

Implementation of Text Mining System Development: A Case Study in Telecommunication Industry

Shelvy RG Elsa Hasibuan¹, Nazaruddin Matondang², Juliza Hidayati³

^{1,2,3}Program Studi Magister Teknik Industri, Fakultas Teknik, Universitas Sumatera Utara, Medan, Indonesia

Abstract. In today's internet age along with the advancement of web technology and its growth, large amounts of data today can be used and utilized to analyze and win the market. Social networking sites such as Twitter are quickly able to get neutral, positive and negative customer opinions and discussions using sentiment analysis methods. In this study the process of classifying sentiments using the Naïve Bayes Classifier method. This method approach provides digital-based surveys and comparative analysis such as machine learning and lexicon-based approaches. The results obtained from research based on data mining surveys there are five main complaints that are the main focus of telecommunications service companies, namely network, convenience, price, internet and services. Sentiment analysis based on test results showed the highest polarity of opinion was in the network at 60.07% and the lowest comfort with a value of 36.63%. The higher the polarity value, the higher the company's chances of making further improvements. The results showed that the development of big data mining systems helped companies to more accurately and effectively meet customer needs and classify sentiment towards products that have been launched.

Keyword: Big Data, Text Mining, Sentiment Analysis, Naïve Bayes Classifier

Abstrak. Pada era internet saat ini beserta kemajuan teknologi web dan pertumbuhannya, sejumlah besar data-data saat ini dapat digunakan dan dimanfaatkan untuk menganalisa dan memenangkan pasar. Situs jejaring sosial seperti Twitter dengan cepat mampu mendapatkan opini maupun diskusi pelanggan yang netral, positif maupun negatif menggunakan pendekatan metode sentimen analisis. Pada penelitian ini proses pengklasifikasian sentimen menggunakan metode Naïve Bayes Classifier. Pendekatan metode ini menyediakan survei berbasis digital dan analisis komparatif seperti pembelajaran mesin dan pendekatan berbasis leksikon. Hasil yang diperoleh dari penelitian berdasarkan survey penambangan data terdapat lima keluhan utama yang menjadi fokus utama perusahaan jasa telekomunikasi yaitu jaringan, kenyamanan, harga, internet dan layanan. Analisis sentimen berdasarkan hasil testing menunjukkan polaritas opini tertinggi terdapat pada jaringan sebesar 60.07% dan terendah kenyamanan dengan nilai 36.63%. Semakin tinggi nilai polaritas maka semakin tinggi peluang perusahaan untuk melakukan perbaikan lanjutan. Hasil penelitian menunjukkan adanya pengembangan sistem penambangan big data membantu perusahaan untuk lebih akurat dan efektif menemukan kebutuhan pelanggan serta mengklasifikasi sentimen terhadap produk yang sudah diluncurkan.

Kata Kunci: Big Data, Text Mining, Sentiment Analysis, Naïve Bayes Classifier

Received 04 November 2021 | Revised 26 November 2021 | Accepted 12 January 2022

*Corresponding author at: Jl. Almamater, Universitas Sumatera Utara, Medan, Indonesia

1. Introduction

In recent years, there have been many new technological advances which have meant that companies should already adopt new technologies for business models that incorporate globalization and use the internet as a promotional tool for products and service [1]. The Internet era has changed the way everyone expresses their opinions and views. Everyone's opinion is currently being done through blogs, app product reviews, and social media. Currently, the number of social media users such as Facebook, Twitter, Instagram, Google, and so on is very large. These social media users express their emotions, opinions, and various views about their daily lives on social media. Social media has become an online community and media that is able to influence others, for example, most people give sentiments in the form of tweets on Twitter, provide reviews, comments and so on.

Moreover, social media provides opportunities for businesses by providing a platform to connect with their customers. Social media can give someone a reference if they want to buy a product or want to use any service, so they first look for reviews online, discuss them on social media before deciding. The amount of user-generated content is too much for the average user to analyze. So, there is a need to automate this, various sentiment analysis techniques are widely used [2]. In general, the news information conveyed is positive, but some is negative. Therefore, sentiment analysis is needed to be able to choose good news. Sentiment analysis is an activity to understand, extract, and process textual data automatically to obtain information [3].

In the past, if a designer wanted to create a new model, customer requirements were gathered through interviews, questionnaires and surveys. At this time with the advancement of internet technology, these customer needs can be obtained through big data Twitter, blogs, and product and service reviews that will describe customer needs. Conventional methods when used will provide a very limited amount as the object of research. However, when compared with conventional survey data methods, big data presents different and contrasting things and provides characteristics. Generally, the characteristics of big data are volume, variety, speed and value which is often referred to as 4V. Exclusively tailored algorithms help business administrators, research engineers and data scientists understand customer needs effectively and efficiently from big data. Success in big data requires computer science algorithms to identify sentimental information from textual data and design knowledge to understand customer needs [4]. This study aims to look at the use of customer big data through Twitter and identify the polarity of product sentiment through creating an information system and change traditional method to gather customer requirement become digital method.

