big Data sets, which provide the basis for Big Data analytics, Machine Learning (ML), and further for Artificial Intelligence (AI) algorithms. Demunter (2017) outlined the many sources of Big Data, including communication systems, the World Wide Web, data created by business processes, sensors, and crowdsourcing, as elaborate in Figure 1.

According to Debortoli et al. (2014), Business intelligence for market research traditionally leans on historical data described in a conventional dashboard, where later, big data analysis successfully transforms into more analytical, insight-focused, and ability for prediction. Alnoukari (2020) explains that business intelligence (BI) enables firms to make better decisions through vital information, data, and knowledge. Therefore, Big Data Analysis (BDA) may be considered a subset of BI (Sun, Zou, & Strang, 2015). Both BI and BDA share specific similar decision-support tools. Also, BI and BDA share an emphasis on the importance of sufficient information, data, and knowledge.

Furthermore, business intelligence is today built on four cutting-edge technological pillars: cloud, mobile, big data, and social technologies. They are all successfully supported by BDA as a service and technology (Passlick, Lebek, & Breitner, 2017). Sun et al. (2015) argue further that BDA is a necessary tool for building BI, at the very least from a technology and data perspective. BDA is a data-driven and business-oriented method that enables firms to make better decisions and increase business intelligence from a technology standpoint. Marasanapalle et al. (2010) study employed typical text-mining algorithms to determine the television program's primary themes and mention Twitter's trending topic as a form of Business Intelligence implementation since Twitter is a Big Data source as the dynamic web was defined by Demunter (2017).

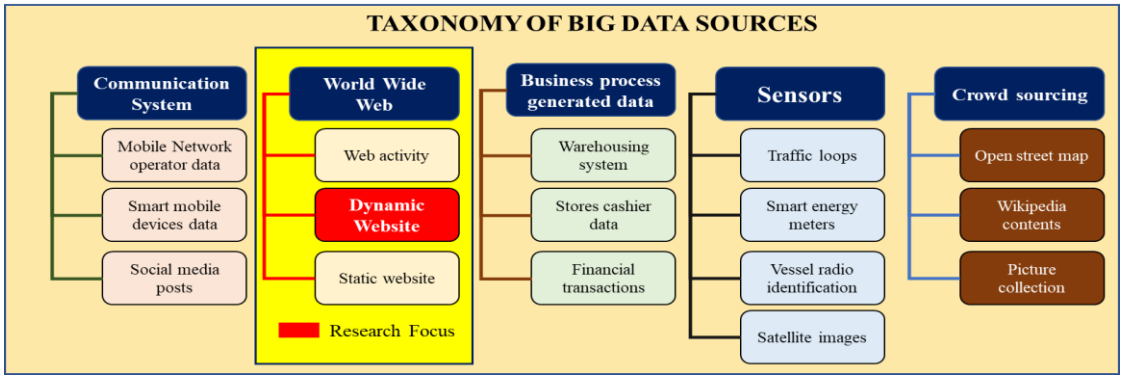


Figure 1. Taxonomy of Big Data (Source: Demunter, 2017)

Legal Issue

Text mining using a web scraping engine to implement Big Data Analytics for market research as part of business intelligence is still debatable for some experts. Saurkar (2018) explain that Web scraping seems to be a manual copy and paste operation. The distinction here is that this task is carried out by a virtual computer agent in an ordered and automated manner. According to Fisher et al. (2010), The Terms of Service (ToS) impose stringent restrictions on data collection. For example, the Facebook ToS mentions that "*You will not collect users' information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our permission.*" Then Hirschey (2014) mentions that to compensate for the absence of explicit doctrinal safeguards in national legislation, data hosts have developed their protections via technical and legal limits on access to their information. Furthermore, Ho (2020) argues that to ensure that web scraping is conducted legally, to examine any limitations to reduce the likelihood of getting banned, there should be a need to determine the structure and robots.txt file of the specified website. Websites extensively use the robots.txt file to connect with web scrapers and specify which locations the web scraper may access, according to Narizhnykh (2018).

Forum Jual Beli Kaskus (Kaskus e-commerce Forum)

Ramaditya (2019) proposes research exploring the brand equity for *kaskus.co.id* after changing the website interface in 2012. Kaskus, being Indonesia's biggest digital community, is always striving to enhance the user experience. Eventhough kaskus is not included in Indonesia's top 5 e-commerce platforms, according to Djufri (2020), this platform has a similar shopping gallery index as other bigger platforms. In addition, the Robots.txt in the Kaskus website allows less restriction than other e-commerce platforms like Tokopedia (Figure 2).

Figure 2. Robots.txt as web scraping restriction list Between Tokopedia (left) and Kaskus (right)

The analysis of the robots.txt list found that during the research process, in fjb.kaskus.co.id website, there is only 28 restriction for web scrapper, compared to more than 40 restriction list in Tokopedia. Since this research was trying to retrieve some information such as username, product title, price, seen, seller location, product description, and URL where it does not appear in the restriction list, then it would be able to achieve the targeted information in kaskus website instead of being an abuse of ToS. Kaskus also employed an active forum of e-commerce, namely Forum Jual Beli, where food production was an essential forum with frequent updates daily. Since this study was aimed to conduct market research legally, then Forum Jual Beli kaskus was chosen for the text mining using web scraping to implement Big Data Analytics for market research as part of Business Intelligence.