There are not so many companies are able to quickly and precisely transform to face today's technological advances. Today's telecommunications industry must be able to quickly translate customer needs by utilizing data sources, including big data on all social media platform. Currently, the translation of customer needs is still done manually, even though there are many data sources available and their use must be maximized. Therefore, with this research, researchers

want to provide input to companies to use and strengthen the use of big data by developing a text mining system.

2. Methodology

The method in this research is divided into 2 phases. The first phase is to build a text data mining system model using Python programming and determine the topics to be analyzed on Twitter. In this study using a research sample, namely customer opinion data from tweets on Twitter big data. The research sample is commentary data used to identify customer needs. Customer opinion tweet data taken from Twitter big data for the period May 2020 to May 2021 with a total number of comments of 84,367 comments. The second phase is the use of sentiment analysis method using the Naïve Bayes Classifier algorithm to identify positive, negative, and neutral tweets. Conceptual framework of this research can be seen in Figure 1.

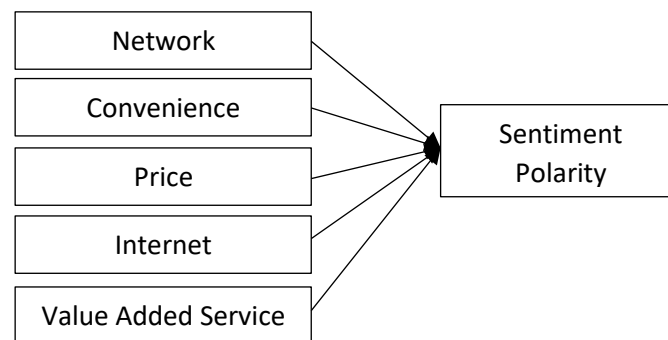


Figure 1. Conceptual Framework

Operational definition of conceptual framework as shown in Figure 1 described as follows:

1. Network Area (Coverage Network)

Guaranteed communication with the minimum failure rate that the company can provide. The company continues to update the network area, evenly distributed signal quality, international roaming, voice call quality, and so on.

2. Convenience

Convenience in telecommunications services, namely customers can comfortably reach the customer care center along with the facilities and services provided at the customer care center. Convenience in this case is defined by the speed of processing customer complaints, as well as friendliness to customers when receiving complaints.

3. Price

The customer's perception of the price of a product or service is determined by supply and demand and the price is a sign of the quality of a product or service. Prices for

telecommunications services in question are telephone service prices, internet service prices, SIM card replacement service prices (Subscriber Identity Module), and other telecommunications product service prices.

4. Internet (Internet Product Availability)

The availability of internet products, both in the form of cards and vouchers, is available to all regions so that customers in rural areas. The company pays attention to the number of sales channels to the smallest areas so that the distribution of internet products for telecommunication services can be used by customers.

5. Service (Value Added Service/VAS)

VAS services are popular in the telecommunications industry. There are two main reasons why VAS is important to create, namely VAS provides additional satisfaction for customers and provides additional benefits for the company. VAS is. VAS is an additional product attached to the main product (basic product), such as digital services, which are currently being developed by the telecommunications industry

6. Sentiment Polarity

Sentiment polarity for an element defines the orientation of the expressed sentiment, it determines if the text expresses the positive, negative or neutral sentiment of the user about the entity in consideration.

2.1. Research Steps

The world of technology is always evolving and will never run out to be discussed. In the world of technology, there are many terms, all of which are related to technology and information. A term that is often used in the world of technology and information is information systems. Where this term is often misunderstood. Research steps will be focused on developing information system using big data Twitter.

The system is a form of interaction between one component and another because the system has different means for each case that occurs in the system. The system has certain characteristics or properties, which have components, system boundaries, system environment, interface, input, output, processor, objective, and goals [5].

2.1.1. Web Crawler

Twitter provides an Application Programming Interface (API) to facilitate data crawling. The API makes it easy for users to retrieve tweet data on a real time basis. The initial purpose of the Twitter API was to gather information on certain communities for their views on current trending topics [6]. In this study, the steps in the web crawler can be seen in Figure 2 with the following command:

Enter the following command "twint -s 'keyword' --until yyyy-mm-dd --since yyyy-mm-dd --lang id --csv -o datax.csv"

Information :

1. -s (*keyword*).
2. --lang (*language*, ex : id, us)
3. --limit (*limit result*)
4. --until (*from date*)
5. --since (*to date*)
6. -o (*save result*)
7. --csv (*output format csv*) / --json (*output format json*)