Research Material and Methodology

This research was conducted by using the latest Workstation PC with 16 GB RAM. The data was collected by a self-programming web scraped engine at Forum Jual Beli Kaskus (Kaskus E-Commerce Platform), particularly in the food and beverages category. The web scrape engine was written on Java Programming Language and run in ECLIPSE IDE application (an integrated development environment used in computer programming), which was previously set up with Apache Maven Project Management Tools and Selenium Web-driver for running the web scraping engine. The web scraping engine collected whole population data included; product URL, product title, product description, Product price tag, product click times (seen), seller name, seller rank in the forum, and seller location, within 2 hours 40 minutes. The data collection process was running under 20 Mbps in fiber optic internet connection, and the number of products that have been stored in the SQL database was 5,191 products.

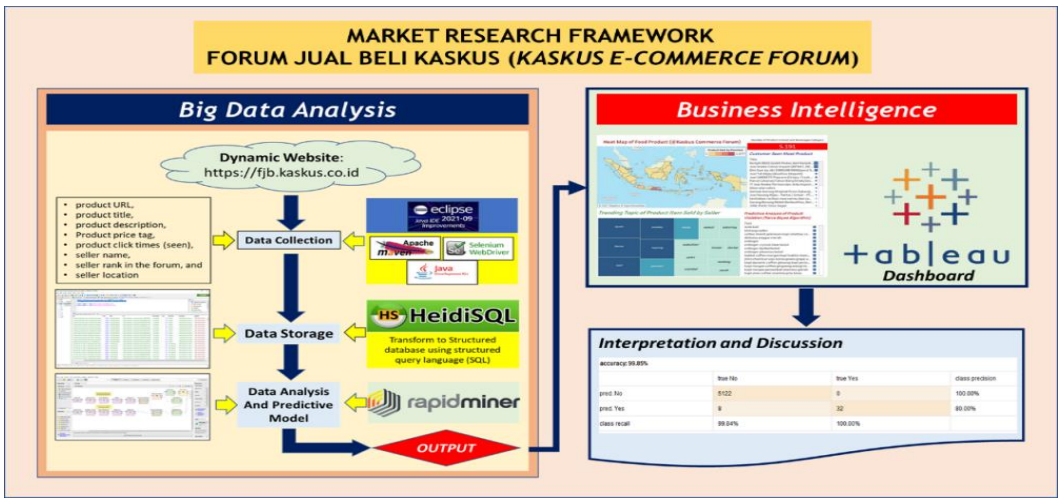


Figure 3. The Research Framework (source: authors, 2021)

After the data collection process was complete, the structured data was converted into .csv format for further analysis in Rapidminer software, used as a data science tool for text processing, analysis and predictive purposes. The license used on this application is for education purposes. The information insight from

Rapidminer, including the model for prediction result, then transform into a dashboard as a business intelligence tool. For this purpose of data visualization, this research utilized Tableau for Desktop with a one-year student license. The result in the dashboard was interpreted according to the current regulation of food and beverages. The complete framework for this research can be illustrated in figure 3 above.

Result and Discussion

Data Collection Method

The first step to begin market research is to assemble the dataset, which uses a web Scrapping Engine. Before developing the programming code for the scraping engine, the web structure analysis must find any location of targeted data that needs to be retrieved for analysis. According to Forum Jual Beli Kaskus website (fjb.kaskus.co.id), the targeted data was placed into a specific location refer to HTML structure as could be seen in table 2 below:

Table 2. Result of Web Element inspection in Forum Jual Beli Kaskus Website

No.	Data need to be found	Class / Subclass in HTML	Column name in SQL database
1.	Product title	link_thread_title	title
2.	Product URL	link_thread_title / "href"	URL
3.	Product Price	price—discounted	price
4.	Product description	entry-body	description
5.	The customer sees the product	data--views	seen
6.	Seller username	username	seller
7	Seller Location	location	seller details
8.	Seller Rank in Kaskus	rank	sellerType

Note: reviewed on October 24, 2021

That kind of web element structure should periodically be checked into correspondence website due to changing web interface made by web owner after establishing the web element location in Table 2. Then web Scraping programming code could be developed, as shown in Figure 4 below.

```
49. List<WebElement> divs = driver.findElements(By.className("item--grid"));                → Find Information at first page 51.
51. String seller = webElement.findElement(By.className("username")).getText();             → get the seller username
52. String url = webElement.findElement(By.className("link_thread_title")).getAttribute("href"); → get the product URL
53. String price = webElement.findElement(By.className("price--discounted")).getText();     → get the product price tag
54. String Seen = webElement.findElement(By.className("data--views")).getText();            → get how many times product were seen
55. String title = webElement.findElement(By.className("link_thread_title")).getText();     → get The Product Title
88. WebElement sellerDetail = driver.findElement(By.className("location"));                 → get The Seller Location
96. WebElement sellerType = driver.findElement(By.className("rank"));                       → get The Seller Rank in Kaskus
103. WebElement description = driver.findElement(By.className("entry-body"));               → get The Product Description
```