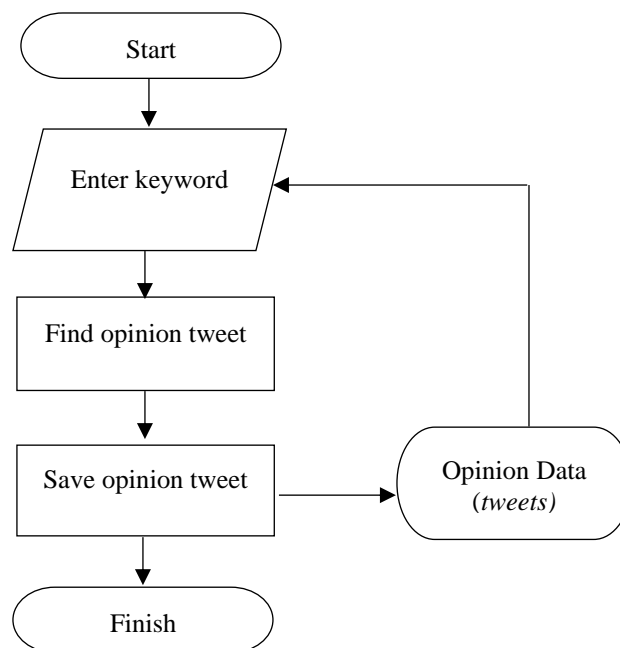


Figure 2. Web Crawler Steps

Crawling in the above step is a process of quickly retrieving large amounts of data on web pages into a local storage area and indexing them based on several keywords [7]. This research also focuses on doing the same way to take advantage of the Twitter API and create PHP-based applications to capture keywords in telecommunications companies and their products.

2.1.2 Preprocessing Process

2.1.2.1. Part of Speech Tagging (POS-Tagging). Part of Speech Tagging (POS-Tagging) is a process that automatically assigns word class labels to a word in a sentence. The steps in POS Tagging can be seen in Figure 3.

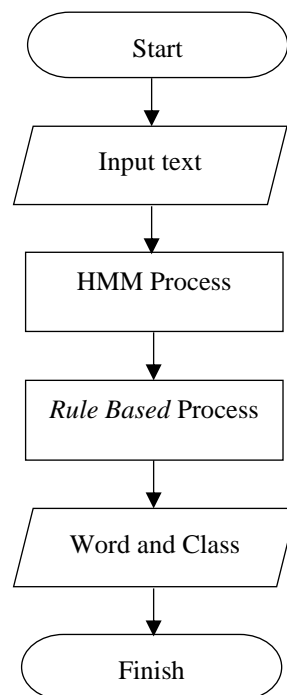


Figure 3. POS Tagging Steps

2.2.1. Scrapping Process

Scraping is done automatically using the Twint library, which data is imported through the Jupiter Lab interpreter using the keywords you want to scrape. The python-based program then scrapes the Twitter website backend by automatically scraping comments. Then the data is collected into a file.csv. Figure 4 shows data tweet that will process into the next step of scrapping process.

	A	B	C
1	created_at	username	name
2	2021-01-01 07:54:00 WIB	akugum_xyx	anak bawang
3	2021-01-01 08:13:24 WIB	eaudeadora	eudora dora dori
4	2021-01-01 08:50:46 WIB	bakol43930358	bakol
5	2021-01-01 09:39:29 WIB	beianjafess	beianjafess
6	2021-01-01 21:02:53 WIB	ybiqb	ybiqb
7	2021-01-01 21:09:03 WIB	fmuhamam	firman muhamam
8	2021-01-01 22:23:07 WIB	n_istikhomah	icetea
9	2021-01-01 22:30:53 WIB	taufiiiiikk	muhammad taufik
10	2021-01-01 23:19:52 WIB	adheliavns	jellyjellylll
11	2021-01-02 00:14:30 WIB	hidayatp20	hidayat
12	2021-01-02 01:33:07 WIB	oreochococreme	doni
13	2021-01-02 01:47:34 WIB	illuminajong	presiden hasil manipulasi
14	2021-01-02 02:28:50 WIB	skyrio999	rio
15	2021-01-02 04:27:35 WIB	gusuran76	gus uran
16	2021-01-02 05:27:56 WIB	binnybunny96	help check pinned
17	2021-01-02 05:40:03 WIB	pradanagara	dita pradanagara
18	2021-01-02 06:00:30 WIB	schleiermarcher	wilhelmdilthey
19	2021-01-02 06:36:54 WIB	benizar	benidzar m. andrie
20	2021-01-02 07:17:18 WIB	reza113333	reza
21	2021-01-02 07:26:10 WIB	lovewhisperer4	with much love
22	2021-01-02 07:52:57 WIB	aruchan23972493	flwhtstrrrr

Figure 4. Tweet Data

Preprocessing is done to process the existing data so that researchers can prevent the occurrence of inconsistent data. The goal is that the output of the classification has a high level of accuracy. The stages of preprocessing include clean html, clean mention, and translating. The explanation of each step is as follows :

1. Clean html

Clean html is the stage of changing all the letters that contain https // or the address of a site.

For example : @PT XYZ mau hadiah pulsa gratis dan quota 5Gb klik <https://www.pulsagratiss.com>

Result: @PT XYZ mau hadiah pulsa gratis dan quota 5Gb klik.

2. Clean Mention

Clean Mention is the stage of removing mentions which are a feature of appointing someone's user on Twitter.

For example : @PT XYZ mau hadiah pulsa gratis dan quota 5Gb klik.

Result: mau hadiah pulsa gratis dan quota 5Gb klik.

3. Translating

Translating is the stage that processes sentences that were previously Indonesian into English.

For example: mau hadiah pulsa gratis dan quota 5Gb klik

Result: want a free credit gift and a 5Gb quota click

3. Result and Discussion

3.1. Sentiment Analysis Using Naïve Bayes

Sentiment Analysis is a research methodology that analyzes the feelings of a given sample, usually from an online digital platform or social network, to find different opinions with different methodological approaches. It has been confirmed that the sentiment analysis methodology can analyze and identify users' feelings and opinions and influence users to make decisions [8]. There are several approaches in using the sentiment analysis method, namely UGC (User Generated Content) such as Naïve Bayes Classifier, Linear Regression, or Deep Learning [9]. This study uses keywords from a topic, semantic meaning and concept, such as hashtags, retweets, and identification points of products and services taken on Twitter web pages [10].

Textblob is a python library for processing text data. Textblob has the basic features of Natural Language Processing (NLP), in this study the features used in the Textblob library are sentiment analysis used to classify training data. The algorithm used is the Naïve Bayes algorithm. The method that is often used for sentiment classification is Naïve Bayes [11].

Combining the probability of words from categories in a particular document which is the basic idea of using the Naive Bayes method. The output of this process is a training data which will then be tested on data testing. The data formed is stored in csv format. The programming script to perform sentiment analysis can be seen as shown in Figure 5 and Figure 6 showing the results of the sentiment analysis programming process using the Textblob algorithm.

```
8]: def clean_tweet(tweet):
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\w+\/\S+)", "", tweet).split())

9]: def analize_sentiment(English):
    analysis = TextBlob(clean_tweet(English))
    if analysis.sentiment.polarity > 0:
        return "positif"
    elif analysis.sentiment.polarity == 0:
        return "netral"
    else:
        return "negatif"

1]: import numpy as np
    from textblob import TextBlob
    import matplotlib.pyplot as plt
    import re
    telkomsel_data['SA'] = np.array([analize_sentiment(tweet) for tweet in telkomsel_data['english']])
    telkomsel_data
```

Figure 5. Function To Perform Sentiment Analysis

Unnamed: 0	created_at	user_id	username	tweet	karakter	cleanHTML	cleanMention	english	SA
0	2021-05-31 06:26:29 SE Asia Standard Time	2.271910e+08	aldosalam	Udah pada tau belum apa itu Telkomsel Orbit ? ...	263	Udah pada tau belum apa itu Telkomsel Orbit ? ...	Udah pada tau belum apa itu Telkomsel Orbit Te...	Already know what it is yet Telkomsel Telkomsel...	positif
1	2021-05-31 05:43:58 SE Asia Standard Time	5.242736e+08	sandy_indigo72	@Telkomsel apa kedepan hp yang bisa menikmati ...	168	@Telkomsel apa kedepan hp yang bisa menikmati ...	apa kedepan hp yang bisa menikmati layanan 5G ...	what hp future could enjoy only hp telkomsel 5...	positif
2	2021-05-31 04:41:11 SE Asia Standard Time	9.494276e+08	heapmemory	@Telkomsel ini layanan *133# kok sedang sibuk ...	114	@Telkomsel ini layanan *133# kok sedang sibuk ...	ini layanan 133 kok sedang sibuk mulu ya Saya ...	This service is really busy 133 mulu yes I wan...	negatif
3	2021-05-31 01:37:56 SE Asia Standard Time	1.386154e+18	warisadijeng	Untuk berubah menjadi planet Neptunus yang bis...	108	Untuk berubah menjadi planet Neptunus yang bis...	Untuk berubah menjadi planet Neptunus yang bis...	To turn into the planet Neptune which can help...	netral
4	2021-05-31 00:24:59 SE Asia Standard Time	1.298256e+18	sobatrejun	Apakah Anda bersedia menerima info layanan ter...	189	Apakah Anda bersedia menerima info layanan ter...	Apakah Anda bersedia menerima info layanan ter...	Are you willing to accept the latest service i...	positif
...
588	2021-05-01 17:41:48 SE Asia Standard Time	8.710422e+08	dinalativa	Min @Telkomsel , beli paket kombo ternyata ada...	164	Min @Telkomsel , beli paket kombo ternyata ada...	Min beli paket kombo ternyata ada include laya...	Min buy a combo package there was a disney ser...	positif
589	2021-05-01 16:56:21 SE Asia Standard Time	6.303768e+07	kang_asep	@Telkomsel layanan telkomsel payah, CSnya ngga...	157	@Telkomsel layanan telkomsel payah, CSnya ngga...	layanan telkomsel payah CSnya nggak pernah res...	Telkomsel service lousy customer response CSnya...	negatif
590	2021-05-01 15:12:12 SE Asia Standard Time	1.587967e+08	odhonindra	Dear @Telkomsel nanti malam pukul 23.00 WIB, l...	112	Dear @Telkomsel nanti malam pukul 23.00 WIB, l...	Dear nanti malam pukul 23.00 WIB layanan maxst...	Dear tonight at 23.00 pm there is a plan maxst...	positif
591	2021-05-01 14:10:23 SE Asia Standard Time	1.260653e+18	rahkuyy	@Telkomsel tolong di respon saya jadi ga nyama...	155	@Telkomsel tolong di respon saya jadi ga nyama...	tolong di respon saya jadi ga nyaman pake Telk...	please in my response so comfortable use Telko...	positif
592	2021-05-01 13:45:31 SE Asia Standard Time	1.189742e+09	nabirenet	Ada FO Cut, Layanan Data Telkomsel Terganggu d...	118	Ada FO Cut, Layanan Data Telkomsel Terganggu d...	Ada FO Cut, Layanan Data Telkomsel Terganggu d...	There FO Cut Telkomsel Data Service Troubled D...	negatif