Figure 4. Part of Web Scraping Code in Java Programming Language (source: authors, 2021)

Data Storage

In this research, there are 240 rows of programming code to run the web scrapping engine. The food and beverage category in Forum Jual Beli Kaskus website has been indexed into several categories: Food, Beverage, Dairy products, Condiment, Cooking materials, and others. After the web scrapping engine retrieves the data, it is time to store it properly. Then the script code, as shown in Figure 5, should be run in SQL application software to arrange the web scraping data into appropriate structure and format.

```
BEGIN
  DECLARE isinserted INT DEFAULT 0;
  SELECT EXISTS(
  select url FROM spyder_kaskus_unique_pangan
  where url=NEW.url) INTO isinserted;
IF isinserted > 0 THEN
UPDATE spyder_kaskus_unique_pangan set
  price=NEW.price,
  seller_detail=NEW.seller_detail
WHERE url=NEW.url;
ELSE
INSERT INTO spyder_kaskus_unique_pangan
( title, price, seller,  url, description, seller_detail)
values
    ( NEW.title, NEW.price, NEW.seller, NEW.url, NEW.description, NEW.seller_detail);
END
```

Figure 5. Part of SQL code to Store the Web Scraping Data (source: authors, 2021)

The data from the product in the whole category was successfully scrapped and inserted into the SQL database, which in this research uses HeidiSQL as a database application. For the result, there are 5,191 rows of product, where each row contains eight elements, then it could collect more than 40,000 relational data within 2 hours 40 minutes (in 20 Mbps internet connection).

Data Analysis

Rapidminer Data Science Tools was used to generate insight for text analysis. The text analysis input was collected from the "title" column in the SQL database, which is the headline of the seller's product. According to de Souza et al. (2018), Products are presented on e-commerce platforms using titles that highlight information about the product. As mentioned above, a trending topic on Twitter is an excellent example of text analysis in big data to get an insight. Imagine this text mining and analysis process

replicating a product's title in e-commerce similar to the Twitter post character. From a thousand list of products, the algorithm will determine the trending topic of the product. Users such as the Government Organization responsible for controlling food products could review this valuable information to set a strategy prior to sampling and testing the food product.

The model of text preprocessing and analysis is described in Figure 6 below. Since the data from text is very noisy, then several methods (operator) were employed to clean the data. It includes a technique for converting textual data to lowercase. The model was utilized to divide the text into words, then eliminate all stop words based list on Bahasa Indonesia Dictionary from the dataset, according to Tala (2013). Thus, it assists in focusing on the actual raw data and improves accuracy. The association rule was implemented to find word interaction that enables finding the meaning (related product) with a minimum of 0.95 confidence level setting.

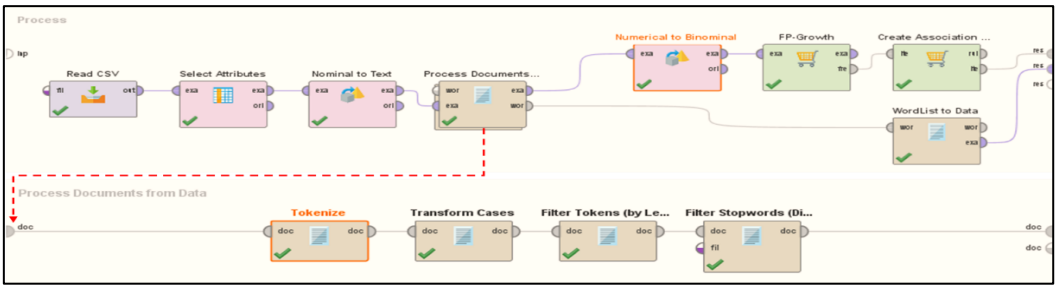


Figure 6. Text Preprocessing Model in Rapidminer (source: authors: 2021)

As the result, the model at Figure 6 was successfully found several keywords. However, the exciting point could be found in Figure 7 which is that several keywords might relate (associate) to each other. For example, the word "Putih" (white) was found to closely related for the word "Permen" (Candy) and "Kayu" then could be describe as "Permen Kayu Putih" (Cajuput Candy) with 0.975 confidence level. Also, the word "putih" closely related to the word "Tepung" (flour), "Beras" (rice), "Rose" and "Brand" which could describe as "Tepung Beras Putih Rose Brand" (Rose Brand White Rice Flour) with 0.979 confidence level.