Figure 6. Sentiment Analysis Results Using Textblob

3.2. Implementation of the Twitter Big Data Text Mining System Interface Development

The application is made based on the attributes of customer needs which is then implemented into an application based on a website dashboard. This application has a system interface that is useful for connecting application users to the application system so that the application can be used [12]. There are 3 main menus in this dashboard, namely the Home menu, Realtime Menu and Static Menu. The following is an explanation of each menu.

1. Home Menu

The Home menu is the initial display when the user first launches the application. In this menu there are several components, namely the dashboard title, the identity of the system maker and there are 3 strip menus consisting of the Home menu, Realtime Menu, Genetic Static menu.

2. Realtime Menu

The Reatime menu is a page on the application that has a function for crawling comments directly from Twitter in realtime which is divided into three processes, namely comment crawling, comment cleaning, and sentiment processing, then it will get results in the form of Twitter user comments where the comments will then be sent. in process. In the first stage of the process, the comment category is selected based on 5 categories, namely PT XYZ network, PT XYZ convenience, PT XYZ internet, PT XYZ service and PT XYZ price. Figure 7. shows the steps for categorizing comments according to customer requirement attributes

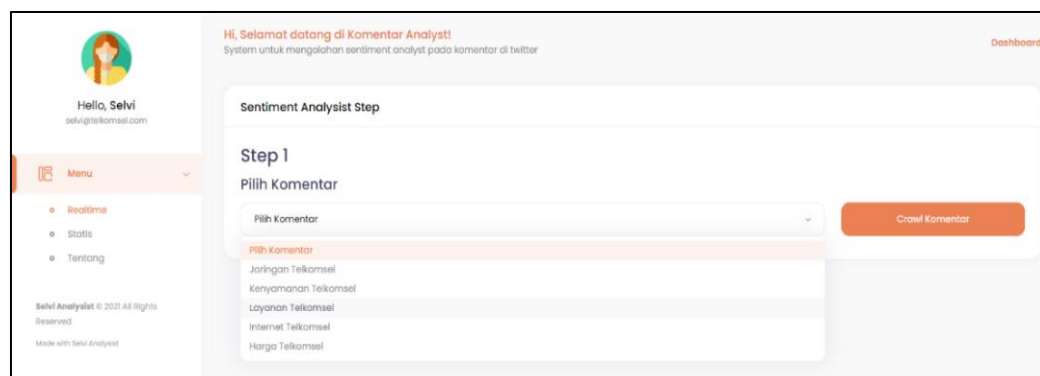
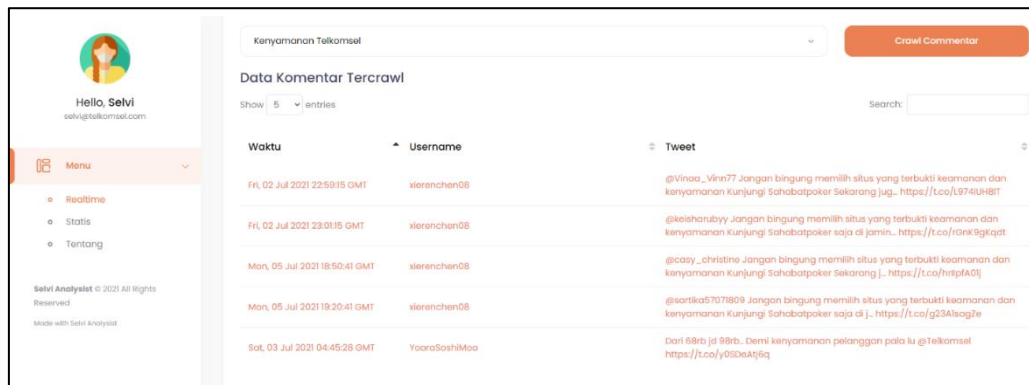


Figure 7. Comment Categorization Steps

The first process the user selects a category will display the results of crawling Twitter comments as shown in Figure 8 below. The second process is that comments that have been successfully obtained will be cleaned of unnecessary data or sentences such as the address of a site and the "@" character and automatically translated into English as shown below. The third process is processing clean data and measuring sentiment on each comment line in the form of negative, positive or neutral comments as in Figure 9. Which will then be visualized through a pie chart as in Figure 10.

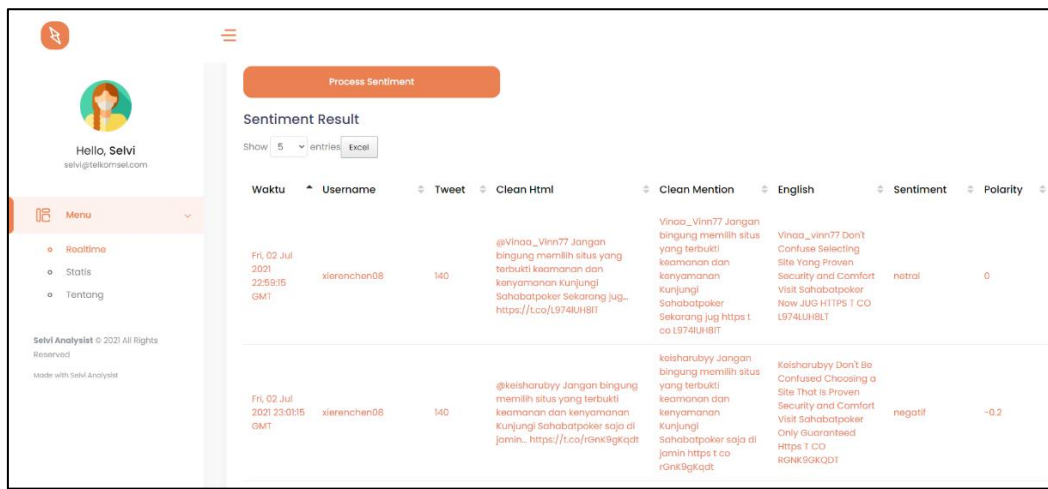


Data Komentar Tercrawl

Show 5 entries

Waktu	Username	Tweet
Fri, 02 Jul 2021 22:59:15 GMT	xierenchao08	@Vinao_Vinn77 Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker Sekarang jug... https://t.co/L974Uu8lt
Fri, 02 Jul 2021 23:01:15 GMT	xierenchao08	@keisharubyy Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker saja di jamin... https://t.co/r0nk3gkqdt
Mon, 05 Jul 2021 18:50:41 GMT	xierenchao08	@easy_christine Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker Sekarang j... https://t.co/httpdA0t
Mon, 05 Jul 2021 19:20:41 GMT	xierenchao08	@artika5707808 Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker saja di j... https://t.co/g23AIsag7e
Sat, 03 Jul 2021 04:45:28 GMT	YoaraSoshiMao	Dari 88rb jd 98rb. Demi kenyamanan pelanggan pala lu @Telkomsel https://t.co/y00DeAt6q

Figure 8. Comment Crawl Results



Process Sentiment

Excel

Sentiment Result

Show 5 entries

Waktu	Username	Tweet	Clean HTML	Clean Mention	English	Sentiment	Polarity
Fri, 02 Jul 2021 22:59:15 GMT	xierenchao08	140	@Vinao_Vinn77 Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker Sekarang jug... https://t.co/L974Uu8lt	Vinao_Vinn77 Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker Sekarang jug https://t.co/L974Uu8lt	Vinao_Vinn77 Don't Confuse Selecting Site Yang Proven Security and Comfort Visit Sahabatpoker Now JUG HTTPS T CO L974Uu8lt	netral	0
Fri, 02 Jul 2021 23:01:15 GMT	xierenchao08	140	@keisharubyy Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker saja di jamin... https://t.co/r0nk3gkqdt	keisharubyy Jangan bingung memilih situs yang terbukti keamanan dan kenyamanan Kunjungi Sahabatpoker saja di jamin https://t.co/r0nk3gkqdt	Keisharubyy Don't Be Confused Choosing a Site That is Proven Security and Comfort Visit Sahabatpoker Only Guaranteed HTTPS T CO R0NK3GKQDT	negatif	-0.2