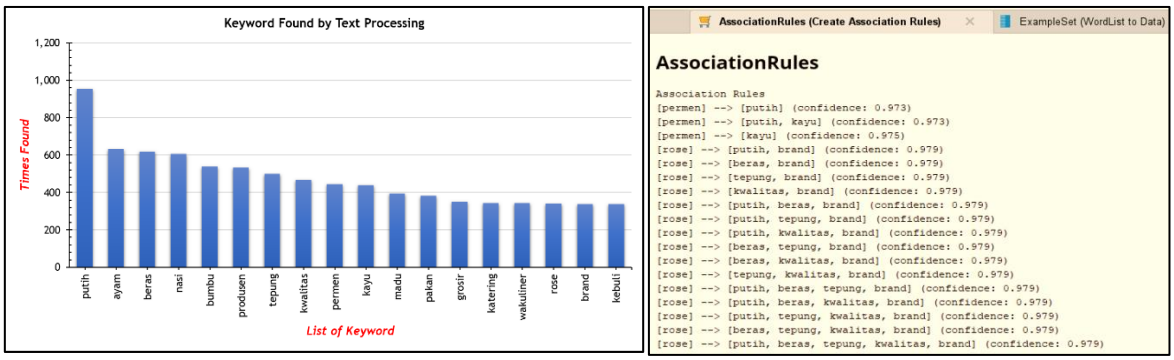


Figure 7. Wordlist of Keyword Found on Text Processing and The Word Relation (source: authors: 2021)

Based on the association rules, the authors could determine the related product which frequent appear at the product gallery. The Cajuput Candy and Rose Brand White Rice Flour were favourable products to sell by most sellers.

Predictive Model

According to the keyword list generated by text preprocessing in figure 6, there are product titles indicated for law violation which are alcoholic beverages and adulterated coffee powder. According to Sandi (2020), there is an increasing trend of alcoholic beverages sold online, which is presumably because distributors and producers have experienced a decline in sales, contributed mainly by offline markets such as bars, hotels, and entertainment venues that have been affected by movement restriction due to the Covid-19 pandemic.

However, Government Regulation No. 69/1999 on Food Labels and Advertisements in article 58 mentions that anybody shall be prohibited from advertising alcoholic drinks in whatever mass media, then this kind of advertisement shall be revoked from the e-commerce list. Nevertheless, some foods describe their product to help for erectile dysfunction. Many of these products were in powder coffee form, and intentionally added the drug substance, which prohibited according to Indonesia Food Law No. 18 the year 2012 in article 75.

As shown in Figure 8, the predictive model was developed to predict if a product with an inappropriate title was shown. It includes the kind of alcoholic beverages and adulterated coffee products. Rapidminer

software was used to implement this model, including machine learning and classification methods. Through the use of training datasets, this model develops its intelligent system. The expert judgement was employed for the training dataset based on web scraping results to determine which product was categorized as the law violation and which are not, while in this research, the alcoholic beverages product and adulterated coffee powder product was being targeted for further action then this product will be included in the attention list.

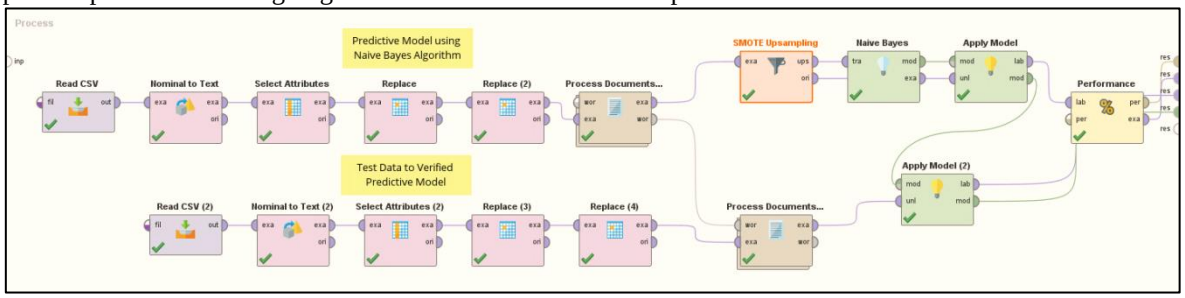


Figure 8. The Predictive Model Run in Rapidminer Software (source: authors 2021)

After training is complete, it will provide output results indicating whether the product is considered attention or not. Attention in this model means that the product has any word indicating violation of regulation (alcoholic beverages and adulterated coffee products). First, a trained dataset was introduced to the model based on expert judgment. However, the population of products that consider attention in training data was only 32 product (minority) against products that are considered not attention. This condition could impact imbalanced data, which affected the prediction model accuracy. Then, resampling data is one of the most common approaches to deal with an imbalanced dataset.

The model above was employed SMOTE (Synthetic Minority Oversampling Technique) operator to overcome the imbalanced data problem. SMOTE is a method for oversampling in which artificial samples are created for the minority class. This approach aids in overcoming the overfitting issue that random oversampling introduces. It concentrates on the feature space in order to produce new examples by interpolating between close positive instances. According to the model training result, the SMOTE operator generates a 10,263 example dataset to train the predictive model approach, which employed the Naïve Bayes classification algorithm to determine the model's correctness. The Naive Bayes algorithm is a classifier that is used to solve classification problems. It is a probabilistic machine learning technique that is based on the Naive Bayes principle, which is expressed in the following formula:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Where: P(A|B) is the probability of event (A) happening in the absence of event (B), P(B|A) is the probability of event (B) happening in the absence of event (A), P(A) is the probability of event (A), and P(B) is the probability of event (B). According to the Bayes theorem above, the classification strategy that presupposes one character's existence in a class does not correlate with the presence of any other feature. Furthermore, according to the training model result shown in Rapidminer, the classification or predictive model achieved 99.85% accuracy and a minimum of 80% for precision as shown in Figure 9.

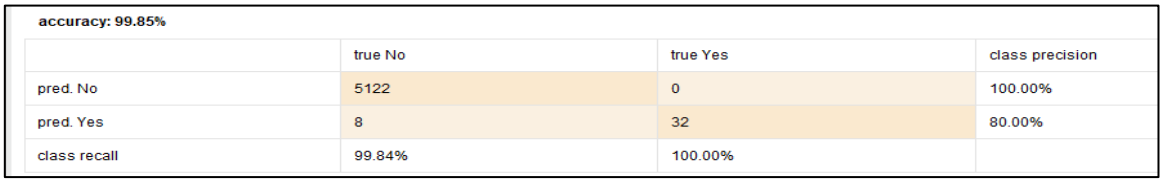
The test dataset is then introduced to the model to verify the predictive model for the next step. Dataset input only consists of 14 dummy product titles which 6 of them were expected to not include in the attention list. The rest of the 8 dummy products title contains words that violate the regulation, as shown in Table 3.

Table 3. The List of Dummy Product Title to Confirm Predictive Model

Original Text for Product Title	English Translation	Decision for Attention		Remarks
		Expected	Prediction Result	
Jual Tuak Minuman Asli Batak Bikin Mabok	Sell: Genuine Bataknesse Liquor (Tuak)	YES	YES	Alcoholic Beverages
Kopi Pecinta Tahan Lama Semalaman	Lover Coffee for Lasting All Night	YES	YES	Adulteration Detected
nasi mandhi pakistan enak	Delicious Pakistani Mandhi Rice	NO	NO	Apropriate Product
Arak China khas Bruce Lee	Bruce Lee's signature Chinese wine (Arak)	YES	YES	Alcoholic Beverages
Kopi kapal api bubuk sachet 300 gr	Kapal Api Brand coffee powder sachet 300 gr	NO	NO	Apropriate Product
Bir rasa fruitty passion merk carlsberg	Carlsberg brand fruitty passion beer	YES	YES	Alcoholic Beverages
mie aceh bawang putih siap santap	Aceh garlic noodles ready to eat	NO	NO	Apropriate Product
Fiesta Chicken Nugget	Fiesta Brand Chicken Nugget	NO	NO	Apropriate Product
Ice Cream Rasa Vanilla Walls	Walls Brand Vanilla Ice Cream	NO	NO	Apropriate Product
Sake Jepang asli dari Duty Free Tokyo	Genuine Japanese sake from Duty Free Tokyo	YES	YES	Alcoholic Beverages
Tequilla Minuman Alkohol 40% v/v	Tequilla Liquor 40% v/v alcohol	YES	YES	Alcoholic Beverages
Kopi bikin alat vital anda tahan lama	Coffee that makes male genital organs last longer	YES	YES	Adulteration Detected
Wine / anggur merah prancis asli 20% v/v	Authentic French Wine 20% v/v	YES	YES	Alcoholic Beverages
Permen asem buat ibu hamil	Sour candy for pregnant women	NO	NO	Apropriate Product

Note: All original dataset tests was successfully achieved the appropriate prediction result

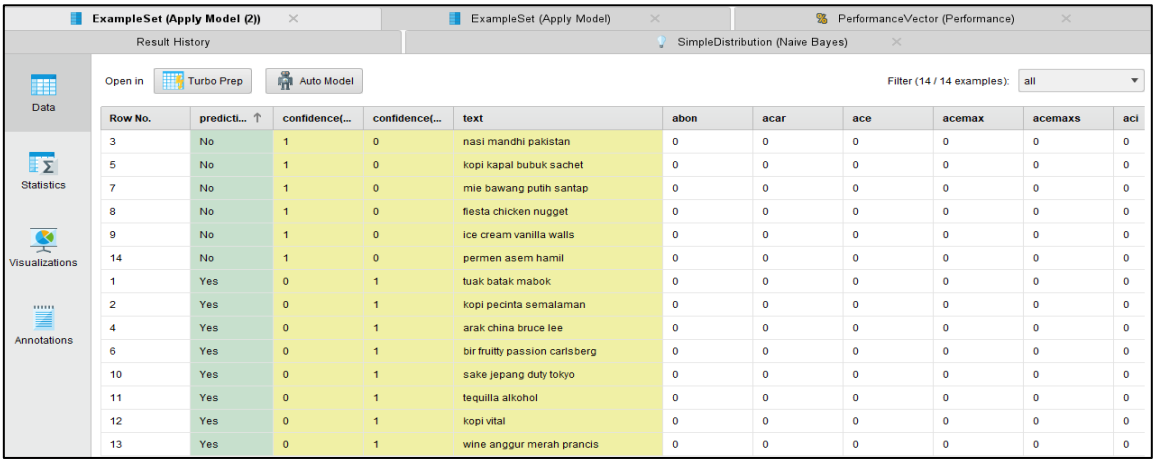
The test result comes with a satisfying result, where all the dummy product title was classified into precise prediction, as shown in Figure 10. Following those results, it could be assumed that the model achieved the best approach for prediction, which would be helpful for authority for further law enforcement action.