Figure 9. Sentiment Process Results

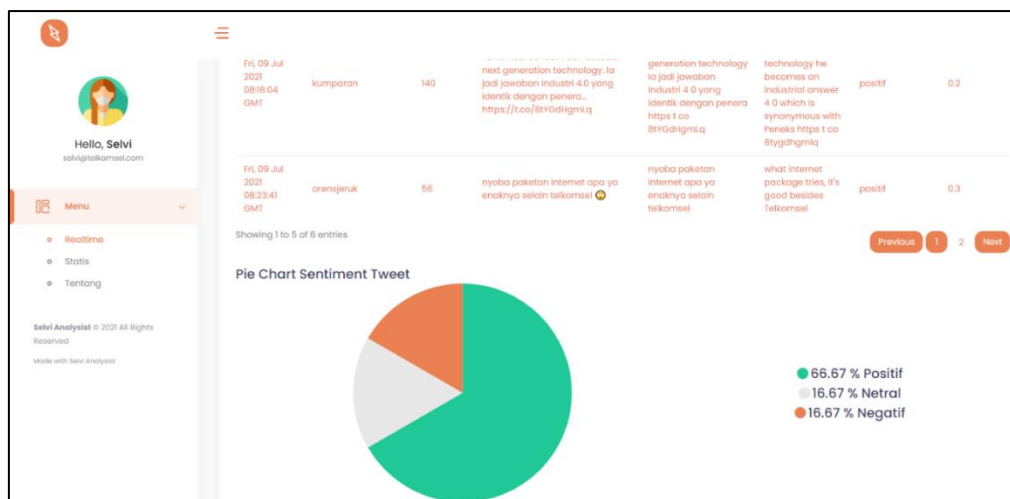


Figure 10. Pie Charts

The results of the summary of the sentiment score (polarity) of the system that has been made can be seen in Table 1.

Table 1. Sentiment Score (Polarity) of The System (a)

No	Period	CR1. Network		CR2 Convinience	
		Positive	Negative	Positive	Negative
1	May-20	45.45%	15.15%	37.10%	30.14%
2	Jun-20	92.92%	4.42%	35.68%	29.40%
3	Jul-20	53.27%	16.82%	30.87%	37.63%
4	Aug-20	52.73%	20.00%	31.59%	30.44%
5	Sep-20	29.09%	45.45%	35.42%	35.46%
6	Oct-20	40.00%	33.33%	38.58%	30.85%
7	Nov-20	47.06%	35.29%	39.92%	27.19%
8	Dec-20	76.92%	15.38%	41.41%	27.97%
9	Jan-21	60.87%	21.74%	37.17%	30.05%
10	Feb-21	66.67%	33.33%	40.96%	29.15%
11	Mar-21	78.95%	10.53%	44.15%	24.51%
12	Apr-21	81.82%	9.09%	34.53%	31.67%
13	May-21	55.17%	13.79%	28.76%	38.22%
Mean Polarity		60.07%	21.10%	36.63%	30.98%

Table 1. Sentiment Score (Polarity) of The System (b)

No	Period	CR3. Price/Tariff		CR4 Internet		CR5. Value Added Service	
		Positive	Negative	Positive	Negative	Positive	Negative
1	May-20	42.20%	28.90%	38.76%	27.50%	41.92%	20.96%
2	Jun-20	46.28%	25.46%	34.91%	27.74%	43.72%	23.43%
3	Jul-20	51.06%	24.53%	35.76%	24.64%	53.15%	22.03%
4	Aug-20	54.69%	16.77%	42.20%	25.99%	52.99%	16.34%
5	Sep-20	65.58%	15.48%	43.18%	25.70%	46.03%	24.73%
6	Oct-20	49.54%	30.85%	41.69%	26.83%	47.44%	23.59%
7	Nov-20	58.84%	15.90%	38.44%	30.08%	51.82%	23.70%
8	Dec-20	58.85%	15.77%	38.33%	31.56%	45.14%	23.88%
9	Jan-21	42.55%	46.01%	31.94%	35.45%	41.96%	17.75%
10	Feb-21	72.99%	10.55%	37.05%	31.81%	42.41%	21.20%
11	Mar-21	81.86%	8.42%	33.19%	34.61%	38.46%	23.08%
12	Apr-21	83.46%	8.53%	40.93%	29.20%	43.35%	18.50%
13	May-21	64.61%	18.66%	45.00%	26.75%	49.07%	10.79%
Mean Polarity		59.42%	20.45%	38.57%	29.07%	45.96%	20.77%

Customer needs can be analyzed further if the positive sentiment score is greater than the negative sentiment score. From Table 1 it can be seen that the positive sentiment score of all attributes > negative sentiment score so that it can be continued to data processing.