accuracy: 99.85%

	true No	true Yes	class precision
pred. No	5122	0	100.00%
pred. Yes	8	32	80.00%
class recall	99.84%	100.00%	

Figure 9. The Model Accuracy According to Trained Dataset (source: authors, 2021)



ExampleSet (Apply Model (2)) | ExampleSet (Apply Model) | PerformanceVector (Performance) | Result History | SimpleDistribution (Naive Bayes)

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model — Filter (14 / 14 examples): all

Row No.	predicti...	confidence(...	confidence(...	text	abon	acar	ace	acemax	acemaxs	aci
3	No	1	0	nasi mandhi pakistan	0	0	0	0	0	0
5	No	1	0	kopi kapal bubuk sachet	0	0	0	0	0	0
7	No	1	0	mie bawang putih santap	0	0	0	0	0	0
8	No	1	0	fiesta chicken nugget	0	0	0	0	0	0
9	No	1	0	ice cream vanilla walls	0	0	0	0	0	0
14	No	1	0	permen asem hamil	0	0	0	0	0	0
1	Yes	0	1	tuak batak mabok	0	0	0	0	0	0
2	Yes	0	1	kopi pecinta semalaman	0	0	0	0	0	0
4	Yes	0	1	arak china bruce lee	0	0	0	0	0	0
6	Yes	0	1	bir fruitty passion carlsberg	0	0	0	0	0	0
10	Yes	0	1	sake jepang duty tokyo	0	0	0	0	0	0
11	Yes	0	1	tequilla alkohol	0	0	0	0	0	0
12	Yes	0	1	kopi vital	0	0	0	0	0	0
13	Yes	0	1	wine anggur merah prancis	0	0	0	0	0	0

Figure 10. Result of Testing Dataset for 14 Dummy Data indicate 100% accurate (source: authors, 2021)

Business Intelligence

According to Ruzgas (2017), Tableau is the golden standard in business intelligence and analytics platforms. This research utilized Tableau for Desktop with a student license to transform the data from Big Data Analysis into a more informative dashboard. The dashboard consists of 4 quadrants which each quadrant represents certain information, as shown in Figure 11; the upper left quadrant was filled with a heat map of the food product. This heat map represents the seller's location, where the darker colour means that those regions have more sellers than any other. According to the heat map, the authority prioritized the region that should be more attention due to regulatory compliance risk. The below-left quadrant represents the keyword's bar chart, representing that the keyword was spoken more than any keyword. This information was collected from text preprocessing, while it represents the product that has more space in listing indexes on the e-commerce platform. The authority could align this keyword within the public policy, particularly for the sampling and testing policy. This real-time database visualization could benchmark a thematic sampling program instead of issuing an annual sampling priority plan.

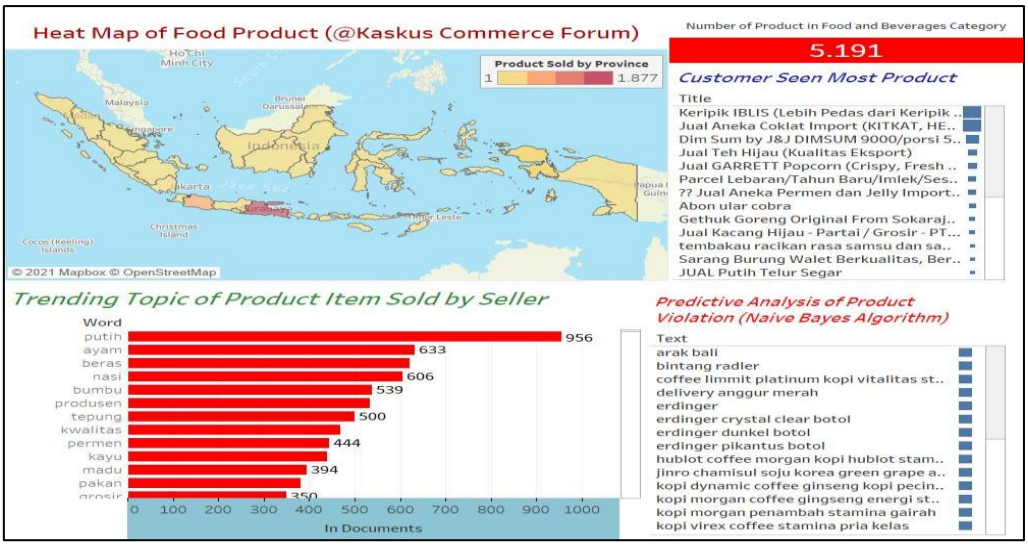


Figure 11. The Output of Business Intelligence using Tableau for Desktop (Source; authors, 2021)

The upper right table showed the list of top listed products clicked by the buyer, where this particular data appeared on the first page of the website product gallery. The clicked advertisement means that the product was attractive in the eyes of consumers and got the highest possibility to be picked up at the basket. The authority could use this detailed data for further understanding of consumption patterns. Finally, the below right quadrant represents the predictive analysis of product violation. This information is essential for authority to make a countermeasure in terms of ensuring public order.

Conclusion

The current Covid-19 pandemic and increasing internet access trend in Indonesia changed consumer behaviour to fulfil their demand, resulting in an unplanned shift from the traditional or physical marketplace into the online market (e-Commerce). It also gives a vast opportunity for entrepreneurs in the food and beverage business new ways to expand their market. The inclusive potential economy offered by the online platform is also a trigger for further economic development that connected the food supply chain, including consumer, producer, logistic or forwarder, and financial technology. Thus, the authority should carefully consider this phenomenon since the conventional way of controlling regulations is no longer enough in the connected or online world. Moreover, with the retail collapsing several years back, there should already be an alarm and confirmation of the changing consumption pattern. Business Intelligence Tools combined with Big Data Analysis has proven successful in providing deep insight into market research from Kaskus E-Commerce Forum. Therefore, the research question easily be answered appropriately.

To answer the first research question, what is the insight from the descriptive result after conducting Big Data Analytics and Business intelligence in the food and beverages category on Forum Jual Beli Kaskus? The authors found that since the data in the dynamic website was able to collect and analyze automatically through Big Data Analysis, the current market researcher could quickly gain enormous insight. Big Data Analysis offers new ways of exploring and interpreting data. Not only provide descriptive analysis, but the scholar has created many new algorithms to improve further prediction precision. For example, according to research text preprocessing output in Forum Jual Beli Kaskus (e-Commerce Forum), the authors could find the current trend of food products sold in Kaskus was dominated by Cajuput Candy and Rose Brand White Rice Flour.

Nevertheless, it is also important to anticipate potential law violations for products sold online due to the advertisement of several unappropriated products. For example, alcoholic beverages were known to be prohibited from being sold online according to government regulation. Also, there is a potential trend of coffee products containing undeclared and controlled harmful drug substances that violate national food laws. The machine learning model using SMOTE (Synthetic Minority Oversampling Technique) operator and Naïve Bayes classification could predict with 99.85% and had 80% class of precision of decision making whether a product could be included in attention list or not according to the regulation. According to the test dataset, this model showed satisfactory results, including 14 rows of the dummy product title.

To answer the second research question, how could the interpretation result benefit the Indonesian Food Control Authority? The authors found that the predictive model was helpful for authority to mitigate the food business violation risk and prepare countermeasure prior incidents occurred. Business intelligence combined with the use of Big Data Analysis shows the state of art data science. This research proves that employing Big Data Analysis is very cost-efficient while the biggest challenge is to prepare the human competency, particularly for data scientist. For future perspective, the government authority could depend on the dynamic

data-driven based policy-making instead of conducting traditional market research while the control object like the physical market was faded due to the shifting behaviour of the consumer.

The current market research model on Rapidminer utilizes only text preprocessing methods for the analysis of machine learning. Furthermore, the prediction model was only tested by 14 dummy product tittle that considers straightforward, so it could not generalize for common results. The machine learning model above could bring further research to be tested with diverse data and complex words to increase accuracy and precision. Also, the model could be developed by implementing the other parameters like product price, seller location, seller rank, and many attributes previously collected by the Web Scrapping engine.

Acknowledgment

This work was funded by BPOM (Indonesian FDA), with project number PB.02.01.1.2.08.20.392, dating on August 28, 2020.

References

Journal articles

Alnoukari, Mouhib. (2020). *From Business Intelligence to Big Data -The Power of Analytics*. https://doi.org/10.4018/978-1-7998-5781-5.ch003.

Chen, Hsiu-chin & Chiang, Roger & Storey, Veda. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact*. MIS Quarterly. 36. 1165-1188. https://doi.org/10.2307/41703503.

De Souza, C., J., Kozielski, M., Mathur, P., Chang, E., Guerini, M., Negri, M., Turchi, M., & Matusov, E. (2018). *Generating E-Commerce Product Titles and Predicting their Quality*. INLG. https://doi.org/10.18653/v1/W18-6530

Debortoli, Stefan & Müller, Oliver & Brocke, Jan vom. (2014). *Comparing Business Intelligence and Big Data Skills*. Business & Information Systems Engineering. 6. 289-300. 10.1007/s12599-014-0344-2.

Djufri, Mohammad. (2020). *Penerapan Teknik Web Scraping Untuk Penggalian Potensi Pajak (Studi Kasus Pada Online Market Place Tokopedia, Shopee Dan Bukalapak)*. Jurnal BPPK : Badan Pendidikan dan Pelatihan Keuangan. 13. 65-75. https://doi.org/10.48108/jurnalbppk.v13i2.636.