Next, translate the attributes of customer needs into initial customer requirements from the results of data crawling using the wordcloud application. An overview of processing using WordCloud can be seen as in Figure 11.



Figure 11. WordCloud Processing

The results of translating attributes into initial customer requirements can be seen in Table 2.

Table 2. Customer Requirement

No	Attribute (CR)	Definition of Need Variables
1	Network (CR1)	The provider has a stable and evenly distributed internet network throughout the location
2	Convenience (CR2)	Complaint handling services spread via offline and online are quickly responded to by the provider through online and offline customer service with accurate information
3	Price (CR3)	Internet package prices are more affordable for all customer segments and packages can be customized by customers themselves (package customize)
4	Internet (CR4)	Products are easy to get and use by customers throughout the region
5	Service (CR5)	Provider adds digital channel services for customers

4. Conclusion and Future Research

The results of the study obtained several conclusions such data collection using text mining method shows that from a total of 2893 complaint tweets, there are 5 main complaint variables that must be corrected by the company. The variables that become the improvement of PT XYZ's services are network, convenience, price, internet, and service. The results of testing on 5 variables carried out on data from May 2020 to May 2021, namely Network 60.07% positive and 21.10% negative, Convenience 36.63% positive and 30.98% negative, Price 59.42% positive and 20.45% negative, Internet 38.57% positive 29.07% negative, Service 45.96% positive and 20.77% negative. 3. In the results of testing for 5 variables, it can be concluded that the highest positive polarity is found in the Network variable 60.07% and for the lowest positive polarity it is found in the Comfort variable 36.63%. Meanwhile, the highest negative polarity is found in the Comfort variable 30.98% and the lowest negative polarity is found in the Price variable 20.45%. This

research contributes a proposal such develop a text mining system to gather and identify customer needs more accurately and effectively, then simultaneously classify and monitor sentiment towards products or services that have been launched.

Suggestions that can be submitted in research are for further research, programming systems can use algorithms other than nave Bayes classification, such as the K-NN algorithm and others and the big data used can use YouTube, Facebook, and so on. Pay attention to the similarity of words that appear so that cleaning can be done because this classification calculates the similarity of words that appear between documents. Add training data for better classification results. When collecting tweet data, it is necessary to better understand the Twint library so that the tweet data obtained is more in line with research needs. Using server in crawling step automatically, using repetition removal, and handling abbreviation and non-formal language translation.

REFERENCES

- [1] S. J. Ramos, "Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining," *International Journal of Sustainability*, vol. 11, no. 917, pp.1–14. 2019.
- [2] S.W. Cho, "Investigating Temporal and Spatial Trend Brand Images Using Twitter Opinion Mining," *International Information Science and Applications*, vol. 10, no. 1109, pp.1–10. 2014.
- [3] V. A. Kharde, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp.1–11. 2016.
- [4] J. Jin, "Understanding Big Consumer Opinion Data for Market-Driven Product Design," *International Journal of Production Research*, vol. 54, no. 10, pp.3019–3041. 2016.
- [5] H. Nguyen, "A Data Driven Study Influences in Twitter," *International Journal of Communications*, vol. 10, no. 1109, pp.1–12. 2014.
- [6] B. Liu, *Web Crawling in Web Data Mining: Exploring Hyperlinks, Content, and Usage Data*. New York: Springer, 2011. [Online] Available: Springer e-book.
- [7] S. Kopera. "Interdisciplinarity in Tech Startups Development–Case Study of Uni Startup Project Sentiment Analysis Using Text Data Mining," *International Journal of Foundation Management*, vol. 10, no. 2478, pp.23–32. 2018.
- [8] J.R. Saura. "Do Online Comments Affect Environmental Management Identifying Factors Related to Environmental Management and Sustainability of Hotels," *International Sustainability Journal*, vol. 9, no. 3016, pp.1–21. 2018.
- [9] L. Nailiang. "Identification of Key Customer Requirement Based on Online Reviews," *Journal of Intelligent and Fuzzy System*, vol. 39, no. 3, pp.3957–3970. 2020.

- [10] F.M. Javed. "An Effective Implementation of Web Crawling Technology to Retrieve Data from The World Wide Web," *International Journal of Scientific and Technology Research*, vol. 9, no. 1, pp.1252–1257. 2020.
- [11] L. Xiangdong. "A Computer Aided Approach for Acquisition and Importance Ranking of Customer Requirement from The Online Comment Mining," *International Computer Aided Journal*, vol. 19, no. 1, pp.132–151.2021.
- [12] S. Siddiqi. "Keyword and Keyphrase Extraction Techniques: A Literature Review," *International Journal of Computer Applications*, vol. 109, no. 2, pp.1–7. 2015.