Fan, S., Lau, R.Y.K., Zhao, J.L., 2015. *Demystifying big data analytics for business intelligence through the lens of marketing mix*. Big Data Res. 2 (1), 28–32. https://doi.org/10.1016/j.bdr.2015.02.006 .

Fisher, D., Mcdonald, D. W., Brooks, A. L., & Churchill, E. F. , 2010, *Terms of Service, Ethics, and Bias: Tapping the social web for CSCW research*, CSCW 2010, February 6–10, 2010, Savannah, Georgia, USA.

Han, S., & Anderson, C. K. (2021). *Web scraping for hospitality research: Overview, opportunities, and implications*. Cornell Hospitality Quarterly, 62(1), 89-104.

Hirschey, J.K. (2014). *Symbiotic Relationships: Pragmatic Acceptance of Data Scraping*. Berkeley Technology Law Journal, 29, 16.

Ho, Hoang Phuong Thao. (2020). *Leveraging web scraping for collecting competitive market data: Case: A case study of an Airbnb rental unit in Helsinki*. LAB University of Applied Sciences: Finland

Kotsanopoulos, Konstantinos & Arvanitoyannis, Ioannis. (2017). *The Role of Auditing, Food Safety, and Food Quality Standards in the Food Industry: A Review: Food safety and quality audits*. Comprehensive Reviews in Food Science and Food Safety. 16. 10.1111/1541-4337.12293.

Marasanapalle, Jayanth & Vignesh, T. & Srinivasan, Praveen & Saha, Angshuman. (2010). *Business intelligence from Twitter for the television media: A case study*. 10.1109/BASNA.2010.5730304.

Milev, P. (2017). *Conceptual Approach for Development of Web Scraping Application for Tracking Information*. Economic Alternatives, 475-485.

Parks, R., Ceccucci, W., & McCarthy, R.V. (2018). *Harnessing Business Analytics: Analyzing Data Analytics Programs in US Business Schools*. Information Systems Education Journal, 16, 15-25.

Passlick, J., Lebek, B., & Breitner, M. H. (2017). *A Self-Service Supporting Business Intelligence and Big Data Analytics Architecture*. Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017) (pp. 1126–1140). Academic Press.

Ramaditya, Muhammad. (2019). *Exploring the Impact of Perception After Rebranding and Customer Satisfaction on Corporate Image (A Case Study: PT. Darta Media Indonesia Kaskus)*. https://doi.org/10.2991/aicmar-18.2019.38.

Ruzgas, Tomas & Bagdonavičienė, Jurgita. (2017). *Business Intelligence for Big Data Analytics. International Journal of the Applications Technology and Research*. 6. 001-008. 10.7753/IJCATR0601.1001.

Saurkar, Anand V., Kedar G. Pathare, Shweta A. Gode. 2018. *An Overview On Web scraping Techniques And Tools*. International Journal on Future Revolution in Computer Science & Communication Engineering (IJFRCSCE), Volume: 4 Issue: 4, April 2018.

Sun, Z., Zou, H., & Strang, K. (2015). *Big Data Analytics as a Service for Business Intelligence*. 14th Conference on e-Business, e-Services and e-Society (I3E), 200-211. 10.1007/978-3-319-25013-7_16

Tala, Fadillah. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation, Universiteit van Amsterdam: The Netherlands

Electronic Book

BPS. (2020). *Statistik E-Commerce 2020*. Sub-directorate of Information and Communication Technology Statistics: Indonesia. Retrieved from https://www.bps.go.id/publication/2020/ (accessed October 29, 2021).

Demunter, C. (2017), *Tourism statistics: early adopters of big data?*, Eurostat, European Union, Luxembourg. Retrieved from https://ec.europa.eu/eurostat/documents/3888793/8234206/KS-TC-17-004-EN-N.pdf (accessed October 29, 2021).

PwC's Global Consumer Insights Survey 2020. (2021). *The consumer transformed: Changing behaviours are accelerating trends along a reinvented customer purchase journey*. Retrieved from https://www.pwc.com/gx/en/consumer-markets/consumer-insights-survey/2020/pwc-consumer-insights-survey-2020.pdf (accessed October 29, 2021).

Online Newspaper

The Jakarta Post (2018, March 25). *Data Scientists In High Demand Amid Talent Shortage*. Retrieved from https://www.thejakartapost.com/life/2018/03/25/data-scientists-in-high-demand-amid-talent-shortage.html. (accessed October 29, 2021).

Narizhnykh, D. 2018. *Is web scraping legal or not?*. Retrieved from: https://medium.com/dataflow-kit/is-web-scraping-legal-or-not-f6c26074584 (accessed October 29, 2021)

Sandi, F. (2020, June 29). *Heboh Bir Hingga Miras Dijual Online, Memang Boleh?.* CNBC Indonesia. Retrieved from https://www.cnbcindonesia.com/news/20200629161804-4-168829/heboh-bir-hingga-miras-dijual-online-memang-boleh. (accessed 29 October, 2021)